928

Music Perception
Winter 1994, Vol. 12, No. 2, 227–255

The Tritone Paradox and the Pitch Range of the Speaking Voice: A Dubious Connection

BRUNO H. REPP
Haskins Laboratories

Deutsch and coworkers (Deutsch, 1991; Deutsch, North, & Ray, 1990) have proposed that individual differences in the perception of the "tritone paradox" derive from listeners' reference to a mental pitch template, acquired through experience with the pitch range of their own voice, as well as with the voice ranges typical of their language community. These authors have reported a correspondence between perceptual results and the upper limit of the individual voice range for a small group of selected subjects, as well as a striking difference in tritone perception between American and British listeners. The present study compared groups of Dutch, British, and American listeners on two tritone tests and also collected voice pitch data for the first two groups in a reading task. There was no within-group correlation of perceptual results with individual differences in voice range. Differences in tritone perception as a function of stimulus characteristics (spectral envelope) were much larger than reported by Deutsch, which casts doubt on the notion of stable individual pitch templates. A significant difference between British and American listeners, with the Dutch group in between, was found in one of the two tritone tests but not in the other. Although the origin of this difference remains unclear, it seems unlikely that it has anything to do with regional differences in voice pitch range.

Introduction

The purpose of the present study was to attempt to replicate and extend the startling and potentially important findings of Deutsch, North, and Ray (1990) and Deutsch (1991) concerning a possible connection between the perception of complex tones and the fundamental frequency range of the speaking voice. The basic theoretical claim is that individuals acquire a stable pitch template from exposure to their own voice and other voices in their language community and that this template determines the perception of relative pitch height in the task that gives rise to the "tritone paradox."

This experimental paradigm uses complex tones of the kind devised by

Requests for reprints may be sent to Bruno H. Repp, Haskins Laboratories, 270 Crown St., New Haven, CT 06511-6695.

Shepard (1964) to demonstrate the independence of pitch height and pitch quality (chroma). They are composed of octave-spaced partials, whose relative amplitudes are determined by a fixed spectral envelope. Shepard showed that, if the frequencies of all partials are increased in small steps of, say, one semitone, listeners perceive successive tones that increase in pitch. These increases continue to be heard indefinitely, even though after 12 steps a tone identical with the starting tone is reached, so that the tones keep going around the "pitch (chroma) circle" without ever increasing in pitch height in any objective sense. When pairs of these tones are formed, the resulting musical interval is perceived as rising or falling according to a principle of pitch proximity; thus, for example, the interval C-E is usually heard as a rising major third, not as a falling minor sixth, whereas the interval C-A is heard as a falling minor third, not as a rising major sixth. The ambiguous interval of a half-octave or tritone, such as C-F♯, is sometimes heard as rising, sometimes as falling. Shepard pointed out that this ambiguity is rarely perceived as such on any given trial, and he referred to reversible figures such as the Necker cube as a visual analogy.[1]

Deutsch's (1986) tritone test consists of a random sequence of such tone pairs, all of which form tritone intervals but start on any of the 12 semitone steps within one octave.[2] Her novel finding, further documented in several subsequent studies (Deutsch, 1987, 1991; Deutsch, Kuyper, & Fisher, 1987; Deutsch et al., 1990), was that individual listeners, rather than perceiving all these intervals sometimes as rising and sometimes as falling, perceive some of them consistently as rising and others (viz., their inverses) consistently as falling. A typical response function of a hypothetical subject is shown in Figure 1. The consistently rising or falling intervals may not be the same for different listeners, so that the same interval may be heard as rising by one subject but as falling by another (the "tritone paradox"). These results suggested to Deutsch that individuals refer to a personal pitch template, which may be portrayed as a particular orientation of the pitch circle (see Figure 1), such that some pitch classes (the ones on top of the rotated circle) are subjectively "higher" than others. Deutsch (1987) showed that this finding is not an artifact of using a fixed spectral envelope

---

[1] Shepard did not consider the possibility that this lack of momentary ambiguity may be due to strong influences of preceding context. Informal observations by the author suggest that the interval C-F♯ is perceived as rising in the context of the interval sequence C-C♯, C-D, C-D♯, . . . , but as falling in the context of C-B, C-A♯, C-A. . . . In fact, the author perceives *all* intervals up to C-B as rising in the first sequence and all intervals down to C-C♯ as falling in the second sequence, which suggests that sequential context effects can completely override the pitch proximity principle discussed by Shepard. (See Appendix A, Sequential Context Effects, for a discussion of sequential effects in the tritone task.)

[2] Deutsch made several changes with respect to Shepard's original stimuli: She used 6 rather than 10 octave-spaced partials, a somewhat differently shaped spectral envelope, longer tone durations (500 ms rather than 120 ms), and no silent intervals between successive tones in a pair. She did not explain the reasons for these changes.

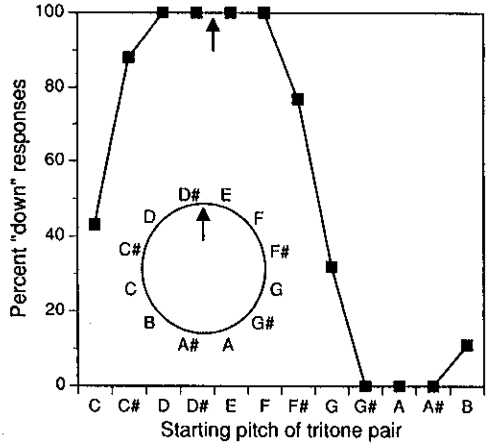

Fig. 1. Response function of a hypothetical subject in the tritone paradigm and inferred orientation of the subject's pitch circle.

centered on a particular frequency: Stimuli with envelopes centered on different frequencies are perceived similarly, although some listeners do show small shifts in their response functions. Moreover, Deutsch et al. (1987) found that, in a group of American listeners, the individually "highest" pitch classes were not randomly distributed: They fell most often between B and D#, but almost never between F# and A.

These interesting observations, which suggested that ordinary listeners possess something like absolute pitch, were followed by two studies in which an explanation of the origin of individual pitch templates was proposed. Deutsch et al. (1990) presented data for a small group of subjects, selected because of their different response functions in the tritone test. Each of these subjects was recorded speaking for about 15 min in an interviewlike situation, and the speech was analyzed to yield an overall fundamental frequency (F_0) distribution. Deutsch et al. then determined for each speaker the octave band that included the largest percentage of the F_0 values. For eight of nine subjects, the pitch classes delimiting this octave band and the pitch classes perceived as "highest" in the tritone test were within 2 semitones of each other, a result that significantly deviated from chance.[3] The authors hypothesized that the individual orientation of the pitch circle derives from experience with one's own speaking voice, and in particular, that the upper limit of the vocal range defines the pitch classes that are perceived as highest.

---

[3] The comparison actually involved pairs of pitch classes: those straddling the limits of the speech octave band and those on top of the pitch circle. The tritone perception results were analyzed in such a way that two adjacent pitch classes emerged as "highest" (e.g., D# and E in Figure 1).

An even more remarkable result was reported by Deutsch (1991). In that study, the tritone perception results of two groups of subjects were compared, one from California and the other from southern England (both tested in California under identical conditions). The distributions of the individual pitch circle orientations within each group were strikingly different: Whereas for the American subjects the highest pitch classes were most often between B and D♯, for the British subjects they were most often between F♯ and G♯. Thus the two distributions were almost complementary. Deutsch concluded that "perception of music can be strongly influenced by the language spoken by the listener" (p. 345). Although she did not spell out what the relevant regional language characteristic was, the obvious implication seems to be that it lies in the F_0 ranges used by American and British speakers.[4] Unfortunately, the study did not contain any speech data.

Differences in intonation between British and American English have been described in some publications, but plots of long-term F_0 distributions of the kind examined by Deutsch et al. (1990) are virtually absent from the phonetic literature. It is known that British English makes use of relatively large pitch excursions compared with languages such as Dutch and German (Collier, 1991; De Pijper, 1983; Willems, Collier, & 't Hart, 1988) and, presumably, American English.[5] The implication of this for the long-term F_0 distribution is that it will be wider and more skewed toward high frequencies. The average F_0 will also be higher. Differences in the average F_0 of speakers from different languages or dialect groups have been reported in several studies, although none included British English. However, they suggest that speakers of general American English have relatively low average F_0 values, compared with speakers of Southern U.S. dialects (Hanley, 1951), Spanish (Hanley & Snidecor, 1967; Hanley, Snidecor, & Ringel, 1966), and Japanese (Hanley & Snidecor, 1967; Hanley et al., 1966; Yamazawa & Hollien, 1992).

Although a difference in average F_0 and/or F_0 range between British and American speakers seems plausible, it is also known that F_0 characteristics vary widely among adults within any language or dialect commuity. The most obvious cause of such differences, the sex of the speaker, can perhaps be disregarded in the present context: Women's voices tend to be roughly one octave higher than men's, so that the average male and female F_0 ranges are fairly similar in terms of musical pitch classes.[6] Among adults of the same

---

[4] Several more recent publications that summarize the same findings (Deutsch, 1992a, 1992b, 1992c) are equally nonspecific about the presumed link between "language" and tritone perception.

[5] It is difficult to find a direct comparison of British and American English in terms of F_0 measurements. The Dutch researchers who have been most active in this area did not include American English in their studies.

[6] Actually, the average difference is less than one octave (see, e.g., Hudson & Holbrook, 1981).

sex, however, there is wide variation in average F_0 (see, e.g., Boë & Rakotofiringa, 1975; Hollien & Jackson, 1973; Horii, 1975), due to anatomic (vocal cord length), physiologic (age, disease), psychological (e.g., personality), linguistic, and other factors. Among male speakers, average F_0 can differ by as much as an octave; for women, the range of interindividual variation (in semitones) is somewhat smaller. This variation implies that, even if there is a difference in average F_0 or F_0 range between two language groups, there will be substantial overlap in the distribution of individual speakers' values. The almost nonoverlapping distributions of the pitch circle orientations of British and American listeners in Deutsch's (1991) study thus seem at variance with the known distributional characteristics of speaking voices.

There are also some potential problems with Deutsch et al.'s (1990) method of capturing intraindividual F_0 distributions within an octave band. Although the average width of speakers' F_0 ranges tends to be close to an octave (Hanley, 1951; Hudson & Holbrook, 1982), individual ranges vary considerably, which is why studies in the speech literature generally use percentiles or standard deviations to characterize F_0 range (see, e.g., Jassem, 1971; Jassem & Kudela-Dobrogowska, 1980). The octave band procedure overestimates the bounds of narrow ranges and underestimates those of wide ranges. Furthermore, Deutsch et al. focus on the upper limit of the F_0 range as a potential correlate of the pitch classes that are perceived as highest in the tritone test. This choice is problematic for two reasons. First, the upper limit of an individual F_0 distribution is not well defined: The distribution has a fairly gradual slope at high values, and the upper limit is likely to be highly situation-dependent. A more stable point of reference is the *lower* limit of the distribution, usually a fairly abrupt cutoff. Many studies of intonation have pointed out that the bottom of a speaker's range is a stable individual characteristic that is usually reached at the end of a complete utterance (Liberman & Pierrehumbert, 1984; Maeda, 1976; Terken, 1993; 't Hart, Collier & Cohen, 1990); no such claim has ever been made about the upper limit of the range, to this author's knowledge. Second, it is not clear why the perceptually highest pitch class should correspond to the highest pitch class in a speaker's octave band, because that pitch class also represents the lower limit of the band. If any pitch class within an (inherently circular) octave range is to be considered "highest," it would have to be one that is several semitones *below* the upper limit, so that it is not only relatively high but also sufficiently removed from the pitch classes at the low end of the range. Therefore, the match between tritone perception and speech production found by Deutsch et al. (1990) may not be so close, after all.

The purpose of the present study was, first, to attempt to replicate the within-group findings of Deutsch et al. (1990) with two separate groups of subjects, one Dutch and the other British, using a sentence-reading task to estimate the upper and lower limits of speakers' F_0 ranges. Second, the between-group differences in tritone perception reported by Deutsch (1991) were reexamined and extended by comparing three groups of listeners: Dutch, British, and American. Initially only Dutch and British subjects (plus a few Americans) were tested, as this study was primarily carried out in Europe; an American group was added after the author's return to the U.S. for comparison on the perceptual test only. Based on what is known about the intonational characteristics of Dutch and British English, it was expected that Dutch speakers would be more like American than like British English speakers in their use of F_0, although considerable overlap of F_0 ranges between language groups was expected. If language influences tritone perception, as Deutsch (1991) conjectured, then a difference between British subjects on the one hand and Dutch and American subjects on the other hand should emerge in the tritone test. Moreover, the results for British and American subjects should match those of Deutsch, with the perceptually highest pitch classes being between F♯ and G♯ for the British and between B and D♯ for the Americans. Third, two sets of tones with different spectral envelopes (plus a third set for Dutch listeners only) were included to verify the crucial prerequisite that individual pitch circle orientations are stable across changes in stimulus characteristics (Deutsch, 1987). Without such stability, it would not make sense to look for correlations between perceptual results and speech characteristics.

Methods

STIMULI

Using software developed at the Institute for Perception Research (IPO, see footnote 18) by W. M. Wagenaars, three sets of 12 complex tones each were synthesized. The first two sets followed the specifications in Deutsch's publications (see, e.g., Deutsch, 1991); these will be called "Deutsch tones" in the following. Each of these tones had six partials spaced at octave intervals.[7] Their frequencies were varied in semitone steps. A different spectral amplitude envelope was used in each set, one centered at 622 Hz (D♯$_5$) and the other at 440 Hz (A$_4$). These two envelopes are illustrated by the right-hand functions in Figure 2.[8] The third set of 12 tones was generated using the specifications in Shepard's (1964) original paper.[9] These "Shepard tones" were included to determine whether they would yield

---

[7] Two of the 12 tones had only five partials, as the first or sixth partial had zero amplitude.

[8] The envelopes were generated by substituting Deutsch's formula for the original Shepard (1964) formula in the program code. They look different from those in Deutsch's figures because they are drawn on a linear scale whereas Deutsch plots them on a decibel scale. Some ripples in the curves are due to the fact that they were drawn by connecting 72 data points (6 partials × 12 stimuli), not by plotting the smooth mathematical function specifying the relative amplitudes.

[9] Except for their longer duration, these tones were identical to those recorded on a well-known compact disk of auditory demonstrations (Houtsma, Rossing, & Wagenaars, 1987).

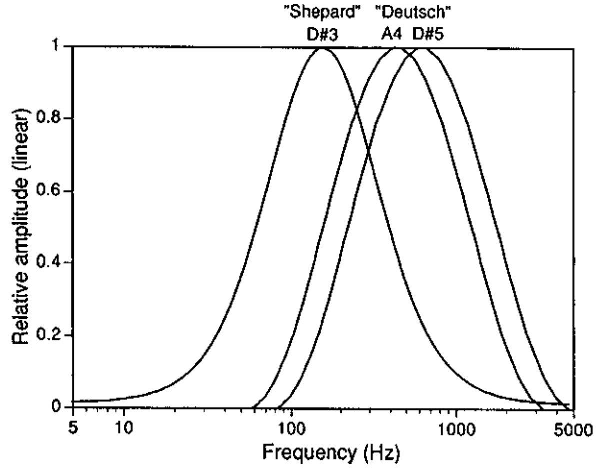

Fig. 2. Fixed spectral envelopes of the Shepard tones (left-hand function) and of two series of Deutsch tones (right-hand functions). Shepard tones had 10 octave-spaced partials under the envelope, Deutsch tones had 6.

perceptual results equivalent to those obtained with the Deutsch tones. (They were presented only to the Dutch group of subjects.) The Shepard tones had 10 octave-spaced partials (the lowest frequencies actually being below the pitch threshold) and a more peaked spectral envelope centered on 156 Hz (D♯$_3$), as shown by the left-hand function in Figure 1. Because of their stronger low-frequency components, the Shepard tones had a fuller, more organlike timbre than the Deutsch tones. All tones were 500 ms in duration, with 10-ms amplitude ramps at onset and offset. Their overall sound levels were very nearly equal within and across series. All partials started in zero phase. The synthesized waveforms were represented with 16-bit precision at a 10-kHz sampling rate.

Twelve tritone pairs (C-F♯, C♯-G, etc.) were constructed from each tone series by concatenating the appropriate waveforms without any intervening silence (as in Deutsch's studies). These tritone pairs were then arranged into three tests (Shepard, Deutsch-A, and Deutsch-D♯), each comprising 12 repetitions of the 12 pairs of one series. The interpair intervals were 2.5 s, with an extra 2.5 s after each block of 12. Each pair occurred once in each block. Because pilot observations had suggested the existence of strong sequential context effects (see Appendix A), the sequence of pairs was such that each pair was preceded once by each other pair, but never by itself.[10] (The initial pair in each block was not scored.) The three 144-item tests used the same context-balanced stimulus sequence, but the order of blocks was different. The test sequences were low-pass filtered at 4.9 kHz and recorded onto digital tape for presentation. Each test was preceded by an ascending ordered sequence of the 12 tones, to mark the beginning and to introduce the subjects to the sound of the tones.[11]

In addition to the listening tests, a set of 10 sentences, both in English and in Dutch translation, was devised for the assessment of F_0 characteristics. The sentences were statements of medium length and had relatively deaccented words at the end. They are listed in Appendix B.

---

[10] This was accomplished by first constructing a 12 × 12 matrix of the appropriate permutations of the numbers 1–12 and then assigning the 12 tone pairs at random to the 12 numbers.

[11] The tones ascended from D♯ to D in the Deutsch-D♯ and Shepard tone tests, and from A to G♯ in the Deutsch-A test. This rather unnecessary deviation from Deutsch's procedures will be discussed further below.

SUBJECTS AND PROCEDURE

Dutch Subjects

These were 15 members of the IPO research staff, 8 men and 7 women, all unpaid volunteers and native speakers of Dutch. They listened to all three tests in a quiet classroom using Sennheiser HD 530 II earphones with circumaural cushions. Presentation was binaural at a comfortable intensity (determined with a sound-level meter and earphone coupler to be approximately 77 dB SPL). Written instructions were given in English. Some examples, taken at random from within the first test, were played for familiarization before starting. Subjects recorded their judgments of whether the pitch went up or down by writing upward- or downward-pointing arrows onto an answer sheet; a forced choice had to be made on each trial. There were breaks of a few minutes between tests. The three tests were presented in three different orders to groups of 5 subjects each.

Immediately after the session or at a later time, each subject was asked to read the 10 Dutch sentences in front of a microphone. The sentences were printed on index cards that were randomly shuffled for each subject. The subjects were asked to familiarize themselves with the sentences and then to read them "in a natural and relaxed fashion." The speech was recorded on digital tape for later analysis.

British Subjects

The British subjects were 10 staff members of the University of Sussex at Brighton, 5 men and 5 women, all natives of southern England, who were paid for their participation, and one additional male volunteer (JS1-m, a native of London) who visited IPO and was tested there. The first 10 subjects were tested individually in a sound-isolated booth in the Laboratory of Experimental Psychology at the University of Sussex. The stimulus tape was played back on a portable Sony DAT recorder, and subjects listened binaurally over a pair of the earphones used at IPO, which had been brought along by the author. The voltage at the earphones was calibrated with a volt meter to equal the value measured at IPO (70 mV). Only the two Deutsch tone tests were presented to the British subjects, with 6 subjects listening in one order and 5 subjects in the other. Otherwise, the procedure of testing and recording was the same as for the Dutch subjects, except that the sentences were read in English.

American Subjects

There were 17 subjects, 8 men and 9 women, all unpaid volunteers. One (GS-m) was a postdoc at the University of Sussex; 4 others (DK-m, JJ-m, SC-f, TM-f) were visiting researchers at IPO. These 5 subjects were tested under the same conditions as the British and Dutch subjects, except that they listened only to the two Deutsch tone tests. They were also recorded reading the English sentences. The remaining 12 subjects were 9 members of the research staff at Haskins Laboratories, 2 additional graduate students, and 1 recent high school graduate.[12] They were tested individually in a quiet room at Haskins Laboratories and used Sennheiser HD 420 SL earphones with on-the-ear cushions. The playback level was similar to (but not calibrated to be identical with) that used in Europe. (See Appendix A concerning effects of playback level.) Only the two Deutsch tone tests were presented, with their order varying across subjects. The ascending tone sequence preceding each test was omitted for these 12 subjects. Also, no speech samples were collected from them, as the IPO software used for the F_0 analysis was no longer available.

---

[12] This subject (MR-f, a violinist) was the only subject in this study known to possess absolute pitch.

DATA ANALYSIS

For each subject and each tritone test, the number of "down" responses to each of the 12 stimulus pairs was tallied (ignoring the first response in each block). Subsequently, the six adjacent stimulus pairs that had the highest number of "down" responses were determined (see Figure 1). Following Deutsch's procedures, the starting pitches of the middle two (i.e., the third and fourth) of these pairs were taken to be the "highest" pitches, regardless of the exact distribution of responses. In most cases, these pairs in fact received the highest number of "down" responses, although adjacent pairs often had equal (maximum) scores. Although the clarity of the "pitch class effect" (i.e., the complementary distribution of "up" and "down" responses across the 12 stimulus pairs) varied, the subjectively highest pitch classes could be determined in this way in all but four instances of apparently random responses.[13]

The recorded speech was digitized at 10 kHz with low-pass filtering at 4.9 kHz, and F_0 contours were determined by using software developed at IPO that uses the method of subharmonic summation (Hermes, 1988). Two F_0 values were measured in each sentence contour: The lowest value near the end (taking care to avoid artifacts due to vocal fry or very low amplitude), and the highest value, which was usually on the first accented syllable. These values could be determined in all but one instance (a missing sentence). In four additional instances, one of the two measurement values was excluded as an extreme outlier compared with the values from the other 9 sentences. One Dutch subject's (JR-m) extremely low F_0 trailed off into creak at the ends of sentences, so only the end of regular voicing could be measured. The measured values of each subject were averaged across the 10 sentences, and standard deviations were calculated (in linear hertz).

Results and Discussion

TRITONE PERCEPTION

Individual Pitch Class Effects

The existence of pitch class effects for individual subjects was amply confirmed by the present results. As already mentioned, there were only four instances in which it was impossible to determine a particular orientation of the pitch circle.[14] No subject failed to show a pitch class effect on all two or three tests, although a few showed weak effects throughout. (See also Appendix A, Sequential Context Effects.) Defining as a "clear" effect any response pattern that showed 0% or 100% "down" responses for at

---

[13] In one exceptional case (Dutch subject LB-m), a reversal of the pitch class effect in the middle of the Shepard tone test was noted. These data were scored according to the first half of the test, which seemed in better agreement with the results of the other subjects. Another Dutch subject (AB-f) gave only "up" responses in the first test (Deutsch-D♯). She was encouraged to vary her responses, and when she repeated the Deutsch-D♯ test at the end of the session, she gave a clear pitch class effect. (See also Appendix A, Listening Strategies.) A few subjects yielded data that suggested a single highest pitch class rather than two.

[14] These instances were Dutch subject JM-f in the Shepard tone test, British subject JS2-m in the Deutsch-D♯ test, and British subject JC-f and American subject DK-m in the Deutsch-A test.

least one stimulus pair, 28 of the 45 Dutch (15 subjects × 3 tests), 14 of the 22 British (11 subjects × 2 tests), and 26 of the 34 American (17 subjects × 2 tests) test results fell in that category. Defining as a "strong" effect any response pattern with 6 or more stimulus pairs receiving extreme scores (as in the example in Figure 1), there were 18 Dutch, 6 British, and 16 American cases in that category. The results also show that reliable pitch class effects can be obtained with the original Shepard tones: Of the 15 Dutch subjects, 8 showed a clear pitch class effect in the Deutsch-D# test, 11 in the Deutsch-A test, and 9 in the Shepard test. However, pitch class effects tended to be more pronounced in the Deutsch-A test than in the Deutsch-D# test: In the former, there were 23 strong, 10 moderate (i.e., clear but not strong), and 8 weak effects overall, whereas in the latter, there were 9 strong, 19 moderate, and 14 weak effects.

Between-Test Differences

To present one extreme but not singular example of individual response functions, Figure 3 shows the results of one Dutch subject (MS-f) who showed clear pitch class effects in all three tests. The results of the three tests were extremely dissimilar, which was quite unexpected given the data reported by Deutsch (1987). Results such as shown in the figure could not possibly reflect a stable individual pitch template that is operative regardless of the spectral characteristics of the tones. Clearly, the spectral envelopes of the tones made a substantial difference.

The tritone test results of all subjects are summarized in Figure 4 in terms of the individually highest pitch classes in each test. It is evident that differences among the three tests were pervasive, systematic, and frequently very large. Of the 40 subjects who yielded at least weak pitch class effects in both Deutsch tone tests, only 3 (all American) showed identical effects, whereas 5 (1 Dutch, 4 American) produced *maximally different* effects (i.e., separated by 6 semitones). Of the remaining 32 subjects, 24 showed a highest pitch class in the Deutsch-D# test that was 1–5 semitones *higher* than that in the Deutsch-A test; this is significant by a binomial test ($p < .005$) and shows the direction of the difference to be systematic. This pattern was especially characteristic of the Dutch (12 out of 14) and British (8 out of 9) subjects, although not of the Americans (4 out of 9). Breaking down the results in yet another way: Of the 40 subjects, 17 (7 Dutch, 4 British, 6 American) showed similar pitch class effects (0–2 semitones difference) in the two Deutsch tone tests, 11 (4 Dutch, 5 British, 2 American) showed moderately different effects (3–4 semitones difference), and 12 (4 Dutch, 8 American) showed highly dissimilar effects (5–6 semitones difference). Thus, fewer than half of the subjects showed the pattern Deutsch's reports would have led one to expect.

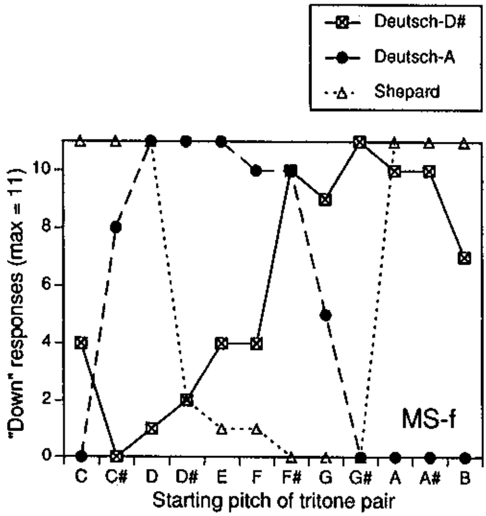

Fig. 3. Response functions of one Dutch subject (MS-f), illustrating large between-test differences.

In addition, it is clear that the Shepard tone test, which was administered only to the Dutch subjects, yielded substantially different results from the two Deutsch tone tests. In only 6 of 28 possible between-test comparisons were the pitch class effects similar (0–2 semitones apart), whereas they were highly dissimilar (5–6 semitones apart) in 12 comparisons. Only one subject (JT-m) showed similar pitch class effects in all three tests. Overall, these results call into question the notion of an individually stable pitch template and suggest strongly that the pitch class effect depends, at least in part, on spectral stimulus characteristics.

One point of concern was that the ascending scale that preceded each test may have biased subjects' responses. Given that preceding context has an effect in these tests (see Appendix A, Sequential Context Effects), it might be argued that the scale primed listeners to consider certain pitch classes as low and others as high, and because the two Deutsch tests were preceded by different scales, this might have contributed to the large between-test differences in results. However, this "priming hypothesis" can almost certainly be dismissed. First, the Deutsch-D# and Shepard tone tests were preceded by the same scale (D#-D) and so should have yielded similar results, which they did not. Second, the priming hypothesis predicts a particular orientation of the pitch circle: For a preceding D#-D scale, the "highest" pitches should be in the B-C# region, where they were very rarely in the Deutsch-D# test (although the prediction is accurate for the Shepard tone results); for a preceding A-G# scale, they should be near F-G, which does not fit the data of the Deutsch-A test. The agreement in

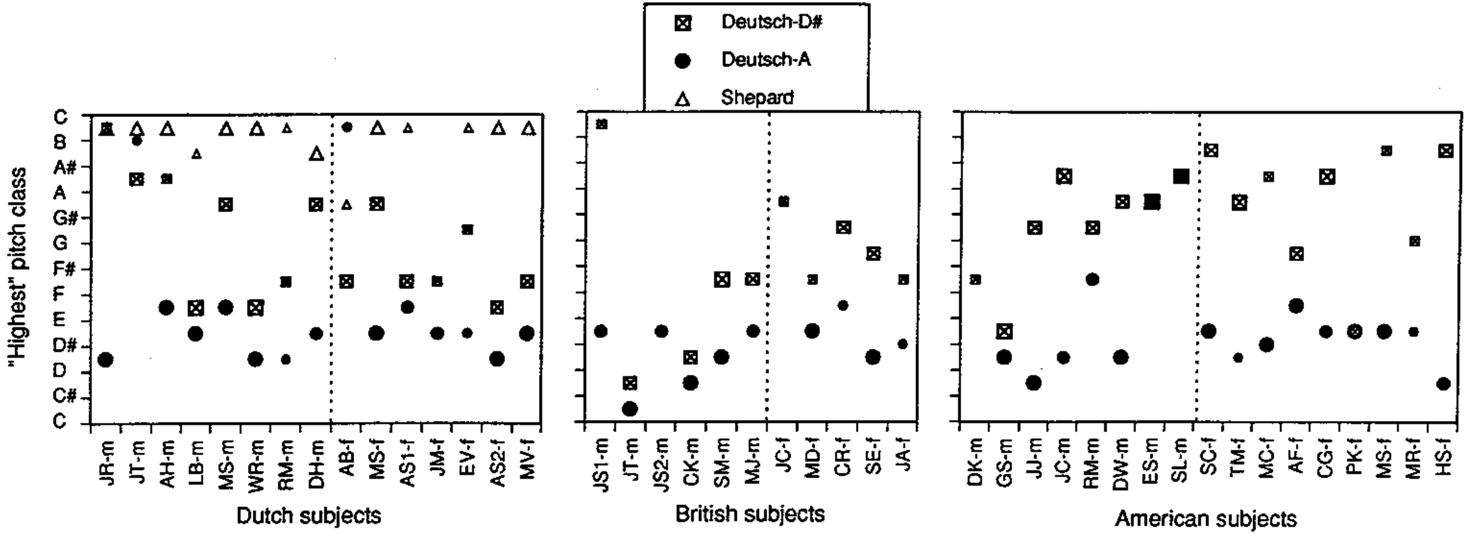

Fig. 4. Summary of tritone test results for 15 Dutch, 11 British, and 17 American subjects. The size of the data points indicates the strength of the effects: strong, moderate, or weak. The subjects within the Dutch and British groups are arranged according to increasing lower limits of their voice range. The first three male and first two female American subjects (tested in Europe) are arranged in the same way; the others are ordered arbitrarily. The dotted vertical lines separate male and female subjects.

the Shepard tone test then is more likely a coincidence. Third, it seems highly implausible that listeners would be able to remember the pitch range of the initial scale for very long, given the interference and the context effects caused by the tritone pairs in the test. It seems equally unlikely that a context effect initiated by the scale would be propagated deterministically through the test sequence. Finally, it will be recalled that the precursor scales were omitted in the tests given to the 12 American subjects tested at Haskins Laboratories. Although three of these subjects showed very similar pitch class effects on the two Deutsch tone tests, others showed very large differences. Both patterns of results were also shown by the American subjects tested in Europe (GS-m, JJ-m, SC-f, TM-f), who did hear the precursor scales. Therefore, the scales probably had little effect, if any, on tritone perception.

Within-Test Differences

There were some striking individual differences within tests, which is consistent with Deutsch's reports. In two of the tests, however, the majority of the subjects showed very similar results, and only a few listeners deviated from this predominant pattern. Individual differences seemed most constrained with the Shepard tones: 13 of the 15 Dutch subjects had their highest pitch classes between A♯ and C, and for 11 of them they were B and C. One of the two subjects who were not scored (JM-f) had an ambiguous response function that could be interpreted as a B-C pitch class effect as well. The single outlier (subject AB-f) was within 3 semitones and represented a weak effect by the definition given earlier. Thus, there was actually very little individual variability in the Shepard tone test.

Similarly, the Deutsch-A test showed considerable consistency among subjects. Of the 41 subjects who showed a pitch class effect, 35 had their highest pitch classes between C♯ and F, 26 between D and E. Two additional subjects were just 1 semitone away. Both of the two Dutch outliers (JT-m, AB-f) showed very weak effects. This leaves only two truly deviant subjects, the Americans ES-m and SL-m, both of whom showed strong pitch class effects that were about 6 semitones away from the dominant region—that is, reversed effects with respect to most other subjects.

Only the Deutsch-D♯ test showed individual variability of the magnitude that Deutsch's results (especially Deutsch et al., 1987) would have led one to expect. Most highest pitch classes (35 out of 42) fell within the half-octave region between E and A♯. Five other subjects were within 1 semitone and the two clearest outliers (Dutch subject JR-m and British subject JS1-m) again represented weak and hence rather unreliable effects.

The unexpectedly restricted within-test variation on the Deutsch-A and Shepard tone tests obviously provides a very poor basis for investigating correlations with voice range. As will be shown below, subjects varied greatly in their voice characteristics, yet they perceived the tritone pairs in these two tests very similarly. Only the Deutsch-D# test results gave any hope of finding a correlation.

Between-Group Differences

It is evident from Figure 4 that the striking difference found by Deutsch (1991) between British and American listeners was not replicated in the Deutsch-A test. In fact, there was high agreement among all three subject groups on this test (notwithstanding the presence of two outliers each in the Dutch and American groups). In the Deutsch-D# test, where there was more variability, the results of the three subject groups also overlapped substantially, but there was a definite tendency for American subjects' results to be higher up on the pitch scale than those of British subjects. This tendency was weakened by the very "high" pitch class effect of British subject JS1-m. However, since the starting pitch class (C) of the circular octave range displayed in Figure 4 is arbitrary, JS1-m could also be imagined just below the abscissa of the graph, in order to maximize the group difference. For the statistical comparison, therefore, the individually highest pitch classes were expressed numerically (C =0, . . . B = 11), but the results for British subject JS1-m as well as for Dutch subject JR-m were coded as −0.5. With the cards stacked in this way, there was indeed a significant difference among the three subject groups in a one-way analysis of variance (ANOVA) [$F(2,39) = 5.47$, $p < .009$], and subsequent pairwise comparisons showed this effect to be due mainly to the difference between British and American subjects [$F(1,25) = 10.81$, $p < .004$], with the difference between Dutch and British subjects being nonsignificant [$F(1,23) = 1.56$], and that between Dutch and American subjects being marginally significant [$F(1,30) = 4.61$, $p < .05$]. By contrast, for the similarly coded results on the Deutsch-A test, there was no effect of language group [$F(2,38) = 1.43$].

The average numerically coded highest pitch classes on the Deutsch-D# test were 6.2 (i.e., just above F#; S.D. = 2.6 semitones) for the Dutch subjects, 4.8 (just below F; S.D. = 2.8 semitones) for the British subjects, and 8.0 (G#; S.D. = 2.2 semitones) for the Americans. Thus the average difference between the British and American groups was 3.2 semitones, which is not inconsistent with the results of Deutsch (1991), although the difference appeared to be larger there. Her British subjects had their pitch classes predominantly between F# and G#, which is consistent with the Deutsch-D# test results, where 7 of 10 British subjects had their highest pitch class between F and G#. However, Deutsch's American subjects fell mostly in the region between B and D#, where none of the present American subjects could be found. Of the 17 Americans, 14 had their highest pitch classes between F# and B in the Deutsch-D# test. The Deutsch-A test results, on the other hand, clash with Deutsch's results for British subjects, whereas they are more compatible for Americans.

SPEECH DATA

Voice Ranges

Figure 5 shows the individual F_0 ranges for the three subject groups, arranged in terms of increasing lower limits within each group. (Speech data were available from only five American subjects, those tested in Europe.) The ranges around the estimates of the lowest and highest F_0 values for each subject represent one standard deviation above and below the mean computed across the 10 sentences; the distance between the two estimates is the individual vocal range. As expected, most speakers showed quite consistent values for the lower limits of their ranges. The estimates of the upper limits were more variable, as they depended more on the linguistic structure of the sentences.

A considerable diversity of vocal ranges was represented, especially among the male speakers. Quite a number of speakers had ranges of roughly one octave. However, there were some speakers with exceptionally wide or narrow ranges, for whom the octave band approach of Deutsch et al. (1990) would not be appropriate. As expected, there was extensive overlap between both the lower and the upper limits of the voice ranges of Dutch and British (and the few American) subjects. Although it

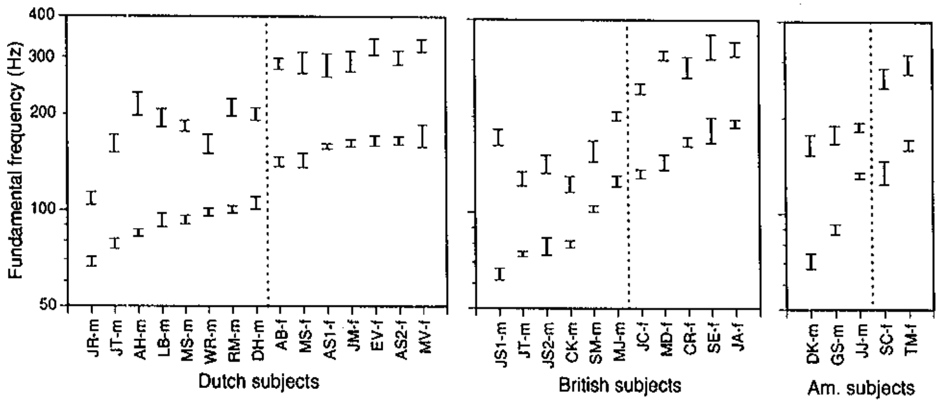

Fig. 5. Estimated values (ranges of one standard deviation above and below the mean) of the lower and upper limits of subjects' speaking ranges. The dotted vertical lines separate male and female speakers. The subjects within nationality groups are arranged according to increasing lower limits of their voice range.

had been predicted that British speakers would exhibit higher upper limits and wider individual ranges than Dutch and American speakers, there was no indication in the data of such a difference. It seems possible that observations in the linguistic and phonetic literature on the wide pitch excursions of British English apply only to "received pronunciation," a traditionally upper-class style of speaking that was not prevalent among the younger generation to which all but one subject belonged. The single middle-aged subject (JS1-m) was the only British speaker who showed an exceptionally wide vocal range.

Two-way ANOVAs were conducted on upper limits, lower limits, and their differences; the factors were language (Dutch vs. English) and speaker's sex. The female F_0 values were first divided by 2, thus lowering them by one octave. There were no significant effects in any of these analyses.

Relationship Between Voice Range and Tritone Perception

The cross-language comparison of vocal ranges offers no clues as to what aspect of the speaking voice might account for the differences between British and American listeners in the Deutsch-D♯ tritone perception test. Admittedly, this aspect of the study is hampered by the absence of speech data for the majority of the American subjects. Note, however, that the perceptually highest pitch classes were *higher* for American than for British subjects (Figure 4), whereas their voice ranges were expected to be lower. This makes it seem very unlikely that the group differences in tritone perception have any F_0 correlate.

Because of their variability, the F_0 data provide a good basis for exploring correlations with individual subjects' perceptual results. Unfortunately, as was mentioned earlier, the results for two of the tritone tests (Deutsch-A and Shepard) do not offer enough individual variability for that purpose. Therefore, the analysis focuses on the Deutsch-D♯ test results.

Deutsch et al. (1990) hypothesized that the upper limit of the vocal range—and, by implication via their octave-band criterion, the lower limit as well—should match the pitch classes on top of the individual pitch circle. This prediction is tested in Figure 6, where the highest pitch classes obtained in the Deutsch-D♯ test are plotted as a function of the individual lower and upper voice limits, respectively, for all subjects that provided speech data. The solid diagonal line represents identity, and the parallel dotted lines delimit a range of ±6 semitones. (These lines coincide on the cylindrical surface that is flattened out in the figure.) It is evident that there was no close correspondence of the perceptual data with either end of the voice range, although in each case there was a slight trend: 18 of 30 data points fell within 3 semitones of the lower limit of the voice range (n.s.), and 20 of 30 fell within 3 semitones of the upper limit ($p = .05$).

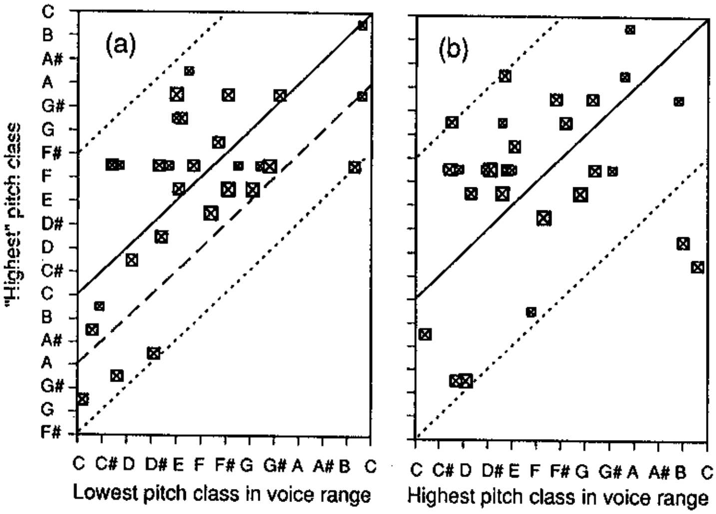

Fig. 6. Mean values of subjects' (a) lowest and (b) highest voice pitches plotted against the perceptually highest pitch classes in the Deutsch-D# tritone test. Results for 15 Dutch, 10 British, and 5 American subjects are combined. The size of the symbols indicates the strength of the pitch class effect (as in Fig. 4). The dotted lines delimit the octave band around the line of equality (solid diagonal). The dashed line indicates the pitch classes 9 semitones above (3 semitones below) the lower limit.

There is a problem with these comparisons, however. As pointed out in the Introduction, it is not clear why either the highest or the lowest pitch class in the voice range should correspond to the highest perceived pitch class. When the voice range equals or exceeds an octave, the highest pitch class(es) is (are) simultaneously the lowest pitch class(es). The truly highest pitch classes in a circular range are those that have no representation at the low end. Allowing for some "smoothing" or uncertainty at each end of the range to avoid abrupt discontinuities, the truly highest pitch class should be about 9 semitones above the lower limit. This more sensible pitch class reference is indicated by the dashed line in Figure 6a. It does not correlate with the perceptual data, as only 16 of 30 data points fall within 3 semitones of it.

General Discussion

Deutsch's findings on the tritone paradox are extremely interesting because they seem to suggest that ordinary listeners without absolute-pitch capacities nevertheless have a stable pitch reference in their heads. Three findings seemed well established before the present study was conducted: (1) Individual listeners show pitch class effects in the tritone task, (2) individual pitch class effects are essentially stable across tests that use tones with different spectral envelopes, and (3) there is large individual variability in pitch class effects.

The present study confirms the first result but unexpectedly challenges the second one and, in part, the third. In the three different tests used here, and in the two Deutsch tone tests in particular, widely divergent results were obtained for many subjects. Moreover, these differences followed a fairly systematic pattern. These results not only suggest that spectral envelope characteristics play an important role in the tritone perception task, but that the assumption of an individually stable pitch template may be incorrect.

The reason for this apparent discrepancy from Deutsch's findings is not known at present. Deutsch (1987) reported data for only four subjects. These subjects listened to 12 different stimulus sets differing in the center frequencies of their spectral envelopes, two of which ($D\#_5$ and A_4) matched the present Deutsch-D# and Deutsch-A sets. For these two tests, the subjects' "highest" pitch classes did not differ by more than 1 semitone, and response functions were generally similar across all 12 tests. In their Figure 7, Deutsch et al. (1987) display results for a single subject whose particularly pronounced pitch class effects were virtually identical in four different tests. In her later studies, Deutsch averaged over the results of several tests, which suggests that the pitch class effect varied only slightly and idiosyncratically as a function of spectral envelope. Whether that was in fact the case in all instances has not been documented. In the present study, however, it would make little sense to average over the Deutsch-D# and Deutsch-A test results; the discrepancies are much too large.

Another unexpected finding was the restricted individual variability in the Shepard and Deutsch-A tests. It appears that tones with certain spectral characteristics are perceived similarly by most or all listeners, which is contrary to the hypothesis that tritone perception reflects diverse individual pitch templates. It could be, however, that spectral and individual determinants of tritone perception are in competition, with the former gaining the upper hand in certain situations. This does not increase one's confidence in the Deutsch-D# test results as pure measures of individual pitch templates; on the contrary, spectral stimulus characteristics probably had an influence on subjects' percepts in that test also.

Deutsch's more recent findings regarding the connection between tritone perception and speech characteristics seemed not as well established and open to criticism even before the present study was conducted (see the Introduction). Her report of a correspondence between the upper limits of subjects' voice ranges and their hypothetical pitch templates (Deutsch et al., 1990) was based on a small group of selected subjects and on a questionable method of estimating the relevant pitch class in the voice range. Although her selection of subjects with very different pitch class effects is methodologically defensible, it does neglect subjects who have similar pitch class effects but different voice ranges—cases that are inconsistent with the hypothesis being tested. Although her most recent study (Deutsch, 1991) presented a striking difference between British and American subjects in tritone perception, it provided neither speech data nor an explicit hypothesis about the way in which tritone perception might depend on "language."

The present study sought to fill some of these gaps but was hampered by the unexpected instability of the pitch class effect across different tests, as well as by the lack of sufficient individual variability within two of the tests. Only the Deutsch-D# test yielded results that could be used to address the issues raised by Deutsch, which naturally weakens the conclusions. There is no reason why the pitch class effect exhibited on that particular test should reflect *the* pitch template of any given listener. Given that caveat, it is perhaps not surprising that there was no clear correspondence between the perceptual results and estimates of the highest pitch classes in subjects' voice ranges.

It could be objected that the current voice range estimates were inaccurate, either because of the small number of speech samples or because of the artificiality of the reading situation. Both objections are probably valid with regard to the upper limit of the range, which is not well defined and changes with speakers' emotional state and other factors. However, as was argued in the Introduction, the lower limit of the range is a more stable reference point, and that parameter was probably estimated with sufficient accuracy in the present study. It was also argued here that the highest pitch class in a subject's range should be defined relative to the lower, not the upper limit. As to the situation dependence of vocal range, studies in the speech literature have consistently reported somewhat higher average F_0 values for reading than for impromptu speaking (e.g., Hanley, 1951; Hanley et al., 1966; Mysak, 1959; Snidecor, 1943), but the difference is negligible at the lower end of the vocal range (Hollien & Jackson, 1973; Hudson & Holbrook, 1982).

Although the present results conflict with Deutsch's findings in several ways, there seems to be agreement with regard to a difference in tritone perception between British and American listeners. This agreement is only superficial, however. First, it derives from the Deutsch-D# test alone; on the Deutsch-A test, some other factor seemed to constrain perception similarly in both groups. Second, although the predominant "highest" pitch classes in the Deutsch-D# test agreed with those found by Deutsch (1991) for British subjects, there was no such match for Americans. It could be argued that this disagreement is due to the fact that Deutsch's subjects were all from California, whereas the present Americans were a heterogeneous group from various parts of the country (including California).[15] Ragozzine and Deutsch (1993) have recently reported some evidence for regional differences in tritone perception within the United States (in fact, within Youngstown, OH), suggesting that the regional origin of a subject's parents needs to be taken into account in interpreting the results. Be that as it may, it is far from clear how such regional differences are to be explained. Corroborative data on corresponding differences in regional speech characteristics have not been presented so far.

Third, the difference between British and American subjects in tritone perception remains puzzling because of the apparent absence of any corresponding difference in voice ranges between the groups. To be sure, not much can be concluded from the present comparison of 11 British speakers with only 5 Americans (those for whom speech data were available). However, by all accounts, Americans should have *lower* voice ranges than speakers of British English, whereas the Deutsch-D♯ test results indicated *higher* "highest" pitch classes for Americans. Thus, it seems extremely unlikely that voice characteristics could account for this difference.

Fourth, the between-group difference occurred in the absence of any within-group correlation with voice characteristics. If the pitch class effect had anything to do with individual voice range, then within-group and between-group correlations between tritone perception and speech production should go hand in hand. If so, however, it should have been much more difficult to find a significant between-group difference, considering the large within-group variability in voice characteristics (within each sex), which would seem to entail a similarly large variability in tritone perception.

Given the discrepancies between the present findings and Deutsch's results, it is perhaps premature to try to come up with alternative explanations of the pitch class effect. There may be unsuspected methodologic factors that explain these discrepancies, and new data may resolve the matter in favor of Deutsch's model. However, alternative models that treat pitch as a linear rather than as a circular dimension might be considered.

Terhardt (1991) has discussed briefly the tritone paradox from the perspective of his well-known virtual-pitch theory (Terhardt, Stoll, & Seewann, 1982a, 1982b). The theory holds that a complex tone evokes a number of competing pitch percepts, some of them spectral (i.e., directly corresponding to partials), others virtual (i.e., not necessarily corresponding to partials that are present, as in the case of a "missing fundamen-

---

[15] The Californians were subjects DK-m, GS-m, SC-f, TM-f, MS-f and HS-f. Other subjects had grown up mainly in the Midwest (JJ-m, JC-m, RM-m, MC-f), the East (ES-m, SL-m, AF-f, PK-f, MR-f), and the South (DW-m, CG-f); some of them had moved around as children. No information on subjects' parents was obtained. The author could not detect any relationship between subjects' regional origin and tritone perception results.

tal"). "Analytic" listeners are more prone to pay attention to spectral pitches (see Appendix A, Listening Strategies), whereas "synthetic" listeners pay more attention to virtual pitches. In the case of Shepard (or Deutsch) tones, the dominant virtual pitches coincide with partials (Terhardt, Stoll, Schermbach, & Parncutt, 1986). Their relative dominance is determined by a spectral weighting function that peaks around 700 Hz. This weighting function effectively gets convolved with the actual spectral envelope of the signal, although Terhardt's model incorporates thresholds below which changes in physical amplitude are assumed to have no perceptual consequences. Terhardt et al. (1986) asked listeners to match pure tones with the perceived virtual pitches of tones with octave-spaced partials; the resulting response distribution ranged from about 200 to 1000 Hz, with a peak around 300 Hz.[16] However, the complex tones in that study had a flat spectral envelope. The pronounced amplitude differences in stimuli with bell-shaped envelopes may well have a significant effect on pitch perception. The perceived pitch of such a tone is likely to be a joint function of the spectral envelope and of the listener's pitch weighting function.[17]

Given a smooth probability distribution of candidate virtual pitches, there must be one Shepard or Deutsch tone in a set of 12 whose two most salient virtual pitches (12 semitones apart) are equally strong, straddling the peak of the function. The tone whose partials are shifted by 6 semitones will then have a single most prominent virtual pitch near the peak of the function, and the pair formed by these two tones will be maximally ambiguous as to the direction of the pitch change. Pairs of other tones in between will be "unbalanced" in the sense that, in tritone pairs, they are perceived as changing pitch in a certain direction (i.e., their strongest pitches are on opposite sides of the maximum of the probability distribution). The occurrence of a pitch class effect is thus predicted by Terhardt's model (cf. Fig. 5 of Terhardt, 1991).

It is also easy to incorporate individual differences in the model by postulating individual variability in the spectral weighting function. Large

---

[16] These results were obtained by having listeners match pure tones to the perceived pitch of single Shepard tones. It would be worthwhile repeating this experiment, asking listeners to match pairs of pure tones to the perceived pitch change in tritone pairs. The author informally tried to match the tritone pitches by imitating them on a digital piano (i.e., with complex harmonic tones) while listening to portions of each test: Consistent with Terhardt's theory, he mapped the Deutsch-D♯ stimuli into the range $G\sharp_3$-A_4 (i.e., 208–440 Hz), and the Deutsch-A stimuli into the only slightly lower range $F\sharp_3$-G_4 (i.e., 185–392 Hz); Shepard tones, however, seemed to have much lower pitches, in the range A_1-B_2 (i.e., 55–123 Hz).

[17] A computational application of Terhardt's algorithm to the parameters of the present stimuli (Richard Parncutt, personal communication) has suggested that the effect of spectral envelope should be slight, in accord with Deutsch (1987). The large differences found in the present study await an explanation.

individual differences in the relative perceptual importance of the partials of complex harmonic tones have been reported by Moore, Glasberg, and Peters (1985). The origin of these differences is not yet clear, however. Terhardt (1991) speculates, like Deutsch et al. (1990), that individual differences may derive from differential exposure to voices, both one's own and those of others. Although he is not specific about the relevant voice characteristics, it should be noted that the spectral weighting function in his model peaks around 700 Hz, which is much higher than the F_0 of adult human speakers. F_0 may thus not be the relevant characteristic; rather, it may be the long-term speech spectrum in the region of the first formant. Very little is known at present about differences in long-term speech spectra among individuals or across languages, and any more specific hypothesis would be pure speculation. However, it should be noted that, once pitch is treated as a linear rather than a circular dimension, sex differences enter the picture, owing to the spectral consequences of women's smaller vocal tracts and higher voices. In the tritone task, however, sex differences are generally absent. This suggests that experience with the sound of one's own voice is not a likely factor.

Some incidental observations made in the course of the present study may provide clues to the nature of the pitch class effect. Appendix A presents some of them; others have been mentioned in footnotes. The reduced individual differences and stronger pitch class effects in the tests that used lower-pitched stimuli (Shepard and Deutsch-A tests) deserve attention. Most intriguing, perhaps, are some hints in the present data that the pitch class effect can be reversed between and within individuals. Although a direct test of strategy-based reversal within listeners failed (see Appendix A, Listening Strategies), there was one subject who spontaneously reversed his responses during the Deutsch-D♯ test (see footnote 13) and another who apparently did so in the last block of the Deutsch-A test. Some of the weak pitch class effects observed may have been due to more frequent reversals during a test, rather than to random responding. (See also Appendix A, Sequential Context Effects.) In the American group of subjects, there were several subjects who showed identical or very similar results on the two tests, although most other subjects showed nearly opposite pitch class effects. The pitch class effects of the American subjects thus seem to be bimodally distributed across the two tests, with the modes being about 6 semitones apart. One is led to wonder how stable the direction of these pitch class effects is. If future studies succeeded in demonstrating that they are reversible within the same listeners, the significance of the "highest" pitch class would be eroded; instead, the *most ambiguous* stimulus pairs would emerge as the common "hinges" of two diametrically opposed pitch class effects. This would be damaging to Deutsch's theory of a stable pitch template. However, there is no convincing evidence at present that individual pitch class effects are in fact reversible.

In summary, the present study has attempted to replicate some of Deutsch's findings and has failed in most respects, although a difference in tritone perception between British and American listeners was confirmed. Further research is needed to clarify the conflicting results and to provide new evidence bearing on Deutsch's provocative theory.[18]

References

Boë, L., & Rakotofiringa, H. A statistical analysis of laryngeal frequency: Its relationship to intensity level and duration. *Language & Speech*, 1975, *18*, 1–13.

Collier, R. Multi-language intonation synthesis. *Journal of Phonetics*, 1991, *19*, 61–73.

De Pijper, J. R. *Modelling British English intonation*. Dordrecht, The Netherlands: Foris, 1983.

Deutsch, D. A musical paradox. *Music Perception*, 1986, *6*, 115–132.

Deutsch, D. The tritone paradox: Effects of spectral variables. *Perception & Psychophysics*, 1987, *41*, 563–575.

Deutsch, D. The tritone paradox: An influence of language on music perception. *Music Perception*, 1991, *8*, 335–347.

Deutsch, D. The tritone paradox: Implications for the representation and communication of pitch structures. In M. R. Jones & S. Holleran (Eds.), *Cognitive bases of musical communication*. Washington, DC: American Psychological Association, 1992a, pp. 115–138.

Deutsch, D. Paradoxes of musical pitch. *Scientific American*, 1992b, *267* (2), 88–95.

Deutsch, D. Some new pitch paradoxes and their implications. *Philosophical Transactions of the Royal Society of London, Series B*, 1992c, *336*, 391–397.

Deutsch, D., Kuyper, W. L., & Fisher, Y. The tritone paradox: Its presence and form of distribution in a general population. *Music Perception*, 1987, *5*, 79–92.

Deutsch, D., North, T., & Ray, L. The tritone paradox: Correlate with the listener's vocal range for speech. *Music Perception*, 1990, *7*, 371–384.

Hanley, T. An analysis of vocal frequency and duration characteristics of selected samples of speech from three American dialect regions. *Speech Monographs*, 1951, *18*, 78–93.

Hanley, T. D., & Snidecor, J. C. Some acoustic similarities among languages. *Phonetica*, 1967, *17*, 141–148.

---

[18] This research was conducted in spring of 1993, while the author spent 3 months as a research fellow at the Institute for Perception Research (IPO) in Eindhoven, The Netherlands. The support of the Technical University Eindhoven and the hospitality of IPO during that period are gratefully acknowledged. Thanks are also due to Adrian Houtsma for providing the software for stimulus generation and the data in Appendix A (Effects of Presentation Level), to Bob Crowder for suggesting the method of constructing the context-balanced stimulus sequences, to Rob Meerding for translating the reading materials into Dutch, to Roel Smits for providing the Sennheiser earphones, to Chris Darwin for making it possible for the author to run subjects at the University of Sussex, to Pennie Smith for scheduling those subjects, to Chris Plack and Twan Aarts for technical help, and to all the colleagues at IPO and Haskins Laboratories who served as unpaid volunteer subjects and often provided useful comments. Thanks are further due to René Collier, Bob Crowder, Diana Deutsch, Bill Hartmann, Adrian Houtsma, Brian Moore, Richard Parncutt, Ani Patel, Jacques Terken, and Dix Ward for helpful comments on earlier versions of the manuscript, to Richard Parncutt for additional extensive discussions of this research, and to Bill Hartmann for diplomacy and advice.

Hanley, T. D., Snidecor, J. C., & Ringel, R. L. Some acoustic differences among languages. *Phonetica*, 1966, *14*, 97–107.

Hermes, D. J. Measurement of pitch by subharmonic summation. *Journal of the Acoustical Society of America, 1988, 83,* 257–264.

Hollien, H., & Jackson, B. Normative data on the speaking fundamental frequency characteristics of young adult males. *Journal of Phonetics*, 1973, *1*, 117–120.

Horii, Y. Some statistical characteristics of voice fundamental frequency. *Journal of Speech and Hearing Research*, 1975, *18*, 192–201.

Houtsma, A. J. M., & Fleuren, J. F. M. Analytic and synthetic pitch of two-tone complexes. *Journal of the Acoustical Society of America*, 1991, *90*, 1674–1676.

Houtsma, A. J. M., Rossing, T. D., & Wagenaars, W. M. *Auditory demonstrations CD.* Acoustical Society of America and Institute for Perceptual Research (IPO), Eindhoven, The Netherlands, 1987.

Hudson, A. I., & Holbrook, A. A study of the reading fundamental frequency of young black adults. *Journal of Speech & Hearing Research*, 1981, *24*, 197–201.

Hudson, A. I., & Holbrook, A. Fundamental frequency characteristics of young black adults: Spontaneous speaking and oral reading. *Journal of Speech and Hearing Research*, 1982, *25*, 25–28.

Jassem, W. Pitch and compass of the speaking voice. *Journal of the International Phonetic Association*, 1971, *1*, 59–68.

Jassem, W., & Kudela-Dobrogowska, K. Speaker-independent intonation curves. In L. R. Waugh & C. H. van Schooneveld (Eds.), *The melody of language.* Baltimore, MD: University Park Press, 1980, pp. 135–148.

Liberman, M., & Pierrehumbert, J. Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oehrle (Eds.), *Language sound and structure.* Cambridge, MA: MIT Press, 1984, pp. 157–233.

Maeda, S. *A characterization of American English intonation.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1976.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, 1985, *77*, 1853–1860.

Mysak, E. Pitch and duration characteristics of older males. *Journal of Speech & Hearing Research*, 1959, *2*, 46–54.

Ragozzine, F., & Deutsch, D. A regional difference within the United States in perception of the tritone paradox. *Journal of the Acoustical Society of America*, 1993, *94*, 1860. (abstract)

Shepard, R. N. Circularity in judgments of relative pitch. *Journal of the Acoustical Society of America*, 1964, *36*, 2346–2353.

Snidecor, J. A comparative study of the pitch and duration characteristics of impromptu speaking and oral reading. *Speech Monographs*, 1943, *10*, 50–56.

Terhardt, E. Music perception and sensory information acquisition: Relationships and low-level analogies. *Music Perception*, 1991, *8*, 217–240.

Terhardt, E., Stoll, G., Schermbach, R., & Parncutt, R. Tonhöhenmehrdeutigkeit, Tonverwandtschaft und Identifikation von Sukzessivintervallen. *Acustica*, 1986, *61*, 57–66.

Terhardt, E., Stoll, G., & Seewann, M. Algorithm for extraction of pitch and pitch salience from complex signals. *Journal of the Acoustical Society of America*, 1982a, *71*, 679–688.

Terhardt, E., Stoll, G., & Seewann, M. Pitch of complex signals according to virtual-pitch theory: Tests, examples, and predictions. *Journal of the Acoustical Society of America*, 1982b, *71*, 671–678.

Terken, J. Baselines revisited: Reply to Ladd. *Language & Speech*, 1993, *36*, 453–459.

't Hart, J., Collier, R., & Cohen, A. *A perceptual study of intonation: An experimental-phonetic approach to speech melody.* Cambridge, U.K.: Cambridge University Press, 1990.

Willems, N., Collier, R., & 't Hart, J. A synthesis scheme for British English intonation. *Journal of the Acoustical Society of America*, 1988, *84*, 1250–1261.

Yamazawa, H., & Hollien, H. Speaking fundamental frequency patterns of Japanese women. *Phonetica*, 1992, *49*, 128–140.

Appendix A
Additional Observations

SEQUENTIAL CONTEXT EFFECTS

Imagine a subject who shows a perfect pitch class effect: 6 adjacent tritone pairs always receive "down" responses, the other 6 (their inverses) always receive "up" responses. This pattern of results (which was closely approximated by some subjects in some tests) implies sequential dependencies of the following sort: When a tritone pair is preceded by another pair whose pitches are 1 semitone higher or lower, it will receive the same response as the preceding pair in 20 of 24 instances. This is so because there are only 4 possible sequences of such pairs that cross the two sharp boundaries between "down" and "up" responses; all other sequences are within response categories. By the same reasoning, tritone pairs separated by 2, 3, 4, and 5 semitones will receive the same response in 16, 12, 8, and 4 instances, respectively, out of 24. Finally, pairs separated by 6 semitones (i.e., each pair followed by its inverse), of which there are only 12 instances, will never receive the same response. These sequential effects may thus be a consequence of the pitch class effect. If so, they should be absent when the pitch class effect is weak or absent.

Nearly all subjects tested showed strong pitch class effects on at least some tests, and those instances where the pitch class effect was weak may have been due to difficulties with the stimuli or the task. There was one listener, however, who consistently refused to show a pitch class effect, at least with the Deutsch tones, while being highly confident of his responses–the author (BHR). He served in extensive pilot runs with various stimuli and did show a weak pitch class effect initially during these runs. However, with the final sets of Deutsch tones, listening under the same conditions as the experimental subjects, he repeatedly gave flat response functions.

Figure A1 shows his results from one listening session. The top panel shows that there was no clear pitch class effect in either test. The bottom panel, however, reveals strong sequential effects—even stronger than would be implied by a perfect pitch class effect (diagonal line): BHR usually gave the same response to the current tritone pair as to the preceding pair when the pitch separation was 1–3 semitones and different responses when the separation was 4–6 semitones. He hardly ever perceived any ambiguity in the direction of pitch change; yet he frequently reversed his responses to the same stimulus pairs. These results are proof, then, that sequential context effects can occur quite independently of the pitch class effect, and they lead one to wonder whether the latter might somehow depend on the former. The exact relation between the two phenomena is not clear at present.

EFFECTS OF PRESENTATION LEVEL

The lowest and highest partials of Deutsch tones are quite weak and probably contribute little to the virtual pitch percept. Still, it is conceivable that they have a disproportionate effect when they cross a threshold from inaudibility to audibility. The sudden appearance of low-frequency partials is actually quite striking as one listens to a regularly ascending sequence of Shepard or Deutsch tones, and this could be linked to the pitch class effect. If so, the pitch class effect should change with presentation level, since, with a fixed spectral envelope, presentation level must affect the pitch at which low-amplitude partials become audible.

Adrian Houtsma kindly provided the following data, which he collected near the end of the author's stay at IPO, using himself (AH-m) and an American colleague (SC-f) as subjects. (Both were also subjects in the present study; cf. Figure 4.) The test contained 10 repetitions of the 12 Deutsch-D♯ tritone pairs at each of five levels: 60, 65, 70, 75, and 80 dB SPL. The sequence was entirely random, and the subjects listened binaurally with Etymotic ER-2 insert earphones.

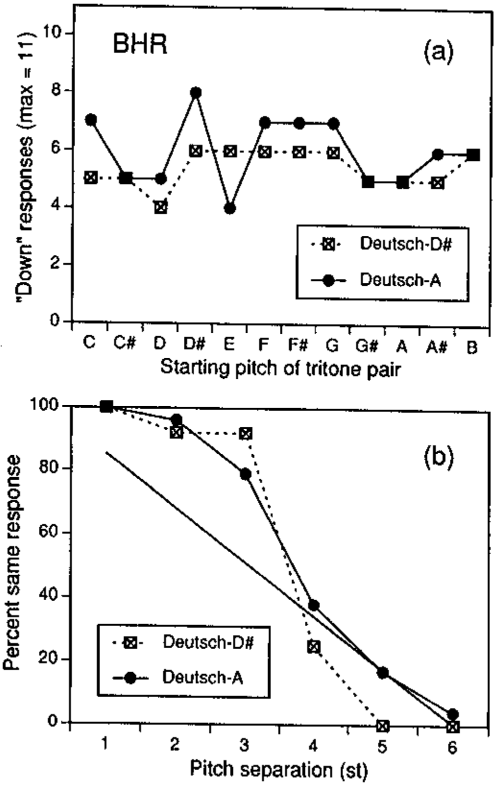

Fig. A1. Results of BHR for the two Deutsch tone tests, from a run conducted at IPO. (a) Response functions. (b) Percentage of identical responses to successive tritone pairs as a function of pitch separation. The diagonal line indicates the percentages implied by a perfect pitch class effect.

Figure A2 shows the results. Subject SC-f (top) showed a very pronounced pitch class effect, with A♯-B being the highest pitch classes; subject AH-m (below) had a less pronounced effect and noisier data, suggesting B-C as the highest pitch classes.[19] There is a suggestion in the data that the pitch class effect becomes weaker as presentation level decreases. Neither subject, however, showed any indication of a systematic shift in the response functions with presentation level. This suggests that low-amplitude partials do not play an important role in the pitch class effect and that presentation level is not a critical variable in experiments on the tritone paradox.

LISTENING STRATEGIES

Most subjects found the tritone tests easy and straightforward. A few, however, complained about ambiguity and claimed to hear occasionally simultaneous changes in oppo-

---

[19] As a regular subject in the Dutch group (77 dB presentation level), AH-m had A-A♯ as his highest pitch classes. Shifts of 1–2 semitones are hardly significant when the pitch class effect is relatively weak. American subject SC-f exactly replicated her earlier test results (cf. Figure 4).

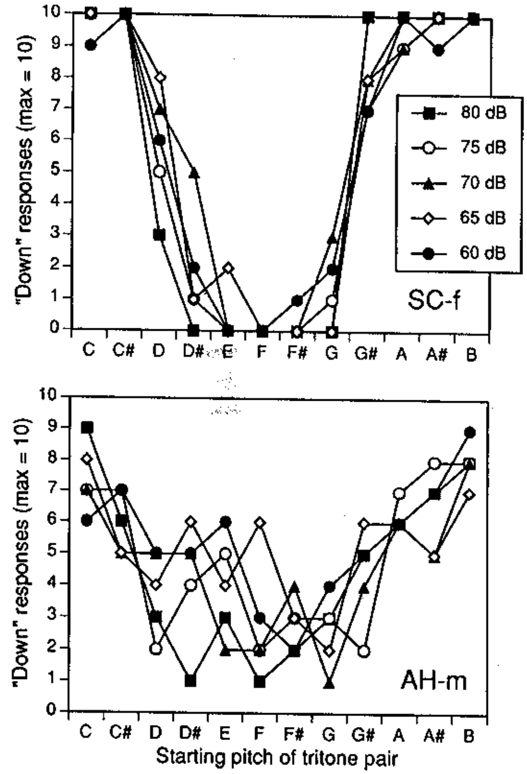

Fig. A2. Results of two subjects in a test that used five different presentation levels (Deutsch-D♯ stimuli). Data courtesy of Adrian Houtsma.

site directions in different frequency regions. Presumably, these subjects were "analytic listeners" (cf. Houtsma & Fleuren, 1991), whereas most others were "synthetic listeners" who perceived only a single dominant pitch and little ambiguity. Each of these analytic listeners nevertheless produced consistent pitch class effects, apparently by adopting a consistent strategy. Their comments suggested, however, that their results might have been different if they had adopted a different strategy. (See also footnote 13.) This possibility was checked out by recalling three Dutch subjects (DH-m, WR-m, AB-f) who had complained about the ambiguity of the tritone pairs. In this follow-up test, they were presented twice with the Deutsch-D♯ test and were asked to listen to the low frequencies the first time and to the high frequencies the second time.

The results of two subjects (DH-m and WR-m) are shown in Figure A3. To the author's surprise, their pitch class effects were not affected by the change in listening strategy. (The very small differences visible in the figure would require more extensive data to be considered significant.) Subject WR-m closely replicated his results of the original Deutsch-D♯ test run, whereas subject DH-m appeared to have shifted his pitch class effect down by 2 semitones. The third subject (AB-f) showed a pitch class effect with the low-frequency strategy which, too, was 2 semitones below her original results, but she gave only "up" responses when listening to high frequencies. (See also footnote 13.) Subsequent questioning revealed that she had been listening to very high frequencies, whereas the other two subjects listened somewhere in the middle region, near the strongest partials. The low-frequency listening strategy apparently focused on the lowest virtual pitches or audible partials of the tone complexes; DH-m called them "the bass notes."

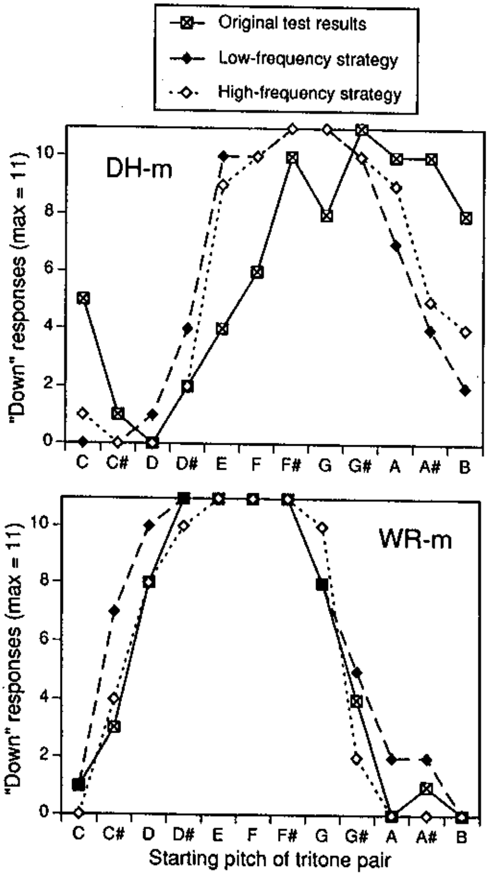

Fig. A3. Results of two subjects who used different listening strategies in the Deutsch-D# test. Their original test results are also shown.

Why, then, did these subjects claim to hear opposite pitch changes in different frequency regions? Presumably, these impressions derived from the ambiguous tritone pairs in the transition zone between "up" and "down" responses. If the responses to these pairs were reversed, this would not change the pitch class effect. The detailed response protocols of subject WR-m were examined to determine the number of response reversals. Between his low-frequency and high-frequency strategy runs, there were 22 reversals (out of 132 scored responses). Between the low-frequency run and the original test, there were also 22 reversals, but between the high-frequency run and the original test there were only 12 reversals. This suggests that WR-m used the high frequency strategy in the original run, which seems plausible. The finding that there were almost twice as many reversals with the low-frequency strategy may then be taken as evidence in support of the hypothesis that responses to ambiguous tritone pairs were affected by listening strategies. Subject DH-m, however, showed only 14 response reversals between his low- and high-frequency runs, and because of his shift in the pitch class effect, a count of reversals with respect to the original test results would not have provided a proper comparison. Therefore, these results remain suggestive at best.

Appendix B
Reading Materials

ENGLISH SENTENCES

1. I really felt sick yesterday, so I left work early and went home to lie down.
2. The three Baltic states gained their independence before the Soviet Union fell apart completely.
3. Foreigners visiting the Netherlands often don't bother to learn the Dutch language.
4. The concert last week was so boring that I dozed off during the performance.
5. After the prince married the princess, they lived happily ever after.
6. Among the endangered species in this world are elephants and tigers.
7. I attempted to call my friend three times, but the line was always busy and so I gave up.
8. Since I got myself a new bicycle, it is much more fun riding to work in the morning.
9. My mother was pleased to find that we had a large supply of paper towels in our cabinet.
10. Now that spring has arrived, the birds are singing and the trees are growing leaves again.

DUTCH TRANSLATIONS

1. Ik voelde mij zo beroerd gisteren, daarom ben ik vroeg naar huis gegaan en op bed gaan liggen.
2. De drie Baltische Staten verkregen hun onafhankelijkheid nog voor de Sovjet Unie volledig uit elkaar viel.
3. Buitenlanders die Nederland bezoeken, nemen vaak niet de moeite de Nederlandse Taal te leren.
4. Het concert van vorige week was zo slaapverwekkend dat ik wegdoezelde tijdens de voorstelling.
5. Nadat de Prins de Prinses had getrouwd, leefden zij nog lang en gelukkig.
6. Tot de bedreigde diersoorten in de wereld behoren de olifanten en tijgers.
7. Ik heb drie keer geprobeerd mijn vriend te bellen, maar hij was steeds in gesprek en dus heb ik het opgegeven.
8. Sinds ik een nieuwe fiets heb, is het veel leuker om 's ochtends naar mijn werk te gaan.
9. Mijn moeder was blij te zien dat wij een grote stapel papieren handdoeken in het keukenkastje bewaarden.
10. Nu de lente eindelijk in het land is, zingen de vogels en lopen de bomen weer uit.