Perception & Psychophysics
2002, 64 (5), 844-854

Schema-based processing in auditory scene analysis

CAROLINE BEY and STEPHEN McADAMS
Laboratoire de Psychologie Expérimentale (CNRS), Université René Descartes, Paris, France and IRCAM-CNRS, Paris, France

What is the involvement of what we know in what we perceive? In this article, the contribution of melodic schema-based processes to the perceptual organization of tone sequences is examined. Two unfamiliar six-tone melodies, one of which was interleaved with distractor tones, were presented successively to listeners who were required to decide whether the melodies were identical or different. In one condition, the comparison melody was presented after the mixed sequence: a target melody interleaved with distractor tones. In another condition, it was presented beforehand, so that the listeners had precise knowledge about the melody to be extracted from the mixture. In the latter condition, recognition performance was better and a bias toward *same* responses was reduced, as compared with the former condition. A third condition, in which the comparison melody presented beforehand was transposed up in frequency, revealed that whereas the performance improvement was explained in part by absolute pitch or frequency priming, relative pitch representation (interval and/or contour structure) may also have played a role. Differences in performance as a function of mean frequency separation between target and distractor sequences, when listeners did or did not have prior knowledge about the target melody, argue for a functional distinction between primitive and schema-based processes in auditory scene analysis.

The ability to segregate sounds produced by distinct sources in the environment is essential for humans (Bregman, 1990, 1993) and other species for whom auditory information is relevant (see Hulse, MacDougall-Shackleton, & Wisniewski, 1997; MacDougall-Shackleton, Hulse, Gentner, & White, 1998). This general skill allows us to follow speech in a noisy environment (Cherry, 1953) and to isolate melodic voices in polyphonic music. What are the processes underlying this auditory scene analysis?

According to Bregman (1990), scene analysis is governed by two different mechanisms, which he refers to as *primitive* and *schema based.* The complex signal arising from various acoustic sources that reaches our ears is decomposed in a preattentive way into independent perceptual entities, called *auditory streams*, which generally correspond to the different sources of the environment. This primitive auditory scene analysis is a sensory partitioning mechanism. To construct streams, the auditory system uses regularly occurring acoustic cues, such as the harmonicity of many relevant sounds of our environment, the asynchrony of independent sources, and smooth change over time of sound properties coming from the same source.

One main argument supporting the hypothesis that auditory stream formation involves preattentive processes is the fact that segregation may occur against listeners' intentions. Indeed, van Noorden (1975) has shown that even if listeners tried to hold together a sequence composed of two tones differing in frequency, there was a frequency/time limit called the *temporal coherence boundary*, above which the sequence split obligatorily into two streams. Other evidence has recently been added by Sussman, Ritter, and Vaughan (1999) in a study of event-related brain potentials. The authors presented two ascending tone triads interleaved in frequency while participants were reading a book in an *inattentive* condition. Tempo was varied so that the sequence was perceived as one stream when it was slow and as two streams when it was fast. In some trials, one ascending triad was replaced by a deviant stimulus forming a descending pattern. A component called *mismatch negativity* (MMN), associated with the automatic detection of stimulus changes, was recorded when the deviant stimulus was present only in the fast-tempo condition. This result suggests that despite their inattentive listening, the participants organized the sequence into two streams. Finally, another reason to think that the primitive process is an unlearned mechanism is given by the results of different studies showing that the ability to organize sensory information into streams is innate and adaptive. Demany (1982) and, more recently, McAdams and Bertoncini (1997) have shown that this skill is apparently present very early in life. MacDougall-Shackleton et al. (1998), as well as Hulse et al. (1997), have shown that other

Preliminary results of the first experiment were presented at the 39th Annual Meeting of the Psychonomic Society (Bey & McAdams, 1998). The authors thank Albert Bregman for enlightening discussions on the issues in this paper and Sandrine Vieillard for help in collecting the data.
Correspondence concerning this article should be addressed to S. McAdams, IRCAM-CNRS, 1 place Igor Stravinsky, F-75004 Paris, France (e-mail: smc@ircam.fr).

species, such as birds—in particular, European starlings—also have this ability.

However, this bottom-up process is not the only one that allows us to access information in a sound mixture. Acquired knowledge about sounds and sound sequences such as music and speech can help us to extract information from a complex scene. This schema-based analysis (a top-down process) is a selection process in Bregman's (1990, chap. 4) conception. The extraction would then be the result of a matching process between the activated knowledge stored in memory and the sensory representation of the incoming signal. This process may be attentive, as is the case when we explicitly try to hear a sound source or a sound sequence in a background mixture. Nevertheless, it can also be preattentive, as occurs, for example, in the common experience of being in a room with many people talking and hearing one's name emerge from the mixture, often erroneously. The auditory representation of one's name would be activated by the sensory representation of the mixed sounds.

Few studies have been conducted on the contribution of top-down processes in auditory streaming (Bregman, 1990, chap. 8). Van Noorden (1975) reported experimental evidence of the involvement of these processes, showing that the frequency difference inducing perceptual fission in a sequence in which high- and low-frequency tones alternate every 150 msec changes depending on what listeners try to hear. If they try to segregate the two sounds, the sequence can be split perceptually down to a difference of 2–3 semitones (STs; 1 ST = 6% difference in frequency), the so-called fission boundary. On the other hand, when they try to perceive the sequence as integrated and the frequency difference is increased, a temporal coherence boundary at about 12 STs is found at this tempo. Dowling (1973; Dowling, Lung, & Herrbold, 1987) found that listeners trained in a task involving detection of a familiar melody interleaved with distractor tones in the same pitch range succeeded if its title was given beforehand. Furthermore, electrophysiological studies have revealed the involvement of attentional components in auditory streaming (Alain & Woods, 1994; Sussman, Ritter, & Vaughan, 1998).

Our general aim in this paper is to learn more about the role and the nature of this schema-based analysis proposed by Bregman (1990, chap. 8). We examined the contribution of knowledge to the perceptual organization of successive sounds by studying the ability to recognize a melody interleaved with distractor tones in three experimental tasks performed by different groups of participants. These tasks were designed so that comparisons among them, as a function of the frequency separation between target melody and interleaved distractor sequence, would shed light on the role of previous knowledge of the to-be-detected target melody and on whether this knowledge was related to the absolute pitches of the melody or the relative pitch relations among notes of the melody.

METHOD

Stimuli

The melody and distractor sequences used in this experiment were those constructed in a previous study (Bey & McAdams, 2002). Thirty-six melodies and 180 distractor sequences, each composed of six notes, were created. The intervals were fixed, but the mean frequency of the sequences varied from trial to trial over a range from −3 to +2 STs.

Each of the 36 melodies had an original and a modified version (Appendix). For the latter version, two notes, the second and fourth or the third and fifth, were changed within a range of ±4 STs. In all cases, the note changes altered the original melodic contour—that is to say, the direction of pitch change between successive notes. This feature is a salient cue for immediate unfamiliar diatonic and nondiatonic melody recognition (Dowling, 1978; Dowling & Fujitani, 1971). These 72 melodies (36 original and 36 modified versions) were composed, for the most part, of ascending and descending pitch intervals, the size of which varied from 1 to 8 STs, but 7 of the modified melodies had repeated notes. All the melodies were played within a one-octave range, their pitch ranges varying from 5 to 11 STs. The mean note was A5 (880 Hz, MIDI note 81). Over the total set of 72 different melodies, 46 were diatonic, and 26 were nondiatonic. Diatonicity refers to the conformity of a melody to a diatonic scale, which corresponds to a specific pattern of intertone intervals in STs (e.g., the interval sequence 2 2 1 2 2 2 1 corresponds to a major scale). Note that the 46 diatonic melodies were not necessarily played in the same key and that the strength of tonality (the sense of having a tonic reference pitch) varied across them as well. These two factors, key and tonality strength induced by a melody, were not studied systematically in this experiment.

Five different distractor sequences were constructed for each of the 36 melody pairs (original and modified). They were all nondiatonic sequences constructed with two constraints: (1) The notes alternated from above to below the frequencies of the target melodies in order to create maximum crossover to camouflage the target (Hartmann & Johnson, 1991); (2) the total range of the distractor sequence exceeded that of the targets at both upper and lower ends when the interleaved sequences were presented at the same mean frequency, with distractors being maximally distant from the two neighboring melody tones by 2 STs.

Melodies and distractor sequences were composed of pure tones of 110-msec duration. The interonset interval was 330 msec for the comparison melody and 165 msec for the mixed sequence composed of 12 tones: 6 target melody tones interleaved with 6 distractor tones. The first note of the mixed sequence was always a target tone. The mixture and comparison sequences were separated by a 1,870-msec silent interval.

Procedure

Two unfamiliar six-tone melodies, one of which was interleaved with distractor tones, were presented successively to listeners, who were required to decide whether the melodies were identical or different. Three experimental tasks were presented to different groups of participants. For the first condition, the comparison melody was presented *after* the interleaved sequence (Figure 1, *After* condition). In this postrecognition task of interleaved melodies developed in a previous study (Bey & McAdams, 2002), listeners did not have precise knowledge about the melody they needed to extract from the composite sequence, so that it would, for the most part, involve primitive auditory scene analysis. For the second condition, the melody to be recognized was presented *before* the sequence to be organized, as with previous studies (Dowling, 1973, Experiment 2; Dowling et al., 1987; Vliegen & Oxenham, 1999), so that listeners had precise knowledge concerning the melody to be extracted from the mixture (Figure 1, *Before* condition). On the basis of the results from these two conditions, we subsequently elaborated a third condition that would test the nature of the schema(s) used by listeners to extract the melody from the mixture. Could they use the absolute pitches of the melody notes, the relations between successive notes (contour and/or intervals), or both? To address this issue, another group of listeners participated in a task in which the comparison melody presented *before* the mixture sequence was randomly transposed upward (*TransBefore*

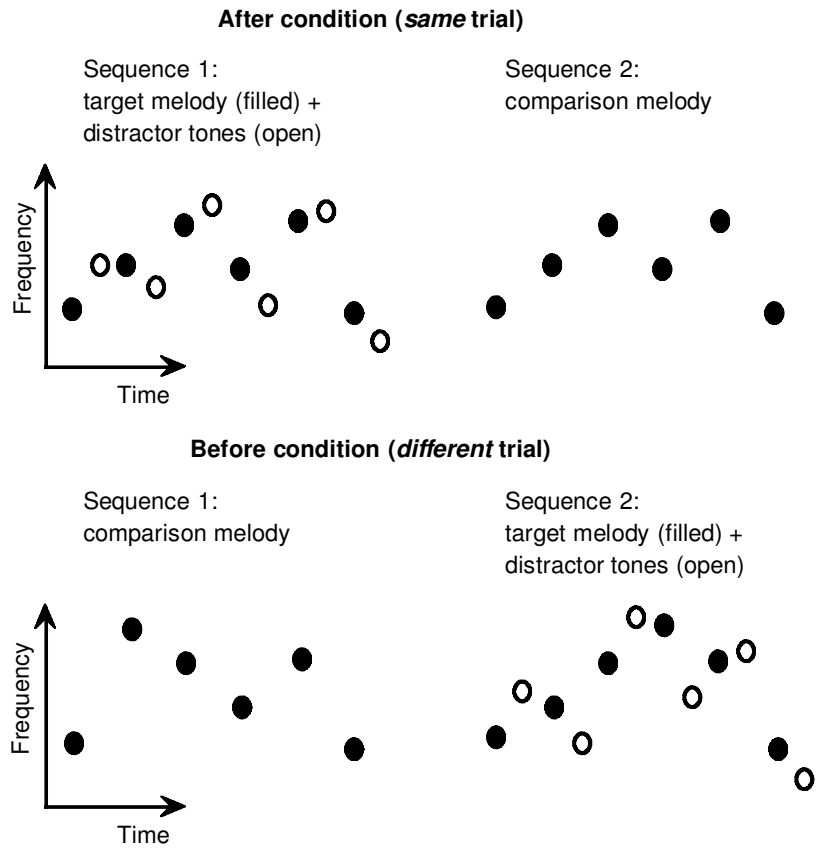

Figure 1. Visual illustration of two of the experimental tasks. In the *After* task, the melody to be compared is presented after the target melody interleaved with distractor tones. In the *Before* task, the melody is presented beforehand. The third task (*TransBefore*) is similar to the before task, except that the first melody is transposed upward.

condition). The amount of transposition exceeded the maximal range of the melodies (that is to say, greater than 11 STs) to ensure that there was no frequency overlap between the melodies to be compared, thus avoiding any possible frequency priming. The amount of transposition was randomly chosen on each trial from among +12, +13, or +14 STs. This trial-to-trial variation was designed to keep listeners from developing a possible (although unlikely) predictive transposition strategy consisting of imagining the notes of the second melody from those of the first one and so preactivating a pitch representation of the notes of the target melody. Also, the melody was shifted up in order not to prime the frequency region of the distractor sequence, which was to be ignored in this task. Note that since exact transposition was used, priming of *relations* between notes (intervals and/or contours) would still be useful. In this TransBefore condition, the melodies were considered *identical* if the sequence of pitch *intervals* between successive notes was identical. Finally, a control condition consisting of a simple melody recognition task without distractor tones was also presented to the participants in order to verify their basic melody recognition ability.

For the three task groups, one target melody and one of the five corresponding distractor sequences were chosen randomly on each trial. The distractor sequence was presented in the same frequency range as the target melody (0-ST mean frequency difference between the melody and the distractor tones) or was transposed toward lower frequencies at a mean frequency difference of 1, 2, 3, 4, 6, 8, 12, or 24 STs. Therefore, the target melody was always presented in the same frequency range; only the mean frequency of the distractor sequence varied. Previous research on the After condition alone had demonstrated that performance was at chance for a 0-ST separation and improved gradually with increasing mean frequency separation between the target and the distractor, reaching an asymptote somewhere between 12 and 24 STs (Bey & McAdams, 2002).

For each of these nine conditions of frequency separation, 24 trials were presented. For half of the trials, the melodies were identical: Two original versions were presented for six trials, and two modified versions were presented for the other six. For the other half, the melodies differed by two notes: The original version followed by the modified one was presented for six trials, and the reverse order was presented for the other six. The nine mean frequency separations between the melody and the distractor sequence and the four trial structures were presented in random order.

The session was composed of one experimental condition (After, Before, or TransBefore, depending on the group of participants) of 216 trials (9 frequency separations × 24 trials), followed by a control condition of 24 trials (simple melody recognition without distractor tones). The trial structure of the Control task was identical to that of the experimental tasks, except that no distractor sequences were presented. For listeners who did the TransBefore condition, an additional Control task was performed, in which the first melody was also shifted up by 12, 13, or 14 STs. The presentation order of the two Control conditions was counterbalanced across participants. The experimental and the control conditions were preceded by familiarization trials, with *same* and *different* trials presented alternately. During the familiarization, feedback concerning the correct response

was provided to the listeners. Ten trials were presented in the experimental task in order of increasing difficulty: two trials in the 24-ST separation and one for each degree of frequency separation in decreasing order. Four familiarization trials were presented in the Control condition. The total duration of the experiment was about 1 h.

The participants were seated in a sound-treated room. They were asked to judge whether the two successive melodies were the same or different. For the TransBefore condition, it was specified that melodies could be the same despite being presented in a different pitch range, as if a man and a woman were singing the same tune in different registers. Responses were made by pressing one of two keys on a computer keyboard, "m" if the target and comparison melodies were identical (*même* in French) and "d" if they were different (*différent* in French). The next trial followed automatically after the response was entered.

Apparatus

The pure tones were synthesized on a Yamaha TX802 FM Tone Generator and were presented diotically at a comfortable level (approximately 76 dBA) over Sennheiser HD 520 II headphones connected directly to the output of the synthesizer. The synthesizer was controlled by a Macintosh SE/30 computer via a Musical Instrument Digital Interface (MIDI). The programs controlling the experiments were written in LISP.

Participants

Seventy-four listeners took part in the experiment. They all reported normal hearing and were paid for their participation. Fourteen participants did not succeed in the experimental task—that is, when the distractor sequence was present (5 in the After, 4 in the Before, and 5 in the TransBefore conditions). They were all nonmusicians except one, who had taken singing lessons in the past. Their correct recognition rates averaged across all frequency separation conditions was .55. Nevertheless, they were able to perform the melody discrimination task in the Control condition without distractor tones: The mean correct recognition rate was .89 when the first melody was not transposed up in frequency and .74 when it was transposed. Further research will be needed to examine why almost 20% of the participants could not succeed in performing the interleaved melody recognition task. A proportion similar to this was found in our previous study (Bey & McAdams, 2002). In the present study, the purpose was to compare the performance obtained in three types of interleaved melody recognition tasks, so the results of these 14 participants were excluded from the analysis.[1]

The results of 60 listeners were thus included in the analysis, 20 in each of the After, Before, and TransBefore conditions. The three groups were relatively homogeneous with respect to age, sex, and musical background. The mean age of the participants was 27 years (range, 19–37 years) in the After condition, 27 years (range, 19–40 years) in the Before condition, and 26 years (range, 21–35 years) in the TransBefore condition. Six women and 14 men performed the After task, 5 women and 15 men the Before task, and 6 women and 14 men the TransBefore task. The numbers of musicians and nonmusicians were equal for each group. Ten had received formal musical training and had been playing a musical instrument for at least 4 years. The other 10 did not have any musical background and did not play an instrument.

RESULTS

Hit rates (proportion of *different* responses when the target and the comparison melodies were different) and false alarm rates (proportion of *different* responses when the two melodies were identical) were computed across melodies for each participant in each stimulus condition. From these values, sensitivity (d') and decision criterion (c_{sd}) indices were determined using a same/different paradigm in a differencing model (Macmillan & Creelman, 1991). To avoid infinite values and given that there were only 12 *same* and 12 *different* trials, we considered a .96 hit rate (11.5/12) and a .04 false alarm rate (0.5/12) as extreme values for both indices (Macmillan & Creelman, 1991, p. 10), giving a maximum d' value of 5.38 for this model. The results for these two indices will be examined separately.

A Sensitivity Difference

A mixed analysis of variance (ANOVA) was performed on d', with repeated measures on mean frequency separation (nine levels: 0–24 STs) and with task (After, Before, and TransBefore) and musical training (musicans, nonmusicians) as between-subjects variables. In this and all subsequent ANOVAs, the Greenhouse–Geisser (1959) correction was applied to compensate for covariance owing to repeated measures. F statistics are cited with uncorrected degrees of freedom. If epsilon (e) is less than one, its value is cited, and the probability is determined with the corrected degrees of freedom.

Figure 2 presents the mean d' values for each task condition at each mean frequency separation, as well as for the control condition. The data are averaged across musical training groups, since the musical training factor did not interact with any other factor, although, on average, the musicians had slightly higher recognition rates than did the nonmusicians [$F(1,54) = 12.0, p < .005$]. In general, sensitivity increased with mean frequency separation in each task condition [$F(8,432) = 98.1, e = 0.74, p < .0001$], but the effect of frequency separation on performance depended on the task condition [$F(16,432) = 3.4, p < .0001$]. For the After condition, there appeared to be plateaus for a range of frequency separations. For the majority of the frequency separations, the means for the task conditions increased from After to TransBefore to Before. For the Control condition, the After and Before conditions were similar, and the TransBefore condition was lower.

Recognition performance obtained in the three tasks was significantly different on average [$F(2,54) = 10.6, p < .0001$]. Tukey's HSD post hoc comparisons confirmed that performance obtained in the Before task was much higher than that in both the After and the TransBefore tasks ($p < .05$), but the global means for the After and the TransBefore tasks were not significantly different. The interleaved melody recognition performance was thus higher when listeners previously heard the melody they had to extract from the mixture. However, this improvement was reduced when the first melody presented was randomly shifted up by 12, 13, or 14 STs (TransBefore condition).[2]

Although the target melody was segregated from the distractor tones at the 24-ST separation, performance remained lower than that obtained without distractor tones (Control condition) for the After task [$t(19) = 3.68, p < .001$]. This result was also found in our previous study (Bey & McAdams, 2002) and suggests that the simultaneously presented distractor tones interfered in the recog-

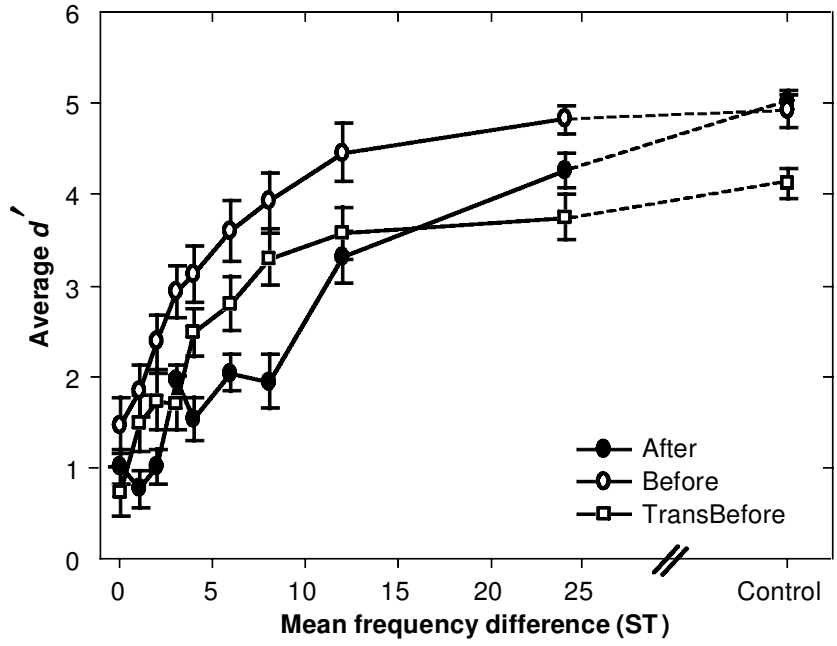

Figure 2. Average d' as a function of the mean frequency difference between target melody and distractor sequence for the After (filled circles), Before (open circles), and TransBefore (open squares) tasks. Each group was composed of 20 participants: 10 musicians and 10 nonmusicians. Performance for the control condition with no distractor is also plotted for comparison. For the TransBefore control condition, the comparison melody was transposed. Vertical lines represent ±1 standard error of the mean. ST, semitone.

nition of the target melody, even if they were perceptually segregated from it. However, we did not find this interference effect in the Before and TransBefore tasks ($p > .20$), indicating that attentional and/or mnemonic processes involved in these two tasks might have been different from those in the After task.

The *t* tests performed on the three task–group pairs revealed that performance obtained in the Control condition, in which the two melodies were presented in the same frequency range and without distractor tones, was equivalent for unhindered melody discrimination in these three independent task groups ($p > .20$). However, the listeners who performed the TransBefore task were less accurate in the transposed Control condition, in which the first melody was shifted up in frequency, than in the nontransposed Control condition [$t(19) = 7.4, p < .0001$].

Two apparent plateaus (0–2 STs and 3–8 STs) can be observed in the After psychometric function that are not present in the Before and Transbefore functions. Selected *t* tests with Bonferroni correction confirm this result. Compare the lowest and highest means within each of these groups of frequency separations for each task: 1 ST versus 2 STs for After [$t(19) = -1.1, p = .29$] and 0 ST versus 2 STs for Before [$t(19) = -2.96, p = .008$] and TransBefore [$t(19) = -3.22, p = .005$] in the 0–2 ST group, and 4 STs versus 6 STs for After [$t(19) = -2.23, p = .038$] and 3 STs versus 8 STs for Before [$t(19) = -3.29, p = .004$] and TransBefore [$t(19) = -6.16, p < .0001$] in the 3–8 ST group. For both regions, the means were not significantly different (corrected for six tests, with $\alpha = .0083$) in the After condition, but they were in the Before and TransBefore conditions. That these plateaus were not simply due to sampling error is supported by the fact that both of them existed for the mean data in both the musician and the nonmusician groups. Furthermore, there was no hint of an interaction between the group factor and mean frequency separation in a mixed ANOVA performed only on the data for the After condition ($F < 1$).

Fisher's protected LSD[3] was computed to perform specific design-related comparisons among task condition means averaged over the musical training factor at each frequency separation. The d' values for the After and Before conditions at each mean frequency separation were significantly different ($p < .01$) for separations of 1–12 STs. The nonsignificant difference found for the 24-ST separation ($p = .14$) can be explained by a performance ceiling effect. There was also no significant difference between means for the 0-ST separation ($p > .20$), in which performance in both tasks was roughly equivalent and slightly higher than chance. This latter result suggests that previous knowledge did not help to extract the melody when no sensory partitioning was possible. The Before and the TransBefore conditions were significantly different at separations of 3, 6, 12, and 24 STs (all $ps < .05$). Only 4–8 ST

separations had significant differences ($p < .05$) between the TransBefore and the After conditions.

A Different Decision Criterion

A mixed ANOVA was performed on c_{sd}, with repeated measures on mean frequency separation, (nine levels: 0–24 STs) and with task (After, Before, and TransBefore) and musical training (musicans, nonmusicians) as between-subjects variables.

Figure 3 presents the mean c_{sd} values for each task condition at each mean frequency separation, averaged across participants. Note that a c_{sd} of 0 means that the participants made omissions and false alarms in the same proportion. Positive values reflect a preponderance of *same* responses, and negative values reflect a preponderance of *different* responses.

The participants' decisions were globally biased toward *same* responses in all tasks, except for the Control task without transposition. However, the decision criterion was significatively different for the three tasks [$F(2,54) = 17.3$, $p < .0001$]. Tukey HSDs revealed that the listeners were more biased in the After condition than in both the Before and the TransBefore conditions ($p < .01$), but the difference between the Before and the TransBefore conditions did not reach significance.

When frequency separation between target melody and distractor sequence increased, mean c_{sd} tended generally to increase [$F(8,432) = 3.3$, $e = 0.59$, $p < .01$]—that is, errors tended to be omissions, rather than false alarms. Furthermore, the difference in strategy developed in the After condition, as compared with the Before and TransBefore conditions, depended on the frequency separation between melody and distractor tones [$F(16,432) = 4.1$, $p < .0001$]. The Before and TransBefore groups were less biased than the After group for small frequency separations. However, for separations higher than 6 STs, the mean c_{sd} of the TransBefore group increased, so that the TransBefore group became as biased as the After group toward *same* responses, whereas the Before group remained less biased. The Fisher protected LSD comparisons among task conditions at each frequency separation confirmed this result. The c_{sd} means of the After group were significantly different from those of the Before group for all the separations, whereas they were different from the TransBefore group only for 0- to 6-ST separations ($p < .05$). Moreover, the difference between the Before and the TransBefore conditions reached significance only for the separations higher than 6 STs—that is, for 8–24 STs ($p < .05$).

Musicians were slightly less biased than nonmusicians [$F(1,54) = 5.6$, $p < .05$], but the musical training factor did not interact with any other factor.

A Posteriori Analysis of the Effects of Diatonicity

Some of the melodies were diatonic, and others were nondiatonic. A diatonic scale is a type of musical structure that is widely used in Western culture (as well as in many other cultures that employ seven-note scales). Western listeners are more familiar with diatonic melodies than with nondiatonic ones. Therefore, we can assume that schemas are stronger for diatonic melodies, perhaps making them easier to recognize and segregate from a mixture than are nondiatonic melodies. An a posteriori analysis was conducted on data for the three task groups to test this hypothesis concerning the recognition of a melody presented in isolation (control condition) or interleaved with distractor tones (experimental conditions). Four trial structures had been presented to the participants: The target and comparison melodies were both diatonic (DD), both were nondiatonic (NN), or one was nondiatonic and the other diatonic in one of two presentation orders (ND and DN). Since *same* and *different* trials are to be separated here, d' cannot be computed. We therefore computed the proportion of correct responses for each trial structure in each task group (After, Before, and TransBefore) for experimental conditions across 0–24 STs separations (since there were not enough observations to compute this score per separation), as well as for the Control condition. Note that these are group scores, so the results described below will remain descriptive.

Performance was better for DD than for NN trials to approximately the same degree for the three experimental task groups (After, DD = .55, NN = .31; Before, DD = .81, NN = .60; TransBefore, DD = .72, NN = .49). However, no diatonicity effect was found in the Control condition (DD = .99, NN = 1.00, ND = .98, DN = .96). Therefore, the superiority of diatonic melody discrimination found in the three experimental tasks cannot be explained in terms of greater memorization of diatonic melodies than of nondiatonic ones. Another alternative is that the difference in diatonicity would be a cue to segregate the target melody from the distractor tones. Indeed, distractors are nondiatonic sequences, so diatonic melodies might be more easily segregated from this nondiatonic background than nondiatonic melodies would be. This hypothesis is weakly supported by the results obtained for the two other trial structures. The DN configuration was slightly better discriminated than ND in the After group (DN = .60, ND = .53), whereas the reverse was observed for the Before (DN = .79, ND = .82) and, especially, the TransBefore (DN = .67, ND = .75) groups. Note, however, that the effect of this asymmetry in the *different* trials was only one fourth the size of the effect on the *same* trials.

DISCUSSION

This study has shown that the ability to recognize interleaved melodies and the decision criteria adopted for doing these tasks were different, depending on whether the listeners did or did not know the melody before hearing the composite sequence. Recognition performance was higher when they had precise knowledge of the target melody to listen for in the mixture (Before condition), as compared with a situation in which they got only general knowledge across trials, such as the pitch range of the melody and its position with respect to the distractor along the pitch dimension (After condition). This improvement decreased

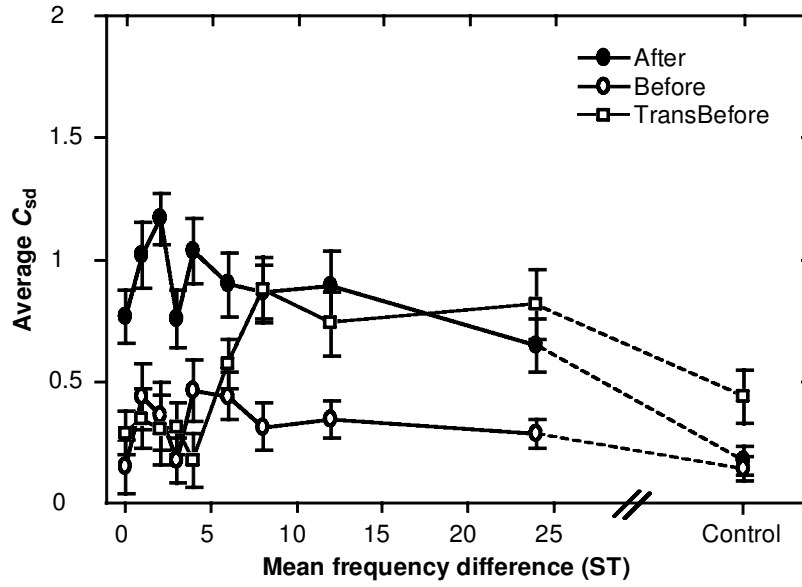

Figure 3. Average c_{sd} as a function of the mean frequency difference between target melody and distractor sequence for the after (filled circles), before (open circles), and transbefore (open squares) task groups. Average c_{sd} obtained in the control condition with no distractor is also plotted for comparison. Vertical lines represent ±1 standard error of the mean. ST, semitone.

when the melody presented beforehand was transposed up in frequency (TransBefore condition), suggesting that it was partly due to frequency or pitch priming (these two alternatives not being distinguishable, since pure-tone signals were used). However, the listeners may also have been able to use other knowledge, such as contour and/or intervals, since performance in the TransBefore condition tended to remain higher, on average, than that obtained in the After condition, although this difference did not attain statistical significance for the majority of frequency separations.

Two explanations might be advanced to explain the sensitivity differences between the After condition and the Before and TransBefore conditions. First, the differences between these two types of tasks might have resulted from the different kinds of knowledge that were available for both groups. Indeed, precise knowledge of the melody could have helped the listeners to extract it from the mixture through the involvement of schema-based analyses (Bregman, 1990, chap. 4). This hypothesis is supported by the result that performance was lower in the TransBefore condition than in the Before condition, suggesting that the schemas used by the listeners involved the tone frequencies or absolute pitches. However, another possible explanation, suggested by A. S. Bregman (personal communication, January 1997), is that differences between the two tasks would also result from the involvement of different memory processes. Indeed, this design required the listeners to compare two successive melodies. So, the first melody presented could have been coded in short-term memory, to be compared with the second one. In the After task, the first melody was presented interleaved with distractor tones and so was coded in memory *with* these distractor tones, whereas in the Before task, it was presented alone. The presence of distractor tones could have interfered with melody encoding in memory and so could have impaired its representation (Deutsch, 1970; Dowling, Kwak, & Andrews, 1995). Thus, poorer performance in the After condition than in the Before condition could be explained also by the interference in memory caused by the distractor tones. One result that could support this hypothesis is that performance obtained for a mean frequency separation between the melody and the distractor tones of 24 STs was lower than that obtained in the Control condition for the after task, whereas it was equivalent to that obtained for the Control condition for the Before and TransBefore tasks.

Performance increased with increases in the mean frequency difference between the melody and the distractor tones for the three tasks. This result is consistent with other studies showing that recognition of interleaved melodies depends on the perceptual organization of the composite sequence (Bey & McAdams, 2002; Dowling, 1973; Hartmann & Johnson, 1991). However, the form of the performance functions depended on the task. We observed plateaus in the After function that were not present in the Before and Transbefore conditions. One possible explanation for this difference is that the two types of psychometric functions reflect the involvement of different processes. This would support Bregman's (1990, chap. 4) theory, which postulates the existence of two different processes, a partitioning (bottom-up) process and a selection (top-down) process. Informal listening to all of the sequences presented to the participants led us to hypothe-

size that the observed plateaus in the After psychometric function could be related to the segregation of only some of the target tones embedded in the mixture sequence. Indeed, the plateaus were found when a constant number of distractor tones segregated from the mixture over a small range of mean frequency separations. A sudden change of performance seems to correspond to the segregation of an additional sound from the mixture sequence, allowing relations between successive melody notes to emerge further. If the plateaus observed in the After curve were effectively due to the number of sounds that segregated, it would suggest that this pattern of performance directly reflects the involvement of a partitioning process. Therefore, the absence of plateaus in the Before condition could have been due to the involvement of an additional process that depended less on the partitioning processes and could select events on the basis of previous knowledge. Informal testing carried out with 5 participants, consisting of counting the number of sounds that segregated from the mixture, did not allow us to confirm clearly this assumption, and further research will be needed to examine this detail of the data more closely.

The improvement in the Before condition, as compared with the After condition, was observed for all the degrees of separation except for 24 STs (owing to a ceiling effect) and 0 ST. This result for the latter separation suggests that previous knowledge of the melody did not help to extract the melody from the mixture if there was no difference in mean frequency and, thus, no segregation on the basis of sensory cues. Contrary to this finding, Dowling (1973, Experiment 3) showed that interleaved familiar nursery rhyme melodies presented in the same frequency range *could* be recognized after an average of 3.6 presentations, if preceded by a congruent verbal prime (the title of the melody). This result has been replicated in another study with participants who were less well trained and with only one presentation of the composite sequence (Dowling et al., 1987, Experiment 1).

Two hypotheses can be advanced concerning these apparently divergent results. First, the ability to extract a melody from a mixture without primitive segregation could depend on the nature of the activated schema. Dowling and his collaborators (Dowling, 1973; Dowling et al., 1987) used nursery rhyme melodies stored in long-term memory, whereas in our study, melodies were unfamiliar and, thus, were coded only in short-term memory. Furthermore, Dowling and his co-authors found in other studies that different melody features were coded according to the time listeners had to memorize the melody (Dowling, 1978; Dowling & Fujitani, 1971; Dowling & Harwood, 1986). They stored contour in short-term memory and intervals in long-term memory. These different features could have an effect on the power of the top-down process. The second hypothesis would hold that the results obtained by Dowling would not be explained by the involvement of top-down processes in auditory streaming. Indeed, the familiar interleaved melodies used by the author may have had local frequency differences despite the fact that they were presented in the same pitch range. These local differences could have induced partial segregation of the melody by a primitive process that could have been sufficient, in many cases, to recognize it.

The presentation of the melody before the composite sequence changed not only the listeners' sensitivity, but also their response strategy. For small frequency separations, the participants who heard the melody beforehand (Before and TransBefore tasks) tended to respond that the melodies were different more often than did listeners who did not know the melody to be extracted from the mixture (After task). However, the TransBefore group, who heard the first melody at a different pitch, changed their strategy when the frequency separation between the target and the distractors was greater than 6 STs, suggesting that the decision criterion had been affected by the frequency priming for these separation degrees. The difference in the decision criteria adopted by the listeners may have resulted from a difference in the precision of the melody representation that they had in these tasks. In the Before condition, the participants had more precise expectations concerning the target melody to listen for than did those in the After and TransBefore conditions. Therefore, when they heard the mixture (for small frequency separations, the melody is difficult, or even impossible, to extract from the distractor tones), the perceptual distance between their expectation and their perception appears to have been greater than that of the listeners who had fewer expectations (Tversky, 1977). Response strategy was also affected by the frequency separation. When frequency separation between the target melody and the distractor sequence increased, mean c_{sd} tended generally to increase. This effect suggests that when the target melody was segregated from the distractor tones, the listeners could fail to detect a difference between the two melodies to be compared but did not detect differences that were not present. On the contrary, when the melody and the distractor tones were close in frequency, the listeners did detect differences that were not present, perhaps owing to the integration of foreign distractor tones into the same stream as the tones of the target melody.

One important question is the way pitch-based schemas—that is, the exact pitches of the notes and the diatonicity of the melodies—are used in the perceptual analysis of the composite sequence. According to Bregman (1990, chap. 4), schema-based analysis is not a partitioning process, but a selection process. This means that top-down processes do not perceptually segregate sounds, creating perceptual units such as auditory streams, but allow us to select information from a mixture by a matching process between schemas stored in memory and a sensory representation. Differences in the performance functions observed between the After condition and the Before and TransBefore conditions led us to assume that two functionally distinct processes would be involved in these two types of tasks. Another issue arising from this research is the relation between these primitive processes and schema-based processes involved in the construction of the auditory scene. How

do these two processes, one of which would construct perceptual entities (streams) and one of which would select information on the basis of activated schemas, operate together? Do they act on sensory representations in an independent way, or do they interact to construct streams?

This study would suggest that both processes are functionally distinct, but not completely independent. First, the global improvement in melody recognition when the listeners knew the melody they had to extract from the mixture appeared to be all the more important if the composite sequence was partially segregated (for a mean frequency difference of 3–8 STs between melody and distractor tones). This suggests that the efficiency of top-down processes depends on primitive segregation. It would be consistent with the assumption that the schema-based process is a matching process, depending on the degree of correspondence between a schema and a sensory representation and, thus, on the segregation of melody notes. Second, top-down processes seem to be operant only when there is a mean frequency difference between melody and distractor tones, since for no difference, recognition performance was equivalent whether listeners knew the melody beforehand or not. The fact that top-down processes did not change the perceptual organization of the mixture when there was no primitive segregation suggests that primitive analysis is at least partially autonomous. This notion is consistent with studies showing that primitive auditory scene analysis is preattentive (Sussman et al., 1999) and innate (Demany, 1982; McAdams & Bertoncini, 1997). It is also consistent with the results of the study conducted by van Noorden (1975, Experiments 1 and 2), who found that there was a limit in the action of attentional set for changing perceptual organization of a cyclical sequence composed of an alternation between high and low tones (fission and temporal coherence boundaries).

Many arguments lead us to conclude that primitive and schema-based analyses operate interactively to construct the auditory scene. The auditory system is endowed with a general mechanism, stemming from the coupling between listeners and their environment, which allows it to construct independent descriptions of distinct auditory sources (Bregman, 1990; Shepard, 1981). Thus, the combination of these bottom-up and top-down processes provides the cognitive system with an optimal adaptation to its environment. Indeed, a system that did not take into account the sensory input would be cut off from the outside world, whereas one that did not use previously acquired knowledge would have a very unstable representation of the changing world. Therefore, taking into account both kinds of information solves the stability–plasticity dilemma, an idea that has been developped by Grossberg in his adaptative resonance theory (see Grossberg, 1999, for a recent overview).

REFERENCES

Alain, C., & Woods, D. L. (1994). Signal clustering modulates auditory cortical activity in humans. *Perception & Psychophysics*, **56**, 501-516.

Bey, C., & McAdams, S. (1998, November). *Implication of top-down processes in auditory streaming*. Paper presented at the 39th Annual Meeting of the Psychonomic Society, Dallas.

Bey, C., & McAdams, S. (2002). *Post-recognition of interleaved melodies as an indirect measure of auditory stream formation.* Revision under review.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: MIT Press.

Bregman, A. S. (1993). Auditory scene analysis: Hearing in complex environments. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp. 10-36). Oxford: Oxford University Press, Clarendon Press.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.

Demany, L. (1982). Auditory stream segregation in infancy. *Infant Behavior & Development*, **5**, 261-276.

Deutsch, D. (1970). Tones and numbers: Specificity of interference in immediate memory. *Science*, **68**, 1604-1605.

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology*, **5**, 322-337.

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, **85**, 341-354.

Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval and pitch recognition in memory for melodies. *Journal of the Acoustical Society of America*, **49**, 524-531.

Dowling, W. J., & Harwood, D. L. (1986). Melody: Attention and memory. In *Music cognition* (pp. 124-152). Orlando, FL: Academic Press.

Dowling, W. J., Kwak, S., & Andrews, M. W. (1995). The time course of recognition of novel melodies. *Perception & Psychophysics*, **57**, 136-149.

Dowling, W. J., Lung, K. M.-T., & Herrbold, S. (1987). Aiming attention in pitch and time in the perception of interleaved melodies. *Perception & Psychophysics*, **41**, 642-656.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, **24**, 95-112.

Grossberg, S. (1999). The link between brain learning, attention and consciousness. *Consciousness & Cognition*, **8**, 1-44.

Hartmann, W. M., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, **9**, 155-184.

Hulse, S. H., MacDougall-Shackleton, S. A., & Wisniewski, A. B. (1997). Auditory scene analysis by songbirds: Stream segregation of birdsong by European starlings (Sturnus vulgaris). *Journal of Comparative Psychology*, **111**, 3-13.

MacDougall-Shackleton, S. A., Hulse, S. H., Gentner, T. Q., & White, W. (1998). Auditory scene analysis by European starlings (*Sturnus vulgaris*): Perceptual segregation of tone sequences. *Journal of the Acoustical Society of America*, **103**, 3581-3587.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge: Cambridge University Press.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, **98**, 185-199.

McAdams, S., & Bertoncini, J. (1997). Organization and discrimination of repeating sound sequences by newborn infants. *Journal of the Acoustical Society of America*, **102**, 2945-2953.

Ott, R. L. (1993). *An introduction to statistical methods and data analysis*. (4th ed.). Belmont, CA: Duxbury.

Shepard, R. N. (1981). Psychophysical complementarity. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 279-341). Hillsdale, NJ: Erlbaum.

Sussman, E., Ritter, W., & Vaughan, J. H. G. (1998). Attention affects the organization of auditory input associated with the mismatch negativity system. *Brain Research*, **789**, 130-138.

Sussman, E., Ritter, W., & Vaughan, J. H. G. (1999). An investigation of auditory streaming effect using event-related brain potentials. *Psychophysiology*, **36**, 22-34.

Tversky, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* Unpublished doctoral dissertation, Eindhoven University of Technology.

VLIEGEN, J., & OXENHAM, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *Journal of the Acoustical Society of America*, **105**, 339-346.

NOTES

1. Note that the outcome of the main analyses conducted in this study was not changed by reincluding the excluded participants, except for the interaction between the effect of musical background and frequency separation on both d' and c_{sd}, which then reached significance [$F(8,544) = 4.8$, $e = 0.69$, $p < .0005$], and $F(8,544) = 4.8$, $e = 0.45$, $p < .005$, respectively]. Without the excluded participants, musically trained listeners were neither more accurate [$F(8,432) = 1.5$, $e = 0.74$, $p = .18$] nor less biased [$F(8,432) = 1.5$, $e = 0.59$, $p = .20$] than nonmusicians, with increasing frequency separation between melody and distractor tones. However, adding 13 nonmusicians for whom performance was near chance at all separations increased the difference between both groups for larger separations.

2. One might ask whether the degree to which the first melody is transposed would have an effect on the recognition of the interleaved target, particularly for 12 STs (i.e., one octave), in which information relative to chroma could be used. We conducted an a posteriori analysis in the Trans-Before group to test the possible effect of the transposition degree in the experimental task of interleaved melody recognition, as well as in the transposed Control condition. We computed group d' and c_{sd} values (Macmillan & Kaplan, 1985) because the number of observations for each separation was insufficient to calculate these statistics individually. We found that performance decreased with an increasing degree of transposition essentially in the transposed control condition, but also slightly in the interleaved melody recognition task for the 12- and 24-ST separations. In the control condition, the decrease was a little bit higher between the transposition of 12 and 13 STs than between 13 and 14 STs (a group d' difference of 1.12 and 0.51, respectively), suggesting that the chroma helped the listeners to perform this melody discrimination task. However, no advantage for the octave transposition was found in the interleaved recognition task. These results suggest that the size of the transposition did not have any effect on the segregation of the target melody but impaired the melody discrimination ability. This impairment was observed for both the musician and the nonmusician groups, but only the musicians changed their response criteria with the size of the transposition. Their judgments were biased toward *same* responses in the 12-ST transposition, not biased for 13 STs, and biased toward *different* responses for 14 STs.

3. For the task condition × frequency separation interaction, only a small number (27) of the 351 possible pairwise comparisons among cell means was tested, and these were derived from the experimental design. We thus felt it appropriate, following Ott (1993, p. 813), to use Fisher's protected LSD, as opposed to Tukey's HSD, to test the differences among task conditions at each frequency separation. The latter test would have estimated a much larger familywise error rate under the assumption that all pairwise comparisons would be examined, which was not the case.

(*Continued on next page*)

APPENDIX
The Melodies

The 72 melodies constructed for the experiments (36 original and their corresponding 36 modified versions with two changed notes) are shown in Table A1. The six notes of each melody are expressed in semitones with respect to the equal-tempered note the closest to the mean frequency of the original melodies. The mean note (zero value) was A_5 (880 Hz, MIDI note 81) varying from trial to trial over a range from −3 to +2 ST ($F\sharp_5$ to B_5).

Table A1
The Full Set of 36 Original and Modified Melodies

Melody No.	Original Melodies						Modified Melodies					
1	2	4	2	1	−1	−3	2	0	2	−2	−1	−3
2	−3	−1	1	2	4	2	−3	−1	0	2	1	2
3	−2	0	−2	0	2	3	−2	0	−1	0	5	3
4	−2	0	2	3	2	3	−2	0	5	3	0	3
5	3	5	3	1	0	−4	3	1	3	−3	0	−4
6	−2	0	2	3	2	−2	−2	0	1	3	−2	−2
7	−3	−1	1	−1	1	4	−3	3	1	−2	1	4
8	3	5	3	1	−2	−4	3	2	3	−2	−2	−4
9	−2	0	2	3	0	−2	−2	0	4	3	−4	−2
10	−3	−1	−3	−1	3	4	−3	0	−3	3	3	4
11	−3	−1	1	−1	3	4	−3	−1	−3	−1	5	4
12	−1	1	3	1	−2	−1	−1	2	3	−3	−2	−1
13	3	5	3	0	−2	−4	3	5	1	0	−6	−4
14	−4	−2	0	3	5	3	−4	−2	3	3	2	3
15	−3	−1	−3	1	3	4	−3	−1	0	1	7	4
16	1	3	1	−2	0	1	1	0	1	−4	0	1
17	−1	1	3	−1	−2	−1	−1	5	3	1	−2	−1
18	−3	−1	1	4	3	4	−3	−1	3	4	6	4
19	1	3	5	1	0	−4	1	3	2	1	−3	−4
20	3	5	1	0	−2	−4	3	5	3	0	−6	−4
21	−2	0	3	2	0	−2	−2	2	3	−2	0	−2
22	−4	−2	2	3	5	3	−4	−2	6	3	7	3
23	−3	−1	3	4	3	4	−3	3	3	2	3	4
24	−4	−2	1	3	5	1	−4	−2	−1	3	2	1
25	−1	1	−3	−1	3	4	−1	1	0	−1	1	4
26	2	4	1	0	−3	−2	2	4	−1	0	0	−2
27	−3	−1	2	4	1	2	−3	−1	6	4	3	2
28	−2	2	3	2	0	−2	−2	2	1	2	−4	−2
29	−4	0	2	3	5	3	−4	−2	2	7	5	3
30	−3	1	−1	1	3	4	−3	−2	−1	2	3	4
31	−1	3	1	−1	−2	−1	−1	3	−3	−1	0	−1
32	−3	1	3	4	3	4	−3	1	6	4	5	4
33	−4	0	1	3	5	1	−4	3	1	2	5	1
34	−3	1	2	4	1	2	−3	4	2	0	1	2
35	−4	0	1	5	3	1	−4	0	5	5	2	1
36	−2	2	3	0	2	3	−2	2	4	0	5	3

Note—The notes are expressed in semitones relative to the equal-tempered note closest to the mean frequency of the melody.

(Manuscript received September 20, 2000;
revision accepted for publication November 12, 2001.)