Speech versus nonspeech in pitch memory

Catherine Semal, Laurent Demany, and Kazuo Ueda[a)]
Laboratoire d'Audiologie Expérimentale et Clinique, BP 63, Université Bordeaux 2, 146 rue Léo-Saignat, F-33076 Bordeaux Cedex, France

Pierre-André Hallé
Laboratoire de Psychologie Expérimentale, URA CNRS 316, Université René Descartes, 28 rue Serpente, F-75006 Paris, France

(Received 14 November 1995; revised 10 April 1996; accepted 18 April 1996)

The memory trace of the pitch sensation induced by a standard tone (*S*) can be strongly degraded by subsequently intervening sounds (*I*). Deutsch [Science **168**, 1604–1605 (1970)] suggested that the degradation is much weaker when the *I* sounds are words than when they are tones. In Deutsch's study, however, the pitch relations between *S* and the *I* words were not controlled. The first experiment reported here was similar to that of Deutsch except that the speech and nonspeech stimuli used as *I* sounds were matched in pitch. The speech stimuli were monosyllabic words derived from recordings of a real voice, whereas the nonspeech stimuli were harmonic complex tones with a flat spectral profile. These two kinds of *I* sounds were presented at a variable pitch distance (Δ-pitch) from the *S* tone. In a same/different paradigm, *S* had to be compared with a tone presented 6 s later; this comparison tone could be either identical to *S* or shifted in pitch by ± 75 cents. The nature of the *I* sounds (spoken words versus tones) affected discrimination performance, but markedly less than did Δ-pitch. Performance was better when Δ-pitch was large than when it was small, for the speech as well as nonspeech *I* sounds. In a second experiment, the *S* sounds and comparison sounds were spoken words instead of tones. The differences to be detected were restricted to shifts in fundamental frequency (and thus pitch), the other acoustic attributes of the words being left unchanged. Again, discrimination performance was positively related to Δ-pitch. This time, the nature of the *I* sounds (words versus tones) had no significant effect. Overall, the results suggest that, in auditory short-term memory, the pitch of speech sounds is not stored differently from the pitch of nonspeech sounds.

PACS numbers: 43.71.An, 43.66.Hg, 43.66.Mk [RAF]

INTRODUCTION

The detection of a pitch difference between two tones separated by a few seconds (a standard tone ''*S*'' and a comparison tone ''*C*'') can be markedly impaired by the presentation of other tones between *S* and *C*. Deutsch (1972) reported that the amount of impairment produced by the intervening (''*I*'') tones depends on their distance in pitch from *S* and *C*. She found that discrimination between *S* and *C* is much better when all the *I* tones are far in pitch from *S* (at least 200 cents removed) than when one of the *I* tones is close in pitch to *S* (about 100 cents removed). In the latter case, presumably, the memory trace of *S* is blurred by the memory trace of the *I* tone close in pitch and this is why discrimination between *S* and *C* is poorer.

In Deutsch's experiment, all the sound stimuli were *pure* tones and thus had similar timbres. What does happen to discrimination performance when the *I* tones are very different in timbre from *S* and *C*? *A priori*, one could think that this should prevent the *I* tones from producing large interference effects, whatever their pitches. However, two of us recently showed that this is not the case (Semal and Demany, 1991, 1993). We found that *I* tones which are very different from *S* and *C* in spectral content or in amplitude envelope still produce poor performance if they are in the pitch vicinity of *S*. We also found that the intensity of the *I* tones was not an important factor. Our experiments indicated that performance depends almost exclusively on the *I* tones' pitches, as if the human brain contained a mnemonic device specifically devoted to the retention of pitch and deaf to any other sound quality. Results supporting this view were also reported by Krumhansl and Iverson (1992).

In the present study, we wished to determine if human listeners retain the pitch of a *speech* sound exactly like the pitch of a nonspeech sound. A contrary hypothesis is that once a sound has been identified as a *speech* sound, the temporary retention of all its perceptual attributes, including its pitch, can take place—or always takes place—in a specific memory store to which nonspeech sounds have no access. A strong version of this ''speech-specificity hypothesis'' is that there are two completely separate pitch stores, one devoted to speech sounds and the other to nonspeech sounds. A weaker version of the same basic hypothesis may be put forth and will be considered later. At this point, let us point out that if the strong version just stated were correct, the results of Semal and Demany (1991, 1993) should not be generalizable to speech *I* sounds: The pitch memory trace of a tone should be systematically less affected by subsequent speech sounds than by subsequent tones, these two kinds of *I* sounds being matched in pitch.

[a)]Now at: Faculty of Letters, Kyoto Prefectural University, Shimogamo, Kyoto 606, Japan.

In her first publication concerning interference phenomena in pitch memory, Deutsch (1970) reported an experiment where both pure tones and spoken words were used as *I* sounds. *S* and *C* were pure tones. Discrimination between *S* and *C* appeared to be much better when the *I* sounds were words than when they were pure tones. This was so even when the *I* words had to be recalled on each trial, whereas the *I* tones had to be ignored. However, the experiment in question did not clearly support the speech specificity hypothesis because the *I* words were not controlled in pitch. The ''speech versus nonspeech'' factor was very probably combined with a pitch distance factor: Presumably, *S* and *C* were much closer in pitch to the *I* tones than to the *I* words;[1] therefore, the good discrimination performance obtained with the *I* words may have been due only to their remoteness in pitch.

The two experiments reported here provide new tests of the speech specificity hypothesis. Basically, they are revised replications of the experiment performed by Deutsch (1970). Their essential novelty lies in a control of the speech sounds' pitches. In both experiments, we compare the interference effects of various speech sounds and nonspeech sounds in a pitch discrimination task requiring only same/different judgments. The two sounds to be compared on each trial, *S* and *C*, were nonspeech sounds in experiment 1 and speech sounds in experiment 2. All the speech sounds (*S*, *C*, and *I*) were meaningful monosyllabic words (numbers, as in the original study by Deutsch) spoken by a natural voice. The nonspeech sounds, on the other hand, were synthetic tones with a flat spectral profile and a flat amplitude envelope. Therefore, whereas some artificial sounds can be perceived either as speech or as nonspeech, depending on the acoustic context and/or attentional biases (see, e.g., Ayres *et al.*, 1979; Neath *et al.*, 1993), the stimuli employed here were quite unambiguous in this regard.

I. EXPERIMENT 1

A. Method

1. Task and conditions

On each trial, subjects had to make a same/different judgment on two tones, *S* and *C*, separated by 6 s (onset-to-onset interval). *S* and *C* were complex tones with the same timbre (harmonic content and amplitude envelope) and the same intensity. Their fundamental frequencies ($F0$'s) were identical or different with equal probability. Each difference in $F0$ amounted to 75 cents (about 4.4%) and was positive or negative with equal probability. More details on *S* and *C* will be provided in Sec. I A 2.

Subjects were run in three conditions. In the ''pretest'' condition, *S* and *C* were separated by a silent interval. In the ''speech'' condition and the ''nonspeech'' condition, four successive *I* sounds were presented between *S* and *C*, in a regular rhythm of one sound per second. The first *I* sound started 1.5 s after the onset of *S* and there was also 1.5 s between the onset of the last *I* sound and the onset of *C*. In the speech condition, each *I* sound could be one of four monosyllabic words (specified later); a random choice between these four alternatives was made before each presentation. In the nonspeech condition, the *I* sounds were complex tones with four possible harmonic contents, among which a random choice was again made before each presentation.

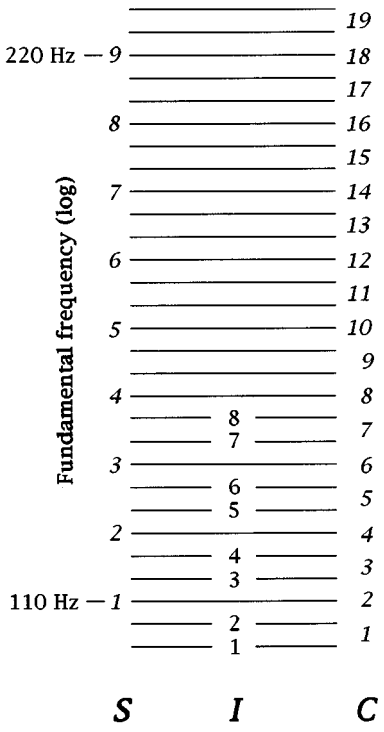

FIG. 1. Pitch levels of the stimuli used in experiment 1 (*S* tones on the left, *C* tones on the right, *I* sounds in the middle). The spacing of the horizontal lines corresponds to 50 cents, i.e., 1/24 octave.

On a given trial, each *I* sound could take, at random, one of four nominal $F0$'s that covered a range of 200 cents. The geometric mean of these four $F0$'s could be (1) 900 cents below the $F0$ of *S*, (2) 450 cents below, or (3) 0 cent below. This defined, for both the speech and the nonspeech conditions, three levels of a factor that we called ''Δ-pitch'': Δ-pitch could be ''large,'' ''medium,'' or ''small.'' For each level of Δ pitch, the four nominal $F0$'s were respectively 50 and 100 cents above and below the mean.

2. Stimuli

The *S* tones and *C* tones had a total duration of 350 ms and were gated on and off with 10-ms linear amplitude ramps. They were composed of three equal-amplitude harmonics, with ranks 1–3, which were added in sine phase. Nine *S* tones ($S1-S9$) were used. Their $F0$'s were regularly spaced by intervals of 150 cents. As shown in Fig. 1, the $F0$ of $S1$ was 110 Hz and $S9$ was one octave above. The 19 *C* tones ($C1-C19$, see Fig. 1) were spaced by intervals of 75 cents. The $F0$ of a given *S* tone, S_i, was equal to the $F0$ of C_{2i}; therefore, S_i could be paired with C_{2i}, C_{2i-1}, or C_{2i+1}.

The *I* tones involved in the nonspeech condition differed from *S* and *C* in duration and timbre. They had a total duration of 250 ms and consisted of the first 6, 9, 13, or 20 harmonics of some $F0$. Like those of *S* and *C*, the harmonics of each *I* tone had equal amplitudes and were added in sine

TABLE I. $F0$ and duration measurements on the original speech recordings.

	File	$F0$ measurements (Hz) Onset	Offset	Minimum	Maximum	Geom. mean	Duration of voiced portion (ms)
Target pitch: low ($S1$)	''7''	119.8	112.1	112.0	119.8	113.7	210
	''9''	111.2	112.9	111.2	114.3	112.7	250
	''10''	112.5	107.1	107.1	113.4	112.2	220
	''15''	119.4	111.7	109.1	119.4	113.5	350
	Mean	115.7	110.9	109.9	116.7	113.0	258
Target pitch: high ($S3$)	''7''	133.2	133.9	129.6	134.2	132.2	160
	''9''	128.1	136.7	128.1	136.7	131.7	250
	''10''	132.6	134.9	131.0	134.9	132.1	220
	''15''	140.1	134.8	129.0	140.1	132.6	310
	Mean	133.5	135.0	129.4	136.4	132.1	235

phase. The I tones had eight possible $F0$'s ($I1-I8$; see Fig. 1). On a given trial, three possible sets of four $F0$'s were used: $[I1-I4]$, $[I3-I6]$, or $[I5-I8]$. Each of these sets was associated with one S tone for each level of Δ-pitch. Thus, $[I1-I4]$ was associated with $S1$ (small Δ-pitch), $S4$ (medium Δ-pitch), or $S7$ (large Δ-pitch). Similarly, $[I3-I6]$ was associated with $S2$, $S5$, or $S8$, and $[I5-I8]$ was associated with $S3$, $S6$, or $S9$.

In the speech condition, the I sounds were derived from recordings of four French words spoken by the second author: ''sept'' (/sɛt/; seven), ''neuf'' (/nœf/; nine), ''dix'' (/dis/; ten), and ''quinze'' (/kɛ̃z/; fifteen). In order to present these words at the eight pitch levels of the I tones, eight sound files were made for each word. These speech files were then associated with the S tones according to the same combination rules as those used for the I tones. Thus, the nominal pitch interval between S and the words varied between -100 and $+100$ cents for $S1-S3$ (small Δ-pitch), between 350 and 550 cents for $S4-S6$ (medium Δ-pitch), and between 800 and 1000 cents for $S7-S9$ (large Δ-pitch).

The recorded words were *spoken* rather than sung, but the speaker endeavoured to produce words with a precise pitch. The eight versions of each word were derived from two original recordings, in which the speaker's intended pitch was respectively the pitch of $S1$ and the pitch of $S3$. Table I presents the results of measurements made on the speaker's original utterances. Note that there was a significant fluctuation of $F0$ within each utterance, as in natural speech. Note also that the voiced portions of the words had a mean duration which was very close to the duration of the I tones (250 ms). All the recordings lasted less than 600 ms. We assessed their actual pitches by a pitch matching experiment: Two subjects (the second and third authors) matched them to a complex tone with the same spectral structure as the S tones (i.e., three harmonics) and an adjustable $F0$. For each recording, the mean of the adjusted $F0$ values was taken as the actual pitch. The sampling rate of the original speech file (20 kHz) was then modified in order to compensate the difference between the actual pitch and the intended pitch. From the resulting file, four other files were finally derived by four further changes in the sampling rate, respectively corresponding to intervals of ± 50 and ± 100 cents. In principle, these final files were exactly matched in pitch to tones at the pitch levels $I1-I4$ (when the source file had been matched to $S1$) or $I5-I8$ (when the source file had been matched to $S3$). Of course, the changes in sampling rate modified the formant frequencies of the original recordings, and thus their timbre; however, the maximum change corresponded to an interval of only 119 cents (7.1%).

All stimuli were presented diotically at roughly the same loudness level (about 65 phons). The nominal sound pressure level of $S5-S9$ and $C10-C19$ was 73.2 dB. For the S and C tones with lower $F0$'s (below 155.6 Hz), the SPL was increased at a rate of 6 dB/oct in order to maintain an approximately constant loudness. This variation of SPL was warranted because the S and C tones possessed only three harmonics. Since the I tones had at least six harmonics, their SPL was not varied as a function of their $F0$. However, in order to compensate the effect of spectral width on loudness, the I tones' power was constrained to be inversely proportional to the number of their harmonics. Thus, the nominal SPL of the I tones with 6 and 20 harmonics was respectively 70.2 and 65.0 dB. Admittedly, our manipulations of SPL did not ensure that all stimuli had exactly the same loudness, but the results of Semal and Demany (1993) indicate that a perfect loudness equalization was unnecessary.

The stimuli were generated via the 16-bit DACs of a DSP card (Oros AU22), passed through antialiasing filters (Kemo VBF/04; cutoff frequency: 8 kHz), and delivered by means of TDH 39 earphones.

3. Procedure and subjects

Subjects were tested individually in a double-walled soundproof booth, where they sat in front of a keyboard connected to the computer containing the DSP card. On each trial, they gave their response (''same'' or ''different'') by pressing one of two labeled keys. There was no feedback concerning response accuracy. Any response initiated the next trial after a 1-s delay. Subjects were instructed to ignore the I sounds, but received no prior information about the nature of the differences between S and C.

Only one experimental session was run for each subject. This session comprised nine blocks of 27 trials: one block in the pretest condition (no I sounds), and then four blocks in both the speech and nonspeech conditions. Within each

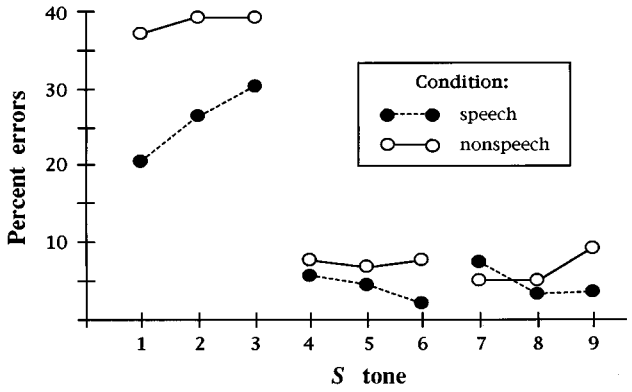

FIG. 2. Error rates measured for each S tone in experiment 1. Δ-pitch was ''small'' for $S1-S3$, ''medium'' for $S4-S6$, and ''large'' for $S7-S9$.

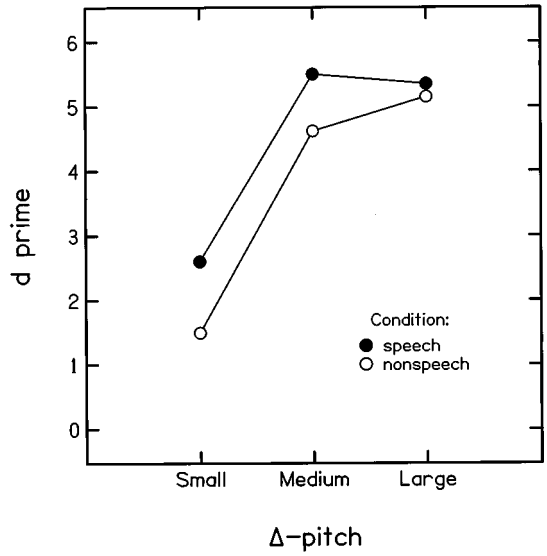

FIG. 3. d' as a function of Δ-pitch and the I sounds' nature (speech or nonspeech), in experiment 1.

block, each of the nine different S tones was used three times; except for this constraint, the successive S tones were selected randomly. The goal of the single block in the pretest condition was to select proficient discriminators, thus reducing the risk of floor effects in the other two conditions. Seventeen potential subjects were discarded because they failed to make less than four errors in the pretest block. Eighteen other listeners, who made less than four errors, were tested in the speech and nonspeech conditions.[2] The four blocks run for both conditions were interleaved; the first one was in the speech condition for half of the subjects, and in the nonspeech condition for the other half.

All subjects but one were native speakers of French. Most of them were in their twenties. Three had previously participated as subjects in another experiment on pitch memory.

B. Results and discussion

Figure 2 shows the error rate obtained for each S tone in the speech and nonspeech conditions. Each data point is based on 216 trials (18 subjects ×4 blocks ×3 trials). Recall that Δ-pitch had the same average value of 0 cent for $S1-S3$, of 450 cents for $S4-S6$, and of 900 cents for $S7-S9$. Statistical analyses were performed in order to determine if, within each of these three groups and for each condition, the error rates differed systematically from each other. Since the distribution of the 18 individual scores measured for a given S tone in a given condition was often markedly asymmetric (with a mode for zero error), we used nonparametric tests, namely Friedman analyses of variance by ranks (Friedman, 1937). No reliable differences were found ($\chi_r^2 \leqslant 2.19$, $P \geqslant 0.33$).

By contrast, similar tests showed that there were highly significant differences *between* the three groups of S tones, for both conditions (speech: $\chi_r^2 = 58.79$, $P < 0.001$; nonspeech: $\chi_r^2 = 71.03$, $P < 0.001$). It can be seen in Fig. 2 that, for each condition, the error rates had high values for $S1-S3$ and abruptly fell to a low plateau for $S4-S9$. The abruptness of this fall is important because it implies that the essential source of variance was Δ-pitch, i.e., the pitch *distance* between S and the I sounds, rather than the pitch of S per se.

The error rates obtained in the speech and nonspeech conditions were compared to each other using sign tests. For $S1-S3$ (small Δ-pitch), subjects made significantly fewer errors in the speech condition than in the nonspeech condition ($P = 0.0012$). This was also true for $S4-S6$ (medium Δ-pitch; $P = 0.033$). However, there was no significant difference for $S7-S9$ (large Δ-pitch; $P = 0.254$).

The statistical tests reported above had to be applied on error rates rather than d' measures (Green and Swets, 1974; Macmillan and Creelman, 1991) because the performance of a given subject for a given S tone and condition could not be assessed in terms of d': The corresponding number of trials (12) was too small. However, we wanted to compare the effect of the I sounds' nature (speech versus nonspeech) to the effect of Δ-pitch, and for this comparison it was appropriate to quantify performance in terms of d' rather than error rate. Thus, after a pooling the 18 subjects' data for $S1-S3$, $S4-S6$, and $S7-S9$, ''group'' d's were computed. In doing so, we assumed that subjects used the ''differencing'' strategy described by Macmillan and Creelman (1991, Chap. 6). The results are shown in Fig. 3. Assuming that the y axis of this figure—d' on a linear scale—provides a valid metric to assess the relative effects of the two independent variables, it can be concluded that, overall, Δ-pitch had a markedly larger effect on performance than the I sounds' nature. Another conclusion is that the two independent variables did not strongly interact: The effect of the I sounds' nature had about the same size for the ''small'' and ''medium'' values of Δ-pitch. For the ''large'' Δ-pitch, the equivalence of the two d's may be considered as the consequence of a ceiling effect. Obviously, the two d's *had* to become similar beyond some value of Δ-pitch since a large Δ-pitch was sufficient to get a nearly perfect performance in the nonspeech condition; indeed, performance was already excellent for the medium value of Δ-pitch.

Overall, the results of this experiment are clearly inconsistent with the *strong* version of the ''speech specificity hypothesis'' stated in the Introduction. If there were two com-

pletely separate pitch stores, one devoted to speech sounds and the other to nonspeech sounds, then *I* words should not affect the pitch memory trace of an *S* tone. In fact, *I* words can produce large interference effects, if they are close in pitch to the *S* tone. However, we found that for both the ''small'' and the ''medium'' values of Δ-pitch, the interference effects of words were somewhat smaller than the interference effects of tones. At first sight, this finding does not tally with the concept of a single ''pitch memorizer'' which would be totally deaf to sound attributes other than pitch. But another interpretation is possible: It may be that the *I* words produced weaker interference effects because of their pitch properties themselves. Semal and Demany (1993) provided evidence that the interference effect of *I* sounds on a pitch memory trace is positively correlated to the precision of the *I* sounds' pitches. Within our *I* words, there were fluctuations of $F0$: The $F0$ contour of natural speech sounds is never flat, and is affected by segmental variations. Such fluctuations were absent from the *I* tones. Therefore, it is reasonable to assume that the pitches of the *I* words were less precise, or less salient, than those of the *I* tones. The results of experiment 2 clarify this issue.

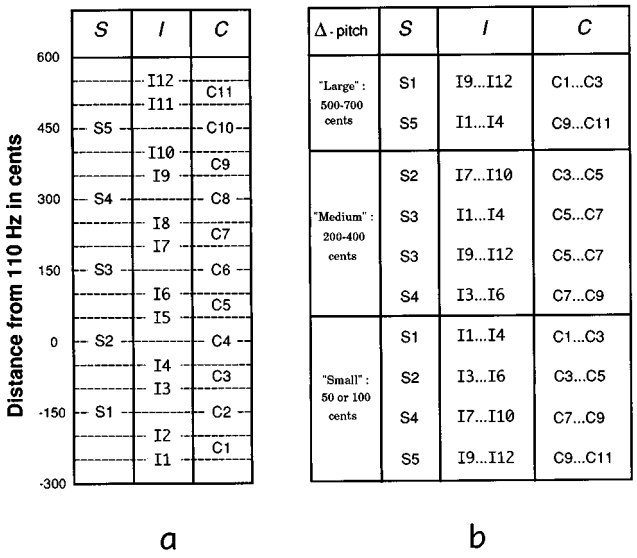

FIG. 4. (a) Pitch levels of the *S*, *I*, and *C* sounds used in experiment 2; (b) Combinations between the pitch levels of the *S*, *I*, and *C* sounds for the three levels of Δ-pitch.

II. EXPERIMENT 2

Experiment 1 discredited an extreme version of the speech specificity hypothesis for pitch memory, but not a weaker and maybe more plausible version of it. Suppose again that there are two pitch stores and that one is devoted to speech exclusively, but this time that the other store operates on both speech and nonspeech. That would be true, for instance, if pitch information extracted from speech sounds was kept initially in a speech-specific store but secondarily transmitted (copied) to a ''universal'' pitch store. Alternatively, these two stores might operate in parallel instead of serially. In each case, anyway, *I* words and *I* tones could produce, as we found, similar interference effects on the pitch memory trace of an *S* tone. However, what will happen if *S* is a word instead of a tone? If the store devoted to speech exclusively is a good pitch memorizer—that is, if it is not poorer than the universal store—then subjects will take advantage of its existence when the *I* sounds are tones, because tones will not produce interference effects in this store. But *I* words should produce interference effects in it, at least in case of pitch proximity. So, one should see a large effect of the *I* sounds' nature on the detection of a pitch difference between *S* and a comparison word *C*. This reasoning was the basis of experiment 2.

A. Method

Essentially, experiment 2 was a replication of experiment 1 with only one crucial change: the replacement of the *S* and *C* tones by *S* and *C* words. However, a number of other methodological details were also different; we describe them below.

1. Stimuli

In experiment 1, the pitches of the *S* sounds covered a range of 1 oct, from 110 to 220 Hz. A different range had to be used in experiment 2, in order to fit the speaker's voice. On a given trial, the pitch of the *S* sound could take five different nominal values. The corresponding set of $F0$'s is displayed in the left panel of Fig. 4. These $F0$'s were again spaced by intervals of 150 cents, but they covered a range of only 0.5 oct and the lowest one ($S1$) was 150 cents below the lowest of experiment 1. The pitch relations between the *S* sounds and *C* sounds were as before, the *C* sounds being again spaced by intervals of 75 cents (see Fig. 4).

Δ-pitch had again three levels, but its ''medium'' level was 300 ± 100 cents instead of 450 ± 100 cents, and its ''large'' level was 600 ± 100 cents instead of 900 ± 100 cents. For each level of Δ-pitch, the right panel of Fig. 4 indicates how the *S*, *I*, and *C* sounds were selected with regard to pitch.

The word stimuli were derived from the eight recordings already used in experiment 1. However, these recordings were processed differently here. First, we reassessed their actual pitches by a revised version of the pitch matching experiment described in Sec. I A 2. The obtained results were very consistent with those found previously (maximum discrepancy: 10 cents). Then, a special implementation of the PSOLA method (Moulines and Laroche, 1995) was used to transpose the recordings at the desired pitch levels by shifts of the $F0$ patterns.[3] An illustration of the transposition procedure is given in Fig. 5. This procedure preserved the durations and formant patterns of the original utterances. Perceptually, therefore, the two words to be compared on each trial never differed from each other in any aspect other than pitch. Since each of the four words (''sept,'' ''neuf,'' ''dix,'' and ''quinze'') had been recorded at two nominal pitch levels corresponding to $S2$ and $S4$ in Fig. 4, the transpositions could be limited to rather small intervals (maximum: 279 cents). Thus, even at the extreme pitch levels ($I1$ and $I12$), the words sounded natural.

The *I* tones used in the nonspeech condition consisted of the first three, four, or six harmonics of some $F0$ (ranging

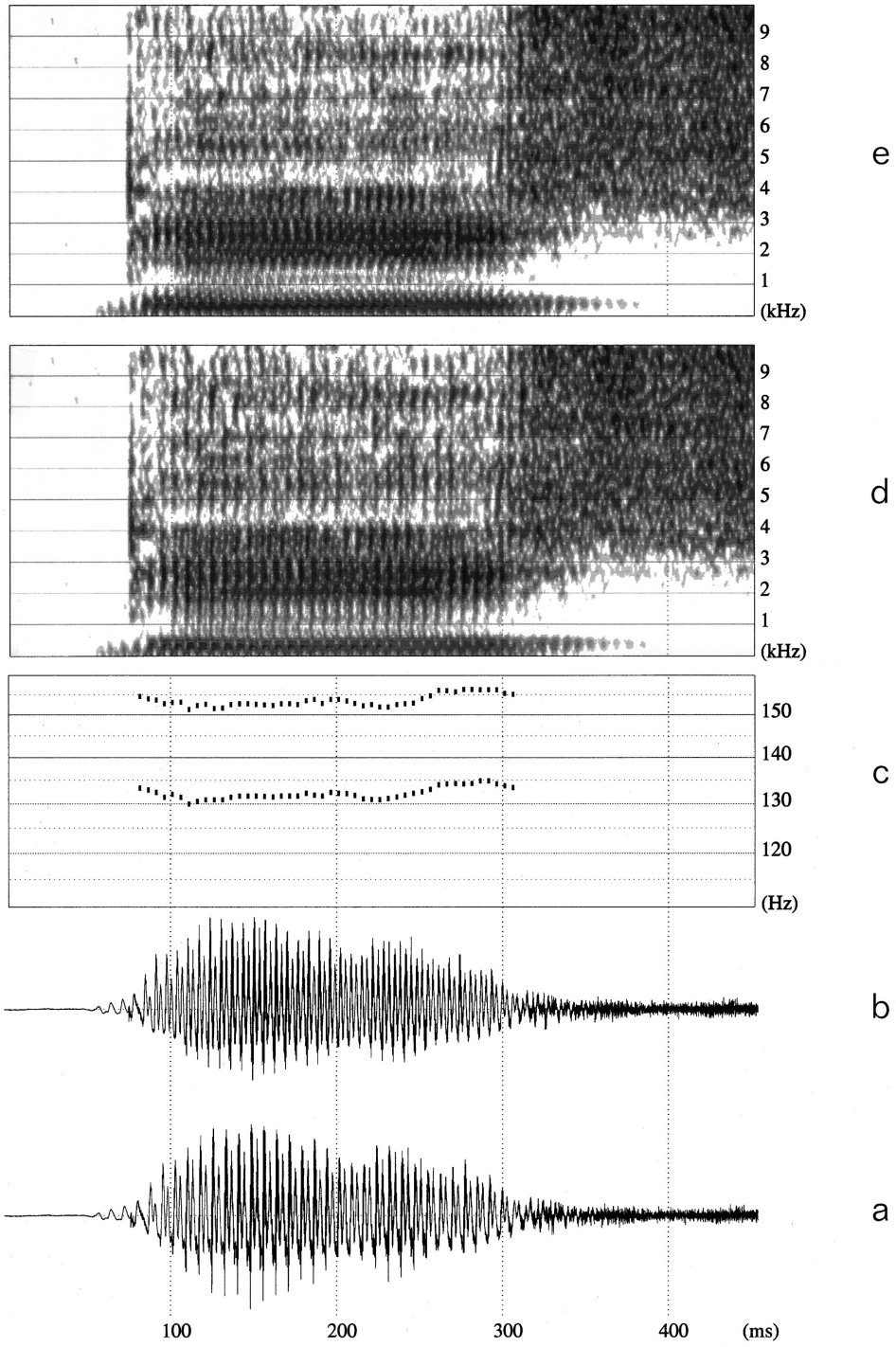

FIG. 5. Illustration of the pitch transpositions made in experiment 2: one of the recorded words—the high ''dix'' (/dis/)—and an upward transposition of it by 245 cents (which corresponds to a frequency increase of about 15%). (a) Waveform of the original signal; only the beginning of /s/ is shown, in order to enhance the voiced portion). (b) Waveform of the transposition; note that there is no change in the duration of the voiced portion; the overall duration was also the same. (c) *F*0 curves of the original signal and the transposition; the ordinate scale is logarithmic; note that the two curves are almost exactly parallel. (d) Wide band spectrogram of the original signal. (e) Wide band spectrogram of the transposition; note that there is no change in the formant frequencies.

from $I1-I12$). The harmonics of each *I* tone had equal amplitudes, inversely proportional to the tone's number of harmonics, and were added in sine phase. As in experiment 1, all the stimuli had roughly the same loudness level, about 65 phons.

On each trial, a four-alterative random choice was made to determine the phonetic identity of the *S* word (''sept,'' ''neuf,'' ''dix,'' or ''quinze''). In the speech condition, the *I* words following the *S* word always differed phonetically from it; the phonetic identity of each *I* word was thus determined by a *three*-alternative random choice. Similarly, in the nonspeech condition, the timbre of each *I* tone was determined by a random choice between the three possible numbers of harmonics.

2. Procedure and subjects

Experiment 1 had been performed on 18 listeners who were tested in a single session. Here, the number of subjects was reduced to 6 but much more data were obtained from each subject: 20 blocks of 24 trials were respectively run in the speech condition, the nonspeech condition, and a third condition, called ''no-*I*,'' where no *I* sound was presented (as in the pretest of experiment 1). A given experimental session consisted of two or three blocks of trials in each of these three conditions (changing from block to block). Within each block, there were eight trials for each of the three levels of Δ-pitch. These three sets of eight trials were randomly intermixed. Within each set, however, the categories of trials listed in the right panel of Fig. 4 were constrained to be equally frequent.

As in experiment 1, a pretest with no *I* sounds was run in order to select proficient discriminators. Five potential subjects were discarded due to their poor performance in this pretest. The six selected listeners had not taken part in experiment 1 or a related experiment. All of them were in their twenties, and all but one were native speakers of French. Until the end of their last session, they were not told anything about the nature of the differences to be detected and the rationale of the experiment.

B. Results and discussion

For each cell of the experimental design (6 subjects×3 conditions×3 levels of Δ-pitch), 160 trials had been run. From these 160 trials, a d' statistic was computed, assuming again that subjects used a ''differencing'' strategy (Macmillan and Creelman, 1991). Figure 6 displays the individual and median results. Notice that two data points (for subjects CL and NK) are surrounded by a small square. These squares mean that d' was arbitrarily set at 6.0 because there was no false alarm at all (i.e., no incorrect ''different'' response).

In the no-*I* condition, Δ-pitch was of course a dummy factor by itself. However, recall that in each condition, the pitch levels of the *S* and *C* sounds varied with Δ-pitch, as shown in the right panel of Fig. 4. The role of this pitch level factor *per se* can be assessed from the results obtained in the no-*I* condition. An examination of Fig. 6 suggests that its role was not significant, and this is confirmed by an analysis of variance [$F(2,10)=2.50$, $P>0.10$]. Not surprisingly, d' was almost always higher in the no-*I* condition than in the speech and nonspeech conditions.

The d's measured in the speech and nonspeech conditions alone were submitted to another analysis of variance. It emerged from this analysis that Δ-pitch had a highly significant effect on d' [$F(2,10)=15.07$, $P<0.001$] and did not interact significantly with the condition factor [$F(2,10)=1.70$, $P>0.10$], which in itself had no significant effect [$F(1,10)=1.05$, $P>0.10$]. The lower panel of Fig. 6 indeed shows very similar results for the speech and nonspeech conditions, and in each case a positive correlation between Δ-pitch and d'.

These results do not fit the idea that the pitch of speech sounds can be memorized in a store devoted exclusively to speech sounds. If that were the case, discrimination performance should have been significantly better in the nonspeech condition than in the speech condition. Instead, our subjects appeared to be deaf to the nature of the *I* sounds, and to be sensitive only to their pitches. In experiment 1, Δ-pitch also appeared to be more important than the nature of the *I* sounds, but the nature of the *I* sounds had a significant effect. It is remarkable that experiment 2 provided *stronger* evidence against the speech specificity hypothesis while its aim was to test a *weaker* version of the hypothesis in question. This can be understood if one assumes that the pitches of our word stimuli were somewhat less precise or salient than the pitches of the tones, as suggested in the discussion of experiment 1. Under this assumption, it could be expected that *whatever the nature of the S sounds*, the *I* words would produce somewhat smaller interference effects than the *I* tones, especially for small values of Δ-pitch. In experiment 2, for the lowest level of Δ-pitch, a slight trend in the corresponding direction was indeed found (see Fig. 6).

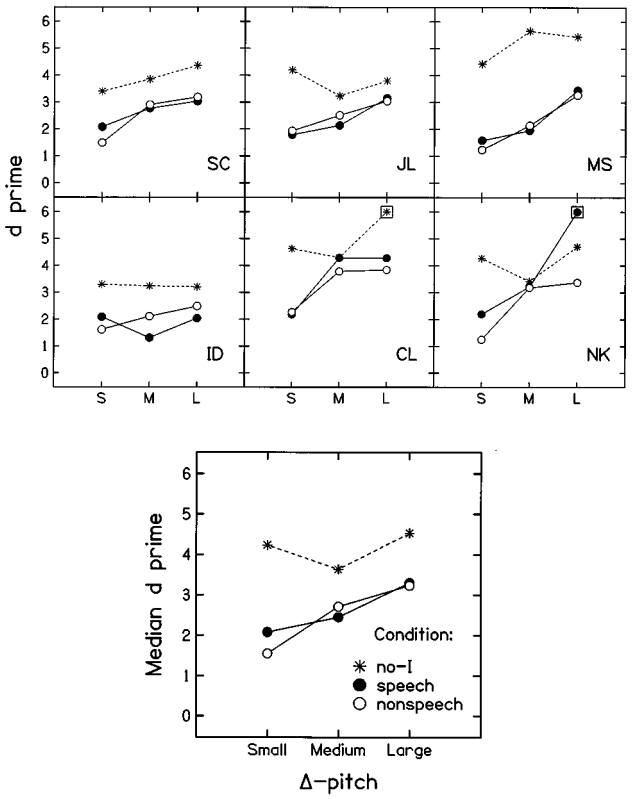

FIG. 6. d' as a function of Δ-pitch and the *I* sounds' nature, in experiment 2. The six upper panels display the individual results and the lower panel displays their medians. In the panels for subjects CL and NK, the small square surrounding one data point means that d' was arbitrarily set at 6.0 because there was no false alarm at all.

III. GENERAL DISCUSSION

It has been repeatedly argued that human listeners process speech sounds in a speech-specific manner. Globally (i.e., without focussing on the case of pitch), this hypothesis is supported by numerous experimental findings. For instance, several investigators of the ''recency'' and ''suffix'' effects in serial recall tasks provided strong evidence that speech sounds and nonspeech sounds are treated differently in auditory short-term memory (Rowe and Rowe, 1976;

Morton *et al.*, 1981; Surprenant *et al.*, 1993). In the experiment performed by Rowe and Rowe (1976), subjects had to recall sequences of either speech or nonspeech sounds. On each trial, the sequence to be recalled was followed by an extraneous suffix consisting also of either speech or nonspeech. The suffix appeared to have a more deleterious effect on recall performance when it was of the same nature (speech or nonspeech) as the sounds to be recalled than when this was not the case.

However, in which respects are speech sounds treated as special entities? One possible thesis would be that *all* the features of a given sound, including its pitch, are analyzed and memorized in a specific manner as soon as the sound in question is identified as a speech sound; pitch retention, then, could take place in a memory store not penetrable by nonspeech sounds. At first sight, the results reported by Deutsch (1970) seemed to support this thesis. They suggested that the memory trace of a tone's pitch is much more affected by subsequent tones than by subsequent words. But the results reported here disproved this suggestion. We did not obtain convincing evidence for the idea that human listeners memorize in a special manner the pitch of speech sounds. Our results are much more consistent with the concept of a single pitch memorizer, deaf to anything but pitch—a concept previously supported by experiments in which only nonspeech sounds were used (for a review, see Semal and Demany, 1993).

According to authors such as Liberman and Mattingly (1985; Mattingly and Liberman, 1988), humans do possess a specialized ''*speech perceiving system*'' but its function is to extract only the *phonetically relevant* aspects of speech sounds. In speech signals, pitch can serve as a phonetic cue to voicing or vowel height, but globally intonation carries little information about the phonetic segments. Mattingly and Liberman (1988, p. 787) even assert that the ''speech perceiving system'' completely ignores ''the laryngeal source signal'', and thus (presumably) pitch. Clearly, our results do not conflict at all with this radical suggestion.

Our results are also consonant with those obtained in a recent brain imaging study by Zatorre *et al.* (1992). These authors measured cerebral blood flow changes in subjects presented with pairs of spoken syllables. Three conditions were run: (1) a ''passive speech'' condition, where the subjects merely listened to the pairs of syllables; (2) a ''phonetic'' condition, where the subjects had to identify the pairs composed of syllables ending with the same consonant; (3) a ''pitch'' condition, in which the pairs to be identified were those forming an ascending musical interval. In order to localize the parts of the brain crucially activated by phonetic judgments and pitch judgments, the cerebral activity measured in the passive speech condition was subtracted from the activities measured in the other two conditions. This revealed that the phonetic judgments activated essentially the left hemisphere whereas the pitch judgments specifically activated the right hemisphere. More precisely, the pitch judgments specifically activated the right frontal lobe. This region of the brain had previously been found to play a significant role in the short-term retention of the pitch of tones (Zatorre and Samson, 1991). Therefore, it seems that the brain regions crucially activated by pitch comparisons are at least partly the same for speech sounds and nonspeech sounds.

Note that there was an important difference between the perceptual judgments required in the pitch condition of Zatorre *et al.* and those required in our second experiment. The subjects of Zatorre *et al.* had to compare syllables which always contained *different* vowels. Thus, in the pitch condition, they were forced to separate pitch from timbre. Our subjects, on the other hand, were not forced to do so: On a given trial, the detection of *any* difference between the two word stimuli to be compared was sufficient for a correct response. Indeed, as mentioned above, our subjects were not even informed initially that the differences to be detected were differences in pitch. Of course, they may have uncovered this by themselves at the beginning of the experiment, and then used this knowledge to perform the task. However, one can imagine that the separation of pitch from timbre in auditory memory is an automatic achievement of the brain rather than an optional process that would be executed only under some specific environmental conditions. It would be worthy to test this hypothesis in further experiments.

ACKNOWLEDGMENTS

This work was supported by the Conseil Régional d'Aquitaine. Part of it was reported in the Proceedings of a meeting of the International Society for Psychophysics (Cassis, France, october 1995). L.D. and P.A.H. are affiliated with the Centre National de la Recherche Scientifique. The stay of K.U. in Bordeaux was made possible by the Ministère de l'Education Nationale (DRED). We are grateful to Neil A. Macmillan, who simplified our computations of d' by sending us useful software, and to Diana Deutsch for preliminary encouragements. We also thank Robert A. Fox, Adrian Houtsma, Bruno Repp, and two anonymous reviewers for their comments on a previous version of the manuscript.

[1]In a private discussion with us, Deutsch supported this conjecture.

[2]Semal and Demany (1991, 1993) also discarded about 50% of listeners following a pretest. The selection of proficient discriminators is probably not an important point since we derive our conclusions from *within*-subject comparisons.

[3]Our goal was to shift the $F0$ contours of the speech recordings while preserving their duration and formant patterns. This was done by a combined pitch-scaling and time-scaling technique whose basic concept is similar to the time domain PSOLA method of speech modification [see Moulines and Laroche (1995), for a synthetic presentation of PSOLA and its variants]. Our technique, however, differs in some respects from the usual implementations of PSOLA. First, the pitch periods of the original speech signal are individuated on the basis of a preliminary $F0$ computation (checked for spurious values) together with waveform inspection. Second, the analysis windows are not centered on pitch pulses. Instead, they each start from a pitch pulse region and leave intact the longest possible portion—given the current pitch scaling factor—of the upcoming signal, thus preserving most of the impulse response found in the individual pitch periods. Original pitch periods are weighted exclusively in the ''overlap-add'' regions with exactly synchronized offset versus onset raised-cosine half-windows: The sum of the weights applied is always one, thereby avoiding unpredictable amplitude modulation and formant weakening. Finally, the dynamic aspect of pitch-scale and time-scale modifications is taken care of. When the two pitch periods made adjacent in the modified speech were apart in the original speech (thus possibly differing in various

respects), an interpolation scheme is applied to ensure a naturally smooth transition.

Ayres, T. J., Jonides, J., Reitman, J. S., Egan, J. C., and Howard, D. A. (**1979**). ''Differing suffix effects for the same physical suffix,'' J. Exp. Psychol.: Human Learn. Memory **5**, 315–321.

Deutsch, D. (**1970**). ''Tones and numbers: Specificity of interference in immediate memory,'' Science **168**, 1604–1605.

Deutsch, D. (**1972**). ''Mapping of interactions in the pitch memory store,'' Science **175**, 1020–1022.

Friedman, M. (**1937**). ''The use of ranks to avoid the assumption of normality implicit in the analysis of variance,'' J. Am. Stat. Assoc. **32**, 675–701.

Green, D. M., and Swets, J. A. (**1974**). *Signal Detection Theory and Psychophysics* (Krieger, Huntington, NY).

Krumhansl, C. L., and Iverson, P. (**1992**). ''Perceptual interactions between musical pitch and timbre,'' J. Exp. Psychol.: Human Percept. Perform. **18**, 739–751.

Liberman, A. M., and Mattingly, I. G. (**1985**). ''The motor theory of speech perception revised,'' Cognition **21**, 1–36.

Macmillan, N. A., and Creelman, C. D. (**1991**). *Detection Theory: A User's Guide* (Cambridge U.P., Cambridge, UK).

Mattingly, I. G., and Liberman, A. M. (**1988**). ''Specialized perceiving systems for speech and other biologically significant sounds,'' in *Auditory Function*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York), pp. 775–793.

Morton, J., Marcus, S. M., and Ottley, P. (**1981**). ''The acoustic correlates of ''speechlike'': A use of the suffix effect,'' J. Exp. Psychol.: General **110**, 568–593.

Moulines, E., and Laroche, J. (**1995**). ''Non-parametric techniques for pitch-scale and time-scale modification of speech,'' Speech Commun. **16**, 175–205.

Neath, I., Surprenant, A. M., and Crowder, R. G. (**1993**). ''The context-dependent stimulus suffix effect,'' J. Exp. Psychol.: Learn. Memory Cognit. **19**, 698–703.

Rowe, E. J., and Rowe, W. G. (**1976**). ''Stimulus suffix effects with speech and nonspeech sounds,'' Memory Cognit. **4**, 128–131.

Semal, C., and Demany, L. (**1991**). ''Dissociation of pitch from timbre in auditory short-term memory,'' J. Acoust. Soc. Am. **89**, 2404–2410.

Semal, C., and Demany, L. (**1993**). ''Further evidence for an autonomous processing of pitch in auditory short-term memory,'' J. Acoust. Soc. Am. **94**, 1315–1322.

Surprenant, A. M., Pitt, M. A., and Crowder, R. G. (**1993**). ''Auditory recency in immediate memory,'' Q. J. Exp. Psychol. **46A**, 193–223.

Zatorre, R. J., Evans, A. C., Meyer, E., and Gjedde, A. (**1992**). ''Lateralization of phonetic and pitch discrimination in speech processing,'' Science **256**, 846–849.

Zatorre, R. J., and Samson, S. (**1991**). ''Role of the right temporal neocortex in retention of pitch in auditory short-term memory,'' Brain **114**, 2403–2417.