A robust method to study stress "deafness"[a)]

Emmanuel Dupoux[b)]
Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS), 54 Boulevard Raspail, 75006 Paris, France

Sharon Peperkamp[c)]
Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS), 54 Boulevard Raspail, 75006 Paris, France and Département de Sciences du Langage, Université de Paris VIII, 2 Rue de la Liberté, 93535 Saint Denis, France

Núria Sebastián-Gallés[d)]
Departamente de Psicologia Bàsica, Universitat de Barcelona, P. de la Vall d'Hebron, 171, 08035 Barcelona, Spain

(Received 11 September 2000; revised 16 April 2001; accepted 30 April 2001)

Previous research by Dupoux *et al.* [J. Memory Lang. **36**, 406–421 (1997)] has shown that French participants, as opposed to Spanish participants, have difficulties in distinguishing nonwords that differ only in the location of stress. Contrary to Spanish, French does not have contrastive stress, and French participants are ''deaf'' to stress contrasts. The experimental paradigm used by Dupoux *et al.* (speeded ABX) yielded significant group differences, but did not allow for a sorting of individuals according to their stress ''deafness.'' Individual assessment is crucial to study special populations, such as bilinguals or trained monolinguals. In this paper, a more robust paradigm based on a short-term memory sequence repetition task is proposed. In five French–Spanish cross-linguistic experiments, stress ''deafness'' is shown to crucially depend upon a combination of memory load and phonetic variability in $F0$. In experiments 3 and 4, nonoverlapping distribution of individual results for French and Spanish participants is observed. The paradigm is thus appropriate for assessing stress deafness in individual participants. [DOI: 10.1121/1.1380437]

PACS numbers: 43.71.Hw, 43.71.Es, 43.71.An [KRK]

I. INTRODUCTION

The way in which we perceive speech sounds depends on the properties of our native language. This phenomenon has been noticed by linguists (Sapir, 1921; Polivanov, 1931), and has been investigated by psycholinguists mostly for segmental categories. For instance, Japanese participants map the English /r/ and /l/ onto a single /ɾ/ category and have trouble discriminating between these English segments (Goto, 1971; Miyawaki *et al.*, 1981). The impact of the mother tongue on the perception of speech segments has attracted considerable attention among psycholinguists, and several models have been proposed to account for it. Some of these models propose that infants learn to focus their attention onto the acoustic cues that are most relevant for their language (Jusczyk, 1993; Nusbaum and Goodman, 1994). Other models postulate the existence of an abstract, language-specific, phoneme detector that segments the continuously varying acoustic signal into discrete categories. According to these models, non-native segments that are close enough to a segment in the native language are assimilated to it; consequently, two non-native segments that are assimilated to the same native category will be very difficult to distinguish (Best, 1994; Flege, 1995; Kuhl, 2000). Several researchers have focused on the age at which this phonological processing level is fixed (Werker and Tees, 1984; Best, McRoberts, and Sithole, 1988; Kuhl *et al.*, 1992; Jusczyk *et al.*, 1993), the extent of individual variability in late learners (Flege, MacKay, and Meador, 1999), and the possible effect of extensive training (Lively, Logan, and Pisoni, 1993; Francis, Baldwin, and Nusbaum, 2000). Hence, this line of research relates to theoretical questions regarding brain plasticity and the existence of a critical period. Moreover, it has practical implications concerning the development of training procedures for second language learning.

Languages differ not only in their repertoire of phonemes, but also in their suprasegmental properties: some use tones (Mandarin), pitch accent (Japanese), length (Finnish), or stress (Spanish) to make lexical distinctions; others do not use any of these suprasegmental properties to distinguish lexical items (French). The impact of this type of variation has been studied less extensively, and its incorporation into models of speech processing and acquisition is still awaited. As to the perception of stress, Dupoux *et al.* (1997) found that native speakers of French, a language with fixed word-final stress, have difficulties with the discrimination of nonwords that differ only in the position of stress (e.g., [vásuma] vs [vasúma] vs [vasumá]). Spanish listeners, by contrast, do not have any difficulties, stress being contrastive in their language. More research has focused on the perception of tone

[a)]Portions of this work were presented in ''Perception of stress by French, Spanish, and bilingual subjects,'' Proceedings of EuroSpeech '99, Budapest, September 1999, Vol. 6, pp. 2683–2686.
[b)]Electronic mail: dupoux@lscp.ehess.fr.
[c)]Electronic mail: sharon@lscp.ehess.fr
[d)]Electronic mail: nsebastian@psi.ub.es

by speakers of nontonal languages. For instance, it has been shown that English listeners have difficulties with the perception of Mandarin Chinese tones (Kiriloff, 1969; Bluhme and Burr, 1971; Wang *et al.*, 1999). Wang *et al.* (1999) showed that American students with one or two semesters in Mandarin correctly identified the four tones in only 69% of the cases. After extensive training, this performance jumped to 90% correct, but still only one out of the eight participants attained native-like performance. Gandour (1983) compared the perception of tone by speakers of English and four tone languages. Compared to the speakers of the latter, the English participants paid more attention to $F0$ and less to contour information in order to identify tones. Finally, Lee and Nusbaum (1993) found that Mandarin but not English listeners are slowed down by an irrelevant change in pitch level when they make a segmental classification. In other words, native speakers of Mandarin perceive pitch and segmental information in an integral fashion, whereas the two dimensions are perceived as orthogonal by native speakers of English.

In this paper, we focus on the perception of stress. Unlike some of the tonal and segmental contrasts, stress contrasts have massive acoustic correlates (duration, $F0$, and energy); it is, therefore, surprising that French participants have any problem at all with the perception of stress. Indeed, informal testing suggests that the acoustic correlates of stress are salient enough for listeners to have little difficulty in an identification paradigm similar to that used in Wang *et al.* (1999) for the perception of tones. Dupoux *et al.* (1997) reported that when tested with a standard AX discrimination task, French participants had *no* detectable problem with the stress contrast. It is only when a more demanding task was used (such as ABX with talker changes) that French participants began to have problems with stress. This suggests two things. First, unlike what happens with the perception of consonants for which within-category discrimination is very difficult, French listeners can use the acoustic stress cues in order to perform standard discrimination tasks flawlessly. Second, French listeners are nevertheless ''deaf'' to stress contrasts at a more abstract processing level, which is revealed only with tasks that are more demanding as far as memory and perceptual resources are concerned. Our aim in this paper, then, is to explore more systematically the effect of these variables in order to build a more robust paradigm to study stress ''deafness.'' Importantly, we require our paradigm to give individual results, such that the study of stress perception in special populations (for instance bilinguals, second language learners, or trained monolinguals) becomes possible.

In the ABX paradigm used in Dupoux *et al.* (1997), participants heard three successive items in three different voices, and had to judge whether the third item was identical to the first or to the second one. This task required a short-term working memory buffer because the decision had to be delayed until the final stimulus was heard. Furthermore, the stimuli A, B, and X were pronounced by three different talkers. This phonetic variability made an acoustically based response strategy more difficult to use than in a standard single-token AX paradigm. In the present study, we set up a short-term memory task and we manipulate both memory load and phonetic variability in order to find the most effective combination for a robust stress deafness effect to arise.

It should be noted that the ABX task used in Dupoux *et al.* (1997) was probably not optimal. First, the observed deafness was far from being total. For instance, in experiment 1, French participants made 20% errors in the stress contrast, whereas Spanish participants made only 4% errors. This difference was highly significant ($p<0.001$ both by items and participants), but still, the French participants performed much better than chance (50%). This might mean either that participants relied on some residual phonological representation of stress, or that the ABX paradigm was not demanding enough and allowed for alternative strategies involving an acoustic level of representation. In this paper, we evaluate whether French participants perform significantly better than chance across the several versions of our new paradigm.

Second, the results showed considerable individual variability, and an inspection of the distribution of individual errors in the stress discrimination task for French and Spanish participants revealed a substantial overlap in the distribution of errors across the two populations. That is, some French participants were as good as typical Spanish participants (three French participants made less than 5% errors), while, conversely, some Spanish participants were as bad as many of the French participants (one Spanish participant made 15% errors). This overlap might be due to the fact that some French participants succeeded in representing stress phonologically, while some Spanish failed; alternatively, it might be due to noise in the experimental method. In the present experiments, we compute the overlap in individual results across populations as well as the reliability of the effects, in order to tease apart the variation due to the participants from the variation due to the method.

To sum up, we propose to study the perception of stress contrasts in French and Spanish participants using a short-term memory task. The same task is used in all experiments. In the experiments, we compare the recall performance of a stress contrast with that of a control phonemic contrast, across different levels of memory load. Experiments differ in the amount of phonetic variability that is introduced in the stimuli tokens. In experiment 1, we use different tokens spoken by the same talker for each item. Experiments 2 and 3 add variability on the mean pitch by using speech resynthesis. Experiment 4 uses tokens from two talkers, and experiment 5 uses, as a control, a single token for each item. In all these experiments, we perform a group analysis in which we compare performance in the stress contrast to performance in the control phonemic contrast as well as to chance performance. We also analyze individual data by evaluating the overlap between the French and Spanish populations. Finally, we assess the reliability of the paradigm by computing a split-half reliability index for each experiment.

II. EXPERIMENT 1

The experiment was divided into two parts. In each part, participants were required to learn two CVCV nonwords that are a minimal pair differing only in one phonological dimen-

sion, i.e., place of articulation of the second consonant or location of stress. In each part, participants were taught to associate the two nonwords to the keys [1] and [2], respectively, of a computer keyboard. After some training with an identification task, participants listened to longer and longer random sequences of the two items, which they were required to recall and transcribe as sequences of [1] and [2]. The phonemic contrast in the first part was meant to be equally easy for speakers of French or Spanish, and was used to establish baseline performance. In order to diminish the likelihood that participants use recoding strategies, the stimuli were short and the tokens in the sequences were separated from one another by a very short interval. Moreover, in order to prevent participants from using echoic memory, every sequence was followed by the word ''OK'' (Morton, Crowder, and Prussin, 1971; Morton, Marcus, and Ottley, 1981).

In this experiment, some phonetic variability was present in that each word was instantiated by one of six acoustically different tokens. Moreover, memory load was manipulated by an increase of the sequence length from two to four and eventually six. We predicted that French participants should make many more errors in the stress than in the phoneme condition, whereas Spanish participants should have a similar performance in both conditions.

A. Method

1. Materials

Two minimal pairs were constructed, one involving a segmental contrast, i.e., [túku, túpu], and the other one involving a stress contrast, i.e., [píki, pikí]. All items are nonwords in both French and Spanish. They were recorded ten times each by a female trained phonetician who is a native speaker of Dutch. Six recordings of each word were selected. Their mean duration was 491 ms. In addition, the word ''OK'' was recorded once by a male talker. All recorded items were digitized at 16 kHz at 16 bits, digitally edited, and stored on a computer disk.

The mean durations of the tokens with the phonemic and the stress contrast were 345 and 351 ms, respectively. As to the tokens with the stress contrast, stressed vowels were on average 20 ms longer than unstressed vowels, a significant difference [$F(1,10)=14.5$, $p<0.003$]. Stressed vowels also had a higher pitch than unstressed vowels; in particular, the maximum $F0$ value of the stressed vowels was on average 45.3 Hz higher than that of the unstressed vowels, corresponding to a significant difference of 3.9 semitones [$F(1,10)=441$, $p<0.001$]. Finally, stressed vowels were on average 1.6 dB louder than unstressed vowels, again a significant difference [$F(1,10)=20.6$, $p<0.001$].

For each minimal pair three experimental blocks were constructed, each containing eight sequences of the two nonwords. The first block contained two-word sequences, the second block contained four-word sequences, and the third block contained six-word sequences. There are four logically possible two-word sequences, which each appeared twice in the first block. In the other two blocks, the eight sequences were all different. Out of the 16 possible sequences of four repetitions, eight among the most varied ones were selected, making up the second block. For instance, 1211, containing one transition from 1 to 2 and another one from 2 to 1, has more variation than 2111, containing only a transition from 2 to 1. There are six possible sequences with two transitions, which were all selected for the second block. Two sequences with one transition were selected to complete this block. Similarly, out of the 64 possible sequences of six repetitions, eight were selected for the third block; four of these contained three transitions and four contained four transitions. (The maximum number of transitions, five, gives rise to the completely regular, hence easy, patterns 121212 and 212121.) All selected sequences are listed in the Appendix. The overall design was $2\times2\times3$: language$\times$contrast $\times$sequence length.

2. Procedure

Participants were first tested on the minimal pair containing the phonemic contrast. Participants were told that they were going to learn two words in a foreign language. They could listen to the various tokens of these two words by pressing the number keys [1] and [2] as many times as they wanted. The nonword [túku] was associated to key [1], while its counterpart [túpu] was associated to key [2]. Pressing each one of these keys resulted in the playing of one token of the corresponding word. Subsequently, it was verified that participants had learned the distinction between the two words as well as the correct association between the words and the number keys. That is, they heard a token of one of the items and had to press the associated key, [1] or [2]. A message on the screen informed participants whether their responses were correct. The message was ''OK!'' or ''ERROR!,'' and was displayed for 800 ms. We defined a success criterion of five correct responses in a row. After having reached this criterion, participants turned to the main experiment.

During the test, participants listened to 24 sequences constituted by repetitions of the two words, divided into three blocks as described above. Their task was to reproduce each sequence by typing the associated keys in the correct order. For each participant, the order of the eight sequences in each block was randomized, and each item was instantiated randomly by one of the six recorded tokens. In order to diminish the likelihood that participants mentally translate the words into the associated numbers while listening to the sequence, the silent period between the items in a sequence was kept very short, i.e., 80 ms. Each trial consisted of a sequence followed by the word OK, and participants could not begin typing their response until they had heard this word. Participants did not receive feedback as to whether their responses were right or wrong. A 1500-ms pause separated each response from the next trial.

The whole procedure was repeated with the minimal pair containing a stress contrast. The nonword with stress on the first syllable, [píki], was associated to key [1], while its counterpart with stress on the second syllable, [pikí], was associated to key [2].

On average, the entire experiment lasted about 15 min. Responses were recorded on a computer disk and classified as follows. Responses that were a 100%-correct transcription

TABLE I. Percent error with phoneme and stress contrast as a function of sequence length for 12 French and 12 Spanish participants in experiment 1.

Sequence length	2	4	6	Mean
		French		
Phoneme	6.3%	11.3%	43.8%	20.5%
Stress	18.8%	38.6%	72.3%	43.2%
		Spanish		
Phoneme	8.3%	14.6%	61.5%	28.1%
Stress	1.0%	16.7%	58.3%	25.3%

of the input sequence were coded as correct; all other responses were coded as incorrect. Among the incorrect responses, those that were a 100%-incorrect transcription—i.e., with each token of the sequence labeled incorrectly—were coded as *reversals*. Participants with more reversals than correct responses in either the phoneme or the stress condition were rejected, the high percentage of reversals suggesting that they might have confused the number key associated with the first item with the one associated with the second item.

3. Participants

Twelve native French speakers, aged between 20 and 38, and 12 native Spanish speakers, aged between 18 and 21, were tested individually. None of the participants had a known hearing deficit. Two additional French participants and one additional Spanish participant were tested and excluded from the results on the basis of the rejection criterion defined above. For both French participants, the reversals outnumbered the correct responses in the case of the stress contrast, while the Spanish participant had too many reversals with the phoneme contrast.

B. Results

Error rates for French and Spanish participants for the phonemic and the stress contrast as a function of sequence length are shown in Table 1.

These data were subjected to an analysis of variance (ANOVA) with the between-participant factor language (French vs Spanish), and within-participant factors contrast (phoneme vs stress) and sequence length (2 vs 4 vs 6). As predicted on the basis of the results in Dupoux *et al.* (1997), there was a significant interaction between language and contrast $[F(1,22)=11.3,\ p<0.003]$. This interaction was due to the fact that there was an effect of contrast for French participants, with stress yielding more errors than phoneme $[F(1,11)=13.7,\ p<0.003]$, but not for Spanish participants $[F(1,11)<1,\ p>0.1]$. *Post hoc* comparisons indicated a significant effect of contrast in French participants for sequence lengths 4 and 6 (Bonferroni-corrected $p=0.024$ and $p=0.03$, respectively), but not for sequence length 2 (Bonferroni-corrected $p=0.27$). However, there was no significant interaction between length and contrast $[F(2,22)=1,\ p>0.1]$. The Spanish participants showed no significant difference between phoneme and stress at any length ($p>0.1$).

For the French participants, we compared the percentage of correct responses with the stress contrast to chance performance at each sequence length. The chance level is defined as the probability of making a *correct* response by responding randomly. According to the binomial law, it is equal to $1/2^n$ with n being the sequence length. The chance level is thus $1/4=25\%$ at length 2, $1/16=6\%$ at length 4, and $1/64=2\%$ at length 6. For the comparisons, we ran one-tailed t-tests. The comparisons were significant at each length (Bonferroni-corrected $p<0.001$). The overall performance across lengths was significantly better than chance $[F(1,11)=62.30,\ p<0.0001]$.

For each participant, the difference score is defined as the percentage of errors with the stress contrast minus the percentage of errors with the phoneme contrast. The *overlap* between populations is defined as the percentage of participants whose difference scores lie in the area that is common to both distributions. In this experiment, the overlap was 83%. Next, we compute the *optimal classification score* as follows. First, we establish an arbitrary separation criterion in the range of the observed difference scores. Second, we classify the individual participants based on their difference scores: all participants with a difference score higher than the separation criterion are classified as French, whereas all participants below the criterion are classified as Spanish. Third, we compute the percentage of participants that are correctly classified according to such a criterion. Fourth, the optimal classification score is now defined as the separation criterion that yields the best classification score. In this experiment, the optimal classification score was 79.2%. That is, if we tried to guess the mother tongue of individual participants based on their results in the experiment, we would correctly classify 79.2% of the participants as either French or Spanish.

Finally, to run a reliability test, we split the data of each participant in two halves in the following way: for each contrast and each sequence length, the first four responses of the participant were put in bin 1 and the final four responses were put in bin 2. We derived a difference score (stress minus phoneme) for these two bins, and ran a correlation analysis across participants. The correlation coefficient r between the two halves was 0.696 ($p<0.001$).

C. Discussion

As predicted, French participants had significant difficulties with the stress contrast compared to the phoneme contrast, whereas Spanish participants had equal performance with the two contrasts. The difficulty with stress in the French was significant only with length 4 and 6, although there was a numerical trend in the same direction at length 2. Note also that the interaction between sequence length and contrast was not significant, which suggests that the three sequence lengths are not qualitatively different. We also found, as in Dupoux *et al.* (1997), that the mean performance with the stress contrast was significantly better than chance; this was true at all sequence lengths.

In order to compare the robustness of the present paradigm with the ones used previously, we reanalyzed experi-

TABLE II. Separation of responses to stress discrimination in French versus Spanish participants in experiment 1 and 3 in Dupoux *et al.* (1997) and in experiment 1 of the present series.

	Stress errors Dupoux *et al.* (1997) Experiment 1	Stress-phoneme score Dupoux *et al.* (1997) Experiment 3	Stress-phoneme score Experiment 1
Number of trials	96	96	24
Number of participants	15	20	24
ANOVA	$F(1,30)=17.1$	$F(1,38)=5.7$	$F(1,22)=11.3$
Overlap of the distributions	46.8%	77.5%	83%
Optimal classification score	81.3%	65%	79.2
Split-half correlation coefficient	$r=0.895$	$r=0.435$	$r=0.696$

ments 1 and 3 of Dupoux *et al.* (1997) that used the ABX discrimination paradigm (see Table II).

As seen in Table II, the robustness of the results in the present experiment was actually weaker than that obtained in experiment 1 in Dupoux *et al.* (1997). Indeed, the overlap of the two distributions of scores for the Spanish vs the French participants was larger in the present experiment than in experiment 1 of Dupoux *et al.* Moreover, the reliability of the present experiment was also smaller. Nevertheless, the results of the present experiment were encouraging for the following reasons:

First, our experiment used only one talker with limited phonetic variability; this may have allowed the French participants to rely on acoustic information to discriminate and classify the two stress patterns. In contrast, experiment 1 and 3 in Dupoux *et al.* used three talkers. In our next experiment, we introduce more phonetic variability by way of pitch manipulation. Second, the present experiment has only 40 data points per participant, whereas experiments 1 and 3 of Dupoux *et al.* (1997) used 96 data points. This might explain the weaker reliability of our experiment. In experiments 3 and 4 of the present series, we increase the number of sequences in order to have a more comparable data set. Finally, the score that we derived in our experiment is relative to a baseline. By contrast, in experiment 1 in Dupoux *et al.*, the scores were absolute error rates for a stress contrast. Absolute scores have intrinsically lower errors of measurement than difference scores, since in the latter the error of measurement appears twice. However, the absence of a baseline induces a potential confound with population sampling biases. Indeed, irrelevant variables such as age, IQ, or motivation can obscure or erroneously increase differences in the mean performance in a task across two populations of participants. Therefore, a within-participant design with a control baseline is preferable. In fact, a comparison between experiment 3 of Dupoux *et al.* (1997), which also uses a baseline condition, and the present experiment reveals much more similar robustness and reliability of the two paradigms. Therefore, we had good reasons to hope that the present paradigm could be modified such as to become more robust and reliable than the ABX discrimination task.

III. EXPERIMENT 2

This experiment was a replication of experiment 1 with only one change: we introduced a variation in the pitch of the tokens through speech resynthesis. These changes were only plus or minus 5% of the original pitch. We expected that this modification would make it difficult for the French participants to rely on an acoustic representation. By contrast, pitch variations alone should not affect the Spanish participants too much, since Spanish speakers represent stress abstractly in the phonological representation. We thus predicted a larger difference between the two populations than in experiment 1.

A. Method

1. Materials

The same nonwords as those in experiment 1 were used. However, pitch variation was obtained in the various tokens by means of a resynthesis algorithm in the waveform editor COOL96,[1] with the percentages 105, 103, 101, 99, 97, and 95, respectively.

2. Procedure

The procedure was the same as in experiment 1, with one modification: within each sequence, a single token could not appear more than once.

3. Participants

Twelve native French speakers, aged between 20 and 40, and 12 native Spanish speakers, aged between 18 and 21, were tested individually. None of the participants had participated in experiment 1, and none had a known hearing deficit. Three additional French speakers were tested and excluded from the results, due to too many reversals among their responses with the stress contrast.

B. Results

Error rates for French and Spanish participants for the phonemic and the stress contrast as a function of sequence length are shown in Table III.

These data were subjected to an ANOVA with the between-participant factor language (French vs Spanish), and within-participant factors contrast (phoneme vs stress) and sequence length (2 vs 4 vs 6). As in the previous experiment, there was a significant interaction between language and contrast [$F(1,22)=17.3$, $p<0.001$]. The interaction was due to the fact that there was an effect of contrast for the French participants [$F(1,11)=77.8$, $p<0.001$], but not for the Spanish participants [$F(1,11)<1$, $p>0.1$]. *Post hoc* comparisons indicated a significant effect of contrast in the French partici-

TABLE III. Percent error with phoneme and stress contrast as a function of sequence length for 12 French and 12 Spanish participants in experiment 2.

Sequence length	2	4	6	Mean
		French		
Phoneme	11.3%	32.5%	71.3%	38.4%
Stress	58.8%	73.8%	97.5%	76.7%
		Spanish		
Phoneme	8.7%	25.0%	63.5%	32.4%
Stress	13.5%	26.9%	74.0%	38.1%

pant at all sequence lengths (Bonferroni-corrected p values: 0.009, 0.001, and 0.009 for sequence lengths 2, 4, and 6, respectively). There was no significant interaction between sequence length and contrast for the French participants [$F(2,22)<1$, $p>0.1$]. The Spanish participants showed no significant difference at any length ($p>0.1$). For the French participants, we ran a t-test at each sequence length to compare the performance in the stress condition to chance performance. The comparison was significant for lengths 2 and 4, but not for length 6 (Bonferroni-corrected one-tailed $p<0.016$, $p<0.0015$, and $p>0.2$, respectively). The overall performance across lengths was significantly better than chance [$F(1,11)=16.51$, $p<0.002$].

The overlap between the two populations was 62%, with an optimal classification score of 71.7%. As for the reliability test, the correlation coefficient r between the two halves of the experiment was 0.566 ($p<0.004$).

C. Discussion

In this experiment, we introduced a change in the pitch of the experimental tokens varying between -5% and $+5\%$. The group results are very similar to those in experiment 1, except that the stress deafness effect for the French participants, i.e., the difference in performance between the stress and the phoneme condition, is numerically larger and now significant even at sequence length 2. Yet, French participants are still better than chance with the stress contrast at sequence lengths 2 and 4. Finally, compared to experiment 1, the distribution of scores between the French and the Spanish participants is more separated. This is shown by the fact that the classification error is divided by 2 and the overlap between the two distributions is reduced. Hence, the introduction of a pitch change was successful in making more difficult an acoustically based strategy for the French participants, thus increasing the size of the language-specific effect.

Note, however, that there was one Spanish participant who made a great number of errors in the stress contrast. This outlier was responsible for the high degree of overlap in the two distributions of scores that remains in this experiment. After an interview with this participant, it turned out that one of the [pikí] tokens was perceived as a little ambiguous in terms of stress. Therefore, we decided to run a replication of this experiment using a novel set of stimuli. Another point worth noting is that the reliability of this experiment was not good ($r=0.566$). In the next experiment, we also added more sequences, in order to get more stable individual data.

IV. EXPERIMENT 3

In this experiment, we introduced a novel set of stimuli that was better controlled for stress. In experiment 2, the minimal pairs consisted of nonwords having identical vowels in the two syllables ([túku, túpu] and [píki, pikí]). This might have made the stimuli confusable in short-term memory. In order to test whether the obtained effects generalize to a situation with a different vowel in the two syllables, we constructed two new minimal pairs, consisting of nonwords having different vowels in the two syllables. We also increased the power of this experiment by adding sequences of length 3 and 5. As in experiment 2, we introduced a $\pm 5\%$ change in pitch in the different tokens.

A. Method

1. Materials

Two minimal pairs were constructed, one involving a segmental contrast, i.e., [kúpi, kúti], the other one involving a stress contrast, i.e., [mípa, mipá]. All items are nonwords in both French and Spanish. They were recorded about ten times by a female talker, the same one who recorded the items used in experiments 1 and 2. The stimuli were judged by a Spanish phonetician and only tokens with unambiguous stress patterns were used. Six recordings of each item were selected. All recorded items were digitized at 16 kHz at 16 bits, digitally edited, and stored on a computer disk.

The mean durations of the tokens with the phonemic and the stress contrast were 439 and 513 ms, respectively. As in experiment 2, more variation was obtained in the six tokens of the four items, in that they had their pitch changed by means of the COOL96 waveform editor, with the percentages 105, 103, 101, 99, 97, and 95, respectively. As to the tokens with the stress contrast, stressed vowels were on average 21 ms longer than unstressed vowels, a significant difference [$F(1,5)=43.20$, $p<0.001$]. Stressed vowels also had a higher pitch than unstressed vowels; in particular, the maximum $F0$ value of the stressed vowels was on average 52.6 Hz higher than that of the unstressed vowels, corresponding to a significant difference of 4.9 semitones [$F(1,10)=521$, $p<0.001$]. Finally, stressed vowels were on average 3.7 dB louder than unstressed vowels, again a significant difference [$F(1,5)=78.20$, $p<0.001$].

Two blocks, containing eight three-word sequences and eight five-word sequences, respectively, were added. Thus, there were five test blocks for each contrast, containing sequences of length 2, 3, 4, 5, and 6. There are exactly eight logically possible three-word sequences, all of which appeared once in the second block. Out of the 32 possible sequences of five repetitions, eight among the most varied ones were selected for the fourth block; half of them contained two transitions; the other half contained three transitions (see the Appendix). As to the remaining blocks with sequences of length 2, 4, and 6, the same sequences as in experiments 1 and 2 were selected. The overall design was $2\times2\times5$: language$\times$contrast$\times$sequence length.

2. Procedure

The procedure was as in experiment 2, with the following modifications. First, the stimuli were presented in a different manner. That is, participants were first asked to press the number key [1], upon which they heard all tokens of the first item. They were then asked to press the number key [2], upon which they heard all tokens of the second item. Subsequently, participants could continue listening to the various tokens of the two items by pressing the associated keys; as in the previous experiments, pressing each one of these keys resulted in the playing of one token of the corresponding item. They could thus hear as many tokens of the two items as they desired. Second, in the training phase, we increased the number of correct responses in the success criterion from five to seven. Third, in order to diminish the amount of noise in the results, participants were warned whenever they entered a sequence with a length that did not correspond to the length of the input string and asked to enter their reply again.

For the phoneme contrast, [kúpi] was associated with key [1] and [kúti] with key [2]. For the stress contrast, [mípa] was associated with key [1] and [mipá] with key [2].

On average, the experiment lasted between 15 and 20 min. Responses were recorded on a computer disk.

3. Participants

Twelve native French speakers, aged between 20 and 42, and 12 native Spanish speakers, aged between 23 and 29, were tested individually. None of the participants had participated in experiments 1 or 2, and none had a known hearing deficit. Two additional Spanish speakers were tested and excluded from the results, due to too many reversals among their responses. For one of these participants this was the case in the phoneme condition, and for the other one it was in the stress condition.

B. Results

Error rates for French and Spanish participants for the phonemic and the stress contrast as a function of sequence length are shown in Table IV.

These data were subjected to an ANOVA with the between-participant factor language (French vs Spanish), and within-participant factors contrast (phoneme vs stress) and sequence length (2 vs 3 vs 4 vs 5 vs 6). As in experiments 1 and 2, there was a significant interaction between language and contrast [$F(1,22)=70.3$, $p<0.0001$]. This interaction was due to the fact that stress yielded significantly more errors than phoneme for French participants [$F(1,11)=71.0$, $p<0.0001$], whereas there was a nonsignificant trend in the other direction for Spanish participants [$F(1,11)=3.7$, $0.1>p>0.05$]. *Post hoc* comparisons indicated a significant effect of contrast in French participants for all sequence lengths (Bonferroni-corrected p values: 0.035, 0.015, 0.001, 0.001, and 0.001 for sequence lengths 2, 3, 4, 5, and 6, respectively). The interaction between sequence length and contrast was significant for French participants [$F(4,44)=7.09$; $p<0.001$]. Spanish participants showed no significant difference at any length ($p>0.1$).

TABLE IV. Percent error with phoneme and stress contrast as a function of sequence length for 12 French and 12 Spanish participants in experiment 3.

Sequence length	2	3	4	5	6	Mean
			French			
Phoneme	2.8%	7.3%	15.6%	18.7%	33.3%	15.4%
Stress	29.2%	28.1%	59.4%	64.6%	86.5%	53.5%
			Spanish			
Phoneme	7.3%	5.2%	12.5%	35.4%	61.5%	24.4%
Stress	0.0%	4.2%	10.4%	32.3%	53.1%	20.0%

For French participants, we ran a t-test at each sequence length to compare the performance with the stress contrast to chance performance. This comparison was significant at each length (Bonferroni-corrected one-tailed $p<0.001$), except at length 6 where the comparison was only marginally significant ($p=0.066$). The overall performance across lengths was significantly better than chance [$F(1,11)=37.04$, $p<0.001$].

The overlap between the two populations was 0%, with an optimal classification score of 100%. The two distributions of points are actually separated by a gap whose size is 9.4% of the total range. As for the reliability test, the correlation coefficient r between the two halves of the experiment was 0.882 ($p<0.001$).

In this experiment, the number of training trials was saved. Note that due to a programming error, the training criterion was seven correct answers in a row for the French but only six for the Spanish participants. For the French participants, the number of training trials was 7.3 for the stress and 7.1 for the phoneme condition, a nonsignificant difference [$F(1,11)<1$, $p>0.1$]. For the Spanish participants, the number of training trials was 6.1 for the stress and 6.7 for the phoneme condition, again, a nonsignificant difference [$F(1,11)=1.31$, $p>0.1$]. In other words, most French and Spanish participants passed the training without any error. In a global ANOVA, there was no significant interaction between contrast and language [$F(1,22)=2.02$, $p>0.1$].

C. Discussion

This experiment replicated experiment 2 using novel and more controlled stimuli, as well as more sequence lengths. The group analysis was very similar to the one in experiment 2: we found a stress deafness effect in the French participants and not in the Spanish participants, and the effect was significant at all sequence lengths, while performance for the stress contrast was better than chance at all sequence lengths (except at length 6, where it was only marginal). In the individual analyses, however, the results were stronger than in experiment 2: there was no overlap in the distributions of the French and Spanish populations and the split-half reliability of the test was quite good.

The fact that the French participants were still able to perform the memory task with the stress contrast better than chance can be interpreted in two ways: either they managed to apply an acoustic strategy with the stress contrast, or they have a residual phonological representation of stress, which would mean that the deafness is not total. In order to distinguish between these two hypotheses, we introduced more

phonetic variability by using two talkers in our next experiment. If the residual capacity to distinguish stress is based on an acoustic strategy, it should be harder to apply such a strategy on tokens with increased phonetic variability, and hence the size of the deafness effect should increase as well.

V. EXPERIMENT 4

This experiment was a replication of experiment 3 with even more phonetic variability, in that we introduced a second, male, talker. The big difference in $F0$ is predicted to make it more difficult for the participants to use acoustic information. For the Spanish participants, this should not be a problem in either the phoneme or the stress condition, since they can use their phonological representations of the target items to perform the task. The same holds for the French participants in the phoneme condition. By contrast, the French participants should have increased problems in the stress condition if the residual performance that we observed in the previous experiments were due at least in part to acoustic strategies. In brief, we predict that if stress discrimination in the French participants is acoustically based, the introduction of a new talker should increase the size of the language-specific effect measured previously, and the difference scores of the two populations should be even further apart than in experiment 3.

A. Method

1. Materials

The same minimal pairs as in experiment 3 were used, but a new recording with a male voice was made. This second talker was a native speaker of French, who imitated the tokens produced by the female talker. A trained phonetician corrected his productions till they were deemed satisfactory. For each of the four test items ([kúpi], [kúti], [mípa], and [mipá]), three tokens from the male talker were selected. These tokens replaced three of the six tokens per item of the female talker used in experiment 3. A total of six tokens per item was thus obtained: three produced by the female talker and three produced by the male talker. Given that the tokens produced by the male talker were shorter than those produced by the female talker, they were stretched such that they matched exactly the length of the tokens from the female talker they replaced.[2]

As to the items for the stress contrast, stressed vowels had a higher pitch than unstressed vowels; in particular, the maximum $F0$ value of the stressed vowels was on average 48.4 Hz higher than that of the unstressed vowels, corresponding to a significant difference of 5.2 semitones [$F(1,10)=304$, $p<0.001$]. Moreover, stressed vowels were on average 6.0 dB louder than unstressed vowels, again a significant difference [$F(1,10)=45.7$, $p<0.001$].

The mean $F0$ of the tokens of the female talker was 181 Hz, and the mean $F0$ of the tokens of the male talker was 144 Hz; hence, there was an $F0$ variation between the two talkers of 20%. As in experiment 2 and 3, $\pm 5\%$ variation in pitch was obtained, in that the three tokens of both sets had their pitch changed by means of the PSOLA algorithm, with the percentages 105, 101, and 95, respectively. The design was the same as in experiment 3.

2. Procedure

The same procedure as in experiment 3 was used.

3. Participants

Twelve native French speakers, aged between 18 and 29, and 12 native Spanish speakers, aged between 23 and 28, were tested individually. None of the participants had participated in the previous experiments, and none had a known hearing deficit. One Spanish participant had to be replaced due to too many complete reversals among his responses in the phoneme condition.

B. Results

Error rates for French and Spanish participants for the phonemic and the stress contrast as a function of sequence length are shown in Table V.

These data were subjected to an ANOVA with the between-participant factor language (French vs Spanish), and within-participant factors contrast (phoneme vs stress) and sequence length (2 vs 3 vs 4 vs 5 vs 6). The interaction between language and contrast was highly significant [$F(1,22)=61.6$, $p<0.0001$]. The interaction was due to the fact that stress yielded significantly more errors than phoneme with the French participants [$F(1,11)=81.1$, $p<0.0001$], whereas there was a nonsignificant trend in the other direction with the Spanish participants [$F(1,11)=3.5$, $0.1>p>0.05$]. *Post hoc* comparisons indicate a significant effect of contrast with the French participants for all sequence lengths (Bonferroni-corrected p values: 0.025, 0.001, 0.001, 0.001, and 0.01 for sequence length 2, 3, 4, 5, and 6, respectively). The sequence length by contrast interaction for the French participants was not significant [$F(4,44)=2.28$, $p=0.075$]. The Spanish participants showed no significant difference at any length ($p>0.1$). For the French participants, we ran a t-test at each sequence length to compare the performance in the stress condition to chance performance. The comparison was significant at lengths 2, 3, and 4 (Bonferroni-corrected one-tailed $p<0.003$, $p<0.003$, $p<0.01$, respectively) but not at lengths 5 and 6 (Bonferroni-corrected one-tailed $p=0.2$ and $p>0.2$, respectively). The overall performance across lengths with the stress contrast was significantly better than chance [$F(1,11)=22.94$, $p<0.001$].

The overlap between the two populations was 0%, with an optimal classification score of 100%. The distributions were actually separated by a gap whose size was 2.7% of the total range. As to the reliability test, the correlation coefficient r between the two halves of the experiment was 0.82 ($p<0.001$).

In this experiment, the number of training trials before criterion was saved. For the French participants, the number of training trials was 23.6 for the stress condition and 11.2 for the phoneme condition, a nonsignificant difference [$F(1,11)=3.10$, $p>0.1$]. The bulk of the difference was due to two participants who had more than 50 trials in the stress condition. For the Spanish participants, the number of training trials was 9.6 for the stress condition and 8.8 for the phoneme condition, again, a nonsignificant difference

TABLE V. Percent error with phoneme and stress contrast as a function of sequence length for 12 French and 12 Spanish participants in experiment 4.

Sequence length	2	3	4	5	6	Mean
			French			
Phoneme	7.3%	5.2%	21.9%	50.0%	63.5%	29.5%
Stress	34.4%	48.9%	75.0%	88.5%	94.8%	68.3%
			Spanish			
Phoneme	6.2%	9.4%	31.2%	56.2%	71.9%	35.0%
Stress	8.3%	13.5%	19.8%	38.5%	57.3%	27.5%

$[F(1,11)<1,\ p>0.1]$. In a global ANOVA, the interaction between contrast and language was only marginal $[F(1,22)=3.4,\ 0.1>p>0.05]$.

C. Discussion

This experiment was identical to experiment 3 except that two talkers were used instead of only one. This introduced a threefold increase in $F0$ variability across tokens, as well as other differences due to timbre, and fine phonetic variations of the segments and of suprasegmental information. Yet, this experiment replicated almost exactly experiment 3. In particular, the stress ''deafness'' effect was not stronger with two talkers than with one talker. The overall interaction between language and contrast, the overlap between the French and Spanish participants, as well as the reliability score, were all very similar in the two experiments. The only difference was that the overall error rate was slightly higher in this experiment than in experiment 3 (40% instead of 28%); this difference was probably due to the fact that the change in talker made the memory task more difficult (see Nygaard, Sommers and Pisoni, 1995). Consequently, the most difficult conditions, i.e., sequence lengths 5 and 6 with the stress contrast for the French participants, were at chance level.

In the last experiment, we tested whether memory load alone, without any phonetic variability, is sufficient to induce a stress deafness effect.

VI. EXPERIMENT 5

In this experiment, we tested whether French participants still display a stress deafness when there is no phonetic variability at all. That is, we used only a single token for each of the test items. With a single token, it is in principle possible to encode a sequence of stimuli in terms of ''same'' and ''different,'' these two categories being definable acoustically. Assuming that participants have access to same/different judgments at the acoustic level, and that they can encode this information in short-term memory in keeping with the presentation rate, we expected that they should be able to perform the task even with contrasts that are nonnative. Hence, we predicted that French and Spanish participants would have equal performance on the task with both the phonemic and the stress contrast.

A. Method

1. Materials

The same minimal pairs as in experiment 4 were used, but only a single token from the female voice for each item was used. The design was the same as in experiment 4.

2. Procedure

The same procedure as in experiment 4 was used.

3. Participants

Twelve native French speakers, aged between 17 and 50, and 12 native Spanish speakers, aged between 23 and 29, were tested individually. None of the participants had participated in one of the previous experiments, and none had a known hearing deficit.

B. Results

Error rates for French and Spanish participants for the phonemic and the stress contrast as a function of sequence length are shown in Table VI.

These data were subjected to an ANOVA with the between-participant factor language (French vs Spanish), and within-participant factors contrast (phoneme vs stress) and sequence length (2 vs 3 vs 4 vs 5 vs 6). The interaction between language and contrast was not significant $[F(1,22)<1,\ p>0.1]$. A separate analysis for the French and the Spanish participants revealed only an effect of sequence length for both groups $[F(4,44)=30.3,\ p<0.001,$ and $F(4,44)=36.0,\ p<0.001,$ respectively]. There was no significant effect of contrast either globally $[F(1,22)<1,\ p>0.1]$ or for individual sequence lengths $[p>0.1]$.

For the French participants, we ran a t-test at each sequence length to compare the performance in the stress condition to chance performance. The comparison was significant at each length (Bonferroni-corrected one-tailed $p<0.001$ at every length). The overall performance across lengths was significantly better than chance $[F(1,11)=131.89,\ p<0.0001]$.

The overlap between the two populations was 95.8%. The optimal classification score was 54.2%. As to the reliability test, the correlation coefficient r between the two halves of the experiment was 0.55 $(p<0.007)$.

In this experiment, the number of training trials before criterion was saved. For the French participants, the number of training trials was 10.1 for the stress and 9.7 for the phoneme condition, a nonsignificant difference $[F(1,11)<1,\ p>0.1]$. For the Spanish participants, the number of training trials was 8.9 for the stress and 8.3 for the phoneme condition, again, a nonsignificant difference $[F(1,11)<1, p>0.1]$. In a global ANOVA, there was no significant interaction between contrast and language $[F(1,22)>1, p>0.1]$.

C. Discussion

This experiment demonstrates that with no phonetic variability, no stress ''deafness'' emerges in the French participants even when the memory load is high. Note that although the performance in this experiment was overall better

TABLE VI. Percent error with phoneme and stress contrast as a function of sequence length for 12 French and 12 Spanish participants in experiment 5.

Sequence length	2	3	4	5	6	Mean
			French			
Phoneme	7.3%	6.2%	12.5%	39.6%	51.0%	23.3%
Stress	8.3%	12.5%	16.7%	38.5%	57.3%	26.7%
			Spanish			
Phoneme	3.1%	1.0%	7.3%	27.1%	47.9%	17.3%
Stress	1.0%	4.2%	12.5%	31.2%	47.9%	19.4%

than in the experiments with greater variability, we cannot explain the lack of a cross-linguistic difference by a ceiling effect. Indeed, even with sequence lengths matched in difficulty with those of the previous experiments (for instance, sequences 4 to 6 in the present experiment versus sequences 3 to 5 in experiment 4), no difficulty with stress emerged in French participants. This replicates and extends the findings in Dupoux *et al.* (1997), where it was reported that the problem of the French participants to discriminate stress disappears in an AX paradigm with no phonetic variability. What the current experiment shows is that the lack of phonetic variability is sufficient to make the stress deafness effect disappear; in particular, the presence of a high memory load alone does not induce the stress deafness effect.

VII. GENERAL DISCUSSION

In a series of five experiments, we demonstrated that the stress deafness effect in French listeners can be replicated with a new paradigm. Moreover, we have substantially improved the methodology, in that the results can be interpreted on an individual basis. In the following discussion we address three issues. First of all, we discuss the effect of memory load and phonetic variability on the group results. We then discuss the residual capacity of French participants to perceive the stress contrast. Finally, we consider the broader implications of our findings regarding the effects of the native language on speech perception.

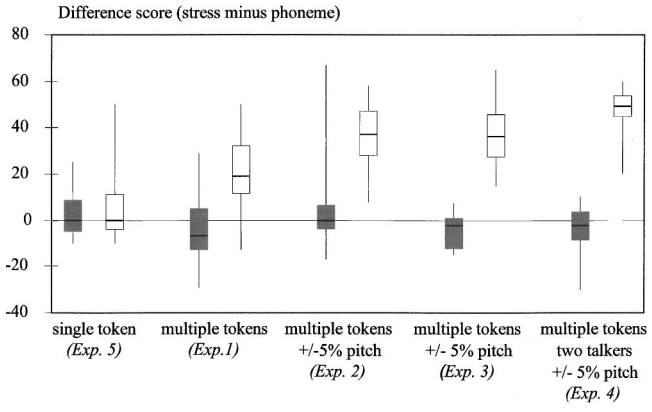

FIG. 1. Distribution of difference scores (phoneme minus stress) as a function of phonetic variability in five experiments. The Spanish difference scores are in gray, and the French in white. The minimum and maximum scores are indicated by the bottom and top of the vertical lines, respectively, the median scores by the thick horizontal lines, and the boxes contain scores that fall within the 25%–75% percentile.

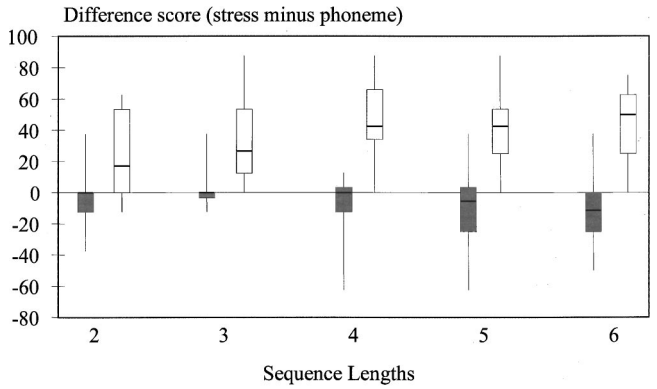

FIG. 2. Distribution of difference scores (phoneme minus stress) in experiments 3 and 4 as a function of sequence length. The Spanish difference scores are in gray, and the French in white. The minimum and maximum scores are indicated by the bottom and top of the vertical lines, respectively, the median scores by the thick horizontal lines, and the boxes contain scores that fall within the 25%–75% percentile.

A. Memory load and phonetic variability

Our findings can be summarized in three points. First, the size of the ''deafness'' effect increases with the amount of phonetic variability. The effect of phonetic variability, however, seems to reach a plateau (see Fig. 1). Specifically, we found that adding $\pm 5\%$ pitch variations on stimuli produced by a single talker (experiments 2 and 3) yielded a similar effect as adding this variation on stimuli produced by two talkers, one male and one female (experiment 4).

Second, although memory load had a strong effect on mean performance in that shorter sequences yielded less errors than longer ones, the stress-phoneme difference score in the French participants across experiments 1 to 4 was found to have roughly the same size at sequence length 2 and at sequence length 6 (means of 28.3% and 34.8%, respectively). In fact, if anything, the effect seemed numerically larger at sequence length 4 (mean of 41.3%). This might be due to the fact that error scores were squeezed by a floor effect at length 2 and a ceiling effect at length 6 (see Fig. 2). Note, however, that with sequences of size 1, as used in the training session, there was no longer a significant difference between stress and phonemes in French participants. In other words, it is the presence of memory load that matters, not the amount of it.

Third, phonetic variability and memory load displayed an interaction. On the one hand, in experiment 5 with zero phonetic variability, no deficit with the stress contrast was found in the French participants, not even with sequences of size 6. On the other hand, in the training sessions (sequence length 1), we found no measurable problem to master the stress contrast, not even if there was a high phonetic variability as in experiment 4 (two talkers, extra pitch variation). Hence, it is only in the presence of both factors that the ''deafness'' effect emerges.

These findings can be interpreted as follows. The sequence repetition task requires participants to encode the information in their short-term memory buffer in order to recall the sequence. Several coding strategies are available according to the task demands and stimulus characteristics.

TABLE VII. Coding strategies available to Spanish and French participants as a function of memory load and phonetic variability.

	Spanish participants	French participants
Sequence 1, no variability (training of Expt. 5)	Explicit categorization Acoustic mismatch Phonological coding	Explicit categorization Acoustic mismatch
Sequence 1, pitch variability (training of Expt. 2–4)	Explicit categorization Phonological coding	Explicit categorization
Sequence 2–6, no variability (Expt. 5)	Acoustic mismatch Phonological coding	Acoustic mismatch
Sequence 2–6, pitch variability (Expt. 2–4)	Phonological coding	No strategy available

We distinguish two acoustic and one phonological strategy (see Table VII). A first coding strategy is based on the fact that stress has massive acoustic correlates. Given enough time, participants can use these cues to explicitly categorize the stimuli, for instance, by comparing them to stored exemplars, by focusing on the melody, or by repeating the token and monitoring for one's own pitch. This strategy is not automated and, therefore, cannot be used in case of long and rapid sequences of stimuli. A second coding strategy is based on the acoustic mismatch signal. This signal has been shown to be automatically generated in the auditory cortex whenever an acoustic signal differs from its predecessor(s) (Näätanen *et al.*, 1997). This acoustic mismatch might be used to recode sequences of stimuli in terms of same/different. The last token can then be explicitly categorized in order to reconstruct the underlying sequence. Of course, such a strategy only works in the absence of phonetic variability. When the tokens are phonetically varied, even sequences of ''same'' tokens give rise to a mismatch signal. A final coding strategy uses the automatic encoding which is provided by the language-specific phonological representation. In the case of stress, such phonological encoding is of course available to the Spanish but not to the French participants participants, since stress is not encoded phonologically in the latter language.

Our analysis of the available strategies has a practical impact regarding how to construct a paradigm that will show maximal cross-linguistic sensitivity. In order to eliminate nonphonological compensatory strategies it is essential to limit the amount of time that subjects have available. In all of our experiments, the duration of the tokens was rather short, and the interstimulus interval, ISI, between them was only 80 ms. Although we have not explicitly varied this variable, informal testing reveals that if participants were given more time—by having either longer items or a longer ISI—they would be able to explicitly recode the tokens (as they do in the training phase, where they have all the time necessary to produce a response). This, then, would have the effect of reducing the size of the cross-linguistic difference.

B. Residual capacity to discriminate stress

In all our experiments, we found that French participants, although they made massively more errors with the stress than with the phoneme contrast, were still significantly better than chance. This was true even for all the individual sequence lengths except 6, which was not significantly different from chance in experiments 2, 3, and 4. Note, however, that it is a matter of experimental power; that is, when these three experiments are pooled together, the performance at sequence length 6 is still better than chance $[t(35) = 2.90, p < 0.006]$.

We would like to discuss three possible interpretations of this residual effect. First, an acoustic strategy might survive both long sequence lengths and phonetic variability. This is somewhat unlikely, since, for instance, sequences of length 6 last for about 4 s, which is longer than the span for the echoic store (Guttman and Julesz, 1963; Crowder and Morton, 1969). Second, there might be a residual encoding of stress within the phonological representation itself. This encoding should be much weaker or less accessible in French than in Spanish participants. One possible reason why French participants would not entirely ignore stress is that they use stress in word segmentation. In French, stress falls on the word's final nonschwa syllable. Items with word-initial stress like [píki] might therefore be perceived as containing a word boundary after the first syllable. Consequently, [píki] and [pikí], although identical in their segmental contents, would differ in their perceived segmentation pattern.[3] Third, there might be another type of representation, i.e., a phonetic representation, which can be used to encode the stimuli in a language-universal fashion, but yet in a more abstract format than the acoustic representation. Goldinger (1998) has shown, for instance, that in an immediate repetition task, participants can imitate certain phonetic details of stimuli that are not used phonologically in their native language. More research is needed to tease apart these various interpretations.

C. Implications for future research

Our results raise a number of issues regarding the influence of the maternal language on speech perception. We found that French speakers have difficulties with distinguishing a stress contrast, but only with phonetically variable stimuli, high memory load, and limited time.

This suggests that, first, the extent of influences of the maternal language might be underestimated when using standard discrimination procedures for which participants have much time to perform the task, and which usually do not introduce phonetic variability. It is, therefore, important to study speech processing under constrained resources, in order to limit the possibility for participants to use nonlinguistic response strategies.

Second, our findings raise the issue of the specificity of suprasegmental features. Indeed, we showed that for French participants, acoustically based response strategies are readily available. This is clearly not the case for the perception of non-native consonantal contrasts (see Goto, 1971). It remains to be investigated whether the availability of acoustically-based response strategies depends upon the size of the acoustic correlates, and whether it extends to other suprasegmental contrasts.

Third, our findings raise the issue of the linguistic and developmental conditions under which stress ''deafness'' arises during language acquisition. Do all languages with noncontrastive stress yield stress ''deafness'' (Dupoux and Peperkamp, in press; Peperkamp and Dupoux, in press)? Conversely, can speakers of languages with contrastive stress exhibit a stress ''deafness'', provided the number of minimal stress pairs in their language is small (Cutler, 1986)?

A final question is whether native speakers of French who learn Spanish can acquire the perception of stress, and whether age of acquisition matters in this respect (Flege, MacKay, and Meador, 1999; Flege, Schmidt, and Wharton, 1996; Pallier, Bosch, and Sebastián-Gallés, 1997; Peperkamp, Dupoux, and Sebastián-Gallés, 1999).

In brief, the existence of stress ''deafness'' raises a number of central issues in speech perception and language acquisition. The availability of an easy to use and reliable methodology allows us to address these issues in the near future.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the Fyssen Foundation to the second author, by the Groupement d'Interet Scientifique: Sciences de la Cognition, and by the Spanish Ministerio de Educación y Cultura, Grant No. PB97-0977. This project benefited from many discussions with Jacques Mehler, who also invented the term stress ''deafness.'' We thank Katherine White and Mar Rodríguez for help in running participants, and Laura Bosch, Anne Christophe, Christophe Pallier, Franck Ramus, and two anonymous reviewers for comments and discussion.

APPENDIX: SEQUENCES USED IN THE EXPERIMENTS

Two-word sequences: (experiments 1–5) 11, 12, 21, 22, 11, 12, 21, 22.

Three-word sequences: (experiments 3, 4, 5) 111, 112, 121, 122, 211, 212, 221, 222.

Four-word sequences: (experiments 1–5) 1121, 1122, 1211, 1221, 2111, 2112, 2122, 2212.

Five-word sequences: (experiments 3, 4, 5) 11121, 12112, 12122, 12211, 21112, 21211, 21221, 22122.

Six-word sequences: (experiments 1–5) 112121, 112212, 121112, 122121, 211221, 212112, 221212, 222121.

[1]Syntrillium Software Corporation (www.synrtillium.com).

[2]We used the compression algorithm in the PRAAT software (www.fon.hum.uva.nl/praat/).

[3]Another possibility stems from the fact that word stress is final in French. Stress-final items could be perceived as more prototypical than stress-initial ones. Prototypicality could then be used to distinguish the two sets of items. However, Dupoux *et al.* (1997) compared discrimination of [vasúma] vs [vasumá] (where the latter is the more prototypical one) with that of [vásuma] vs [vasúma] (where the two are equally nonprototypical), and failed to find a difference.

Best, C., McRoberts, G., and Sithole, N. (**1988**). ''Examination of perceptual reorganization for non-native speech contrasts; Zulu click discrimination by English-speaking adults and infants,'' J. Exp. Psychol. Hum. Percept. Perform. **14**, 345–360.

Best, C. (**1994**). ''The emergence of native-language phonological influence in infants; A perceptual assimilation model,'' in *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, edited by J. Goodman and H. Nusbaum (MIT Press, Cambridge), pp. 167–224.

Bluhme, S., and Burr, R. (**1971**). ''An audio-visual display of pitch for teaching Chinese tones,'' Studies Ling. **22**, 51–57.

Crowder, R.G., and Morton, J. (**1969**). ''Precategorical acoustic store (PAS),'' Percept. Psychophys. **5**, 365–373.

Cutler, A. (**1986**). ''*Forbear* is a homophone: Lexical prosody does not constrain lexical access,'' Lang. Speech **29**, 201–220.

Dupoux, E., Pallier, C., Sebastián-Gallés, N., and Mehler, J. (**1997**). ''A destressing 'deafness' in French?'' J. Memory Lang. **36**, 406–421.

Dupoux, E., and Peperkamp, S. (**in press**). ''Fossil markers of language development: Phonological 'deafnesses' in adult speech processing,'' in *Cognitive Phonology*, edited by J. Durand and B. Laks (Oxford University Press, Oxford).

Flege, J.E. (**1995**). ''Second language speech learning: Theory, findings, and problems,'' in *Speech Perception and Linguistic Experience*, edited by W. Strange (York, Baltimore), pp. 233–272.

Flege, J., MacKay, I., and Meador, D. (**1999**). ''Native Italian speakers' perception and production of English vowels,'' J. Acoust. Soc. Am. **106**, 2973–2987.

Flege, J., Schmidt, A., and Wharton, G. (**1996**). ''Age of learning affects rate-dependent processing of stops in a second language,'' Phonetica **53**, 143–161.

Francis, A., Baldwin, K., and Nusbaum, H. (**2000**). ''Effects of training on attention to acoustic cues,'' Percept. Psychophys. **62**, 1668–1680.

Gandour, J.T. (**1983**). ''Tone perception in Far Eastern languages,'' J. Phonetics **11**, 149–175.

Goldinger, S.D. (**1998**). ''Echoes of echoes? An episodic theory of lexical access,'' Psychol. Rev. **105**, 251–279.

Goto, H. (**1971**). ''Auditory perception by normal Japanese adults of the sounds 'l' and 'r','' Neuropsychologia **9**, 317–323.

Guttman, N., and Julesz, B. (**1963**). ''Lower limits of auditory periodicity analysis,'' J. Acoust. Soc. Am. **35**, 610.

Jusczyk, P.W. (**1993**). ''From general to language-specific capacities: the WRAPSA Model of how speech perception develops,'' J. Phonetics **21**, 3–28.

Jusczyk, P.W., Friederici, A.D., Wessels, J.M.I., Svenkerud, V.Y., and Jusczyk, A.M. (**1993**). ''Infants' sensitivity to the sound patterns of native language words,'' J. Memory Lang. **32**, 402–420.

Kiriloff, C. (**1969**). ''On the auditory perception of tones in Mandarin,'' Phonetica **20**, 63–67.

Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., and Lindblom, B. (**1992**). ''Linguistic experience alters phonetic perception in infants by 6 months of age,'' Science **255**, 606–608.

Kuhl, P.K. (**2000**). ''Language, mind, and brain: Experience alters perception,'' in *The New Cognitive Neuroscience*, edited by M. Gazzaniga (MIT Press, Cambridge, MA), pp. 99–114.

Lee, L., and Nusbaum, H. (**1993**). ''Processing interactions between segmental and suprasegmental information in native speakers of English and Mandarin Chinese,'' Percept. Psychophys. **53**, 157–165.

Lively, S.E., Logan, J.S., and Pisoni, D.B. (**1993**). ''Training Japanese listeners to identify English /r/ and /l/. II. The role of phonetic environment and talker variability in learning new perceptual categories,'' J. Acoust. Soc. Am. **94**, 1242–1255.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A., Jenkins, J., and Fujimura, O. (**1981**). ''An effect of linguistic experience; The discrimination of /r/ and /l/ by native speakers of Japanese and English,'' Percept. Psychophys. **18**, 331–340.

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

Morton, J., Crowder, R., and Prussin, H. (**1971**). ''Experiments with the stimulus suffix effect,'' J. Exp. Psychol. **91**, 169–190.
Näätanen, R., Lehtokovski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R., Luuk, A., Allik, J., Sinkkonen, J., and Alho, K. (**1997**). ''Language-specific phoneme representations revealed by electric and magnetic brain responses,'' Nature (London) **385**, 432–434.
Nusbaum, H., and Goodman, J. (**1994**). ''Learning to hear speech as spoken language,'' in *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*, edited by H. Nusbaum and J. Goodman (MIT Press, Cambridge, MA), pp 299–338.
Nygaard, L.C., Sommers, M.S., and Pisoni, D. (**1995**). ''Effects of stimulus variability on perception and representation of spoken words in memory,'' Percept. Psychophys. **57**, 989–1001.
Pallier, C., Bosch, L., and Sebastián-Gallés, N. (**1997**). ''A limit on behavioral plasticity in speech perception,'' Cognition **64**, B9–B17.
Peperkamp, S., Dupoux, E., and Sebastián-Gallés, N. (**1999**). ''Perception of stress by French, Spanish, and bilingual subjects,'' Proceedings of Euro-Speech '99, Vol. 6, 2683–2686.
Peperkamp, S., and Dupoux, E. (**in press**). ''A typological study of stress 'deafness','' in *Laboratory Phonology VII*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin).
Polivanov, E. (**1931**). ''La perception des sons d'une langue étrangère,'' Travaux du Cercle Linguistique de Prague **4**, 79–96.
Sapir, E. (**1921**). *Language* (Harcourt Brace Jovanovich, New York).
Wang, Y., Spence, M., Jongman, A., and Sereno, J. (**1999**). ''Training American listeners to perceive Mandarin tones,'' J. Acoust. Soc. Am. **106**, 3649–3658.
Werker, J., and Tees, R. (**1984**). ''Cross language speech perception: Evidence for perceptual reorganization during the first year of life,'' Infant Behav. Dev. **7**, 49–63.