Language identification with suprasegmental cues: A study based on speech resynthesis

Franck Ramus and Jacques Mehler
Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS), 54 boulevard Raspail, 75006 Paris, France

(Received 4 April 1997; revised 30 June 1998; accepted 8 September 1998)

This paper proposes a new experimental paradigm to explore the discriminability of languages, a question which is crucial to the child born in a bilingual environment. This paradigm employs the speech resynthesis technique, enabling the experimenter to preserve or degrade acoustic cues such as phonotactics, syllabic rhythm, or intonation from natural utterances. English and Japanese sentences were resynthesized, preserving broad phonotactics, rhythm, and intonation (condition 1), rhythm and intonation (condition 2), intonation only (condition 3), or rhythm only (condition 4). The findings support the notion that syllabic rhythm is a necessary and sufficient cue for French adult subjects to discriminate English from Japanese sentences. The results are consistent with previous research using low-pass filtered speech, as well as with phonological theories predicting rhythmic differences between languages. Thus, the new methodology proposed appears to be well suited to study language discrimination. Applications for other domains of psycholinguistic research and for automatic language identification are considered. [S0001-4966(98)04512-3]

PACS numbers: 43.71.Hw [WS]

INTRODUCTION

The predicament of the newborn having to learn a language seems quite difficult by itself. But things become even more complicated when the infant is raised in a bilingual or multilingual environment. If the child has no means to separate input utterances according to source languages, great confusion ought to arise. Such confusion, however, is not supported by informal observation. We will explore one possible strategy that infants may adopt to organize their linguistic environment.

To begin with, let us emphasize that bilingual environments are more than a remote possibility. Bilingualism is, in fact, more widespread than is usually acknowledged. Bilinguals may represent more than half the world's population (Hakuta, 1985; MacKey, 1967). Moreover, bilingual children do not show any significant language-learning impairment or retardation due to possible confusion between languages. What is interpreted as confusion by monolingual parents is usually code-switching, a common feature of the bilingual's linguistic system (see Grosjean, 1982, 1989).

Children's proficiency at learning multiple languages simultaneously suggests that they should have some way to discriminate languages, prior to learning any of them. Early language discrimination has indeed been demonstrated by a growing number of researchers. Mehler *et al.* (1986, 1988), Bahrick and Pickens (1988), Jusczyk *et al.* (1993), Moon *et al.* (1993), Bosch and Sebastián-Gallés (1997), and Dehaene-Lambertz and Houston (1998) have found that very young children, including newborns, are able to discriminate native from non-native utterances. Moreover, Nazzi *et al.* (1998) recently demonstrated that newborns also discriminate utterances from two unknown languages, e.g., English and Japanese for French subjects (see also Mehler *et al.*, 1988 as reanalyzed by Mehler and Christophe, 1995). However, this result does not extend to any pair of languages, which will be discussed below.

What cues are available to achieve such precocious discrimination? The adult bilingual may rely upon lexical knowledge, but such information is not available to infants. Therefore, the speech signal must contain some prelexical cues that enable language discrimination. The most obvious cues that can be thought of are the following:

(i) Phonetic repertoire. It is well-known that different languages use different sets of phonemes (see Maddieson, 1984 for an inventory). For example, an English speaker should have no trouble discriminating between French and Arabic, since Arabic makes use of very characteristic pharyngeal consonants, which don't exist in French.

(ii) Phonotactic constraints. In every language, there are constraints on the structural distribution of phonemes. In Japanese, for instance, a liquid (r) can never follow a stop consonant (p,b,k...), unlike in English or French.

(iii) Prosody. The term prosody collectively refers to the suprasegmental features of speech, mostly captured by the notions of rhythm and intonation. Since Pike (1945) and Abercrombie (1967), it has been acknowledged that languages can have different rhythms. English, as with all Germanic languages, has been described as stress-timed, while French and other Romance languages have been described as syllable-timed. Furthermore, Ladefoged (1975) has proposed a third rhythmic class consisting of mora-timed languages, such as Japanese. Although Nespor (1990) warns that these rhythmic differences might be better described as a continuum than as classes, they certainly can serve as reliable cues for language discrimi-

nation (Nazzi *et al.*, 1998). Finally, let us note that languages can also have different melodic properties, and therefore, intonation can be expected to play a role in language discrimination as well, as suggested by Maidment (1976, 1983), Ohala and Gilbert (1979), Willems (1982), and de Pijper (1983).

Obviously, all of these prelexical cues could be of interest for language discrimination. However, they may not all be relevant for discrimination by newborns. Mehler *et al.* (1988) and Nazzi *et al.* (1998) have shown that language discrimination is not hindered when utterances are filtered (low-pass, 400 Hz): newborns can perform the task equally well when segmental cues are removed. This led these authors to favor the *rhythm hypothesis*, i.e., that newborns can discriminate two languages if, and only if, they belong to different rhythmic classes, as defined above. In order to clarify the *rhythm hypothesis*, we reformulate it as follows:

(1) There are groups of languages that share a number of phonological properties.
(2) Rhythm is one these phonological properties, or alternatively, it is the outcome of some of them.
(3) By paying attention to rhythm, newborns are able to discriminate languages which have different phonological properties.

This hypothesis has been tested and confirmed by Nazzi *et al.* (1998) by showing that French newborns can discriminate filtered English and Japanese sentences (stress- versus mora-timed), but not English and Dutch ones (both stress-timed) under the same conditions. Moreover, infants can discriminate groups of languages, but only if these groups are congruent with rhythmic classes, e.g., they can discriminate English+Dutch from Spanish+Italian (stress- versus syllable-timed), but not English+Italian from Spanish +Dutch (incoherent groups). Thus, Nazzi *et al.*'s findings are in perfect agreement with the *rhythm hypothesis*.

However, we feel that the case for the *rhythm hypothesis* still needs to be bolstered for at least two reasons:

(1) The range of languages explored is insufficient. For example, Nespor (1990) questions the dichotomy between syllable-timed and stress-timed languages by presenting languages that share phonological properties of both types (Polish, Catalan, Portuguese). For such languages, one would like to know whether they can be discriminated from syllable-timed languages, or stress-timed languages, or both, or neither. The *rhythm hypothesis*, in its current formulation, would hold only if they clustered along with one or the other language group. Recent work by Bosch and Sebastián-Gallés (1997) suggests that Catalan is discriminable from Spanish (with low-pass filtered speech). Thus, either Catalan should not be considered as a syllable-timed language, as it has often been, or the *rhythm hypothesis* is wrong.
(2) Low-pass filtering is not an ideal way to degrade utterances with the aim of deleting segmental information and preserving prosody. Basically, filtering does not allow one to know which properties of the signal are eliminated and which are preserved. As a first approximation, segmental information should be eliminated because it is mainly contained in the higher formants of speech, and pitch should be preserved because it rarely rises higher than 400 Hz. But this is only an approximation. Listening to filtered speech makes it obvious that *some* segmental information is preserved (sometimes words can even be recognized), and pitch *does* sometimes rise higher than 400 Hz, especially for female voices.[1] The proportion of energy preserved is also problematic because it differs from phoneme to phoneme: for example, an /a/ vowel has a lot more energy in the low frequencies than an /i/ vowel, not to mention other segments like stop consonants. Low-pass filtering thus gives an unwarranted amplification to /a/. Consequently, there is no guarantee that filtered speech really preserves rhythm, at least from an acoustical point of view. From a perceptual point of view, it seems that the alternation between consonants and vowels is essential to the notion of syllabic rhythm, and there is no reason to believe that this is preserved either. Finally, Mehler *et al.*'s and Nazzi *et al.*'s results leave open another interpretation, one that we could call the *intonation hypothesis*: the idea being that discrimination may have been performed on the basis of intonation and not rhythm. Filtering, once again, does not make any distinction between intonation and rhythm, and much information would be gained by separating these two components of the speech signal.

In the remainder of this paper, we will concentrate on this second point by putting forward a new experimental paradigm to better assess the relative importance of the different components of prosody. The first point will not be addressed here, but it is quite clear that if one is to investigate the discrimination of more language pairs, one would first want to control more precisely the acoustic cues made available to subjects.

I. SPEECH RESYNTHESIS

A. General principles

The difficulties with low-pass filtering we mentioned above indicate that speech rhythm is an ill-defined concept. The cues that make us perceive rhythm in the speech signal are not well understood. Perceived speech rhythm could emerge from the succession of syllables, vowels, stresses, pitch excursions, energy bursts within a certain range of frequencies, or whatever occurs repeatedly in speech that the human ear can perceive. In this paper, we propose a methodology that can be used to explore the perception of rhythm under most of the above interpretations.

The main hypotheses that guided our search for better controlled stimuli can be stated as follows:

(i) what the newborn actually perceives and analyzes is a sequence of vowels or syllables, where the syllables are signaled by the energetic and spectral prominence of vowels.
(ii) if rhythm can be said to be a cue to language discrimi-

nation, it is in the sense that rhythm is the perceptual outcome of the succession of syllables and their organization.

(iii) if one wants to test rhythm as a potential cue to discriminate between two languages, one should have stimuli that preserve as much as possible the organization of sequences of syllables and degrade as much as possible all alternative cues.

To this end, we explored a new technique, namely *speech resynthesis*, to determine the perceptual cues relevant to language discrimination and to test the *rhythm hypothesis*. Speech resynthesis was first developed at IPO at Eindhoven, and it has been used for delexicalization purposes by Pagel *et al.* (1996) and Guasti *et al.* (1998). It amounts to:

(i) measuring all relevant acoustic components of the speech signal;

(ii) using these measures and an appropriate algorithm to resynthesize the spoken material.

The distinctiveness of our approach rests in the selection of the acoustic components used for resynthesis. This allows us to eliminate or preserve at will different dimensions of the speech signal, such as the nature of phonemes, rhythm, or intonation. See below for a description of signal treatment.

In order to explore the validity and usefulness of this technique, we limited the present study to adult subjects and to two languages whose discrimination was highly predictable: English and Japanese. Sentences were recorded by native speakers of each language and resynthesized in order to preserve various levels of information. In a first condition, intonation, rhythm, and broad phonetic categories were preserved in order to evaluate the technique with a maximum amount of information for discrimination. In a second condition, only intonation and rhythm were preserved. In a third condition, only intonation, and in a fourth condition, only rhythm was preserved. In all the experiments, French native speakers were trained and tested on a language categorization task.

B. Construction of the stimuli[2]

1. Source sentences

The stimuli used were taken from the set of sentences recorded by Nazzi *et al.* (1998). They consisted of 20 sentences in Japanese and 20 sentences in English (see list in Appendix) read by four female native speakers per language, and digitized at 16 kHz. Sentences were all declarative, and speakers read them as adult-directed utterances. They were matched in mean number of syllables (16.2 syllables per sentence in both languages), and in mean-fundamental frequency (229 Hz (s.d. 15.3) for English, 233 Hz (s.d. 15.9) for Japanese). However, the mean length of the sentences was not perfectly matched between the two languages: 2752 ms (s.d. 219) for English, 2627 ms (s.d. 122) for Japanese. It will be argued later that this difference had no consequence on the results observed.

2. General treatment

The following treatment was applied to each sentence:

(1) The fundamental frequency was extracted every 5 ms, using the Bliss software, by John Mertus;

(2) The beginning and end of each phoneme was marked by an experimenter, using both auditory and visual cues;

(3) The two types of information were merged into a text file including, for each phoneme of the sentence, its duration, and its pitch contour points;

(4) In this text file, a transformation was applied to the phonemes and/or to the pitch contour points, depending on the condition (see below).

(5) The resulting file was fed into the MBROLA software (Dutoit *et al.*, 1996) for synthesis by concatenation of diphones, using a French diphone database. The French (rather than Japanese or English) diphone database was chosen in order to remain neutral with respect to the language discrimination task.

3. Transformations applied

(i) The first kind of transformation, which we named ''*saltanaj*,'' consisted of replacing all fricatives with the phoneme /s/, all stop consonants with /t/, all liquids with /l/, all nasals with /n/, all glides[3] with /j/, and all vowels with /a/. These phonemes were chosen because they were the most universal in their respective categories (Maddieson, 1984; Crystal, 1987). Thus new sentences were synthesized, preserving the following features of the original ones:

(1) Global intonation;
(2) Syllabic rhythm;
(3) Broad phonotactics.

However, all nonprosodic lexical and syntactic information was lost. Exact phonetic and phonotactic information was lost as well, both because of the substitution of the phonemes before synthesis, and because the phonemes used by the software were French.

(ii) The second kind of transformation, named ''*sasasa*,'' consisted of replacing all consonants with /s/, and all vowels with /a/. The consonant /s/ was selected because its continuant character enabled transformation of consonant clusters into something sounding like a single (but long) consonant. Thus, in this condition, only syllabic rhythm and intonation were preserved.

(iii) The third kind of transformation, named ''*aaaa*,'' consisted of replacing all phonemes with /a/. It was ensured that the synthesized sentences did not sound like a weird succession of /a/s with noticeable onsets. Instead, they sounded like one long /a/, varying continuously in pitch (fundamental frequency was interpolated over unvoiced portions of the sentences). Here, only the intonation of the original sentences was preserved.

(iv) As for the fourth kind of transformation, named ''*flat sasasa*,'' the phonemes were substituted as in the *sasasa* transformation, but all sentences were synthesized with a constant fundamental frequency at 230 Hz (i.e., approximately the mean $F0$ measurement of

the original sentences). Thus, the only cue for language discrimination was syllabic rhythm.

II. EXPERIMENTAL TEST OF RESYNTHESIZED STIMULI

A. Method

The experimental protocol was programmed on an IBM-compatible computer using the EXPE language (Pallier *et al.*, 1997). Subjects read instructions indicating that they would be trained to recognize ''acoustically modified'' sentences of two languages, Sahatu and Moltec. The instructions were written in such a way as to make the subjects believe that the sentences belonged to two real and exotic languages, rather than to languages that they might know. Subjects heard the sentences through headphones. After the experiment, they were asked to explain which strategies they had used to perform the task.[4]

The 40 sentences were divided into two arbitrary sets of 20 sentences, each containing 10 sentences in each language, pronounced by two different speakers per language. They were called the training set and the test set. This was done to assess if what the subjects learned in the training phase was due only to particular sentences' or speakers' characteristics, or to more general properties of the two languages.

At the beginning of the training phase, one sentence of each language was selected at random from the training set, and served as a preliminary example. Then, all the sentences from the training set were presented in random order. After each sentence, the subject was required to enter S or M on the keyboard for Sahatu and Moltec and was given immediate feedback on the answer. After hearing the 20 sentences, the subjects who scored 70% or more correct responses went on to the test phase, while the others went through another training session with the same 20 sentences. Subjects were allowed to undergo a maximum of three training sessions, after which they were given the test session irrespective of their scores.

In the test phase, subjects heard the 20 sentences of the test set in a random order and answered as in the training phase. They were given feedback as well.

B. Participants

Sixty-four students participated voluntarily, without payment. They were all French native speakers with a mean age of 22 years. They were tested in their own rooms with a portable PC. There were four experimental conditions, corresponding to the four types of transformations mentioned above. They were run sequentially with the first 16 participants in the *saltanaj* condition, the next 16 in the *sasasa* condition, then the *aaaa* condition, and finally, the *flat sasasa* condition. Participants in the four experiments were drawn from the same pool of students, and the order in which they were tested was random. Besides the nature of the stimuli, the only thing that differed among the conditions was the minimum training score required to switch directly to the test phase: originally it was 75% for the *saltanaj* condition, but it was then lowered to 70% for the other conditions to allow more successful subjects to complete the experiment quickly.

TABLE I. Mean-percent scores during the different sessions of each condition (chance is 50%). In parentheses: number of subjects.

	Training 1	Training 2	Training 3	Test
saltanaj	61.8(16)	59.6 (14)	61.2(12)	66.9 (16)
sasasa	54.2(16)	63.1 (13)	66.1 (9)	65.0 (16)
aaaa	49.7(16)	55 (14)	54.1(11)	50.9 (16)
flat sasasa	62.5(16)	55.5 (10)	55.6 (8)	68.1 (16)

C. Results

A summary of the raw data, session by session, is presented in Table I. As can be seen, the number of subjects decreases during the training phase due to the fact that the most successful ones are allowed to skip training sessions 2 or 3. The scores correspond to total hit rates of all the answers.

In order to assess which general properties of the two languages the subjects have learned, independently of the characteristics of particular sentences or speakers, we restricted the statistical analyses to the test session. Indeed, scores during the test session measure the ability of subjects to generalize what they have learned during training sessions to novel sentences produced by new speakers. Therefore, it would be very difficult to interpret the results as showing that the subjects have learned individual characteristics of certain sentences or speakers. Test-session scores thus represent a conservative measure of language discrimination.

Moreover, we converted our test-session scores into hit rates and false-alarm rates (in the sense of signal detection theory) in order to perform an analysis of discrimination, taking into account any response biases that subjects may have had. We used as hit rates the percentage of Japanese sentences correctly recognized, and as false alarms, the percentage of English sentences incorrectly labeled as Japanese. Table II presents, for each condition, mean hit rates, false-alarm rates, discrimination scores (A') and response bias measures (B''_D)[5] (see Donaldson, 1992).

A Kolmogorov test for normality ensured that the distributions of A' scores could be considered as normal (all p values >0.36). A t-test was computed to compare A' scores to chance level (0.5). Discrimination scores were found to be significantly above chance in the *saltanaj* [$t(15)=4.47$, p

TABLE II. Mean-hit rates, false-alarm rates, discrimination scores, and response bias measures, in each condition during test session. A' is compared to 0.5 (chance level) and B''_D to 0 (no bias).

	Hit rates	False alarms	A'	B''_D
saltanaj	0.69	0.35	0.71[a]	−0.11
sasasa	0.65	0.35	0.68[b]	−0.02
aaaa	0.56	0.54	0.52	−0.18[c]
flat sasasa	0.70	0.34	0.72[a]	−0.18

[a] $p<0.001$.
[b] $p<0.01$.
[c] $p<0.05$.

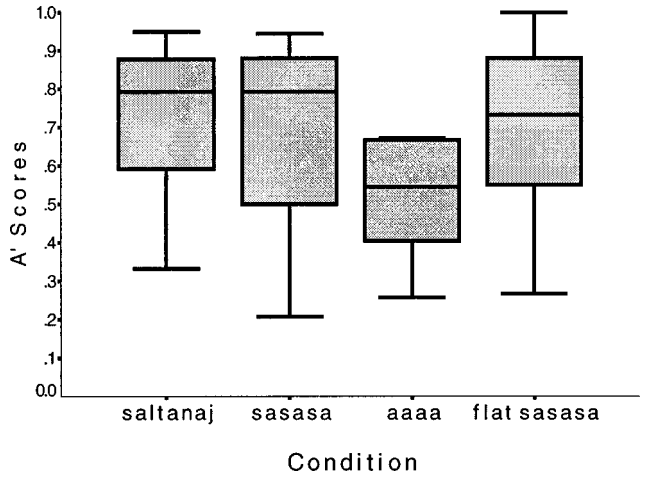

FIG. 1. Distribution of A' scores in each condition. Horizontal bars represent the medians, boxes the central half of the data, and whiskers the whole range of the data.

$=0.000\,4]$, *sasasa* $[t(15)=3,\ p=0.009]$, and *flat sasasa* $[t(15)=4.15,\ p=0.000\,9]$ conditions, but not in the *aaaa* condition $[t(15)<1]$.

The results presented in Table II seem to be quite clear-cut: the two sets of sentences were discriminable in all but the *aaaa* condition. To further evaluate the four conditions, the distribution of A' scores in each condition is shown in Fig. 1. Multiple comparisons of the four conditions with a Bonferroni correction showed that the *aaaa* condition was significantly different from both the *saltanaj* ($p=0.002$) and *flat sasasa* ($p=0.004$) conditions. No other differences showed significance, but there was a tendency for the *aaaa* condition to be different from the *sasasa* condition as well ($p=0.026$), which was offset by the Bonferroni correction. It is thus reasonable to say that the *aaaa* condition was different from all the others.

Finally, B''_D scores show that the subjects did not have any particular bias, except in the *aaaa* condition, where they were slightly liberal ($p=0.046$); that is, they tended to answer ''Japanese'' more often than ''English.'' This isolated and modest effect does not seem to us to require any particular interpretation or attention.

D. Discussion

1. Acoustic cues available to the subjects

In the *saltanaj* condition, the manner of articulation, the duration, and the place of each phoneme was preserved. Since the overall structure and organization of the syllables was preserved, syllabic rhythm certainly was as well. In addition, global intonation was also preserved. Thus, subjects had many available cues for discriminating utterances. Possibly the most salient one was the presence of numerous consonant clusters in English, with almost none in Japanese.

In the *sasasa* condition, in contrast, the identity of the phoneme classes, their respective distributions, and their arrangement was lost. Only the intonation and gross syllabic information was preserved. More precisely:

(i) the consonant/vowel distinction and temporal ratio were preserved;

(ii) the weight of the syllables was also preserved, since consonant clusters of the original stimuli were converted into long consonants (indeed, /s/ of the same duration as the corresponding clusters);

(iii) the broad temporal organization of the syllables was preserved as well;

(iv) finally, the rendering of the fundamental frequency conveyed information about both the global intonation of the sentences and, more locally, stress and pitch-accent, i.e., stressed or accented syllables were detectable, at least with partial cues (intensity cues were not available, for instance).

The subjects' ability to discriminate the two sets of *sasasa* sentences has an interesting implication, namely that suprasegmental cues are sufficient to allow for discrimination of the two languages. In this respect, our results are quite similar to those of Ohala and Gilbert (1979), who showed discrimination between several languages with stimuli that also preserved rhythm and intonation (although in their experiment, rhythm was that of the envelope of the signal, rather than of the syllables).

In the *aaaa* condition, the only remaining cue was the global intonation of the sentences, as resynthesized from the $F0$ data. Local intonational cues were probably of little use since they were not aligned with any syllable. Therefore, this condition simply explored whether melody could serve to discriminate English from Japanese. It seems that it cannot, as subjects behaved in a way that looked like guessing.

This result can be viewed as being at odds with some of the few previous studies on the role of intonation in language discrimination (Maidment, 1976, 1983; Willems, 1982; de Pijper, 1983). However, these experiments differ from ours in at least two respects: first, they compared English with Dutch and French, but not with Japanese; second, the native language of the subjects was always pitted against another language, and the subjects were aware of this fact. This must have made the task considerably easier. Indeed, when hearing a sentence, the subjects had to judge whether it met the intonational pattern of their native language, and did not have to forge new categories from scratch. This strategy would not be possible for an infant who has not yet acquired a native language. Given that one of our aims was to explore language acquisition, we wanted to place the adult subjects in a similar situation. Thus, our findings are not in contradiction with previous studies. However, it is not yet clear whether our subjects failed because English and Japanese intonations are not different enough, or because our stimuli were too degraded, or because the subjects were not native speakers of either of the two languages presented.

To further explore this question, we recruited 16 native English speakers (ten Americans, four English, one other, and one unknown), with a mean age of 29 years. Most of them were paid for their participation. They were tested on the *aaaa* stimuli under the same conditions as the French subjects, except that they were told that the languages were English and Sahatu, and that they were to recognize them by their intonation. The task thus was as close as possible to the previous studies cited above. The average A' score was 0.61

(s.d. 0.14), which was significantly above chance [$t(15) = 3.25$, $p = 0.005$]. There was no response bias [$B''_D = 0.09$, $t(15) < 1$]. Thus, it seems that English and Japanese intonations are sufficiently dissimilar to be differentiated, and that the *aaaa* stimuli are not too degraded or uninteresting for the task to be performed. However, the task seems to be feasible only when subjects have a certain knowledge of one of the languages and of the task itself.

Finally, the success of our subjects in discriminating between the two sets of sentences in the *flat sasasa* condition shows that they could easily do without any intonation, and that syllabic rhythm was a robust cue for discrimination. Indeed, this finding seems surprising, given the disembodied nature of speech uttered with a flat intonation. But at the same time, this points out the remarkable robustness of the cues present in the *flat sasasa* stimuli. As we mentioned above, these cues are related to the temporal organization of consonants and vowels within the sentence. Since there are very few consonant clusters in Japanese and many in English, large differences may persist between the two languages. *Flat sasasa* English sentences were characterized by longer consonants, heavier syllables, a greater variety of syllable types, weights, and durations, and thus a more irregular temporal organization of syllables than Japanese sentences. These cues are indeed thought to be the main constituents of syllabic rhythm (see Dauer, 1983, 1987; Nespor, 1990).

In conclusion, syllabic rhythm was shown to be both necessary and sufficient for the discrimination task. Indeed, its presence was sufficient in the *flat sasasa* condition, and its absence was an obstacle in the *aaaa* condition. This is not to say that this is always the case; as we mentioned above, intonation can be of greater interest to native speakers. It could also be a crucial cue for other pairs of languages, like tonal languages. Conversely, one can also imagine situations where rhythm may not be sufficient, possibly English and Dutch, or Spanish and Italian. This is a matter for future research, where speech resynthesis methodology should be of further use.

2. Possible problems and improvements

Before drawing more general conclusions, we will now turn to more methodological questions concerning this particular study and the general procedure.

First, one might be concerned with the fact that, in this study, the length of the sentences was not perfectly matched between the two languages. Indeed, as the English sentences were on average about 5% longer than the Japanese ones, it could be argued that the discrimination observed had nothing to do with rhythm, but rather with a strategy relying on sentence length. If this were the case, then we would expect a similar result in the *aaaa* condition, where the sentences were exactly the same length as in the other conditions. The results obtained in the *aaaa* condition clearly show that subjects were unable to use average sentence length to perform the task, and therefore this interpretation must be ruled out, unless one is prepared to argue that the length information was unusable only in the *aaaa* condition.

As regards the methodology itself, one might want to argue that the discriminability of the two sets of resynthesized sentences could be an artefact of the synthesis itself. However, since all the stages in the resynthesis process were performed in a similar fashion for both languages, it seems unlikely that some artefact or artificial difference was introduced for one language and not the other. At any rate, as we have already noted, there are differences between English and Japanese that we expected subjects to use in the task.

An aspect of our results that can seem surprising is the relatively low level of average discrimination scores (68%–72%), when the two languages studied seem so different. Doesn't this suggest that the technique lacks sensitivity? This would be consistent with the fact that scores are not higher in the *saltanaj* than in the *sasasa* condition, despite the additional information provided to perform the task. Indeed, a more sensitive task that would allow us to detect more subtle effects would be desirable. However, we have several reasons to think that discrimination scores would not be dramatically higher. As the stimuli are quite impoverished, they are not particularly interesting for the subjects. In addition, since they unfold over three seconds, the task demands sustained attention and an unusual effort to extract regularities. Likewise, the source sentences themselves show great variability, and the acoustic cues do not allow for a definite determination of their origin, i.e., what is true of the prosody of English sentences in general is not necessarily true of the prosody of every English sentence, and there can be an overlap between the prosodies of English and Japanese sentences.

To confirm this intuition, we ran an item analysis on the *sasasa* sentences used in the test phase. Scores for individual sentence recognition ranged from 38% to 88% (chance = 50%), and an ANOVA (analysis of variance) using the logistic generalized linear model (Hosmer and Lemeshow, 1989) showed a significant effect of the sentence factor, i.e., some sentences yielded scores that were significantly different from others. In brief, some sentences were not very good exemplars of their language, at least in the sense of the acoustic cues preserved under the different conditions. For instance, the three sentences yielding the worst scores (38%, 44%, and 50%) were English sentences (respectively #20, 16, 17, see Appendix) that have few consonant clusters. Indeed, they were found to have a higher vowel/consonant temporal ratio (respectively, 0.49, 0.44, 0.45) than most other English sentences (average 0.4 over our 20 sentences, s.d. = 0.05), thus getting closer to the Japanese prototype (average 0.53, s.d. = 0.03). This confirms that syllabic complexity is a critical cue in the English/Japanese discrimination. This might also explain why subjects tended to respond slightly more ''Japanese'' than ''English'' overall: English sentences can occasionally have mostly simple syllables like Japanese ones, but the phonology of Japanese forbids the reverse situation. As infants are confronted with similarly noisy input, it seems only fair to test adults under the same conditions, rather than with sentences selected for their prototypicality. Lower discrimination scores are thus to be expected.

The great complexity of the stimuli and their variability within one language may also explain why more information does not seem to improve necessarily our subjects' perfor-

mance. In the *flat sasasa* condition, we claim that subjects are provided with the most reliable cue, i.e., syllabic rhythm. If intonation is irrelevant to the task, or at least if it is a less reliable cue than rhythm, then the presence of intonation in the *sasasa* and *saltanaj* conditions may not necessarily help subjects; it could even disturb them by distracting them from the most relevant cue. The same can be said of broad phonotactics.

Finally, a possible way to improve the subjects' scores might be to incorporate a measure of amplitude in the synthesis. This has not been done in the present work simply because the MBROLA software doesn't take amplitude as a possible input. Thus, in our resynthesized stimuli, stress was signaled only by pitch excursions and duration, not by amplitude. As there is reason to think that stress is an important component of rhythm, adding a cue such as amplitude could make the perception of rhythm more accurate, and would furthermore make it possible to analyze separately the respective roles of rhythm due to the succession of syllables and rhythm due to amplitude.

How dependent are our results on the maternal language of our subjects, and on the language chosen as a diphone database (French)? As mentioned above, being a native speaker of one of the target languages helps, at least when one is aware of it. More generally, the subjects' native language may influence performance in the tasks we proposed. Indeed, speech perception is often said to be biased by one's maternal language. This is particularly true for phonemic perception, but also for more abstract phonological processing. For instance, French native speakers are quite poor at perceiving stress (Dupoux *et al.*, 1997; see also Dupoux *et al.*, 1999, for another example). Granting that English has stress and Japanese has pitch-accent, and if one accepts that these cues remained present in the resynthesized stimuli (possibly in the *saltanaj* and *sasasa* conditions), it is possible that French subjects were unable to detect this type of information. If so, this could actually account for the lack of a difference in performance between the intonated and *flat sasasa* conditions, in which the presence or absence of intonation seemed to make no difference to the subjects. We hope to test speakers of other languages in order to assess whether they do better in the *sasasa* condition. Nonetheless, considering performance in the *flat sasasa* condition, we find no similar reason to believe that the perception of syllabic rhythm alone would be any better or worse for speakers of languages other than French. Therefore, we think that our main conclusion, that syllabic rhythm is enough to allow for discrimination of English and Japanese, should hold across speakers of any other languages.

Another point worth mentioning is that our subjects were much more familiar with English than with Japanese. English is learned at school in France, not Japanese. However, subjects were told that the languages were Sahatu and Moltec. Moreover, sentences were delexicalized, providing subjects with no obvious way to detect the presence of English. As a matter of fact, *a posteriori* reports revealed that none of them guessed that Moltec was English. Moreover, no response asymmetries were observed (such as a tendency to recognize Moltec sentences more often), so there is no reason to believe that the subjects' greater familiarity with English had an influence on the results.

Finally, the influence of the French diphone database could be relevant for the *saltanaj* condition only, as *sasasa* or *aaaa* sentences would hardly have sounded any different if we had used another diphone database. For the *saltanaj* condition, the number of phonemes used was low, and the chosen phonemes (s, a, l, t, n, j) exist in both Japanese and English. We checked that the transposition of the phonemes did not produce illegal sequences in either language. All the resulting diphones were legal in French, which enabled a correct diphone synthesis. Occasionally the phoneme transposition led to a slight change of syllabification. For example, the English phrase ''the truck'' was transformed into /satlat/. /tl/ is a legal phoneme sequence in English, but only across a syllable boundary (as in ''butler''). The same is true for French. Thus, the transformation of ''the truck'' into /satlat/ shifted the perceived syllable boundary to fall between /t/ and /l/. If one is concerned with the precise contribution of phonotactics for language discrimination, such effects could indeed be a problem, and one should then choose the phonemes accordingly. In the present case, where the discrimination was made possible by massive differences in syllable weight and the presence or absence of consonant clusters, such minor effects must have been negligible.

III. GENERAL DISCUSSION

In this study, we have put forward a new method, speech resynthesis, to explore the discrimination of languages on the basis of prosodic cues. We used this method to construct stimuli that preserved different possible levels of prosodic information in both English and Japanese sentences, and we tested discrimination of these two sets of stimuli by French subjects. Our results show that syllabic rhythm is clearly sufficient to allow for discrimination between English and Japanese.

This finding is consistent with both phonological theories and past experimental studies. Indeed, the contrasting rhythmic patterns of languages such as English and Japanese have been noticed by linguists (Pike, 1945; Abercrombie, 1967; Ladefoged, 1975), leading them to classify languages into different rhythmic classes. Mehler *et al.* (1996) and Nazzi *et al.* (1998) have, moreover, hypothesized that discrimination should be possible between languages belonging to different rhythmic classes. Our results not only confirm that this is true of English and Japanese, but also demonstrate that syllabic rhythm is, as predicted, a relevant parameter.

In this respect, we think that the scope of our work goes beyond past studies upholding the role of prosody for language discrimination. Indeed, previous studies have relied on only one type of degradation of the speech signal at any one time. Ohala and Gilbert (1979), for instance, explored the joint role of intonation and rhythm, whereas Maidment (1976, 1983), Willems (1982) and de Pijper (1983) explored the role of intonation alone. Likewise, in their studies on infants, Mehler *et al.* (1988), Nazzi *et al.* (1998), Bosch and Sebastián-Gallés (1997) and Dehaene-Lambertz and Houston (1998) relied on low-pass filtering to isolate gross prosodic

cues. In all those studies, however, the different levels of prosodic information were not separated and compared.

We thus view our main contribution as having (1) provided a methodology allowing to separate and analyze different components of prosody in a systematic fashion, (2) isolated the prosodic component of interest to the *rhythm hypothesis*, that is, syllabic rhythm, (3) shown that this component is, as expected, an excellent and possibly the best prosodic cue for the discrimination of languages that are said to differ in rhythm.

Let us now turn to the possible future applications of this new methodology. To further test the *rhythm hypothesis*, the *flat sasasa* stimuli provide a better tool than low-pass filtering. For example, a replication of Nazzi *et al.*'s (1998) experiments with such stimuli would allow us to rule out the alternative-intonation hypothesis. Indeed, even though our present results on adults strongly suggest that their rhythm-based interpretation was right, extrapolation of results from the adult state to the initial state is not warranted.

More language discrimination experiments on adults and infants using *flat sasasa* stimuli would also be needed to evaluate whether languages actually tend to congregate into rhythmic classes, or whether, as Nespor (1990) suggests, they form a rhythmic continuum.

Studying the prosodic properties of languages using speech resynthesis may also influence research on automatic language identification. Indeed, much of the research in this domain has concentrated on modeling the short-term acoustics of the speech signal. Prosodic features have rarely been taken into account, and with relatively low success (for a review, see Muthusamy *et al.*, 1994). Even though one should not expect to discriminate all pairs of languages using prosodic cues only, prosody could still be used as a first-order classifier, thus restraining the problem space for analysis with other cues. In this respect, we feel that language-discrimination studies using speech resynthesis might be a practical way to establish a taxonomy of the world languages along different prosodic dimensions, and such a taxonomy could be a first step towards the elaboration of a prosodic classifier.

Outside the range of the *rhythm hypothesis*, one can imagine various applications of the speech resynthesis paradigm. When studying the perception of prosody, it is often desirable to cancel possible lexical and/or segmental influences. This has sometimes been done in the past by using reiterant speech, that is, by asking speakers to produce nonsense syllables (like ''mamama'') while imitating the prosody of a natural sentence (Larkey, 1983; Liberman and Streeter, 1978). In our view, resynthesis provides a way to create such reiterant stimuli in a more controlled and systematic manner, without having to rely on speakers producing nonspeech, which is quite an unnatural task.

A possible application is the study of prosodic correlates of word boundaries. For instance, de Pijper and Sanderman (1994) delexicalized whole sentences and asked subjects to judge word and phrase boundaries. In the authors' opinions, their stimuli proved quite painful to listen to, so similar work would benefit from using speech resynthesis (see Pagel *et al.*, 1996 for a first approach).

Finally, higher-level prosodic cues can also be studied using speech resynthesis. For instance, the head-direction parameter in syntax is said to have a prosodic correlate, namely prosodic phrase prominence (Nespor *et al.*, 1996). By carefully resynthesizing their sentences to control the acoustic cues preserved, Guasti *et al.* (1998) showed that such prominence is perceived by adults and infants, and could thus serve to set the head-direction parameter early on.

To conclude, we think that the use of speech resynthesis goes beyond the need, evident in the above studies, for a practical delexicalization tool. Its flexibility authorizes countless ways to selectively preserve or eliminate cues, of which the present paper has proposed only a few. For other purposes yet to be defined, one could also decide to preserve the place rather than the manner of articulation of phonemes, or to blur function words while preserving content words and prosody, or vice versa. We leave it to the reader's imagination to invent other interesting manners to manipulate speech resynthesis.

ACKNOWLEDGMENTS

This work was supported by the Délégation Générale pour l'Armement and the Human Frontiers Science Program. We would like to thank Emmanuel Dupoux and Anne Christophe for their help and suggestions, and Peter Jusczyk and Thierry Nazzi for comments on a previous version of this paper.

APPENDIX

English sentences

Speaker 1

1. The next local elections will take place during the winter.
2. A hurricane was announced this afternoon on the TV.
3. The committee will meet this afternoon for a special debate.
4. This rugby season promises to be a very exciting one.
5. Artists have always been attracted by the life in the capital.

Speaker 2

6. My grandparents' neighbor is the most charming person I know.
7. The art gallery in this street was opened only last week.
8. The parents quietly crossed the dark room and approached the boy's bed.
9. Nobody noticed when the children slipped away just after dinner.
10. Science has acquired an important place in western society.

Speaker 3

11. Much more money will be needed to make this project succeed.
12. This supermarket had to close due to economic problems.

13. The first flowers have bloomed due to the exceptional warmth of March.
14. The last concert given at the opera was a tremendous success.
15. Finding a job is difficult in the present economic climate.

Speaker 4

16. The local train left the station more than five minutes ago.
17. In this famous coffee shop you will eat the best donuts in town.
18. The young boy got up quite early in order to watch the sunrise.
19. In this case, the easiest solution seems to appeal to the high court.
20. The library is opened every day from eight A.M. to six P.M.

Japanese sentences

Speaker 1

1. Oono shigo ni machi no saiken ga hajimatta.
2. Noomin no sonchoo ni taisuru fuman ga tamatta.
3 Totemo kichoona kaiga ga saikin nusumareta.
4. Kochira no kata wa keiseigeka no senmonka desu.
5. Tsugino chihoosenkyo wa kondo no harugoro deshoo.

Speaker 2

6. Monku wa shihainin ni iuno ga tettoribayai.
7. Nihon no tabemononara mazu teni hairu.
8. Operaza no saigo no konsaato wa seikoodatta.
9. Kaikakusuishinha ga kenchoomae de demokooshinshita.
10. Bakayooki no seide hayakumo hana ga saiteiru.

Speaker 3

11. Noomin no sonchoo ni taisuru fuman ga tamatta.
12. Haru no koozui de zuibun ookina higaiga deta.
13. Konshuu mo terebibangumi o mirujikan ga nai.
14. Tsugino chihoosenkyo wa kondo no harugoro deshoo.
15. Tsugi no gekijooshiizun wa totemo kyoomibukaidaroo.

Speaker 4

16. Hachiji no nyuusu de jiken ga hoodoosareta.
17. Kinyoobi no gogo wa ginkooga hayaku shimaru.
18. Konopanya no keiki wa konokaiwai de hyoobanda.
19. Bakayooki no seide hayakumo hana ga saiteiru.
20. Kusuriya no kamisan wa moosugu kaimononi deru.

[1]In experiments on infants, female voices are used almost exclusively.
[2]Samples of all the types of stimuli described in this article can be heard on http://www.ehess.fr/centres/lscp/persons/ramus/resynth/ecoute.htm
[3]At this point, the ambiguous status of glides should be mentioned. The following rule was applied: pre- and inter-vocalic glides were marked as consonants, post-vocalic glides (in diphthongs) were marked as vowels. Therefore, pre- and inter-vocalic glides were transformed into /j/ in the *saltanaj* condition and /s/ in the *sasasa* condition, whereas postvocalic glides were transformed into /a/ in both conditions.
[4]Subjects' reports were not found to be consistent nor informative and are therefore not reported here.
[5]We are grateful to Dr. Strange for suggesting this type of analysis.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Aldine, Chicago).
Bahrick, L. E., and Pickens, J. N. (**1988**). ''Classification of bimodal English and Spanish language passages by infants,'' Infant Behav. Dev. **11**, 277–296.
Bosch, L., and Sebastián-Gallés, N. (**1997**). ''Native language recognition abilities in 4-month-old infants from monolingual and bilingual environments,'' Cognition **65**, 33–69.
Crystal, D. (**1987**). *The Cambridge Encyclopedia of Language* (Cambridge U.P., Cambridge).
Dauer, R. M. (**1983**). ''Stress-timing and syllable-timing reanalyzed,'' J. Phon. **11**, 51–62.
Dauer, R. M. (**1987**). ''Phonetic and phonological components of language rhythm,'' in *XIth International Congress of Phonetic Sciences* (Tallinn), pp. 447–450.
de Pijper, J. R. (**1983**). *Modelling British English Intonation* (Foris, Dordrecht).
de Pijper, J. R., and Sanderman, A. A. (**1994**). ''On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues,'' J. Acoust. Soc. Am. **96**, 2037–2047.
Dehaene-Lambertz, G., and Houston, D. (**1998**). ''Faster orientation latency toward native language in two-month old infants,'' Language Speech **41**, 21–43.
Donaldson, W. (**1992**). ''Measuring recognition memory,'' J. Exp. Psychol. **121**, 275–277.
Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., Fitneva, S., and Mehler, J. (**1999**). ''Epenthetic vowels in Japanese: a perceptual illusion?'' J. Exp. Psychol. Hum. Percept. Perform. (to be published).
Dupoux, E., Pallier, C., Sebastian, N., and Mehler, J. (**1997**). ''A destressing 'deafness' in French?'' J. Memory Lang. **36**, 406–421.
Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. (**1996**). ''The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes,'' in *ICSLP'96* (Philadelphia, PA), pp. 1393–1396.
Grosjean, F. (**1982**). *Life With Two Languages: An Introduction to Bilingualism* (Harvard U.P., Cambridge, MA).
Grosjean, F. (**1989**). ''Neurolinguists, beware! The bilingual is not two monolinguals in one person,'' Brain and Language **36**, 3–15.
Guasti, M. T., Nespor, M., Christophe, A., and van Ooyen, B. (**1998**). ''Prelexical setting of the head-complement parameter through prosody,'' in *Signal to Syntax II*, edited by J. Weissenborn and B. Höhle (to be published).
Hakuta, K. (**1985**). *Mirror of Language: The Debate on Bilingualism* (Basic Books, New York).
Hosmer, D. W., and Lemeshow, S. (**1989**). *Applied Logistic Regression* (Wiley, New York).
Jusczyk, P. W., Friederici, A., Wessels, J., Svenkerud, V., and Jusczyk, A. (**1993**). ''Infants' sensitivity to the sound pattern of native language words,'' J. Memory Lang. **32**, 402–420.
Ladefoged, P. (**1975**). *A Course in Phonetics* (Harcourt Brace Jovanovich, New York).
Larkey, L. S. (**1983**). ''Reiterant speech: An acoustic and perceptual validation,'' J. Acoust. Soc. Am. **73**, 1337–1345.
Liberman, M. Y., and Streeter, L. A. (**1978**). ''Use of nonsense-syllable mimicry in the study of prosodic phenomena,'' J. Acoust. Soc. Am. **63**, 231–233.
MacKey, W. F. (**1967**). *Bilingualism as a World Problem/Le Bilinguisme: Phénomène Mondial* (Harvest House, Montreal).
Maddieson, I. (**1984**). *Patterns of Sounds* (Cambridge U.P., Cambridge).
Maidment, J. A. (**1976**). ''Voice fundamental frequency characteristics as language differentiators,'' Speech and hearing: Work in progress, University College London, 74–93.
Maidment, J. A. (**1983**). ''Language recognition and prosody: further evidence,'' Speech, hearing and language: Work in progress, University College London **1**, 133–141.
Mehler, J., and Christophe, A. (**1995**). ''Maturation and learning of language during the first year of life,'' in *The Cognitive Neurosciences*, edited by M. S. Gazzaniga (Bradford Books/MIT, Cambridge, MA), pp. 943–954.
Mehler, J., Dupoux, E., Nazzi, T., and Dehaene-Lambertz, G. (**1996**). ''Cop-

ing with linguistic diversity: The infant's viewpoint,'' in *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, edited by J. L. Morgan and K. Demuth (Erlbaum, Mahwah, NJ), pp. 101–116.
Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. (**1988**). ''A precursor of language acquisition in young infants,'' Cognition **29**, 143–178.
Mehler, J., Lambertz, G., Jusczyk, P., and Amiel-Tison, C. (**1986**). ''Discrimination de la langue maternelle par le nouveau-né,'' Comptes-rendus de l'Académie des Sciences de Paris **303**, **Série III**, 637–640.
Moon, C., Cooper, R. P., and Fifer, W. P. (**1993**). ''Two-day-olds prefer their native language,'' Infant Behav. Dev. **16**, 495–500.
Muthusamy, Y. K., Barnard, E., and Cole, R. A. (**1994**). ''Reviewing automatic language identification,'' IEEE Signal Process. Mag., 33–41 (Oct. 1994).
Nazzi, T., Bertoncini, J., and Mehler, J. (**1998**). ''Language discrimination by newborns: towards an understanding of the role of rhythm,'' J. Exp. Psychol. Hum. Percept. Perform. **24**, 756–766.
Nespor, M. (**1990**). ''On the rhythm parameter in phonology,'' in *Logical Issues in Language Acquisition*, edited by I. M. Roca (Foris, Dordrecht), pp. 157–175.
Nespor, M., Guasti, M. T., and Christophe, A. (**1996**). ''Selecting word order: The rhythmic activation principle,'' in *Interfaces in Phonology*, edited by U. Kleinhenz (Akademie, Berlin), pp. 1–26.
Ohala, J. J., and Gilbert, J. B. (**1979**). ''Listeners' ability to identify languages by their prosody,'' in *Problèmes de Prosodie*, edited by P. Léon and M. Rossi (Didier, Ottawa), pp. 123–131.
Pagel, V., Carbonell, N., and Laprie, Y. (**1996**). ''A new method for speech delexicalization, and its application to the perception of French prosody,'' in *ICSLP'96* (Philadelphia, PA).
Pallier, C., Dupoux, E., and Jeannin, X. (**1997**). ''EXPE: An expandable programming language for on-line psychological experiments,'' Behav. Res. Methods Instrum. Comput. **29**, 322–327.
Pike, K. L. (**1945**). *The Intonation of American English* (University of Michigan Press, Ann Arbor, Michigan).
Willems, N. (**1982**). *English Intonation from a Dutch Point of View* (Foris, Dordrecht).