Discriminating speech rhythms in audition, vision, and touch

Jordi Navarra[a], Salvador Soto-Faraco[b,c], & Charles Spence[d]

[a] Fundació Sant Joan de Déu, Parc Sanitari Sant Joan de Déu, CIBERSAM, Spain.

[b] Institució Catalana de Recerca i Estudis Avançat (ICREA)*, Spain.*

[c] Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Spain.

[d] Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, UK.

RE-SUBMITTED TO: *Acta Psychologica*

('Temporal Processing Within and Across Senses' special issue*)*

(March 16th, 2013)

CORRESPONDENCE TO: Jordi Navarra, Hospital de Sant Joan de Déu, Edifici Docent, C/ Santa Rosa, 39-57, planta 4ª, 08950 Esplugues – Barcelona, Spain. TEL: +44-1865-271307; FAX: +44-1865-310447

E-MAIL: jnavarrao@fsjd.org

ABSTRACT

We investigated the extent to which people can discriminate between languages on the basis of their characteristic temporal, rhythmic information, and the extent to which this ability generalizes across sensory modalities. We used rhythmical patterns derived from the alternation of vowels and consonants in English and Japanese, presented in audition, vision, or touch. Experiment 1 confirmed that discrimination is possible on the basis of auditory rhythmic patterns, and extended it to the case of vision, using 'aperture-close' mouth movements of a schematic face. In Experiment 2, language discrimination was demonstrated using visual and auditory materials that did not resemble spoken articulation. In a combined analysis including data from Experiments 1 and 2, a beneficial effect was also found when the auditory rhythmic information was available to participants. In a final experiment, we demonstrate that the rhythm of speech can also be discriminated successfully by means of vibrotactile patterns delivered to the fingertip. The results of the present study therefore demonstrate that discrimination between language's syllabic rhythmic patterns is possible on the basis of visual and tactile displays. However, despite the fact that discrimination can be achieved using vision alone, auditory performance is nevertheless better.

KEYWORDS: Speech Perception; Speechreading; Rhythm; Audition; Vision; Touch; Discrimination.

1. **Introduction**

In 1940, Lloyd James suggested an intriguing classification of spoken languages based on their rhythmic properties, as having either 'machine-gun' (e.g., Spanish) or 'Morse-code' (e.g., Dutch) rhythms (see also Pike, 1945). Modern reformulations of this original idea have proposed that languages can be roughly classified according to their different temporal patterns in stress-, syllable-, or mora-timed (e.g., Nazzi, Bertoncini, & Mehler, 1998). The ability to parse the rhythmic properties of the speech input is thought to be critical for young infants in order to discriminate between the languages that are present in their environment (see Mehler, Dupoux, Nazzi, & Deahene-Lambertz, 1996). This is an important ability since, speaking globally, bilingual communities are more numerous than monolingual ones (see de Bot & Kroll, 2002; Brutt-Griffler & Varghese, 2004). Newborns seem to be remarkably sensitive to temporal properties of the acoustic signal that discriminate between languages belonging to different rhythmic classes, but are seemingly unable to discriminate between languages that belong to the same rhythmic class until much later in life (see Nazzi et al., 1998). These findings have been extended to non-human animals such as monkeys (Ramus et al., 2000) and even rats (Toro, Trobalón, & Sebastián-Gallés, 2003).

In a seminal study conducted with adult humans, Ramus and Mehler (1999) demonstrated that information about speech rhythm alone (i.e., based on the temporal organization of consonants and vowels) is sufficient to discriminate between different languages. In their study, Ramus and Mehler used a transformation of the spoken signal called *flat sasasa* that preserves syllabic rhythm while filtering out other linguistic cues relating to the segmental content. They used a set of spoken sentences in English and Japanese in which all of the consonant segments were digitally replaced with the sound /s/ and all of the vowel segments with /a/ (all of the stimuli were also shifted to a constant fundamental frequency of 230 Hz). In this way, while the temporal distribution of consonants and vowels of English and Japanese was preserved, other cues such as phonetics, phonotactics, and intonation contour were removed completely (Ramus & Mehler, 1999; see also Grabe & Low, 2002). English is, for example, characterized by a more irregular temporal organization than Japanese. The presence of longer (and more variable in

duration) consonant intervals in English (due to the fact that English has many consonant clusters), and the existence of weak vs. strong syllable alternation (i.e., with short vs. long vowels or diphthongs, respectively), and more diverse syllable types in English contrasts with the relatively constant rhythmical characteristics of Japanese. Therefore, the temporal differences between these two languages may well explain why it is that people can discriminate between their associated *flat sasasa* patterns auditorily.

The goal of the present study was therefore to investigate whether the rhythm obtained from the temporal distribution of vowels and consonants could also lead to successful discrimination in non-acoustic stimuli, through visual (Experiments 1 and 2) and somatosensory patterns (Experiment 3). Obtaining alternative ways to improve a speech signal may ultimately be relevant in technological domains such as telephony (e.g., to facilitate the comprehension of spoken messages in phone conversations in noisy environments) or visual/tactile aids for hearing-impaired individuals. The real-time presentation of specific rhythmic cues (by means of bone conduction) that may help to understand degraded speech is a technological advance that has already been used in mobile phones (e.g., in the Pantech A1407PT model).

Many studies conducted over the last few decades have repeatedly shown that linguistic information can be retrieved not only from the acoustic signal, but also from the visual speech signal (e.g., McGurk & MacDonald, 1976; Ross et al., 2007; Sumby & Pollack, 1954). For example, the kinematics involving the language articulators (the jaw, the cheeks, and the mouth), as well as head movements, can provide information concerning certain acoustic properties of the signal, such as the fundamental frequency or the voice of the speaker (Kamachi, Hill, Lander, & Vatikiotis-Bateson, 2003; Vatikiotis-Bateson et al., 1996; Yehia, Kuratate, & Vatikiotis-Bateson, 2002), and even more complex information (e.g., lexical stress, syntactic boundaries, and pragmatics; see Hadar et al., 1983, 1984; Munhall et al., 2004; Risberg & Lubker, 1978). Soto-Faraco and his colleagues (2007) have demonstrated that adults can successfully discriminate the facial movements associated with different languages, even when those languages differ only minimally. Strikingly, these discrimination abilities generalize to very young infants of less than 6 months of age

(Weikum et al., 2007). The rhythmic (temporal) characteristics of the languages are, according to Ronquest, Levi, and Pisoni (2010), one of the available cues to identify a particular language visually. Importantly, prior experience with particular languages seems to reduce the ability of an individual to use non-native supra-segmental cues such as stress (e.g., Dupoux et al., 1997) or pitch in a tonal language (Wang et al., 1999) auditorily. An interesting question regards whether these effects of linguistic experience can also be observed in the perception of syllabic rhythm or not. Recent evidence suggests that linguistic experience with one particular language hampers the visual (lip-reading) discrimination of unfamiliar and non-native languages in both infancy (Weikum et al., 2007; in press) and adulthood (Soto-Faraco et al., 2007). However, we still do not know which linguistic cues (differences at the visemic level, stress patterns, syllabic rhythm…) can be modulated by native experience and which of them cannot.

Research on sensory substitution systems for deaf and deaf-blind individuals has shown that many different kinds of linguistic information can be delivered, within certain limits, by means of patterns of vibrotactile stimulation (see Summers, 1992, for a review). In the present study, we also addressed the question of whether or not the rhythmic information present in speech can be extracted from visual (Experiments 1 and 2) and tactile temporal patterns (Experiment 3). The *flat sasasa* manipulation was used here as a tool with which to investigate the possible contribution of rhythmic information to speech perception through different modalities (vision, audition, and touch).

2. EXPERIMENT 1

Experiment 1 was designed to address the question of whether visual information suffices to discriminate between languages on the basis of rhythm alone. By including samples of participants from different linguistic backgrounds (native and non-native speakers of English), we were also able to investigate the possible role of prior experience in language discrimination through rhythm. To this end, we created a visual version of Ramus et al.'s (1999) *flat sasasa* materials, consisting of a schematic face articulating the phonemes /s/ and /a/[1]. We included an acoustic version of the stimuli in order to replicate

the main conditions tested in Ramus et al.'s previous study, and an audiovisual condition in order to test whether or not the combination of auditory and visual information might lead to any improvement in performance during the discrimination task (see Navarra et al., 2007). In contrast with some complementarities observed between audition and vision in, for example, phonetic perception, where auditory and visual inputs might sometimes carry different aspects of the speech information (see Summerfield, 1987), the redundancy between modalities is almost complete for rhythm (the alternation of consonants and vowels periods). Bearing this in mind, it is unclear whether or not bimodal presentation would necessarily be expected to lead to a multisensory gain with respect to the unimodal presentations (of visual or auditory stimuli in isolation).

2.1. Methods

2.1.1. Participants

Thirty-one naïve participants (25 female, mean age of 22 years) took part in Experiment 1. Eleven of the participants were English native speakers and 20 were Spanish native speakers who also spoke Catalan. All of the participants reported having normal hearing and normal or corrected-to-normal vision and received course credit (the Spanish group) or a £5 gift voucher (the English group) in exchange for their participation. The experiments were conducted in accordance with the Declaration of Helsinki.

2.1.2. Materials

2.1.2.1. Auditory and visual speech re-synthesis. The "*flat sasasa face*". In order to isolate the syllabic rhythm from any other possible cues, the sentences corresponding to the *flat sasasa* condition in Ramus et al.'s (1999) study were used in the present study. In that study, auditory recordings from another previous study (Nazzi et al., 1998) were employed. These recordings were obtained, in Nazzi et al.'s (1998) study, from 4 English and 4 Japanese speakers, who read 5 different sentences in one of the languages. The use of sentences from different speakers was crucial in order to minimize the possible effects of speakers' particularities in terms of delivering undesired segmental and suprasegmental

cues for discrimination (e.g., a speaker producing the same vowel with different average duration in English and in Japanese).

In the *flat sasasa* manipulation, all of the vowels were digitally re-synthesized as free-of-intonation /a/ and all of the consonants as /s/. Low-level discriminative cues other than syllabic rhythm were not kept in the final re-synthesized version of the sentences. The use of flat digitally-resyntesized versions of /s/ and /a/ allowed us to condensate the vowel and consonant intervals of the sentences (see also Ramus & Mehler, 1999). Therefore, the differences between the English and Japanese materials only existed in temporal-rhythmic dimensions (e.g., the temporal intervals of /s/ and /a/ being more variable in English than in Japanese).

The auditory stimuli (mean-fundamental frequency of 230 Hz) lasted 2640 ms on average (the English and Japanese sentences were 2720 ms and 2560 ms in duration, respectively) and were presented at 68 dB(A), as measured from the participant's head position, via two loudspeakers, one located on either side of the computer screen (Labtec LCS 1050, China, for the Spanish group; Dell Multimedia Speaker A215, China, for the English group). An 80-ms ramp (from 0 to 100% sound intensity) and a 120-ms ramp (from 100% back down to 0% sound intensity) were introduced in the onset and the offset of all the stimuli, respectively. The fade out of the offset was followed by 80 ms of silence.

The use of a mean fundamental frequency of 230Hz to re-synthesize the sentences used in Ramus et al.'s (199) and Nazzi et al.'s (1998) studies allowed as to deliver auditory streams that did not contain any prosody or intonation. The average number of syllables was matched for all of the sentences (16.2 syllables per sentence in both languages). Due to video-frame length constraints (1 frame = 40 ms), the consonant and vowel intervals were modified slightly in accordance with the following rules: All of the (vowel and consonant) intervals that lasted for up to 60 ms were replaced with a 40-ms interval, the intervals that lasted between 60 ms and 100 ms were replaced with a 80-ms interval, the intervals that lasted between 100 ms and 140 ms were replaced with a 120-ms interval and so on[2]. All the sentences (with the new, although very similar, timings) were subsequently re-synthesized using the MBROLA software (Dutoit et al., 1996).

Insert Figure 1 about here

The visual stimuli were created using two different outline pictures (see Figure 1), based on pictures of a real person articulating the phonemes /s/ and /a/. The acoustic sine wave of the sentences was taken as a model for the correct concatenation of the visual /s/ and /a/ images. We used Adobe Premiere 6.0 to achieve this. Finally, the auditory and visual streams were mixed using the same software to create the audiovisual videoclips.

The experiment involved 40 sentences (20 in English and 20 in Japanese; see Ramus & Mehler, 1999, for details). There were 3 different versions of each sentence: auditory, visual, and audiovisual (giving rise to 120 sentences in total). The videoclips were presented in a rectangle (14cm horizontally x 9cm vertically) at the center of a black screen (X-black, LCD, 18 x 28.5cm) at a rate of 25 frames per second. All of the clips started with a 80 ms fade-in from black during the onset of the utterance, and ended with a 120 ms fade-out to black that led to the last 2 frames of each sentence, that were presented in black to avoid possible confounds, in the interpretation of the results, due to the fact that 95% of the Japanese sentences (but just 30% of the English sentences) ended in a vowel (see Ramus & Mehler, 1999, Appendix).

2.1.3. Procedure

The experiment was conducted in a dark, sound-attenuated booth. The participants sat 50 cm from the computer screen. Each participant completed 3 blocks of 40 trials (auditory, visual, and audiovisual; with the order of presentation counterbalanced across participants). Each trial followed the same sequence: A sentence in English or in Japanese was presented. After a 1000-ms interstimulus interval (black screen), a second sentence (with the same number of syllables) in one of the two languages was played back. The second sentence was immediately followed by the question “Same or Different?”, presented on the screen as a prompt for participants to respond (pressing “S” or “D”, respectively, on a

computer keyboard). There was a 3000 ms response deadline. The participants were encouraged to find any possible difference between the two “Martian” languages being presented and to respond even if they were unsure of the correct answer. The next trial started 1000 ms after the participant’s response on the preceding trial. The stimulus pairs were randomly assembled during the experiment, with only two restrictions: (1) If one sentence appeared in the first position during the block, it was presented in the second position on its second appearance (or vice versa); (2) There were a minimum of 2 trials between the first and the second presentation of each sentence, in order to minimize any possible effect of recency from trial to trial. The percentage of same/different trials in each block was, in both cases, 50%.

2.2. Results and discussion

First of all, a correlation analysis was performed between overall sentence duration and performance to check whether participants might have used duration (i.e., the fact that the English sentences were, on average, 160 ms longer than the Japanese sentences) as a potential cue to aid discrimination. The basic idea behind this analysis was to see whether the duration of the sentences correlated positively with accuracy in English sentences (i.e., the longer the sentence, the better the performance) and negatively in Japanese sentences (i.e., the shorter the sentence, the better the performance). Only the data from those blocks of trials in which the participants performed at a level that was clearly above chance (that is, including data exclusively from the upper quartile of the distribution) were used in this analysis. No correlation was found between sentence duration and performance [R^2 = .001, $F < 1$; R^2 = .005, $F < 1$; R^2 = .022, $F < 1$ between duration and % correct for the auditory, the visual, and the audiovisual English sentences, respectively; and R^2 = .065, F = 1.24, p = .28; R^2 = .023, $F < 1$; R^2 = .011, $F < 1$ between duration and percent correct for the auditory, the visual, and the audiovisual Japanese sentences, respectively]. The same analysis was also performed with the data from Experiments 2 and 3, leading to very similar results, that is, no trace of any correlation between sentence duration and participants’ performance.

The percentage of correct discrimination responses was calculated for each participant and modality of presentation (visual, auditory, and audiovisual). Individual t-tests revealed that the participants responded at significantly above chance levels in all three presentation modes [$t(30) = 5.3$, $p = .00001$; $t(30) = 3.59$, $p = .001$; and $t(30) = 4.5$, $p = .0001$, for the auditory, visual, and audiovisual conditions, respectively], indicating that discrimination between the transformed English and Japanese sentences was possible in all of the conditions (see Figure 2A). Performance in all three conditions (auditory, visual, and audiovisual) was significantly better than chance level for both the English speakers [$t(10) = 4.9$, $p = .001$; $t(10) = 2.3$, $p = .04$; $t(10) = 3.2$, $p = .009$, respectively] and for the Spanish speakers [$t(19) = 3.3$, $p = .004$; $t(19) = 3$, $p = .007$; $t(19) = 3.3$, $p = .004$, respectively].

Insert Figure 2 about here

Non-parametric analyses (one-sample Wilcoxon signed rank test) confirmed that the whole group of participants (including English and Spanish speakers) discriminated English and Japanese at levels that were significantly above chance auditorily ($z = 3.994$, $p < .0001$), visually ($z = 3.141$, $p=.002$), and audiovisually ($z = 3.603$, $p < .0001$). One-sample Wilcoxon tests performed for the English and the Spanish groups separately also confirmed that both groups were able to discriminate the sasasa patterns in the auditory ($z = 2.706$, $p = .007$ and $z = 2.809$, $p = .005$ for the English and the Spanish group, respectively), visual ($z = 1.983$, $p = .047$ and $z = 2.631$, $p = .009$ for the English and the Spanish groups, respectively), and audiovisual conditions ($z = 2.446$, $p = .014$ and $z = 2.730$, $p = .006$ for the English and the Spanish groups, respectively)[3].

In order to verify whether the results found in the visual condition were not due to the fact that some participants had previous experience with the same materials presented auditorily and/or audiovisually, we compared the discrimination performance of those participants who received the visual condition at the beginning of the experiment (i.e., before

the auditory and audiovisual conditions; 8 participants) with those who received the visual condition at the end of the experiment (12 participants). A Mann-Whitney U test revealed no statistical difference, in terms of discrimination, between these two groups ($U = 46.5$; $p = .910$).

The possible effects of prior experience with one of the languages used in the experiment were explored by comparing the discrimination performance of English and Spanish speakers. A subsequent repeated-measures ANOVA including the factors modality (visual, auditory and audiovisual) and participants' L1 (English vs. Spanish) revealed a marginally-significant effect of modality ($F(2,58)=2.463$, $p=.094$). No interaction was found between modality and participants' L1 ($F(2,58)=2.3$, $p=.112$).

Insert Figure 3 about here

In summary, the results of Experiment 1 show that it is possible to discriminate between different spoken languages using nothing more than the sight of the rhythmic movements of digitally-generated schematic articulators. This result would perhaps help to decipher the processes that may underlie the ability that humans show, even at just 4 months of age, to visually discriminate languages. Since other segmental (e.g., phonetics) or supra-segmental (e.g., intonation contour) linguistically-relevant cues had been removed from our materials, it can be concluded that the discrimination ability exhibited by the participants in Experiment 1 was based on rhythmic cues (i.e., the temporal distribution of consonant and vowel intervals in the speech signal; see Introduction). According to our results, the participants' prior experience with one of the languages did not increased rhythm discrimination of English and Japanese significantly.

An interesting question to emerge from the results of Experiment 1 concerns whether the mechanisms underlying the discrimination of speech rhythm are specific to language or else rely on more general abilities that also emerge while perceiving non-linguistic stimulus

patterns. This question was motivated by previous evidence suggesting that the brain mechanisms that are involved in the processing auditory stimuli such as speech or as non-speech may not be the same (e.g., see Vouloumanos, Kiehl, Werker, Liddle, 2001). We addressed this issue in Experiment 2 by presenting the same rhythmic patterns as in Experiment 1, but using simpler stimuli in a non-linguistic context.

3. EXPERIMENT 2

In Experiment 2, we investigated whether the discrimination between the rhythms associated with English and Japanese can be successfully achieved even when the patterns presented do not resemble speech articulations in any obvious way. In this experiment, the vowel intervals of Experiment 1 were replaced with 500Hz pure tones (auditory modality) and with a black flickering circle (in the visual modality) instead of the schematic face. Any reference to the linguistic origins of the rhythmic patterns and to speech was removed from the instructions that were given to the participants. Considering both the results of Experiment 1 and previous evidence suggesting that the linguistic background does not have strong effects on the ability to discriminate languages visually (see also Soto-Faraco et al., 2007), we decided to concentrate more on general/across-language discrimination of syllabic rhythm than on cross-linguistic differences.

3.1. Methods

The method (and experimental settings) was the same as in Experiment 1 with the following exceptions.

3.1.1. Participants

Eleven naïve participants[4] (8 female; mean age of 22 years) took part in Experiment 2. All of the participants received a £5 gift voucher in exchange for their participation.

3.1.2. Materials

The timings of all the stimuli were the same as in Experiment 1. The auditory /a/ intervals in the *flat sasasa* sentences were replaced by 500Hz pure tone intervals presented at 68 dB(A). The /s/ consonant intervals (basically short noise bursts, which could easily be

conceived of as being produced by a non-human agent) were kept as in Experiment 1. The onsets and offsets of all the tones presented had a 5-ms ramp (from 0 to 100%, and from 100 to 0% sound intensity, respectively). In order to create the 'non-linguistic' visual materials, the pictures corresponding to /s/ and /a/ in Experiment 1 were replaced by a picture of a small circle (2 mm in diameter) and another picture of a bigger circle (6 cm in diameter), respectively. The videoclips were presented on a grey rectangle (with the same dimensions as in Experiment 1) at the center of a black screen (X-Black, LCD, 18 x 28.5cm), with a rate of 25 frames per second.

In order to ensure that the materials used in this experiment were not perceived as speech, a brief test was performed at Hospital Sant Joan de Déu, in which 10 participants (7 female, mean age of 25 years) judged whether 10 randomly-selected audiovisual streams (5 in English and 5 in Japanese) resembled human speech or not. The participants were not informed about the use of the test and performed their judgment after being presented with all of the 10 sentences to avoid any bias towards the "similar to speech" response. None of the participants found any resemblance between the audiovisual streams and language/speech. However, 4 of them spontaneously reported to perceive similarities between our materials and other communication systems such as the Morse code.

3.1.2. Procedure

The participants were told that the (auditory, visual, or audiovisual) patterns originated from two different "mechanical devices", and they were instructed to find any possible difference between the two kinds of patterns being presented (each one originated from a different "mechanical device"). The task was to decide whether the two patterns in each pair corresponded to the same "machine" or not. As in Experiment 1, the participants did not receive any practice before taking part the experiment.

3.2. Results and discussion

The percentage of correct responses was calculated for each participant and modality of presentation, just as in Experiment 1. T-tests for each presentation modality revealed that performance was, again, above chance levels in all three modalities [$t(10)$ =

3.9, p = .003; t(10) = 3.6, p = .005; and t(10) = 6.3, p = .00009], for the auditory, visual, and audiovisual conditions, respectively; see Figure 2B]. A one-sample Wilcoxon signed rank non-parametric test confirmed these results (z = 2.558, p = .011; z = 2.388, p = .017; and z = 2.823, p= .005; for the auditory, visual and audiovisual conditions, respectively). According to the present results, the discrimination between the rhythmic patterns derived from the English and the Japanese sentences was possible even when the appearance of the stimuli bore no relation to speech (mean = 59.3%, 53.9% and 60% correct; median = 60%, 55% and 62.5% correct, respectively).

A subsequent repeated-measures ANOVA including the variable 'modality of presentation' (with three levels: auditory, visual, and audiovisual) revealed a significant main effect of this variable [F(2, 20) = 5.8, p = .01]. Further analyses determined that the audiovisual condition led to significantly better discrimination than the visual condition (t(10) = -3.418, p = .007). A marginally significant difference was also found between the auditory and visual conditions (t(10) = 2.185, p = .054). Non-parametric analyses confirmed these results (z = -2.567, p = .01; and z = 2.388, p = .074; for the visual-audiovisual and the visual-auditory comparisons, respectively).

Further analyses were carried out in order to compare the performance of participants in Experiments 1 and 2. More specifically, another repeated-measures ANOVA, carried out with the data from these two experiments and including the within-participants factor of 'modality of presentation' (auditory, visual, and audiovisual) and the between-participants factor 'linguistic context' (language, vs. non-language, in Experiments 1 and 2, respectively), revealed only a significant main effect of modality [F(2, 80) = 4.583, p = .013]. Non-parametric analyses (Wilcoxon signed rank test) revealed a difference between the visual and the auditory conditions (z = -2.214, p = .027), a difference between the visual and the audiovisual conditions (z = -2.689, p = .007), but no difference between the auditory and the audiovisual conditions (z = -.416, p = .677).

As in Experiment 1, a Mann–Whitney U test suggested that the participants who received the auditory and/or the audiovisual conditions before the visual condition in

Experiment 2 had no advantage in the visual condition relative to those participants who received the visual condition at the start of the experiment ($U = 4.5$; $p = .323$).

In summary, the outcome of Experiment 2 suggests that our participants were able to discriminate between English and Japanese on the basis of simplified temporal patterns, even if these patterns bore no direct resemblance to speech articulations (i.e., a circle was presented instead of a face and non-human digital sounds). The results of Experiment 2 also suggest that, despite the fact that visual information is more than sufficient for discrimination, the addition of auditory information leads to a significant improvement in performance. This hypothesis is in line with the well established idea that the auditory modality is better suited to represent time than the visual modality (see Lhamon & Goldstone, 1974). Since we cannot be completely sure that the participants did not subconsciously perceive the audiovisual materials of Experiment 2 as speech-like (see Spence & Deroy, 2012), the plausible conclusion that the discrimination of the temporal characteristics of syllabic rhythm could be based on general (rather that language-specific) mechanisms is still open to further research. However, our results so far indicate a rather abstract encoding of rhythmic information, so that rhythmic patterns can be discriminated with relative independence of sensory modality and presentation format. Therefore, we thought it important to address whether these rhythmic attributes could also be encoded and discriminated via a different sensory modality such as touch. Several studies have already demonstrated that some language cues, such as intonation and stress, can be extracted, within certain limits, from tactile vibrations originating from speech (e.g., Auer et al., 1998; Gick & Derrick 2009), so we decided to test the discriminability of consonant-vowel speech rhythm in the tactile domain as well.

4. EXPERIMENT 3

As in Experiment 1, the participants were instructed to perform a discrimination task that consisted of deciding whether pairs of sentences, presented through a vibrotactile stimulator, had been presented in the same or different language. Considering the strong weight given to training and practice in previous studies that have looked at the efficacy of

tactile aids (see Summers, 1992, for reviews of the methodological aspects of testing tactile speech aids), we decided to provide the participants in Experiment 3 with some practice on the discrimination of rhythm through tactile vibrations. Consequently, the first half of the experiment (2 blocks of trials) included feedback regarding the participant's performance. Feedback was not provided in the second half of the experiment (the last 2 blocks).

4.1. Methods

4.1.1. Participants

15 naïve participants[5] (10 female, mean age of 23 years), with normal hearing, with no history of reported somatosensory deficits and normal or corrected-to-normal vision (by self-report) took part in this study. The participants were given a £5 gift voucher in return for taking part in the study.

4.1.2. Materials

Syllabic rhythm was once again isolated from any other cues by presenting the sentences corresponding to the *flat sasasa* condition of Ramus et al.'s (1999) study. In contrast with Experiments 1 and 2, no further transformations were applied to the original Ramus and Mehler (1999) *flat sasasa* sentences in Experiment 3 (there were no temporal restrictions due to the use of video frames). One bone-conducting vibrator (Oticon-A, 100Ω; Hamilton, UK) with a 1.6 cm by 2.4 cm vibrating surface was used to deliver the vibrotactile stimuli. The vibrator was attached to an amplifier (Optimus SA-155, China) that received the input from a PC computer controlling the experiment. The vibrotactile stimulator was attached to the participant's right index fingertip. The intensity of the vibratory signal was adjusted until the participants could clearly perceive the vibratory intervals corresponding to /a/ in *flat sasasa*, while presenting the intervals that corresponded to /s/ as nearly imperceptible noise. Prior to conducting Experiment 3, the experimenter ensured that all of the vibrotactile intervals were fully detectable (and not perceived as continuous at the fastest rates). It is worth noting that all of the intervals between vibrations were longer than 20 ms and therefore perfectly detectable (see Gescheider, Bolanowski, & Chatterton, 1979). White

noise was presented continuously over headphones (Beyerdynamic DT-531, Germany) at 70 dB(A) in order to mask the sound elicited by the operation of the tactor.

4.1.3 Procedure

The experiment was conducted in a dark, sound-attenuated booth. The participants sat approximately 50 cm from the computer screen, on which the instructions were presented. As in Experiment 1, the task involved the participant having to decide on each trial whether the two sentences had been presented in the same or different 'Martian' language. The instructions also encouraged the participants to find "something" in the tactile sequences that would allow them to differentiate between the two languages. Each participant completed 4 blocks of 40 trials, receiving feedback regarding their performance ('correct' or 'wrong' response, presented on the screen shortly after their response) in the first 2 blocks of trials. All the other details of the procedure used in Experiment 3 were kept exactly as in Experiments 1 and 2.

4.2. Results and discussion

The percentage of correct responses was calculated for each participant and modality of presentation, as in Experiments 1 and 2. A t-test, collapsing the data for all four blocks revealed that as a whole the group of participants performed the task significantly above chance level [$t(14) = 3.2$, $p = .006$], indicating that discrimination between the vibrotactile sequences derived from the English and Japanese sentences was possible (see Figure 2C). A more detailed analysis revealed that performance was not statistically different in the blocks containing feedback (i.e., the first two) and the blocks without feedback [$t(14) = -.59$, $p = .57$]. This result was confirmed non-parametrically ($z = -.44$, $p = .66$).

Further analyses (repeated-measures ANOVA) comparing the performance of participants in Experiments 1 and 3 (both involving language discrimination but in different sensory modalities) did not yield any significant difference [$F(1, 45) = 2.2$, $p = .15$, $F(1, 45) < 1$, $F(1, 45) = 2.09$, $p = .16$, for the comparison between the tactile and the auditory, the tactile and the visual, and the tactile and the audiovisual conditions, respectively].

The results of Experiment 3 provide the first empirical evidence to demonstrate that people are able to discriminate vibrotactile sequences corresponding to the rhythmic alternation (vowel and consonant intervals) directly derived from speech in different languages. The extraction of linguistically-reliable time-based information seems to be, at least within certain limits, quite independent of the sensory modality in which that information happens to be presented.

5. General Discussion

Our results suggest that the discrimination between languages from different 'rhythmic' groups is possible not only auditorily but also visually and via the sense of touch, on the basis of the temporal distribution of consonants and vowels in the speech signal. In Experiment 1, we extended previous results suggesting that visual discrimination between different silent movement patterns is possible (e.g., Mendelsson, 1986) to the case of the perception of more complex temporal patterns derived from natural speech. These results are in line with previous evidence (obtained with other manipulations of the visual speech; see Ronquest et al., 2010; Soto-Faraco et al., 2007) suggesting that the processing of the temporal patterns associated with each language is one of the cues that allow for its correct identification. It is worth highlighting that, unlike in previous research in the field, our participants did not have the opportunity to familiarize with the task and had no previous experience with the *sasasa* materials prior to the experiment itself. Despite this clear disadvantage (relative to participants in previous studies), the participants performed at a level that was significantly above chance.

In Experiment 2, we demonstrated that people can tell the difference (visually, auditorily, and audiovisually) between the rhythm derived from English and Japanese even when the rhythmic patterns themselves are not explicitly associated with speech, such as when presenting a flickering circle and a beeping 500-Hz tone. This experiment also suggested that rhythm discrimination may be better in the auditory and the audiovisual domain than in the visual domain. This result is, once again, in line with previous research pointing to the idea that the auditory domain is preferential for the processing of temporal

information (see Lhamon & Goldstone, 1974; Recanzone, 2003; Wada, Kitagawa, & Noguchi, 2003; Welch, DuttonHurt, & Warren, 1986). However, an alternative explanation might be that the visual materials used in Experiments 1 and 2 were not as suitable as the auditory materials to convey the temporal dynamics of consonants and vowels. More research is therefore needed in order to further clarify the possible superiority of audition over vision to deliver syllabic rhythm.

The results of Experiment 3 demonstrated that discrimination is also possible when the English and Japanese rhythmic patterns are presented by means of tactile vibrations delivered on the fingertip.

Altogether, our results help interpret previous findings of visual discrimination between languages such as French and English by infants (Weikum et al., 2007), and by adults (between Catalan and Spanish; i.e., two languages with a very similar phonemic repertoire and few subtle suprasegmental differences; Soto-Faraco et al., 2007). In particular, the results of the present study suggest that visual discrimination is probably not only based on visemic differences (corresponding to the segmental level of the speech signal), but also on rhythmic differences (corresponding to the supra-segmental level). It will be interesting in future research to address the question of whether syllabic rhythm also provides a useful cue with which to help discriminate languages that are closer in terms of their rhythmic properties (e.g., Catalan vs. Spanish, or English vs. Dutch; see Nazzi et al., 1998; Ramus, Nespor, & Mehler, 1999).

The present results also suggest that extensive previous exposure to the *flat sasasa* materials, or to at least one of the languages the patterns were derived from, is not necessary for people to successfully use rhythm in the visual domain (see also Ramus & Mehler, 1999). This is apparently at odds with the results reported by Soto-Faraco et al. (2007), where previous linguistic experience with at least one of the test languages proved critical for observers to discriminate visually Spanish from Catalan. It is, though, worth stressing that in the present study, the observers were confronted with a task that involved the discrimination between two languages belonging to different rhythmic classes, and that rhythm was about the only cue left in the signal. By contrast, Soto-Faraco et al. presented

unfiltered visual speech in Spanish and Catalan, two languages belonging to the same rhythmic class. Therefore, it is likely that the observers who were tested in Soto-Faraco et al.'s study had to resort to more language specific strategies in order to perform the task. If this had in fact been the case, then the modulation by linguistic experience would have been more likely.

Another important aspect of the present findings is that, as opposed to previous research that has measured different aspects of the speech signal (e.g., Molholm, Ritter, Javitt, & Foxe, 2004; Navarra et al., 2007), no significant benefit was observed when the auditory and visual speech streams were presented together, as compared to when they were presented in isolation (see Ross et al., 2007; Sumby & Pollack, 1954). This difference in results probably reflects two facts. First, the visual and auditory rhythmic patterns seem to be more redundant than complementary, as opposed to the well-known complementary nature of the vision and audition in terms of segmental content (see Summerfield, 1987; for possible examples about visemic and phonetic complementarities). Second, the auditory information is probably much more precise at transmitting temporal patterns than the visual modality (as the slightly superior performance in the auditory and the audiovisual conditions than in the visual conditions in Experiments 1 and 2 suggests), just making it unlikely to observe any benefit from the combination of the two from a computational point of view. It would be interesting to investigate whether information concerning the rhythm of speech, presented through the visual (or the tactile) channel can perhaps help to understand speech in less favorable, noisy environments (see Sumby & Pollack, 1954).

As Liberman and his colleagues suggested several decades ago (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1968), acoustic speech cues carry information about successive phonemic segments (due to coarticulation) and this allows our perceptual systems to decode the (fast) speech signal more easily. Unfortunately, however, tactile stimuli seem to be, a priori, a quite limited format for carrying all of this complex information (i.e., coarticulated phonemes being presented at extremely fast rates). Over the years, researchers working on the development of tactile aids for deaf and deaf-blind individuals have tried to solve this fundamental problem by: 1) slowing down the speech signal in order

to allow sufficient time to process the speech utterances as discrete units (e.g., Keidel, 1968; Newman, 1960; Rothenberg & Molitor, 1979); or 2) by presenting as much speech information as possible by means of different channels, by means of what was called a "vocoder" (see Kirman, 1973); that is, delivering qualitatively different information to different skin sites. The results of many such studies have shown that, even when perceivers find these sensory substitution devices very difficult to get used to, a remarkable improvement in the perception and comprehension of, for example, phonemes, syllables, words and/or sentences can be observed after some hours of training (see Kirman, 1973). Unfortunately, however, as many authors have pointed out (Brown, Nibarber, Ollie, & Solomon, 1967, Picket & Picket, 1963; Sherrick, 1964), the use of tactile aids that stimulate more than one location of the body surface is probably limited by the (initially unforeseen) effects of tactile phenomena such as tactile simultaneous masking (see Gallace & Spence, 2008, 2014, for reviews).

Over the last few decades, a number of researchers have proposed that several aspects of prosody can be effectively transmitted to the perceiver through the use of tactile devices (see Summers et al., 1997, for a review). What is more, vibrotactile information is certainly helpful when it complements the incoming information from other sensory modalities (i.e., vision and audition) rather than when it is used solely as a substitute for visual or auditory speech signals (see Cowan et al., 1990; Eberhardt et al., 1990; Yuan et al., 2005). It is in this context (and also in interpersonal communication technologies; e.g., in noisy environments) where signal manipulations such as the tactile version of *flat sasasa* reported here may prove most useful. Our findings show that some information concerning the temporal patterns (i.e., suprasegmental information) of language can be extracted from relatively simple sequences of tactile vibrations. From the point of view of the present study, one of the possible keys to help decode the (highly-informative) temporal patterns of speech using visual or tactile aids could reside in reducing the complexity of the speech signal by grouping vowels and consonants as in fact happens in the *flat sasasa* manipulation.

According to some authors (e.g., Jusczyk, 1997; Mehler et al., 1996; Mehler & Nespor, 2004), the suprasegmental-prosodic information (e.g., rhythm) provides the

grounding for pre-lexical infants to segment speech in speech units to learn (e.g., words), in a process called “prosodic bootstrapping”. Since the segmentation of the speech signal into meaningful units (words) may constitute one of the first and most important problems that we encounter during speech perception under certain conditions (e.g., such as in noisy environments), it is not unreasonable to expect that a simultaneous (or slightly advanced; see Stekelenburg & Vroomen, 2007; van Wassenhove, Grant, & Poeppel, 2005) presentation of the visual or the tactile equivalent of *flat sasasa* might provide some additional information regarding, for example, the boundaries between certain phonemes, syllables, or even words, thus facilitating the segmentation of the speech signal. Ongoing research in our laboratory is currently testing this possibility.

Finally, it is interesting to note that recent studies have shown that vibrotactile information can activate auditory cortex (Auer, Bernstein, Sunqkarat & Sinqh, 2007; Caetano & Jousmaki, 2006; Schürmann et al., 2006; see Kitagawa & Spence, 2006, for a review). Considering such evidence (as well as other evidence of activity in auditory cortex during the perception of visual speech; Calvert et al., 1997; Sams et al., 1991; though see Bernstein et al., 2002), it is possible that the ability of people to discriminate rhythm demonstrated in the present study may have been mediated by an early tactile-to-auditory or visual-to-auditory recoding of the original input. The nature of the discrimination patterns observed therefore needs to be clarified in future research. In any case, the possibility that vibrotactile stimuli can be translated into some sort of auditory code has to be accommodated in the literature on sensory substitution and tactile aids for speech perception.

AUTHOR NOTES

This research was supported by grants PSI2012-39149, PSI2009-12859 and RYC-2008-03672 from *Ministerio de Economía y Competitividad* (MINECO, Spain) to J. Navarra, the European COST action TD0904, and by grants PSI2010-15426 and CDS00012 from MINECO (Spain), 2009SGR-292 from *Comissionat per a Universitats i Recerca del DIUE-Generalitat de Catalunya,* and European Research Council (StG-2010 263145) to S. Soto-Faraco.

REFERENCES

Auer, E.T., Jr., Bernstein, L.E., & Coulter, D.C., 1998. Temporal and spatio-temporal vibrotactile displays for voice fundamental frequency: An initial evaluation of a new vibrotactile speech perception aid with normal-hearing and hearing-impaired individuals. *Journal of the Acoustical Society of America*, *104*, 2477-2489.

Auer, E.T., Jr., Bernstein, L.E., Sunqkarat, W., & Sinqh, M., 2007. Vibrotactile activation of the auditory cortices in deaf versus hearing adults. *Neuroreport*, *18*, 645-648.

Bernstein, L.E., Auer, E.T., Jr., Moore, J.K., Ponton, C.W., Don, M., & Sinqh, M., 2002. Visual speech perception without primary auditory cortex activation. *Neuroreport*, *13*, 311-315.

Brown, R.L., Nibarber, D., Ollie, G., & Solomon, A., 1967. A differential comparison of two types of electropulse alphabets based on locus of stimulation. *Perceptual and Motor Skills*, *24*, 1038-1044.

Brutt-Griffler, J. and Varghese, M. (2004). Introduction. In J. Brutt-Griffler, and M. Varghese (Eds.), *Bilingualism and language pedagogy* (pp. 1-9). New York, NY: Multilingual Matters Limited.

Caetano, G., & Jousmaki, V., 2006. Evidence of vibrotactile input to human auditory cortex. *Neuroimage*, *29*, 15-28.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., & David, A.S., 1997. Activation of auditory cortex during silent lipreading. *Science, 276*, 593-596.

Cowan, R.S.C., Blamey, P.J., Galvin, K.L., Sarant, J.Z., Alcántara, J.I., & Clark, G.M., 1990. Perception of sentences, words, and speech features by profoundly hearing-impaired children using a multichannel electrotactile speech processor. *Journal of the Acoustical Society of America*, *88*, 1374-1384.

de Bot, K., & Kroll, J. F., 2002. Psycholinguistics. In N. Schmitt (Ed.), *Introduction to applied linguistics* (pp. 133-149). New York: Arnold Publishers.

Dupoux, E., Pallier, C., Sebastián-Gallés, N., & Mehler, J., 1997. A destressing "deafness" in French? *Journal of Memory and Language*, *36*, 406-421.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & 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).

Eberhardt, S.P., Bernstein, L.E., Demorest, M.E., & Goldstein, M.H., Jr., 1990. Speechreading sentences with single-channel vibrotactile presentation of voice fundamental frequency. *Journal of the Acoustical Society of America*, *88*, 1274-1285.

Gallace, A., & Spence, C., 2008. The cognitive and neural correlates of "tactile consciousness": A multisensory perspective. *Consciousness and Cognition*, *17*, 370-407.

Gallace, A., & Spence, C. (2014). *In touch with the future: The sense of touch from cognitive neuroscience to virtual reality*. Oxford: Oxford University Press.

Gescheider, G.A., Bolanowski, S.J., & Chatterton, S.K., 2003. Temporal gap detection in tactile channels. *Somatosensory & Motor Research*, *20*, 239-247.

Gick, B., & Derrick, D. 2009. Aero-tactile integration in speech perception. *Nature*, 462, 502-504.

Grabe, E., & Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology* (pp. 515-546). Berlin: Mouton de Gruyter.

Hadar, U., Steiner, T.J., Grant, E.C., & Rose, F.C., 1983. Head movement correlates of juncture and stress at sentence level. *Language and Speech, 26*, 117-129.

Hadar, U., Steiner, T.J., Grant, E.C., & Rose, F.C., 1984. The timing of shifts in head posture during conversation. *Human Movement Science, 3*, 237-245.

Jusczyk, P.W., 1997. *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E., 2003. "Putting the face to the voice": Matching identity across modality. *Current Biology*, *13*, 1709-1714.

Keidel, W.D., 1968. Electrophysiology of vibratory perception. In W.D. Neff (Ed.), *Contributions to sensory physiology* (Vol. 3, pp. 1-79). New York: Academic Press.

Kirman, J.H., 1973. Tactile communication of speech: A review and an analysis. *Psychological Bulletin*, *80*, 54-74.

Kitagawa, N., & Spence, C., 2006. Audiotactile multisensory interactions in information processing. *Japanese Psychological Research*, *48*, 158-173.

Lhamon, W.T., & Goldstone, S., 1974. Studies of auditory-visual differences in human time judgment: 2. More transmitted information with sounds than lights. *Perceptual and Motor Skills*, *39*, 295-307.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M., 1968. Why are speech spectrograms hard to read? *American Annals of the Deaf*, *113*, 127-133.

Lloyd James, A., 1940. *Speech signals in telephony*. London, UK: Pitman and Son.

McGurk, H., & MacDonald, J., 1976. Hearing lips and seeing voices. *Nature*, *265*, 746-748.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G., 1996. Coping with linguistic diversity: The infant's viewpoint. In J.L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 101-116). Mahwah, NJ: Erlbaum.

Mehler, J., & Nespor, M., 2004. Linguistic rhythm and the development of language. In A. Belletti (Ed.), *Structures and beyond: The cartography of syntactic structures, Vol. 3* (pp. 213-222). Oxford, UK: Oxford University Press.

Mendelson, M.J., 1986. Perception on the temporal pattern of motion in infancy. *Infant Behavior and Development*, *9*, 231-243.

Molholm, S., Ritter, W., Javitt, D.C., & Foxe, J.J., 2004. Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, *14*, 452-465.

Munhall, K., Jones, J.A., Callan, D.E., Kuratate, T., & Vatikiotis-Bateson, E., 2004. Head movement improves auditory speech perception. *Psychological Science, 15*, 133-137.

Navarra, J., & Soto-Faraco, S., 2007. Hearing lips in a second language: Visual articulatory information enables the perception of L2 sounds. *Psychological Research*, *71*, 4-12.

Nazzi, T., Bertoncini, J., & Mehler, J., 1998. Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 756-766.

Newman, R., 1960. *The feasibility of speech transmission using the skin as a sensor*. Paper presented at the Air Research and Developmental Command Seventh Annual Science and Engineering Symposium, Boston, November, 1960.

Picket, J.M., & Picket, B.M., 1963. Communication of speech sounds by a tactile vocoder. *Journal of Speech and Hearing Research*, *6*, 207-222.

Pike, K.L., 1945. *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Ramus, F., Hauser, M.D., Miller, C., Morris, D., & Mehler, J., 2000. Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, *288*, 349-351.

Ramus, F., & Mehler, J., 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, *105*, 512-521.

Ramus, F., Nespor, M., & Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*, 265-292.

Recanzone, G.H., 2003. Auditory influences on visual temporal rate perception. *Journal of Neurophysiology*, *89*, 1078-1093.

Risberg, A., & Lubker, J., 1978. Prosody and speech-reading. *STL Quarterly Progress and Status Report*, *4*, 1-16.

Ronquest, R.E., Levi, S.V., & Pisoni, D.B. (2010). Language identification from visual-only speech signals. *Attention, Perception and Psychophysics*, *72*, 1601-1613.

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., & Foxe, J.J., 2007. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*, 1147-1153.

Rothenberg, M., & Molitor, R.D., 1979. Encoding voice fundamental frequency into vibrotactile frequency. *Journal of the Acoustical Society of America*, *66*, 1029-1038.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O.V., Lu, S.T., & Simola, J., 1991. Seeing speech: Visual information form lip movements modifies activity in the human auditory cortex. *Neuroscience Letters*, *127*, 141-145.

Schürmann, M., Caetano, G., Hlushchuk, Y., Jousmäki, V., & Hari, R., 2006. Touch activates human auditory cortex. *Neuroimage*, *30*, 1325-1331.

Sherrick, C.E., 1964. Effects of double simultaneous stimulation of the skin. *American Journal of Psychology*, *77*, 42-53.

Soto-Faraco, S., Navarra, J., Weikum, W., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J., 2007. Discriminating languages by speechreading. *Perception & Psychophysics*, *69*, 218-231.

Spence, C., & Deroy, O. 2012. Hearing mouth shapes: Sound symbolism and the reverse McGurk effect. *i-Perception*, *3*, 550-552.

Stekelenburg, J. J., & Vroomen, J., 2007. Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*, 1964-1973.

Sumby, W., & Pollack, I., 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212-215.

Summerfield, Q., 1987. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-51). London, UK: LEA.

Summers, I.R., (Ed.) 1992. *Tactile aids for the hearing impaired (practical aspects of audiology)*. London, UK: Wiley.

Summers, I.R., Cooper, P.G., Wright, P., Gratton, D.A., Milnes, P., & Brown, B.H., 1997. Information from time-varying vibrotactile stimuli. *Journal of the Acoustical Society of America*, *102*, 3686-3696.

Toro, J.M., Trobalon, J.B., & Sebastián-Gallés, N., 2003. The use of prosodic cues in language discrimination tasks by rats. *Animal Cognition*, *6*, 131-136.

van Wassenhove, V., Grant, K.W., & Poeppel, D., 2005. Visual speech speeds up the neural processing of auditory speech. *Proceedings of the Natural Academy of Sciences USA, 102*, 1181-1186.

Vatikiotis-Bateson, E., Munhall, K.G., Kasahara, Y., Garcia, F., & Yeshia, H., 1996. Characterizing audiovisual information during speech. *Proceedings of the 4th International Conference on Language Processing (ICSLP)*. Available online at: http://www.asel.udel.edu/icslp/cdrom/vol3/1004/a1004.pdf

Vouloumanos, A., Kiehl, K.A., Werker, J.F., & Liddle, P.F. 2001. Detecting sounds in the auditory stream: Event-related fMRI evidence for differential activation to speech and non-speech. *Journal of Cognitive Neuroscience*, *13*, 994-1005.

Wada, Y., Kitagawa, N., & Noguchi, K., 2003. Audio-visual integration in temporal perception. *International Journal of Psychophysiology*, *50*, 117-124.

Wang, Y., Spence, M., Jongman, A., and Sereno, J., 1999. Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, *106*, 3649-3658.

Weikum, W.M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J.F., 2007. Visual language discrimination in infancy. *Science*, *316*, 1159.

Weikum, W.M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J.F. 2013. Age-related sensitive periods for visual language discrimination in adults. *Frontiers in Systems Neuroscience*, *7*, 86.

Welch, R.B., DuttonHurt, L D., & Warren, D.H., 1986. Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics*, *39*, 294-300.

Yehia, H.C., Kuratate, T., & Vatikiotis-Bateson, E., 2002. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*, 555-568.

Yuan, H., Reed, C.M., & Durlach, N.I., 2005. Tactual display of consonant voicing as a supplement to lipreading. *Journal of the Acoustical Society of America*, *118*, 1003-1015.

FOOTNOTES

1. A schematic face was used instead of a real face in order to avoid the semantic conflict of pairing an unnatural sound with a human face in the audiovisual condition. We tried to use a real face during the preparation of the stimulus materials, but the result of matching the audio and the “more realistic” video streams was less than optimal (and even distracting, according to some pilot testing).

2. The difference between the duration of the original vocalic and consonantal intervals in Ramus and Mehler’s (1999) study and the duration of the same intervals in our materials was, on average, 9.97ms (SD = 5.8). The temporal distribution of the vowels and consonants was kept nearly as in the original sentences. More importantly, the results of all 3 experiments reported in the present study suggest that this subtle temporal modification applied to the original sentences did not remove the rhythmic information present in Ramus and Mehler’s original materials.

3. Considering previous results suggesting a role of linguistic experience in unimodal (visual) language discrimination (Soto-Faraco et al., 2007; Weikum et al., 2013), more specific analyses were performed. A non-parametric Mann-Whitney U test corroborated the observation that the English speakers were significantly better than the Spanish speakers at discriminating the Japanese and the English temporal consonant-vowel patterns in the auditory condition ($U = 53.5$, $p = .018$), but not in the visual ($U = 137.5$, $p = .26$) or the audiovisual ($U = 112$, $p = .95$) conditions.

4. The sample for Experiment 2 included 6 English and 5 native speakers of other languages including Hebrew, Spanish, and German. In order to avoid any reference to language before or during this experiment, the participants were questioned about their native language after the experiment. Considering that cross-linguistic differences were

found in Experiment 1 only in the auditory condition, the goal of Experiment 2 was not to compare different linguistic groups but rather to see whether discrimination is possible when reducing the linguistic and human appearance of the stimuli as much as possible. As a consequence, we considered that trimming the linguistic heterogeneity of our participants (students at the University of Oxford) was unnecessary. Non-parametric analyses comparing performance by English native and non-native speakers showed no difference between these 2 groups in terms of discrimination between the artificial versions of English and Japanese rhythmic patterns ($U = 23$, $p = .177$; $U = 20$, $p = .429$ and $U = 19.5$; $p = .429$, for the auditory, visual, and audiovisual conditions, respectively). However, these results should not be taken as a test of the effects of linguistic background in general due to a lack of power in this respect.

5. For the same reasons as in Experiment 2 (see Footnote 4) our sample included some linguistic variability (10 native speakers of English and 5 native speakers of other languages including Indonesian, Hebrew, Russian, Chinese, and German who spoke English as an L2). Just for completeness, we compared the performance of the English native speakers and non-native speakers. As in Experiment 2, no differences were found, in terms of performance, between these 2 linguistic groups [$U = 40$, $p = .075$ and $U = 25$; $p = 1$; for the feedback and no-feedback conditions, respectively].

FIGURE CAPTIONS

Figure 1. A virtual emulation of the human language articulators, producing the phonemes /s/ and /a/, was used in Experiment 1. In order to isolate syllabic rhythm from other linguistic cues, the sentences corresponding to the *flat sasasa* condition in Ramus et al.'s (1999) study were used to create new video (and auditory) streams of *robot sasasa*. Due to video-frame length constraints (25Hz, 1 frame = 40 ms), the consonant and vowel intervals were slightly adjusted (see the Methods section for details).

Figure 2. Mean percentage of correct responses in Experiments 1 (A), 2 (B), and 3 (C). Panels A and B show that discrimination between the transformed English and Japanese sentences was possible (i.e., performance was significantly better than chance; see asterisks) in all conditions (auditory, visual, and audiovisual) in Experiments 1 (language discrimination) and 2 (discrimination of 'non-linguistic' patterns). These results demonstrate that the discrimination between sentences spoken in different languages is possible using not only the auditory rhythm, but also the visual information from the rhythmical movements of the mouth. A cross-experiment analysis revealed that participants' performance was overall better in the auditory and the audiovisual conditions than in the visual condition. The results of Experiment 3 (C) revealed that the discrimination between the English and Japanese rhythms is also possible when using only vibrotactile information delivered to the fingertip. However, in this case, performance was significantly above chance level only after the participants had acquired some experience in the discrimination task (where feedback about their performance was provided). The error bars indicate the standard error of the mean.

Figure 1.

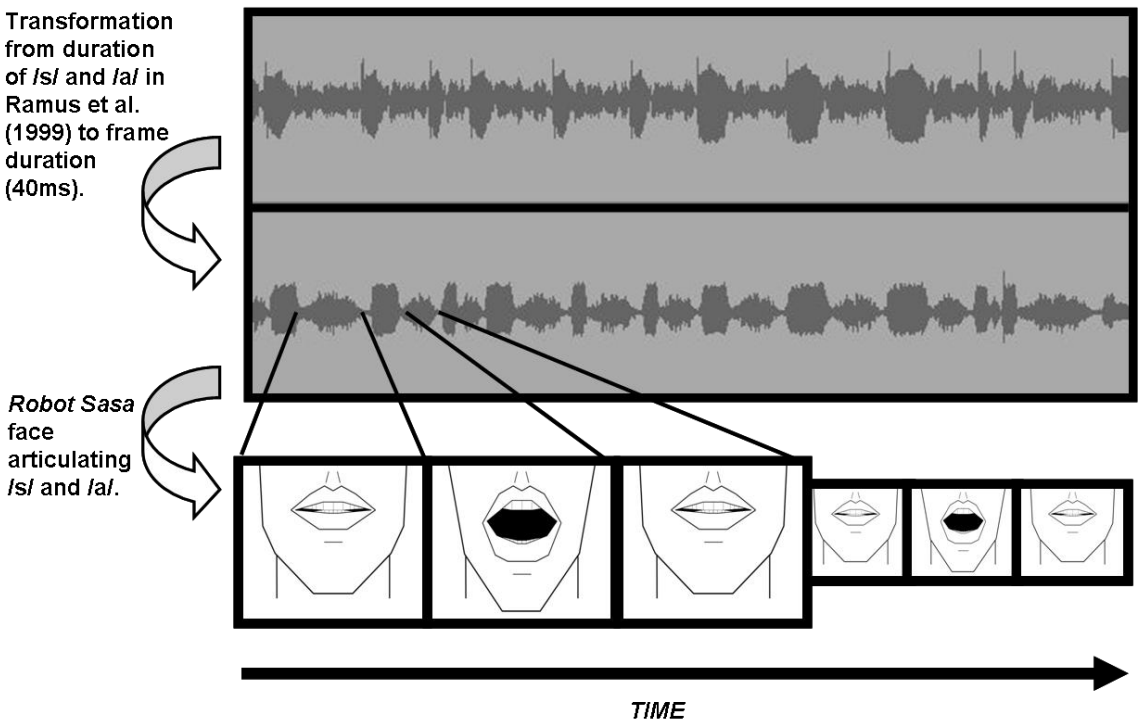

Transformation from duration of /s/ and /a/ in Ramus et al. (1999) to frame duration (40ms).
Robot Sasa face articulating /s/ and /a/.
TIME

Figure 2.

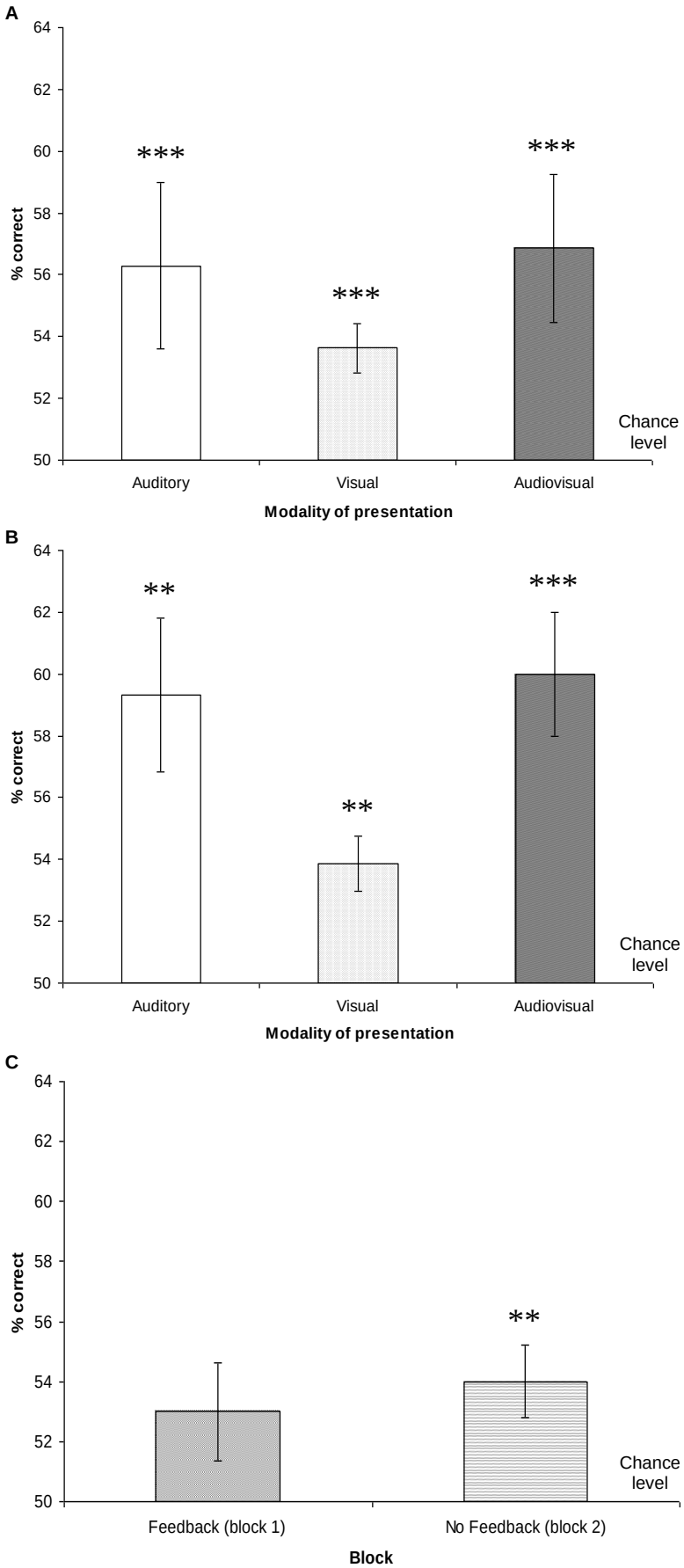
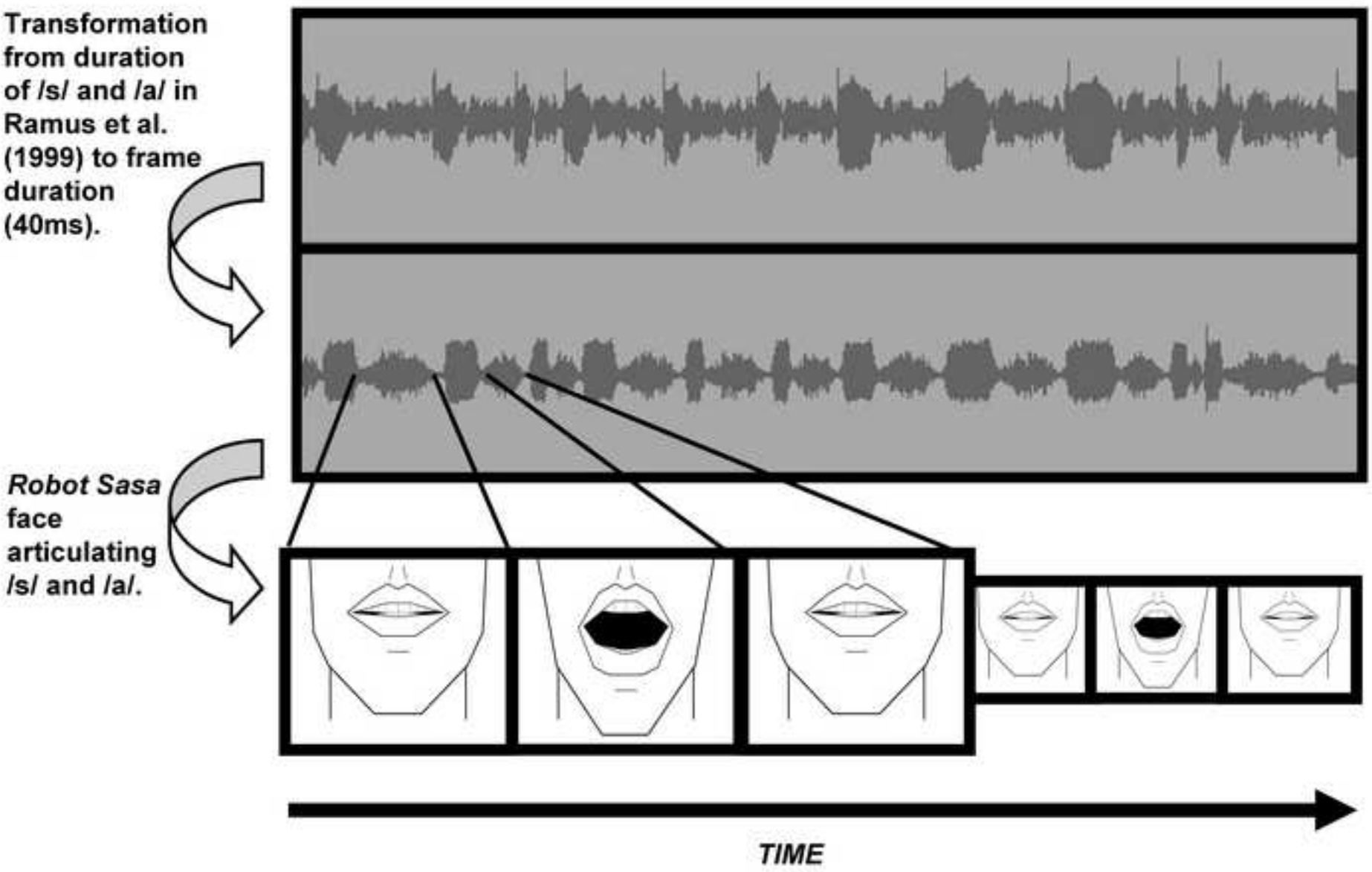

A
64
62
60
58
56
54
52
50
% correct
***
***
***
Chance level
Auditory
Visual
Audiovisual
Modality of presentation
B
64
62
60
58
56
54
52
50
% correct
**
**
***
Chance level
Auditory
Visual
Audiovisual
Modality of presentation
C
64
62
60
58
56
54
52
50
% correct
**
Chance level
Feedback (block 1)
No Feedback (block 2)
Block

Figure
Click here to download high resolution image



Figure

Click here to download high resolution image