Acta Psychologica 151 (2014) 197–205

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

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, Parc Sanitari Sant Joan de Déu, CIBERSAM, Spain
[b] Institució Catalana de Recerca i Estudis Avançats (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

ARTICLE INFO

Article history:
Received 14 January 2013
Received in revised form 17 March 2014
Accepted 7 May 2014
Available online 19 July 2014

PsycINFO classification:
2300 Human Experimental Psychology
2320 Sensory Perception
2323 Visual Perception
2326 Auditory & Speech Perception

Keywords:
Speech rhythm
Speechreading
Audition
Vision
Touch
Discrimination

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, both audition and vision at the same time, 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. Despite the fact that discrimination could be achieved using vision alone, auditory performance was nevertheless better. 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.

1. Introduction

Lloyd James (1940) 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, & Dehaene-Lambertz, 1996). This is an important ability since, speaking globally, bilingual communities are more numerous than monolingual ones (see Brutt-Griffler & Varghese, 2004; de Bot & Kroll, 2002). 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, Hauser, Miller, Morris, & Mehler, 2000) and even rats (Toro, Trobalon, & 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

* Corresponding author at: Hospital de Sant Joan de Déu, Edifici Docent, C/ Santa Rosa, 39-57, planta 4ª, 08950 Esplugues, Barcelona, Spain. Tel.: +34 936009751; fax: +34 936 00 97 71.
E-mail address: jnavarrao@fsjd.org (J. Navarra).

http://dx.doi.org/10.1016/j.actpsy.2014.05.021
0001-6918/

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 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, Saint-Amour, Leavitt, Javitt, & Foxe, 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, Munhall, Kasahara, Garcia, & Yeshia, 1996; Yehia, Kuratate, & Vatikiotis-Bateson, 2002), and even more complex information (e.g., lexical stress, syntactic boundaries, and pragmatics; see Hadar, Steiner, Grant, & Rose, 1983, 1984; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Risberg & Lubker, 1978). Soto-Faraco et al. (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 4 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, Pallier, Sebastián-Gallés, & Mehler, 1997) or pitch in a tonal language (Wang, Spence, Jongman, & Sereno, 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 other unfamiliar and non-native languages in both infancy (Weikum et al., 2007, 2013) 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, Nespor, and Mehler'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 & Soto-Faraco, 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 syllabic rhythm. 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-resynthesized 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 average duration, respectively) and were presented at 68 dB(A), as measured from the participant's head position,

[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).