Six-month-old infants discriminate voicing on the basis of temporal envelope cues (L)

Josiane Bertoncini, Thierry Nazzi, Laurianne Cabrera, Christian Lorenzi

Six-month-old infants discriminate voicing on the basis of temporal envelope cues (L)

Josiane Bertoncini[a] and Thierry Nazzi
Laboratoire de Psychologie de la Perception, Université Paris Descartes, CNRS 45 rue des Saints Pères, 75006 Paris, France

Laurianne Cabrera
Laboratoire de Psychologie de la Perception, Université Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France

Christian Lorenzi
Laboratoire de Psychologie de la Perception, Université Paris Descartes, Ecole Normale Supérieure 29 rue d'Ulm, 75005 Paris, France

(Received 21 September 2010; revised 30 November 2010; accepted 8 March 2011)

Young deaf children using a cochlear implant develop speech abilities on the basis of speech temporal-envelope signals distributed over a limited number of frequency bands. A Headturn Preference Procedure was used to measure looking times in 6-month-old, normal-hearing infants during presentation of repeating or alternating sequences composed of different tokens of /aba/and /apa/ processed to retain envelope information below 64 Hz while degrading temporal fine structure cues. Infants attended longer to the alternating sequences, indicating that they perceive the voicing contrast on the basis of envelope cues alone in the absence of fine spectral and temporal structure information. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3571424]

PACS number(s): 43.71.Es, 43.71.Ft, 43.71.Ky [MSS] Pages: 2761–2764

I. INTRODUCTION

Within the cochlea, speech sounds are decomposed by the "auditory filters" into a series of narrowband signals, each evoked at a different place on the basilar membrane. Each signal can be considered as a "carrier"—the temporal fine structure (TFS), which is determined by the dominant frequencies in the signal that fall close to the center frequency of the band—and the temporal envelope (E), which corresponds to the relatively slow fluctuations in amplitude superimposed on the carrier (Smith *et al.*, 2002). Both the E and TFS information are represented in the pattern of phase-locking in auditory-nerve fibers. However, for most mammals, the accuracy of phase-locking to TFS is only constant up to about 1–2 kHz (Johnson, 1980), whereas phase-locking to E remains accurate for the carrier frequencies beyond 6 kHz (Joris and Yin, 1992).

When speech sounds are presented in quiet, E cues alone are sufficient for adults to readily recognize the speech input at different levels (phonemes, words, and sentences), even with a limited number of frequency bands (4–16), and after a brief exposure to these stimuli (Shannon *et al.*, 1995). Indeed, voicing, nasality, place, and manner of articulation are signaled by various E cues relatively widespread or located in the high-frequency range (Rosen, 1992). In comparison, TFS cues seem to play a little role in speech identification, when speech is intact and presented in quiet (Shannon *et al.*, 1995). However, the recent studies suggest that TFS cues may play a specific role in conveying phonetic information regarding voicing and nasality, when E cues are severely degraded by acoustic distortions (Sheft *et al.*, 2008). Consistent with this notion, the main segmental cues to voicing and nasality are restricted to the low-mid frequency range and are well represented in the pattern of phase-locking in auditory-nerve fibers.

During the last decade, profoundly deaf children have been fitted with the cochlear implants (CI) at a younger and younger age with reasonable success (Holt and Svirsky, 2008; Miyamoto *et al.*, 2003). The CI speech processors deliver reliably E cues over a small number of independent frequency channels (i.e., about eight channels; Friesen *et al.*, 2001), but degrade severely TFS cues. This nevertheless allows most CI users to perform well in quiet situations. However, the oral language level attained by the infant CI receivers is not consistently within the "normal" range. Most studies show that a better outcome is principally correlated with an early implantation, the initial period of the first 2 years (or 18 months) being generally recommended when possible. In addition, several other factors have been shown to play a significant role such as parental support, education level, or socioeconomic status. But what is largely ignored is how congenitally deaf infants could learn the properties of their native language by processing the E information transmitted by the CI device in the absence of TFS cues.

One way to evaluate this issue is to turn to early language acquisition, and examine the potential role of E cues in early phonological acquisition. The early typical development of the segmental side of speech processing has

[a]Author to whom correspondence should be addressed. Electronic mail: josiane.bertoncini@parisdescartes.fr

received much interest between 1980 and 2000. Many experimental studies focused on phoneme discrimination and categorization by the infants (Kuhl, 1991; Werker and Tees, 1984). Later, the role of supra-segmental prosodic cues (rhythm and pitch) was assessed in different periods of language acquisition (Bertoncini *et al.*, 1995; Nazzi *et al.*, 1998a, 1998b; Jusczyk *et al.*, 1993). Taken together, these studies suggest that the early acquisition of segmental and supra-segmental properties could be driven by E cues or at least, that having access to E cues only, as for the CI infants, would provide enough information for such acquisition. However, since pitch perception is considered to be related to TFS processing, the role of TFS cues during the acquisition of tonal languages by normal-hearing (NH) and CI children remains an open question (Xu and Pfingst, 2003).

One first step to start exploring this aspect of speech perception is to present speech signals processed by noise- or tone-excited vocoders (as in Shannon *et al.*, 1995) to near-term fetuses, and the NH children and infants, to verify whether they are sensitive to those E cues related to the phonetic information, i.e., when TFS cues are removed, and the stimuli are presented in quiet. A first study suggests that as early as 38 weeks of gestational age, fetuses perceive the E cues of sentences processed by a single band noise-excited vocoder (Granier-Deferre *et al.*, 2011). A second study conducted with children (5–12 yr) demonstrated that the recognition of speech processed by multi-band noise-excited vocoders to retain mainly E cues develops before the age of 7, and becomes adult-like around the age of 10 (Eisenberg *et al.*, 2000). In a third study, 5- to 7-yr-old children were shown to be able to discriminate non-sense syllables varying in voicing, place, manner, and nasality, when those syllables were processed by a 16-band tone-excited vocoder (Bertoncini *et al.*, 2009). In addition, the response correctness and latencies demonstrated by the children were found to be similar to that of young adult controls. The discrimination scores with the vocoded signals were very high ($d' \geq 2$) and minimally reduced if compared to those obtained with the intact speech. Like adults, 5- to 7-yr-old children did not receive any training before testing. These results suggest that even at an early age, the NH children are able to rapidly adapt and assimilate vocoded signals to (degraded) speech sounds.

The aim of the present study is to start a new line of research on the development of speech perception by presenting different vocoded speech signals to NH infants, at an early stage of development. As a first step in this direction, we present here one experiment indicating that 6-month-old infants successfully discriminate a voicing contrast on the sole basis of E cues.

II. METHOD

A. Participants

Twenty 6-month-old infants (8 girls, 12 boys) with NH (based on newborn hearing screening and parental report) were tested and their data included in the analyses (mean age = 6.2 months; range = 5.8–6.6 months; and standard deviation = 0.25 months). The data of six additional infants were not taken into account, due to fussiness or crying (four) and to extreme mean looking times leading to outlier differences between the two kinds of stimulus series (two). Families were informed about the goals of the current study, and provided written consent before the participation of their children.

B. Stimuli

Eight exemplars of each category /aba/ and /apa/ were selected from a set of vowel–consonant–vowel (VCV) nonsense syllables uttered by a French female speaker. The stimuli were recorded in a quiet room, and digitized via a 16-bit analog-to-digital converter at a 44.1-kHz sampling rate.

The 16 selected stimuli were then processed to degrade TFS cues, while preserving E cues at the output of a bank of analysis filters using the same speech-processing technique used by Bertoncini *et al.* (2009). Speech was filtered into 16 adjacent 0.35-oct wide frequency bands spanning the range of 0.08–8.02 kHz. The temporal envelope was extracted in each frequency band, using the Hilbert transform followed by lowpass filtering with a zero-phase, sixth-order Butterworth filter (cutoff frequency = 64 Hz). The filtered envelope was used to amplitude modulate a sine wave with a frequency equal to the center frequency of the band, and with a random starting phase. The 16 amplitude-modulated sine waves were summed over all frequency bands. The 16 processed stimuli were finally equated in root mean square (rms) power. The effects of signal processing are illustrated in Fig. 1, showing spectrograms of the intact (left panels) and vocoded (right panels) versions of a given /aba/ (top panels) and /apa/ (bottom panels) stimulus.

The stimuli were presented into series of 24 items constituted of three different orders of the eight vocoded signals, in succession. There were four resulting series, two repeating (REP) (/apa apa apa …/ and /aba aba aba …/) series and two alternating (ALT) series that differed only by the first presented item (/apa aba apa aba …/ and /aba apa aba apa …/). The stimuli were stored in digitized form on the computer, and were delivered at a sound pressure level of 70 dB by the loudspeakers via an audio amplifier.

C. Procedure

The head-turn preference procedure (HPP) used here was adapted for studying discrimination as in Nazzi *et al.* (2009). Each infant was held on a caregiver's lap, in the center of the test booth, facing a green light. Before each trial, the infant's attention was re-centered by blinking the central green light. As soon as the infant was correctly face oriented, the green light was extinguished and the red light started to flash either on the right or the left side of the booth. When the infant made a correct turn in that direction, a stimulus series began to be delivered by the loudspeaker located behind the flashing red light. Each series was played to completion, or stopped as soon

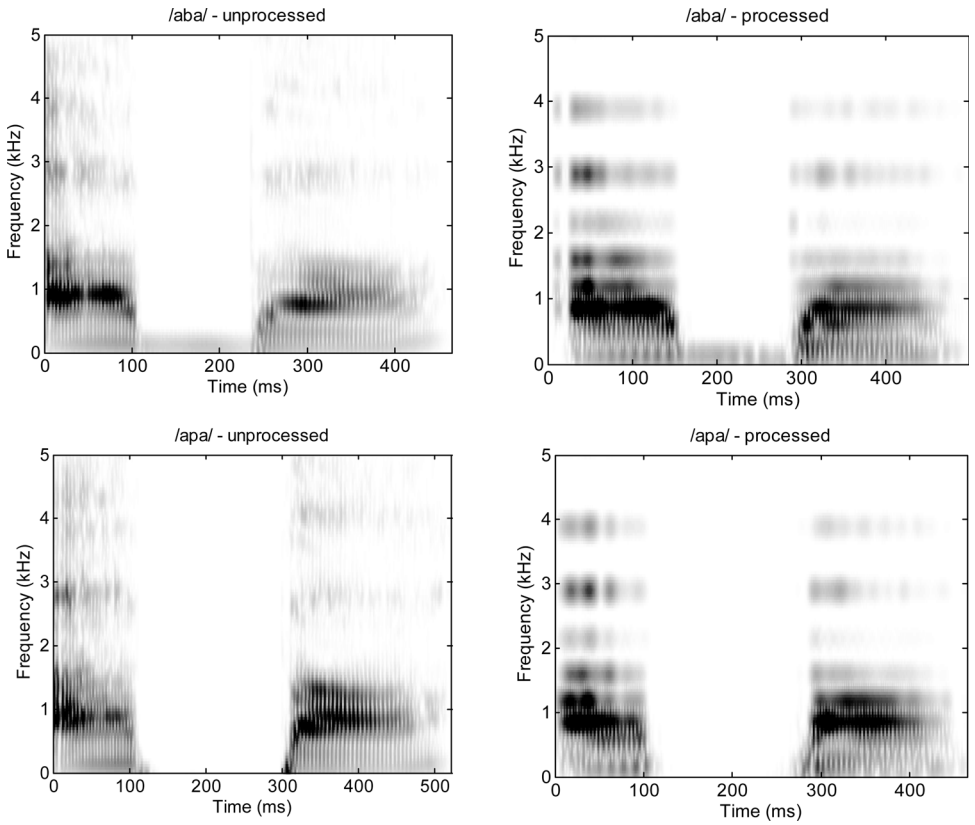

FIG. 1. Spectrograms of the intact (left panels) and vocoded (right panels) versions of a /aba/ (top panels) and /apa/ (bottom panels) stimulus.

as the infant looked away for two consecutive seconds. The total duration of looking times was recorded into a personal computer, via a response box controlled on line by the experimenter who was out of the booth and unaware of the stimulus series presented (both the caregiver and the experimenter listened to masking music).

For each infant, the experimental session began with two musical trials, one on each side (randomly ordered) to give the infants an opportunity to practice one head-turn to each side before the test itself. The test phase consisted of two test blocks (in each of which the two REP and the two ALT series were presented). The mean duration of looking times was calculated for each type of series, over the first and second blocks of test trials.

In this HPP paradigm, discrimination is attested when infants demonstrate a differential response to the two kinds of stimuli they are presented with. Here, infants might show longer looking times during listening either to ALT series or to REP ones if they are sensitive to the difference between ALT and REP series. Furthermore, a "preference" indexed by longer looking times might be observed for the ALT series, because the alternation (if perceived) might facilitate perceptual comparison, and maintain attention for a longer while (Best and Jones, 1998).

III. RESULTS

The mean results are shown in Fig. 2. An analysis of variance (ANOVA) for repeated measures was performed with blocks and series as within-subject factors. The group of 20 infants presented longer mean looking times during ALT series than during REP series on the two blocks of trials [$F(1,19) = 4.75$, $p = 0.042$]. Fourteen infants among 20 demonstrated a bias toward listening longer to ALT series compared to REP ones. The duration of looking times decreased from the first to the second block of test trials [$F(1,19) = 11.63$, $p = 0.003$]. Although the interaction between blocks and series was not significant [$F(1,19) = 2.29$, $p = 0.15$], the general trend in looking times to decrease as experiment is progressing affected the mean size of the effect [2.88 s in the first block ($t(19) = 2.29$, unilateral $p = 0.017$), and 0.26 in the second

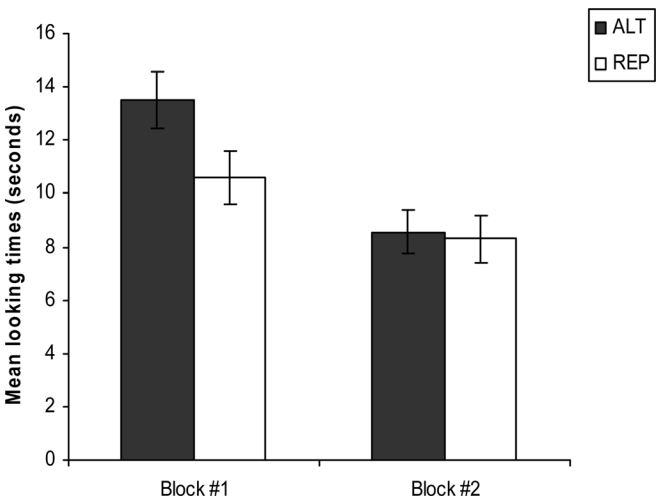

FIG. 2. Mean looking times (in seconds) to the ALT series versus the REP series. The error bars indicate the standard error of the mean.

block ($t(19) < 1$)]. Looking times were also found to decline in the subjects who, nonetheless, showed a "preference" for ALT series (5.97 and 2.99 s in the first and second block, respectively).

IV. CONCLUSIONS

NH 6-month-old infants presented with degraded syllables /aba/ and /apa/ are sensitive to the difference between the E cues associated with the voiced and voiceless French phonemes /b/ vs /p/. Without any pre-exposure, in a paradigm favoring immediate comparison between two contrasting categories (remember that each category was represented by eight different exemplars), infants were able to pick up the difference between /apa/ and /aba/ on the basis of E cues only. These data reveal that for 6-month-old infants, as for adults, fine spectral and temporal structure information, although potentially important for musical pitch and localization of sounds in space (Smith *et al.*, 2002), is not essential for speech discrimination, at least in quiet. They also demonstrate that the auditory temporal resolution (i.e., the ability to detect E fluctuations) is sufficient to support the phonetic discrimination at 6 months of age. Finally, the fact that the preference toward ALT series over REP ones was relatively immediate, but did not last over few minutes suggests the action of low-level auditory mechanisms.

This first study is also promising because, by using vocoded speech sounds, it opens a new way of studying how auditory (peripheral) mechanisms might be engaged in speech processing during language acquisition. Obviously, further studies must be conducted to determine whether E information could be used by the infants not only to discriminate but also to categorize speech syllables, or to extract word form patterns from the speech stream. As an example, if the infants' sensitivity to E cues was shown to be modulated by diverse linguistic input, it would support the idea that E perception is a built-in part of speech processing.

Regarding possible follow ups in relation with the spoken language acquisition by young CI users, it will be useful to extend the present study by testing NH infants on different phonetic contrasts (manner, place), while using a lower frequency resolution (e.g., four or eight analysis filters), and comparing their performance with the CI infants' perception of the same contrasts. This will allow us to evaluate whether the NH infants' discrimination of phonetic contrasts on the basis of E cues is a reliable simulation of how young CI users discriminate the same contrasts during development.

ACKNOWLEDGMENTS

The authors wish to thank X. Li for preparing stimuli and testing infants.

Bertoncini, J., Floccia, C., Nazzi, T., and Mehler, J. (**1995**). "Morae and syllables: Rhythmical basis of speech representations in neonates," Lang. Speech **38**, 311–329.

Bertoncini, J., Serniclaes, W., and Lorenzi, C. (**2009**). "Discrimination of speech sounds based upon temporal envelope versus fine structure cues in 5- to 7-year-old children," J. Speech Lang. Hear. Res. **52**, 682–695.

Best, C. T., and Jones, C. (**1998**). "Stimulus-alternation preference procedure to test infant speech discrimination," Infant Behav. Dev. **21**(Suppl.1), 295.

Eisenberg, L. S., Shannon, R. V., Schaefer Martinez, A., Wygonski, J., and Boothroyd, A. (**2000**). "Speech recognition with reduced spectral cues as a function of age," J. Acoust. Soc. Am. **107**, 2704–2710.

Friesen, L. M., Shannon, R. V., Başkent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Granier-Deferre, C., Ribeiro, A., Jacquet, A.-Y., and Bassereau, S. (**2011**). "Near-term fetuses process temporal features of speech," Dev. Science **14**, 336-352.

Holt, R. F., and Svirsky, M. A. (**2008**). "An exploratory look at paediatric cochlear implantation: Is earliest always best?," Ear Hear. **29**, 492–511.

Johnson, D. H. (**1980**). "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," J. Acoust. Soc. Am. **68**, 1115–1122.

Joris, P. X., and Yin, T. C. (**1992**). "Responses to amplitude-modulated tones in the auditory nerve of the cat," J. Acoust. Soc. Am. **91**, 215–232.

Jusczyk, P. W., Cutler, A., and Redanz, N. J. (**1993**). "Infants' preference for the predominant stress patterns of English words," Child Dev. **64**, 675–687.

Kuhl, P. K. (**1991**). "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not," Percept. Psychophys. **50**, 93–107.

Miyamoto, R. T., Houston, D. M., Kirk, K. I., Perdew, A. E., and Svirsky, M. A. (**2003**) "Language development in deaf infants following cochlear implantation," Acta Oto-Laryngol. **123**, 241–244.

Nazzi, T., Bertoncini, J., and Bijeljac-Babic, R. (**2009**). "A perceptual equivalent of the labial-coronal effect in the first year of life," J. Acoust. Soc. Am. **126**, 1440–1446.

Nazzi, T., Bertoncini, J., and Mehler, J. (**1998a**). "Language discrimination by newborns: Towards an understanding of the role of rhythm," J. Exp. Psychol. Hum. Percept. Perform. **24**, 356–366.

Nazzi, T., Floccia, C., and Bertoncini, J. (**1998b**). "Discrimination of pitch contour by neonates," Infant Behav. Dev. **21**, 779–784.

Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London Ser. B **336**, 367–373.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Sheft, S., Ardoint, M., and Lorenzi, C. (**2008**). "Speech identification based on temporal fine structure cues," J. Acoust. Soc. Am. **124**, 562–575.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90.

Werker, J. F., and Tees, R. C. (**1984**). "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," Infant Behav. Dev. **7**, 49–63.

Xu, L., and Pfingst, B. E. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception," J. Acoust. Soc. Am. **114**, 3024–3027.