Infants discriminate voicing and place of articulation with reduced spectral and temporal modulation cues

Laurianne Cabrera, Christian Lorenzi, Bertoncini Josiane

Infants discriminate voicing and place of articulation with reduced spectral and temporal modulation cues

Cabrera Laurianne*
Laboratoire de Psychologie de la Perception
CNRS, Université Paris Descartes
45 rue des saints Pères, 75006 Paris, France
Lorenzi Christian
Laboratoire des systèmes perceptifs, CNRS,
Institut d'Etude de la Cognition, Ecole normale supérieure,
Paris Sciences et Lettres Research University
29 rue d'Ulm, 75005 Paris, France
Bertoncini Josiane
Laboratoire de Psychologie de la Perception
CNRS, Université Paris Descartes
45 rue des saints Pères, 75006 Paris, France

* Corresponding author: Laboratoire de Psychologie de la Perception, CNRS-UMR 8158, Université Paris Descartes, 45 rue des saints pères, 75006, Paris, France.
Tel: +33 1 42 86 43 20
E-mail address: laurianne.cabrera@gmail.com

Submitted on May 2014, JSLHR

Running title: speech modulation cues and phonetic discrimination

Abstract

Purpose: This study assessed the role of spectro-temporal modulation cues in the discrimination of two phonetic contrasts (voicing and place) for young infants.

Method: A visual-habituation procedure was used to assess the ability of French-learning 6-month-old infants with normal hearing to discriminate voiced *versus* unvoiced (/aba/-/apa/) and labial *versus* dental (/aba/-/ada/) stop consonants. The stimuli were processed by tone-excited vocoders to degrade frequency-modulation (FM) cues while preserving: 1) amplitude-modulation (AM) cues within 32 analysis frequency bands, 2) slow AM cues only (< 16 Hz) within 32 bands, and 3) AM cues within 8 bands.

Results: Infants exhibited discrimination responses for both phonetic contrasts in each processing condition. However, when fast AM cues were degraded, infants required a longer exposure to vocoded stimuli to reach the habituation criterion.

Conclusions: Altogether, these results indicate that the processing of modulation cues conveying phonetic information on voicing and place is "functional" at 6 months. The data also suggest that the perceptual weight of fast AM speech cues may change during development.

Key words: Speech perception, Amplitude modulation, Frequency modulation, Infants

I. INTRODUCTION

A large number of studies have investigated separately auditory and speech perception in infants (for a review see Kuhl, 2004, and Saffran, Werker, & Werner, 2006), but knowledge about the auditory capacities involved in the typical early development of speech processing is still lacking. A better characterization of these early auditory capacities is important because accurate sensory coding of acoustic cues is required for robust speech perception in real-life listening conditions (*e.g.*, Moore, 2007). The development of the basic auditory capacities involved in the extraction of these acoustic cues has been assessed extensively with non-linguistic sounds such as pure tones, complex tones or noises. These capacities appear to be "adultlike" by 6 months of age (*e.g.,* Levi & Werner, 1996; Spetner & Olsho, 1990; Werner, Folsom, Mancl, & Syapin, 2001), although some continue to develop until late into childhood (see*,* Burnham & Mattock, 2010; Saffran *et al.*, 2006). Recently, the low-level spectro-temporal auditory abilities involved in the perception of speech signals have been investigated for adults using speech-processing algorithms called vocoders (see Shamma & Lorenzi, 2013 for a review). The present study extended this investigation to the discrimination of phonetic contrasts for infants.

Speech signals are commonly assumed to be decomposed by the cochlea into a series of narrowband signals (usually described as 32 independent frequency bands) each with a passband equal to one "equivalent-rectangular bandwidth" ($\mathrm{ERB_N}$ which is an approximation of the auditory filter bandwidth given by the equation: $\mathrm{ERB} = 24.7(4.37F + l)$ with F in kHz corresponding to the center frequency of the filter) (Glasberg & Moore, 1990; Moore, 2003). Each 1-$\mathrm{ERB_N}$ wide band may be viewed as a sinusoidal carrier with superimposed amplitude modulation (AM, or relatively slow modulations in amplitude over time) and frequency modulation (FM or relatively fast fluctuations in instantaneous frequency over time; *e.g.*, Drullman, 1995; Shamma & Lorenzi, 2013; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Sheft, Ardoint, & Lorenzi, 2008; Smith, Delgutte, & Oxenham, 2002; Zeng *et al.*, 2005). In a seminal review, Rosen (1992) systematically described how AM (temporal envelope) and FM (temporal fine structure) cues signal phonetic contrasts. The slowest AM cues (i.e., <16 Hz) are mostly related to speech rhythm and syllabicity and the fastest AM cues and FM cues are mostly related to formant transitions (i.e., ~30 Hz) and voice-pitch information (*i.e.*, periodicity cues, between 50-500Hz).

Vocoders are signal-processing algorithms that extract (i.e., compute) the AM and FM cues of speech signals from different analysis frequency bands (Dudley, 1939). Using vocoders, one can therefore manipulate the relative strength of the AM and FM components and the fine spectral cues[1] conveying the phonetic information of the vocoded speech. Over the last decades, vocoder studies have repeatedly demonstrated the importance of AM and FM cues in speech perception for adults (*e.g.*, Sheft *et al.*, 2008; Smith *et al.*, 2002; Zeng *et al.*, 2005). Normal-hearing adults showed accurate speech recognition in quiet when listeners are presented with syllables or sentences vocoded to preserve only the slowest AM cues (< 16 Hz) extracted (*i.e.*, computed) from four broad frequency bands (*e.g.,* Shannon *et al.,* 1995). Additional work suggested that faster and fine spectro-temporal modulations (that is, FM cues and fine spectral cues) are required for robust speech recognition in adverse listening conditions such as when speech is interrupted or masked by noise (*e.g.,* Eaves, Summerfield, & Kitterick, 2011; Gnansia, Péan, Meyer, & Lorenzi, 2009; Nelson & Jin, 2004; Qin & Oxenham, 2003). Recent studies suggest that the high recognition performance demonstrated by adults when listening to vocoded speech retaining AM cues only may be due to their phonological and lexical skills (*e.g.*, Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Sohoglu, Peelle, Carlyon, & Davis, 2012).

The present study explored the extent to which 6-month-old infants, whose perception of segmental cues is not yet tuned to their native language (*e.g.*, Kuhl, 2004) are able to use the AM, FM and fine spectral cues of speech when discriminating phonetic contrasts.

To the best of our knowledge, only a few studies have assessed the ability of infants and children to use these spectro-temporal modulation cues in discrimination and identification tasks using noise or tone-excited vocoded stimuli. As for children, Newman and Chatterjee (2013) showed that English-learning 2-year-old toddlers accurately identify vocoded words when AM cues are extracted and preserved from 8 frequency bands. For older children, Eisenberg, Shannon, Martinez, Wygonski and Boothroyd (2000) found that English-learning 5- to 7-year-old children require a greater number of frequency bands than adults to identify vocoded words and sentences. More recently, Bertoncini, Serniclaes and Lorenzi (2009) showed that French-learning 5-year-old children discriminate nonsense bisyllables as well as older children and adults when only the relatively slow (< 64 Hz) AM cues are preserved within 16 frequency bands.

Identification and discrimination task do not involve the same auditory and linguistic processes. For instance, sentence identification tasks require speech segmentation and lexical access. These studies suggest that, for tasks requiring elaborate or higher-level (cognitive/linguistic) processing such as tasks requiring to identify words and sentences, younger children may require greater redundancy in the speech signal (that is, a greater number of frequency bands, and therefore fine spectral cues) than older ones or adults. When the task is less demanding, as in the case of discrimination of isolated phonemes or syllables, younger children may be able to make accurate responses based on limited sensory cues, and this is why performance is similar across age groups.

Less information is available regarding infants. Bertoncini, Nazzi, Cabrera and Lorenzi, (2011) studied the ability of French-learning 6-month-olds to discriminate a /apa/-/aba/

voicing contrast when the relatively slow (< 64 Hz) AM cues were preserved within 16 frequency bands using a tone-excited vocoder. Such a tone-excited vocoder degrades the fine spectro-temporal modulations by reducing the spectral resolution (i.e., only 16 frequency bands are used) and by replacing the FM cues by pure tones in each frequency band. As in Bertoncini *et al.* (2009), the speech AM cues were low-pass filtered at 64 Hz, attenuating the fast periodic AM cues related to formant transitions, bursts and periodic fluctuations produced by the vocal folds at the fundamental frequency (F0) rate (see Rosen, 1992). A head-turn preference procedure was used to assess preference for sequences composed of alternated *versus* repeated /apa/ and /aba/ stimuli. The head-turn preference procedure consists in the presentation of sound sequences while infants look at a blinking light on their right or left side. The results showed that infants listened longer to the alternating vocoded stimuli, providing evidence that fast AM cues (>64 Hz) and fine spectro-temporal modulations (FM cues and fine spectral cues) are not required to discriminate voicing by the age of 6 months. Recently, Cabrera, Bertoncini and Lorenzi (2013) showed that infants are able to discriminate the voicing contrast even when speech signals contain only the slowest (< 16 Hz) AM cues in 32 frequency bands and when AM cues are preserved within 4 broad frequency bands. Thus, fast AM cues, FM cues and fine spectral cues are not required for voicing discrimination at this early age. Interestingly, the results suggested that the vocoded-speech conditions required extra processing effort compared to the condition where speech was unprocessed. Indeed, the head-turn procedure used in this study included a familiarization phase of one or two minutes to a given phonetic category processed using a specific vocoder. Then, the infants' looking time for the blinking lights was recorded during 8 trials: 4 trials presenting a "novel" phonetic category, and 4 trials presenting the familiar one processed with the same vocoder as used in the familiarization phase. Results showed that familiarization time had to be increased by a factor 2 for infants to discriminate the vocoded-speech contrast (1 *versus* 2 min). In addition, in the different vocoder conditions, discrimination (at the group level) was revealed either by a classical preference for novelty (when AM and FM cues were intact), or by a preference for the familiar stimuli (when FM and AM cues were both reduced).

The present study aimed to investigate further the importance of spectro-temporal modulation cues in phonetic discrimination at 6 months and the impact of exposure time to the impoverished speech sounds. Previous work on infants was restricted to the perception of voicing, which is known to be robust for normal-hearing adults in the sense that it is resistant to the effect of filtering and masking noise (*e.g.,* Miller & Nicely, 1955). The present study attempted to extend this investigation to another phonetic feature, that is, *place of articulation*, a feature known to be more susceptible to signal distortions such as those produced by filtering, noise or vocoding (*e.g.*, Gilbert & Lorenzi, 2006; Gnansia, Jourdes, & Lorenzi, 2008; Gnansia *et al.*, 2009; Miller & Nicely, 1955; Shannon *et al.*, 1995). In his seminal review, Rosen (1992) suggested that the perception of voicing is conveyed by both temporal-envelope, periodicity and temporal fine structure cues (that is, slow AM, fast AM and FM features, respectively), whereas the perception of place is more dependent on fine spectro-temporal cues (that is, fine spectral and FM features). The present study therefore tested whether infants show this pattern of dependency on AM, FM and fine spectral cues when discriminating voicing and place.

To address this issue, four tone-excited vocoders[2] were designed to evaluate the respective role of FM cues, fast AM cues and spectral resolution (that is, fine spectral cues) in voicing and place discrimination. Thus, a total of eight independent groups of 6-month-old infants were tested in the four vocoder conditions and with the two phonetic contrasts. It was expected that discrimination of place would be more affected than discrimination of voicing in French-learning infants when target syllables were vocoded to selectively degrade fine spectral and FM cues. Instead of using the head-turn preference procedure and a preset amount of familiarization to the vocoded-speech sounds as in Cabrera *et al.* (2013), the current study introduced an infant-controlled habituation phase using the visual habituation procedure (see Colombo & Mitchell, 2009; Werker *et al.*, 1998). This procedure has mostly been used to assess speech contrast discrimination in infants. This procedure consisted in the presentation of a visual display on a screen together with sequences of repeated sounds. In each vocoder condition, several sequences of the same sound category were played during the habituation phase. When the infant's looking time for the visual display decreased and reached a criterion preset by the experimenter, the test phase started. The habituation criterion corresponded to the most commonly used one in the literature: it was equal to a looking-time decrement of 50 % averaging on 3 consecutive trials and compared to the 3 highest trials. In the test trials, sequences of familiar *versus* novel sounds were presented to measure the infants' discrimination abilities (*i.e.*, longer looking times for a given sound category). In the present study, the absence of longer looking times for novel sound sequences in the test phase was used to reveal that 6-month-old infants have difficulty in discriminating the vocoded speech stimuli. Moreover, the amount of habituation time required by infants to switch to the test phase was also used to evaluate processing difficulty. Hunter and Ames (1988) suggested that habituation times required to lead to a novelty preference in infants reflect the interaction between several factors such as stimuli complexity and processing difficulty. Vocoded-speech signals are reduced (that is, less complex) versions of the original speech signals. For this reason, it could be expected that infants would require a shorter habituation time for the most impoverished speech sounds. This is the case: i) when both FM and fast AM cues are reduced (the "32-band AM<16Hz" condition described below), and ii) when both FM and fine spectral cues are reduced (the "8-band AM" condition described below). However, if infants take longer to reach the habituation criterion for one among the two conditions mentioned above, this would reveal that fast AM cues and fine spectral cues do not have equivalent effects on speech processing, and that one among these two spectro-temporal modulation components of the speech signal has greater importance for efficient processing of phonetic cues.

II. METHOD

A. Participants

Six-month-old infants were recruited from the university database of birth announcements in Paris, France. All families were informed about the goals of the current study and provided a written consent before their participation in accordance with the current French ethical requirements. Data from 160 infants from French-speaking families (20 infants x 2 phonetic contrasts x 4 vocoder conditions) were analyzed in this experiment (87 girls; age range: 5 months 27 days - 7 months 17 days; mean = 6 months and 12 days; standard deviation (SD) = 10 days). All infants had normal hearing (based on parental report of newborn hearing screening results). The data from 155 additional infants were not included for the following reasons: fussing and crying (n=116), looking time shorter than 1000 ms for one trial (in habituation and test phase, n=12), failure to reach the habituation criteria (n=27, see section D). In the conditions corresponding to the most degraded speech sounds (*i.e*, when degrading FM cues, fast AM cues and fine spectral cues, see section B), the attrition rate was higher (~50%) compared to conditions corresponding to the "intact" speech sounds (32-band AM+FM; ~40% attrition rate). Although high attrition rates (~40%) are usually observed with the habituation criterion used in the present study (see Narayan, Werker, & Beddor, 2010), the high attrition rate found here may be mostly related to fussiness and inherent to the vocoded-speech signals as such (*i.e.*, unfamiliar and distorted speech sounds).

B. Stimuli

Eight exemplars of each category /aba/, /apa/ and /ada/ were selected from a set of vowel–consonant–vowel (VCV) nonsense bisyllables produced by a French female speaker who was asked to speak clearly. The F0 was estimated at 242 Hz using the YIN algorithm (de

Cheveigné & Kawahara, 2002). The stimuli were recorded in a soundproof room, and digitized via a 16-bit analog-to-digital converter at a 44.1-kHz sampling rate. The stimuli were not significantly different in duration (mean=634 ms, SD=68.8 ms) for /aba/, mean=632 ms, SD=47.5 ms for /apa/, and mean=622 ms, SD=68.4 ms for /ada/). For each phonetic category, four different sequences were created. Each sequence was composed of four tokens of the same phonetic category, repeated four times in a different random order. Two sequences were used for the habituation phase and two sequences were used for the test phase. The tokens used in the test phase for each category were different from the ones used in the habituation phase. The inter-stimulus interval varied randomly in the 16-item sequences, between 600 and 1300 ms. This variation was introduced to make small variations in duration between items irrelevant within and between categories. All the sequences had the same duration (26 s).

The stimuli were processed by vocoders to alter their spectro-temporal modulations. Four different vocoder conditions were designed. In the first condition (called "32-band AM+FM speech" or "intact" condition), the original speech signal was passed through a bank of 32 2nd-order gammatone filters (Gnansia *et al.*, 2009; Patterson, 1987), each 1-ERB wide with center frequencies (CFs) uniformly spaced along an ERB scale ranging from 80 to 8,020 Hz. The Hilbert transform was then applied to each bandpass filtered speech signal to compute the AM component and FM carrier. The AM component was low-pass filtered using a zero-phase Butterworth filter (36 dB/octave rolloff) with a cutoff frequency set to $\mathrm{ERB_N}/2$. The final narrow-band speech signal was obtained by multiplying each sample of the FM carrier by the filtered AM function. The narrow-band speech signals were finally added up and the level of the wideband speech signal was adjusted to have the same root-mean-square value as the input signal. Thus, the vocoded speech signals retained the original AM and FM speech cues within each of the 32 analysis frequency bands.

In the second condition (called “32-band AM speech”), the same signal processing scheme was used, except that the FM carrier was replaced by a sine wave carrier with frequency at the CF of the gammatone filter, and with random starting phase in each analysis frequency band. Thus, the resulting vocoded speech signal retained AM speech cues within 32 bands, but discarded the original (within-channel) FM speech cues.

In the third condition (called “32-band AM<16Hz speech”), the same signal processing scheme was used as in the “32-band AM speech” condition, except that the AM component was low-pass filtered with a cutoff frequency of 16 Hz for each of the 32 bands in order to remove the fast AM cues. Thus, the resulting vocoded speech signal retained mainly the slowest (< 16 Hz) AM speech cues within 32 bands, and discarded the original FM speech cues.

In the last condition (called “8-band AM speech”), the same signal processing scheme was used as in the “32-band AM speech” condition, except that AM cues were extracted and preserved from only 8 broad (4-ERB$_N$ wide) frequency bands. Thus, the original FM speech cues were discarded, and AM cues were distorted substantially compared to the original AM speech cues. It is important to note that this kind of vocoder reducing both the original FM cues (i.e., by replacing them by a noise or tone carrier) and spectral auditory resolution simulates the sound processing achieved by current cochlear implant (CI) sound processors (*e.g.,* Friesen, Shannon, Baskent, & Wang, 2001; Shannon *et al.*, 1995; Fu & Nogaki, 2005).

Figure 1 shows the spectrograms of one exemplar of /aba/ stimuli in each experimental condition.

-Figure 1 about here-

C. Material and Apparatus

Infants were seated on their caregiver’s lap in a sound-attenuated room. The caregiver was instructed not to speak and not to point at the screen and wore headphones delivering masking music. Infants faced a 61-cm LCD television screen positioned approximately 1.5 m from the infant. The audio stimuli were presented through two Fostex loudspeakers located on each side of the screen playing the auditory stimuli at a level of approximately 70 dB SPL. A black and white checkerboard was presented on the screen during habituation and test trials. At the beginning of each trial, a silent flashing ball video was played to attract the infants'attention to the screen. The infant's looking time was monitored online *via* a video camera positioned 30 cm below the screen and linked to the observer's monitor in the adjacent room. The observer, blind to the audio file presented, recorded the duration of the infant's looking time by a key press and controlled stimuli presentation using Habit X.10 (Cohen, Atkinson, & Chaput, 2000).

D. Procedure

A visual habituation method was used (Mattock, Molnar, Polka, & Burnham, 2008; Werker *et al.*, 1998) in which sound sequences were presented contingently with the infants' look at the black and white checkerboard displayed on a screen.

Auditory and visual presentations continued until the infant looked away for 2 s (automatically calculated by the computer based on the experimenter's key press) or at the end of the sound sequence (maximum 26 s). At the end of the sequence, the checkerboard disappeared and a more attractive display appeared in order to draw the infant's attention to the TV monitor. Once the infant looked at the screen, the experimenter initiated the next sequence. In each vocoder condition, the experiment began with a habituation phase, during which infants heard several sequences of the same sound category. The habituation phase ended when the mean looking time on three consecutive sequences decreased by 50% compared to the longest three consecutive trials from a sliding window. Infants were excluded a posteriori in the analysis only if they habituated with less 50s-cumulated looking time and showed extreme cumulated looking times (more than the group mean cumulated looking time

+ 2 SD in each condition). The test phase followed immediately. During the test phase infants heard 4 novel (N) and 4 familiar (F) sequences in alternation with order counterbalanced across subjects. Infants who did not reach the end of the test phase because of fussiness were excluded from the analyses.

Four independent groups (n=20) were tested for the voicing contrast (one group per vocoder condition): half of the subjects were habituated with /aba/ stimuli, and the other half with /apa/. Four independent groups (n=20) were tested for the place of articulation contrast (one group per vocoder condition): half of the subjects were habituated with /aba/ stimuli, and the other half with /ada/.

The total looking time to reach habituation criterion and the mean looking times in the test phase for the 4 novel and the 4 familiar test trials were recorded and analyzed in each condition.

III. RESULTS

Discrimination data

Figure 2 shows the mean looking time in the 8 groups of infants (2 phonetic contrasts x 4 vocoder conditions) for both novel and familiar sequences. In all groups, infants showed longer looking times for the novel sequences during the test phase. The discrimination was assessed by comparing the looking times for novel and familiar sequences in the test phase and the effect of Vocoder condition and Phonetic contrast. A 4 (Vocoder conditions: “32-band AM+FM”, “32-band AM; “32-band AM<16Hz”; “8-band AM”) x 2 (Phonetic contrasts: place *versus* voicing) X 2 (Sequence type: familiar *versus* novel) repeated-measure analysis of variance (ANOVA) of looking time was run, with Sequence type as a within-subject factor. This analysis revealed a main effect of Sequence type [mean novel=7.5 s, SD=0.81 s *versus* mean familiar=6.1 s, SD=0.91 s; F(1,152)=40.11, p<.001, $\eta^2 = 0.21$]. There was no effect of Vocoder condition [F(1,152)=2.01, p=.12] or Contrast [F(1,152)=1.67, p=.20], and no significant interaction between factors. The same results have been observed when the analyses were restricted to the first novel and familiar trials or the first 2 novel and familiar trials. Thus, 6-month-olds discriminated voicing and place contrasts regardless of the vocoder condition.

- Figure 2 about here-

Habituation data

Figure 3 shows the mean looking times for the first five habituation trials in each condition collapsed across speech stimuli. The number of habituation trials needed to reach criterion (Figure 4A) and the mean looking times across habituation (Figure 4B) were calculated for each condition collapsed across speech stimuli. Infants required a higher number of habituation trials in the "32-band AM<16Hz speech" condition compared to the other conditions. Moreover, the habituation time was longer in this "32-band AM<16Hz speech" condition [mean=133.9 s; SD=57.4 s] compared to the "8-band AM speech" [mean=108.3 s; SD=47 s] and the "32-band AM speech" [mean=101.6 s; SD=41.8 s] conditions. The habituation time in the "32-band AM+FM speech" condition [mean=96.1 s; SD=44 s] was shorter than that in the other three conditions.

-Figure 3 about here-

-Figure 4 about here-

A 4 (Vocoder conditions) x 3 (Stimuli:/aba/, /apa/ or /ada/) factorial ANOVA was conducted on the number of habituation trials. A main effect of Vocoder condition was observed [F(3;148) = 3.86; p=0.012; η^2 = .07]. *Post-hoc* Tukey tests revealed that the "32-band AM<16Hz speech" condition led to a higher number of habituation trials compared to the "32-band AM+FM speech" and "32-band AM speech" conditions. No significant effect of Stimuli [F(2;148) = 1.46; p=.24] or interaction [F(6,148) = 0.71; p=.64] were observed.

The same results were observed on the mean cumulated habituation times. A 4

(Vocoder conditions) x 3 (Stimuli:/aba/, /apa/ or /ada/) factorial ANOVA on the mean cumulated habituation times revealed only a main effect of Vocoder condition [F(3,148) = 4.2, p=.007; η^2 = .08]. *Post-hoc* Tukey tests indicated that habituation times were significantly longer in the "32-band AM<16Hz speech" condition compared to the "32-band AM+FM speech" and "32-band AM speech" conditions. The remaining comparisons were not statistically significant. The analysis also showed no significant effect of Stimuli [F(2;148)=2.46, p=.09] and no significant interaction between Vocoder condition and Stimuli [F(6,148)=1.54, p=.17].

Thus, the analyses of the habituation data reveal that infants require more habituation trials and a higher habituation time to the stimuli (/aba/, /ada/ and /apa) in the "32-band AM<16Hz speech" condition.

IV. DISCUSSION

The present study assessed the ability of French-learning 6-month-old normal-hearing infants to discriminate spectro-temporally degraded speech signals. This study replicated and extended previous results obtained by Bertoncini *et al.* (2011) and Cabrera *et al.* (2013) for 6-month-old infants learning French.

Discrimination data

The present results showed that 6-month-old infants discriminated the degraded speech signals in all processing conditions. This was manifested by a clear-cut novelty preference. Infants did not require FM, fast (> 16 Hz) AM, or fine spectral speech cues to perceive variations in voicing and place of articulation. Altogether, these results indicate that as early as 6-months, the slowest AM cues extracted from a limited number of broad frequency bands are sufficient to discriminate these two phonetic contrasts. In adults, voicing perception is robust to such degradations (*e.g.*, Gilbert & Lorenzi, 2006; Gnansia *et al.*, 2008; 2009; Miller & Nicely, 1955; Shannon *et al.*, 1995) however, for place of articulation, the spectral details and FM cues (Başkent, 2006; Shannon *et al.,* 1995) were found to influence identification responses. Our results indicate that the *discrimination* of place of articulation–at least the difference between French plosives /b/-/d/–remains possible even with reduced FM cues at 6 months of age. However, it is important to note that adults were tested using an identification task, whereas infants were tested using a discrimination task. It is possible that robust speech identification requires greater redundancy (and thus, finer spectral and temporal details) than discrimination. Furthermore, it is possible that the visual habituation procedure used here is not sensitive enough to reveal a difference in the discrimination abilities of infants for voicing and place.

Habituation data

Still, differences appeared in the number of habituation trials and habituation-time data across vocoder conditions. It was initially expected that infants would require a shorter habituation time to the most *impoverished* (that is, less complex) speech stimuli. In the present experimental design, the most impoverished conditions corresponded to the two following forms of speech processing: i) the "32-band AM<16Hz" condition where FM and fast AM cues were degraded, and ii) the "8-band AM" condition where the FM and fine spectral cues were degraded. Results showed that infants took longer to reach the habituation criterion for the "32-band AM<16Hz" speech condition compared to the "32-band AM+FM" and "32-band AM" speech conditions. The longer habituation time observed for this speech condition not only shows that infants are sensitive to the fast AM cues ordinarily present in speech signals; it also reveals a specific difficulty when processing temporally-smeared-speech sounds, and indicates that fine temporal resolution (fast AM cues) has greater importance than fine spectral resolution (fine spectral details) for efficient processing of phonetic cues in infants.

This is consistent with the study of Cabrera *et al.* (2013) showing that with two minutes of familiarization, 6-month-olds exhibited a longer looking time for familiar sounds rather than novel sounds when the speech signals contained only the slowest AM cues. According to the conceptual framework proposed by Hunter and Ames (1988) and Holt (2011), these results suggest that the attenuation of fast AM speech cues increases the processing time required to adequately represent phonetic cues. Thus, the longer habituation required for temporally-smeared speech signals indicates that the fast AM cues corresponding to bursts, formant transitions and periodic F0-related fluctuations may indeed play some role in the accurate/efficient phonetic processing.

The reasons for a greater role of fine temporal cues–that is fast AM cues–in phonetic discrimination for infants are unclear. However, this result is consistent with previous studies emphasizing the role of prosodic cues related to voice-pitch (F0) in infants' speech perception and language learning (*e.g.*, Kemler Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989; Mehler *et al.*, 1988). Previous investigations of auditory development indicate that auditory temporal resolution is mature by 6 months of age (*e.g.*, Levi & Werner, 1996). It is thus unlikely that the greater role of fast AM cues reflects a specific limitation in the sensory encoding of the slow or fast AM components of speech sounds for infants. It is conceivable that greater experience with the native language–and thus, with the acoustic cues related to syllabic rate–may gradually improve the ability to make efficient use of the slow AM speech cues in phonetic perception and/or speech segmentation. Consistent with this idea, psychoacoustic studies indicate that auditory sensitivity to AM improves significantly with training for adults (*e.g.*, Füllgrabe & Moore, 2007). This may explain why the slow AM cues are found to play the most important role in phoneme and sentence identification for adults (*e.g.,* Drullman, Festen, & Plomp, 1994; Shannon *et al.*, 1995; Stone, Füllgrabe, & Moore, 2008).

This raises the possibility that the processing difficulty and perceptual weight of slow and fast AM speech cues may change during development. Nevertheless, it is still unclear whether fast AM rates play a role in phonetic discrimination in quiet for adults. For instance, previous work focused on identification tasks, and did not report any measure of listening effort (*e.g.*, reaction times) in these tasks. Future studies may investigate this point further by comparing infants (of different ages) and adults' abilities to use the fast *versus* slow AM cues in phoneme discrimination tasks assessing both accuracy and listening effort.

Further work is required to investigate directly the mutual dependence between spectro-temporal auditory processes on one hand and speech processes on the other hand at an early age of development. The progressive maturation of the auditory system should affect the development of speech perception knowing that, at least for adults, resolving fine spectral and temporal details in speech sounds is indispensable to speech perception (especially, in adverse listening conditions). Despite this, infants demonstrate exquisite abilities to perceive speech sounds and learn the properties of their native language by exposure and social interactions during their first six to twelve months of life (Kuhl, 2004). Fine manipulations of the speech signal are now possible using psychoacoustic tools and will participate to reveal the contribution of low-level auditory mechanisms to speech perception during early development.

Clinical implications

In the present experiment, the vocoder extracting AM cues from a small number of frequency bands (8 broad bands) simulates speech processing in CI devices (Friesen *et al.*, 2001; Fu & Nogaki, 2005; see also Strydom & Hanekom, 2011). Our results indicated that at 6 months, normal-hearing infants are able to use the highly impoverished speech information transmitted by CI processors to discriminate phonetic contrasts such as voicing or place when the language is syllable-timed (*e.g.,* French). The present results therefore suggest that CI devices deliver sufficient information to the (typical) central auditory system in development for phonetic discrimination in quiet. Consistent with these results, relatively good syllable identification capacities are reported in French children wearing CIs (*e.g.,* Bouton, Bertoncini, Serniclaes, & Colé, 2011; Bouton, Serniclaes, Bertoncini, & Colé, 2012). However, these studies showed that place of articulation was more difficult to identify than voicing, in contrast with the present study showing no difference between voicing or place discrimination responses in the vocoder condition simulating CI processing. This discrepancy may point to an important limitation in the use of vocoders to simulate CI hearing. Alternatively, the behavioral method used in the present study may fail to demonstrate graded discrimination responses between voicing and place features. Nevertheless, vocoder studies simulating speech perception under CI conditions in normal-hearing infants should help to improve some aspects of early rehabilitation in deaf infants and may help to identify major difficulties in speech discrimination with the current CI processors.

V. CONCLUSION

The current study investigated the role of speech modulation cues in phonetic discrimination for young normal-hearing infants. The results showed that the discrimination of voicing and place of articulation is possible in the absence of FM and fast (> 16 Hz) AM cues when the spectral detail in speech signals is preserved and when spectral information is severely reduced to 8 broad frequency bands.

These results demonstrate that the slowest AM cues are sufficient for phonetic discrimination in infants. However, when the fast AM cues were attenuated, infants required a longer time to habituate to the degraded stimuli, suggesting that fast AM cues may contribute to efficient phonetic processing in infants.

ACKNOWLEDGMENTS

Christian Lorenzi was supported by grants from investissements d'avenir / ANR: ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*. Josiane Bertoncini was supported by ANR-12-ISH2-0001-01. The authors would like to thank the parents and their kids for their participation. The authors also thank Lynne Werner for detailed comments on this article.

FOOTNOTES

1. "Fine spectral cues" refer to the spectral cues conveyed by stimuli processed by a vocoder using a *high* spectral resolution (32 analysis bands, that is a spectral resolution mimicking that of the normal human ear; e.g., Glasberg & Moore; 1990). Consequently, a vocoder using a much poorer spectral resolution (e.g., a 8-band vocoder) reduces the fine spectral cues.

2. In the present study, tone-excited vocoders were used instead of noise-excited vocoders (Cabrera *et al.*, 2013; Eisenberg *et al.*, 2000; Newman & Chatterjee, 2013; Shannon *et al.,* 1995) because they were found to distort speech AM cues less. Kates (2011) compared the effect of five different procedures used to preserve AM cues and replace the original FM cues. This study showed that for each processing scheme (e.g., tone- and noise-excited vocoders), the original AM cues are substantially distorted; however, the most accurate (i.e., less distortive) processing is performed by the tone vocoder. This is due to the fact that noise carrier convey intrinsic random fluctuations in amplitude, that interfere with the auditory processing of the target speech AM cues (see also Dau, Kollmeier, & Kohlrausch, 1997; Stone, Füllgrabe, & Moore, 2012).

REFERENCES

Başkent, D. (2006). Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels. *The Journal of the Acoustical Society of America*, *120*(5 Pt 1), 2908–2925.

Bertoncini, J., Nazzi, T., Cabrera, L., & Lorenzi, C. (2011). Six-month-old infants discriminate voicing on the basis of temporal envelope cues. *The Journal of the Acoustical Society of America*, *129*(5), 2761–2764.

Bertoncini, J., Serniclaes, W., & Lorenzi, C. (2009). Discrimination of speech sounds based upon temporal envelope versus fine structure cues in 5-to 7-year-old children. *Journal of Speech, Language, and Hearing Research*, *52*(3), 682–695.

Bouton, S., Bertoncini, J., Serniclaes, W., & Colé, P. (2011). Reading and reading-related skills in children using cochlear implants: prospects for the influence of cued speech. *Journal of Deaf Studies and Deaf Education*, *16*(4), 458–473.

Bouton, S., Serniclaes, W., Bertoncini, J., & Colé, P. (2012). Perception of speech features by French-speaking children with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *55*(1), 139–153.

Burnham, D., & Mattock, K. (2010). Auditory development. *The Wiley-Blackwell Handbook of Infant Development*, *1*, 81–119.

Cabrera, L., Bertoncini, J., & Lorenzi, C. (2013). Perception of Speech Modulation Cues by 6-Month-Old Infants. *Journal of Speech, Language, and Hearing Research*, *56*(6), 1733–1744.

Cohen, L. B., Atkinson, D. J., & Chaput, H. H. (2000). *Habit 2000: A new program for testing infant perception and cognition*. Austin: The University of Texas.

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, *92*(2), 225–234.

Dau, T., Kollmeier, B., & Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. *The Journal of the Acoustical Society of America*, *102*(5 Pt 1), 2906–2919.

De Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *The Journal of the Acoustical Society of America*, *97*(1), 585–592.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, *95*, 2670.

Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, *11*, 169–177.

Eaves, J. M., Summerfield, A. Q., & Kitterick, P. T. (2011). Benefit of temporal fine structure to speech perception in noise measured with controlled temporal envelopes. *The Journal of the Acoustical Society of America*, *130*(1), 501–507.

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., & Boothroyd, A. (2000). Speech recognition with reduced spectral cues as a function of age. *The Journal of the Acoustical Society of America*, *107*(5 Pt 1), 2704–2710.

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, *110*(2), 1150–1163.

Fu, Q.-J., & Nogaki, G. (2005). Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology*, *6*(1), 19–27.

Füllgrabe, C., & Moore, B.C.J. (November, 2007). A perceptual-learning investigation of auditory amplitude-modulation detection: Testing the existence of frequency-selective mechanisms in the temporal-envelope domain. 37th Annual Meeting of the Society for Neuroscience. San Diego, USA.

Gilbert, G., & Lorenzi, C. (2006). The ability of listeners to use recovered envelope cues from speech fine structure. *The Journal of the Acoustical Society of America*, *119*(4), 2438–2444.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1-2), 103–138.

Gnansia, D., Jourdes, V., & Lorenzi, C. (2008). Effect of masker modulation depth on speech masking release. *Hearing Research*, *239*(1-2), 60–68.

Gnansia, D., Péan, V., Meyer, B., & Lorenzi, C. (2009). Effects of spectral smearing and temporal fine structure degradation on speech masking release. *The Journal of the Acoustical Society of America*, *125*(6), 4023–4033.

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460.

Holt, R. F. (2011). Enhancing Speech Discrimination Through Stimulus Repetition. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1431–1447.

Hunter, MA.; Ames, EW. A multifactor model of infant preferences for novel and familiar stimuli. In: Lipsitt, LP., editor. Advances in child development and behavior. New York:

Academic Press; 1988. p. 69-95.Kates, J. M. (2011). Spectro-temporal envelope changes caused by temporal fine structure modification. *The Journal of the Acoustical Society of America*, *129*(6), 3981–3990.

Kemler Nelson, D. G., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. *Journal of Child Language*, *16*(1), 55–68.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews. Neuroscience*, *5*(11), 831–843.

Levi, E. C., & Werner, L. A. (1996). Amplitude modulation detection in infancy: Update on 3-month-olds. *Association for Research in Otolaryngology*, *19*, 142–150.

Mattock, K., Molnar, M., Polka, L., & Burnham, D. (2008). The developmental course of lexical tone perception in the first year of life. *Cognition*, *106*(3), 1367–1381.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27*, 338–352.

Moore, B. C. J. (2003). Speech processing for the hearing-impaired: successes, failures, and implications for speech mechanisms. *Speech Communication*, *41*(1), 81–91.

Moore, B. C. J. (2007). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*. John Wiley & Sons.

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*, *13*(3), 407–420.

Nelson, P. B., & Jin, S.-H. (2004). Factors affecting speech understanding in gated interference: cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *115*(5 Pt 1), 2286–2294.

Newman, R., & Chatterjee, M. (2013). Toddlers' recognition of noise-vocoded speech. *The Journal of the Acoustical Society of America*, *133*(1), 483–494.

Patterson, R. D. (1987). A pulse ribbon model of monaural phase perception. *The Journal of the Acoustical Society of America*, *82*(5), 1560–1586.

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *The Journal of the Acoustical Society of America*, *114*(1), 446–454.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects.

Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, *336*(1278), 367–373.

Saffran, J. R., Werker, J. F., & Werner, L. A. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In *Handbook of child psychology* (D. Kuhn & R. Siegler., Vol. 2, pp. 58–108).

Shamma, S., & Lorenzi, C. (2013). On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system. *The Journal of the Acoustical Society of America*, *133*(5), 2818–2833.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Sheft, S., Ardoint, M., & Lorenzi, C. (2008). Speech identification based on temporal fine structure cues. *The Journal of the Acoustical Society of America*, *124*(1), 562–575.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, *416*(6876), 87–90.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience*, *32*(25), 8443–8453.

Spetner, N. B., & Olsho, L. W. (1990). Auditory frequency resolution in human infancy. *Child Development*, *61*(3), 632–652.

Stone, M. A., Füllgrabe, C., & Moore, B. C. (2008). Benefit of high-rate envelope cues in vocoder processing: Effect of number of channels and spectral region. *The Journal of the Acoustical Society of America*, *124*, 2272–2282.

Stone, M. A., Füllgrabe, C., & Moore, B. C. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, *132*(1), 317–326.

Strydom, T., & Hanekom, J. J. (2011). The performance of different synthesis signals in acoustic models of cochlear implants. *The Journal of the Acoustical Society of America*, *129*(2), 920–933.

Werker, J. F., Shi, R., Desjardins, R., Pegg, J. E., Polka, L., & Patterson, M. (1998). Three methods for testing infant speech perception. In *Perceptual development: Visual, auditory, and speech perception in infancy* (A. Slater., pp. 389–420). East Sussex, United Kingdom: Psychological Press.

Werner, L., Folsom, R. C., Mancl, L. R., & Syapin, C. L. (2001). Human Auditory Brainstem Response to Temporal Gaps in Noise. *Journal of Speech, Language & Hearing Research*, *44*(4), 737–750.

Zeng, F.-G., Nie, K., Stickney, G. S., Kong, Y.-Y., Vongphoe, M., Bhargave, A., … Cao, K. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2293–2298.

Figures

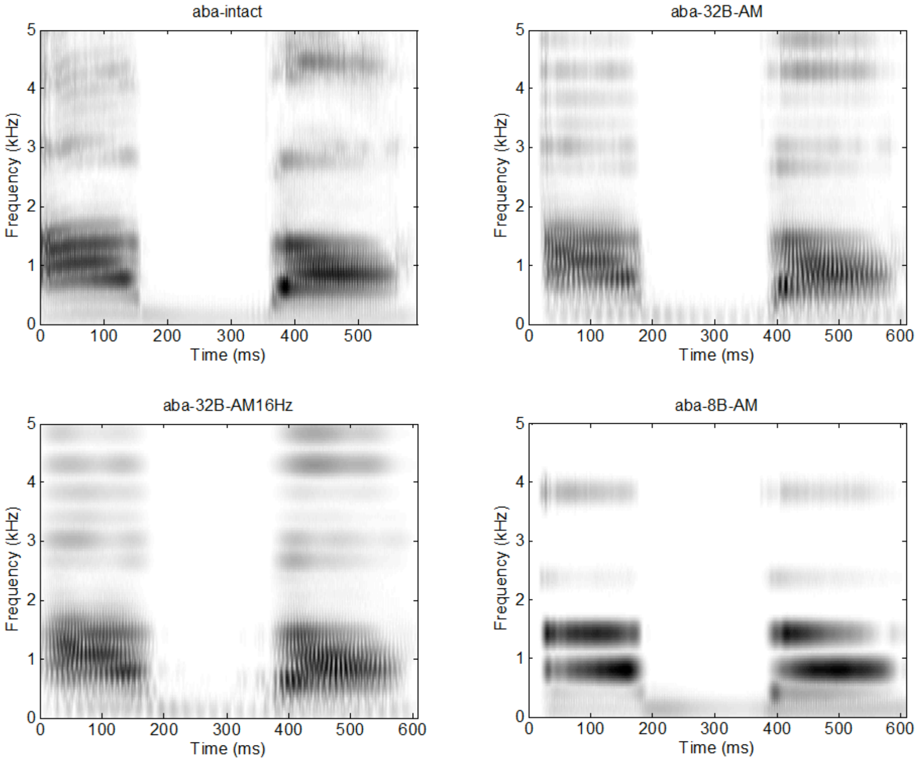

Figure 1. Spectrograms of /aba/ stimuli in each speech-processing condition. Upper left panel: intact condition (“32-bands-AM+FM”); upper right panel: “32-bands-AM”; lower left panel “32-bands-AM<16Hz”; lower right panel “8-bands-AM” speech conditions.

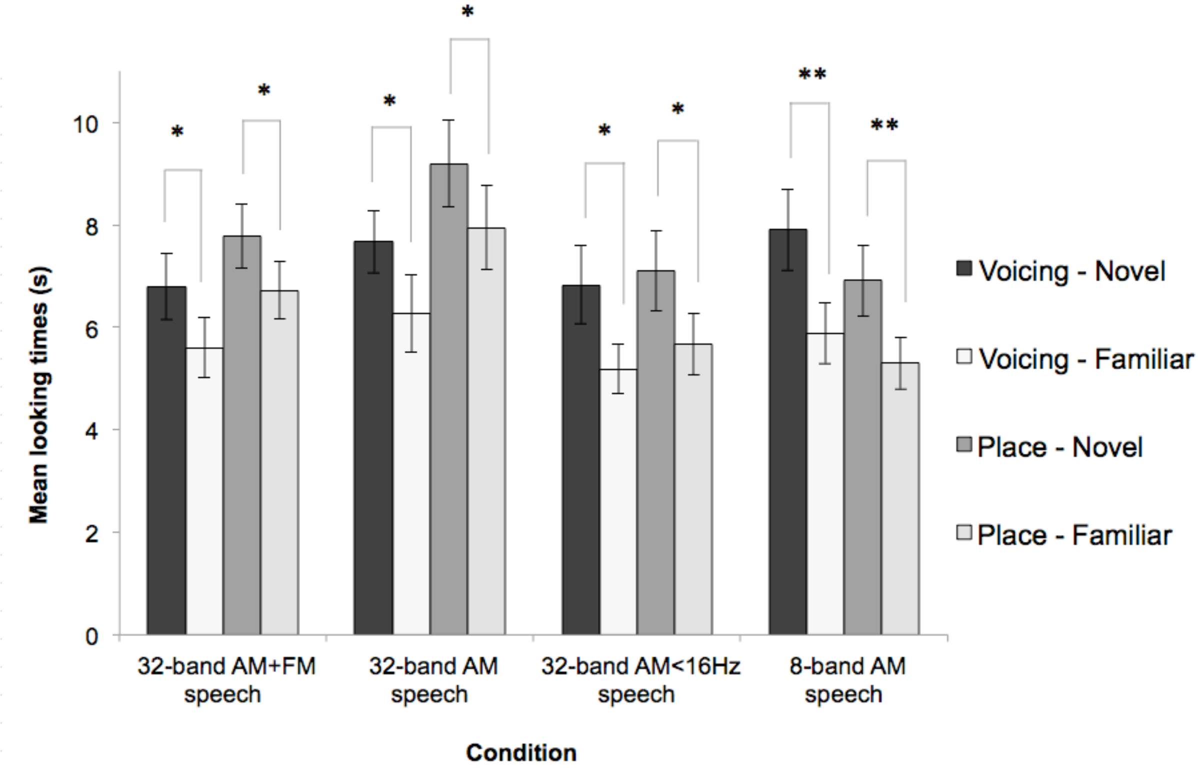

Figure 2. Mean looking times (s) for familiar and novel stimuli during the test phase, for voicing and place contrasts in each speech-processing condition: 32-band AM+FM speech, 32-band AM speech, 32-band AM<16Hz speech, 8-band AM speech. Errors bars indicate standard error.

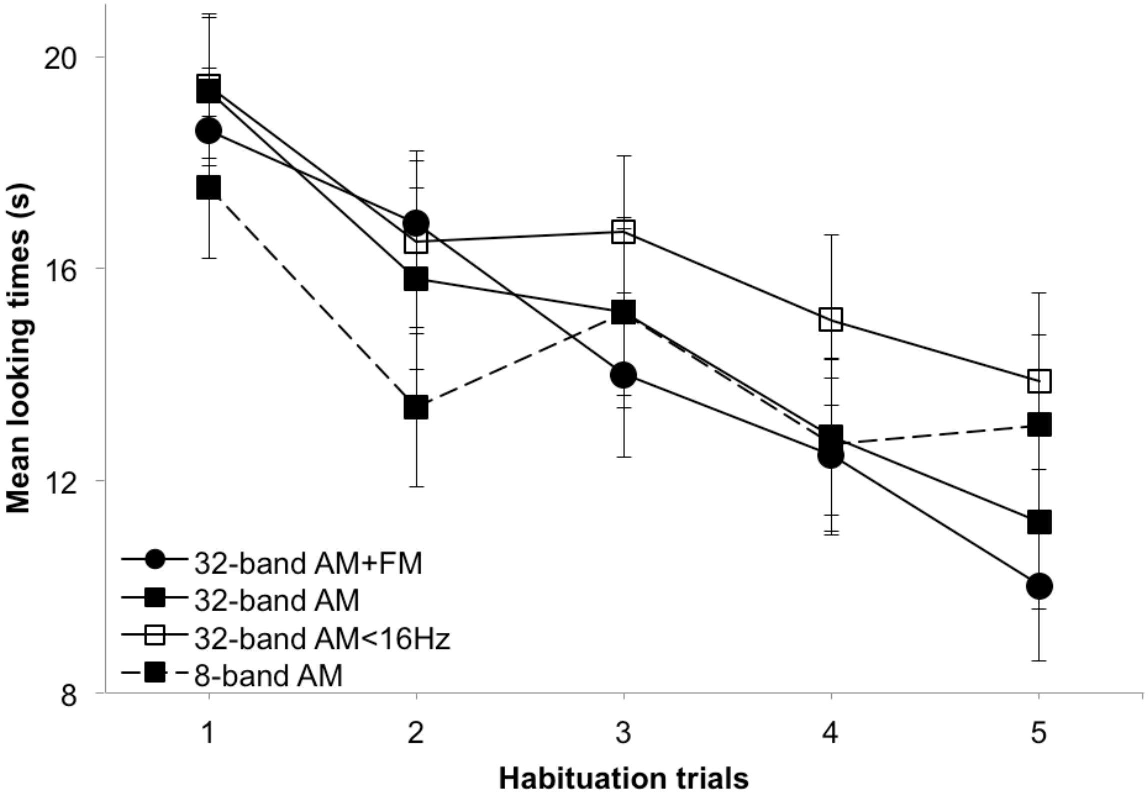

Figure 3. Mean looking time (s) in the four vocoder conditions collapsed across contrast across the first 5 habituation trials. Error bars indicate standard error.

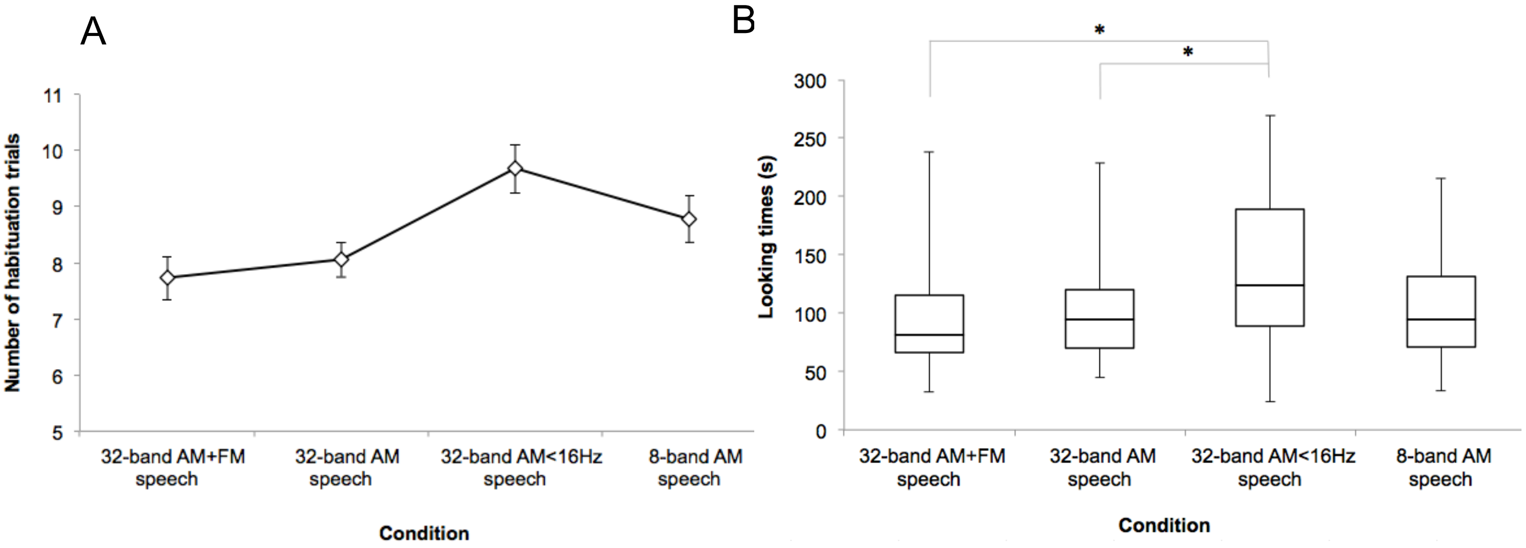

Figure 4. A. Mean number of habituation trials for the 40 infants in each vocoder condition (32-band AM+FM speech, 32-band AM speech, 32-band AM<16Hz speech, 8-band AM speech), collapsed across contrast. Errors bars indicate standard error.

B. Total habituation time (s) for infants in each vocoder condition (32-band AM+FM speech, 32-band AM speech, 32-band AM<16Hz speech, 8-band AM speech). The three horizontal lines represent the 25th, 50th and 75th percentiles, respectively, and the ends of the vertical bars minimum and maximum habituation times.