Animal Cognition (2021) 24:843–854
https://doi.org/10.1007/s10071-021-01478-4

ORIGINAL PAPER

A chimpanzee recognizes varied acoustical versions of sine-wave and noise-vocoded speech

Lisa A. Heimbauer[1] · **Michael J. Beran**[2] · **Michael J. Owren**[3]

Received: 6 June 2020 / Revised: 1 January 2021 / Accepted: 11 January 2021 / Published online: 8 February 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

Abstract

Previous research demonstrated that a language-trained chimpanzee recognized familiar English words in sine-wave and noise-vocoded forms (Heimbauer et al. Curr Biol 21:1210–1214, 2011). However, those results did not provide information regarding processing strategies of the specific acoustic cues to which the chimpanzee may have attended. The current experiments tested this chimpanzee and adult humans using sine-wave and noise-vocoded speech manipulated using specific sine-waves and a different number of noise bands, respectively. Similar to humans tested with the same stimuli, the chimpanzee was more successful identifying sine-wave speech when both SW1 and SW2 were present – the components that are modeled on formants F1 and F2 in the natural speech signal. Results with noise-vocoded speech revealed that the chimpanzee and humans performed best with stimuli that included four or five noise bands, as compared to those with three and two. Overall, amplitude and frequency modulation over time were important for identification of sine-wave and noise-vocoded speech, with further evidence that a nonhuman primate is capable of using top-down processes for speech perception when the signal is altered and incomplete.

Keywords Chimpanzee · Speech perception · Acoustic cues

Introduction

There has been much discussion regarding whether the capability to perceive speech, which involves many levels of processing—from the auditory input to the comprehension of lexical meaning—is uniquely human. Lenneberg (1967) claimed that speech production and speech perception are uniquely human adaptations—a view later termed "Speech is Special" (SiS) by Liberman (1982). While the SiS view proposed that humans possess a specialized cognitive module for speech perception (Mann and Liberman 1983), the "Auditory Hypothesis" (Kuhl 1988) suggested spoken-language evolution took advantage of existing auditory-system capabilities.

Historically, numerous experiments have investigated human and nonhuman speech perception to evaluate these opposing views. For example, evidence proposed to support the SiS approach includes that humans are able to recognize meaningful speech in a number of fundamentally altered, synthetic forms (Davis et al. 2005; Remez 2005; Trout 2001). In contrast, early studies with nonhumans revealed that some animals are able to discriminate and categorize phonemes—the smallest unit of speech sounds—much as humans do (Kluender et al. 1987; Kuhl and Miller 1975; Kuhl and Padden 1982, 1983). More recent experiments with a language-trained chimpanzee (*Pan troglodytes*) named Panzee demonstrated that she recognized synthetic speech in some highly reduced forms (i.e., sine-wave and noise-vocoded speech) that humans have been tested with (Heimbauer et al. 2011). Therefore, it may be that auditory processing in humans and nonhumans is similar, as originally proposed by Kuhl (1988). In this view, a common evolutionary ancestor of humans and other mammals possessed latent speech-processing capabilities that predated speech itself. Evolution of human speech-production capabilities would then have taken advantage of existing auditory processing and cognitive abilities.

Michael J. Owren deceased (2014).

✉ Lisa A. Heimbauer
lisa.heimbauer@gmail.com

[1] State University of New York at Delhi, Delhi, NY, USA

[2] Georgia State University, Atlanta, GA, USA

[3] Emory University, Atlanta, GA, USA

As evidenced by her language abilities, Panzee was a unique animal model for investigating speech perception. She was raised by human caregivers from the age of eight days old and exposed to language during the first few weeks of her life (Brakke and Savage-Rumbaugh 1995a, b). Panzee learned about the visual, lexigram-based, symbolic language that she used to communicate in the context of everyday life, similarly to how a human infant would learn language skills. As a result of this human-like method of language acquisition, she was the only *Pan troglodytes* trained to use lexigrams that understood spoken English words. Over a period of 10 years (from 1999 through 2008), Panzee's English comprehension averaged 125 words and demonstrated consistency for 107 of those words over at least 8 of the 10 years. Not only did this show a savings measure of 85% for meaningful words, but it also revealed a significantly higher ability of identifying lexigrams from spoken words than from photo samples (Beran and Heimbauer 2015). Additionally, Panzee was tested with whispered speech and with speech by a variety of talkers. Her comprehension abilities for whispered words matched that of the same words in natural (un-whispered) form, 82.3% correct compared to 83.6%, respectively (Heimbauer et al. 2018). This also was consistent with her historically 75% to 85% correct performance on natural words (Heimbauer et al. 2011). Across a variety of talkers, known and unknown to her, there was no difference in her comprehension performance, revealing her ability to solve the lack-of-invariance problem that human listeners are presented with in infancy (Heimbauer et al. 2018). In fact, research has shown that learning talker-specific characteristics can improve linguistic processing (Nygaard and Pisoni 1998; Nygaard et al. 1994; Pisoni 1995), which may have been the case with Panzee as well. In all of these experiments, Panzee was tested with the same method and recognition task used in the current experiments.

Although Panzee's demonstrated natural and synthetic speech word-recognition abilities provided evidence for the Auditory Hypothesis view (Heimbauer et al. 2011), those abilities did not, however, allow for an unequivocal conclusion that Panzee's speech processing is fundamentally similar to human perception. It could have been that Panzee was able to recognize her relatively small number of familiar words utilizing more holistic cues, such as word length or general sound impressions. However, additional analyses conducted by Heimbauer et al. (2011) and provided in the Supplemental Experimental Procedures (under *Syllable number and duration*), argued against this possibility. In the previous experiments, it was expected that when Panzee made errors she would have been choosing foils that corresponded to words whose overall duration or syllable count were similar to those of the target word. However, she did not demonstrate either of these strategies when errors on sine-wave and noise-vocoded words were analyzed. Therefore, in an attempt to acquire specific evidence about detailed aspects of the acoustic cues Panzee may have been attending to when identifying synthetic, altered words, we conducted the current experiments.

Current experiments

Studies using fundamentally altered speech forms have been invaluable for understanding how the human cognitive system organizes acoustic elements of speech for meaningful language comprehension. One important characteristic is the fundamental frequency (F0), which corresponds to the basic rate of vibration of the vocal folds in the larynx, and typically is the lowest prominent frequency visible in a speech spectrogram (Fig. 1a). Regular vibration also produces energy at integer multiples of F0, referred to as "harmonics" (e.g., H2 and H3; see Fig. 1a), and this energy is filtered via resonances that occur as it passes through the vocal tract. These resonances, termed "formants," strengthen the energy in some frequency regions while weakening it in others. These effects are visible in a spectrogram as larger, dark bands of energy, as can be seen for the lowest three formants of the natural word "tickle" in Fig. 1a (i.e., F1, F2, and F3).

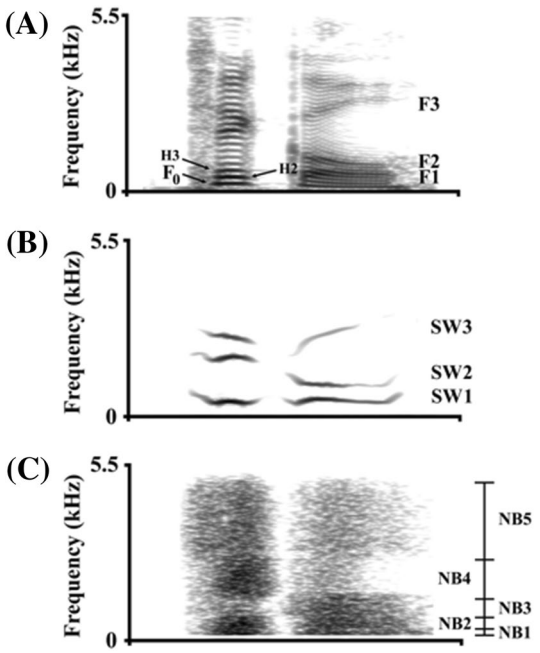

Fig. 1 Spectrographic word examples. The spectrograms were created using a sampling rate of 44,100 Hz and 0.03 s Gaussian analysis window. **a** The natural word "tickle," showing its fundamental frequency (F0), next higher harmonic (H2 and H3), and lowest three formants (F1, F2, and F3). **b** The word "tickle" in sine-wave form, with individual sine waves (SW) marked. **c** The word "tickle" in noise-vocoded form, made with five noise bands (NB)

In the current experiments, two forms of altered speech were of particular interest: sine-wave and noise-vocoded. Both synthetic forms lack many of the acoustic features traditionally considered crucial to speech perception, including F0 and formants (see Fig. 1b and c). Sine-wave and noise-vocoded speech are synthetic versions that reduce normal speech acoustics to small sets of sine waves and noise bands, respectively. When humans identify words in these forms, they may be attending to the spectro-temporal cues produced by amplitude and frequency modulations over time and not necessarily phonemic components of the acoustic signal (for a review see Remez et al. 2013). Based on Panzee's previous performance and the fact that we had preliminary evidence that she was not attending to more holistic cues (Heimbauer et al. 2011), we hypothesized that she would rely on the same acoustic parameters as humans when hearing these reduced speech forms (Remez et al. 1981; Shannon et al. 1995). Therefore, to test this hypothesis, Panzee and human participants were presented with varying versions of sine-wave and noise-vocoded speech.

More specifically, in Experiment 1 we examined the acoustic cues that Panzee and human participants may be attending to when hearing English words in three different sine-wave forms. In Experiment 2 we addressed this question using words in five noise-vocoded forms, again testing Panzee and human participants. As a result, the synthetic words included differing degrees of time-varying amplitude and frequency cuing of a kind previously shown to systematically affect human performance and characterized as critical spectro-temporal patterning in each natural word preserved in synthetic versions (Remez et al. 1981; Shannon et al. 1995). Our rationale was that if performance by Panzee and humans was similarly compromised or facilitated across the various synthetic stimuli, the two species could be inferred to be attending to similar elements of the sounds.

General methods

Subject

The subject was a female chimpanzee, Panzee, who was 25 years old at the time the experiments began. This animal was socially housed with three conspecifics at the Language Research Center at Georgia State University. Panzee had daily access to indoor and outdoor areas, unlimited access to water, and was fed fruits and vegetables three times a day. She participated in testing on a voluntary basis and was able to choose not to participate or to stop responding during a session by leaving the test area and moving to indoor or outdoor areas of her choosing. Panzee used a symbol (lexigram)-based communication system to request items throughout the day and often during experimental situations. In addition to language-comprehension testing using lexigrams and photographs (e.g., Beran and Heimbauer 2015), this animal also had experience with numerous, computer-based protocols (Beran 2010; Beran et al. 2004; Rumbaugh and Washburn 2003). In the current experiments, she participated in three to four, 20- to 30-min sessions per week, working for requested food items. She was tested in an indoor area of her daily living space, which was adjacent to other chimpanzee areas. During test sessions, other chimpanzees could be indoors or outdoors, with the option of moving between those areas at will.

Participants

Human participants were Georgia State University undergraduate students, 18–55 years old, who were recruited via an online experiment participation system. Twelve participants (eight females) were tested in Experiment 1, and 12 participants (eight females) were tested in Experiment 2. All participants reported having no hearing problems and were Native English speakers.

Apparatus

Computer programs used to test the chimpanzee were written in Visual Basic Version 6.0 (Microsoft Corp., Redmond WA) and run on a Dell Dimension 2400 personal computer (Dell USA, Round Rock TX). A Samsung Model 930B LCD monitor (Samsung Electronics, Seoul, South Korea), a Realistic SA-150 stereo amplifier (Tandy Corp., Fort Worth TX), and two ADS L200 speakers (Analog & Digital Systems, Wilmington, MA) were connected to the computer. The chimpanzee registered her choices using a customized Gravis 42111 Gamepad Pro video-gaming joystick (Kensington Technology Group, San Francisco CA). Human participants heard experimental stimuli through Sennheiser HD650 headphones in a sound-deadened room. The experiments were controlled via a computer from an adjacent room, and sounds were presented via TDT System II modules (Tucker-Davis Technologies Alachua, FL). Audio recording was conducted with a Shure PG14/PG30-K7 head-worn wireless microphone system (Shure Inc., Niles, IL), and either a Realistic 32–12,008 stereo mixing console (Tandy Corp., Ft. Worth TX) and Marantz PMD671 Professional Solid-State Recorder (Mahwah, New Jersey), or a MacBook Pro laptop computer (Apple Inc., Cupertino CA). Acoustic processing was conducted using a MacBook Pro laptop, Praat Version 5.1.11, acoustics software (Boersma and Weenink 2008), and custom-written scripts (Owren 2010).

Stimuli

Natural word stimuli were recorded at 44,100 Hz with 16-bit word-width and filtered to remove any 60-Hz, AC contamination and DC offset. Words were spoken by an adult male researcher who was very familiar to Panzee and were chosen from a list of approximately 130 words that she has consistently identified in a decade of annual word-comprehension testing (Beran and Heimbauer 2015). Individual words were isolated by cropping corresponding segments at zero crossings, with 100 ms of silence then added to the beginning and end of each file. Finally, each waveform was rescaled so its maximum amplitude value coincided with the maximum representable value. Twenty-four natural processed words were resynthesized into three sine-wave forms in Experiment 1, and 24 were resynthesized into four noise-vocoded forms in Experiment 2, with some word overlap between the two experiments.

Chimpanzee procedure

Panzee was tested using a general procedure employed for her annual word-comprehension testing (Beran and Heimbauer 2015). She initiated a trial by using the joystick to move a cursor from the bottom of the LCD screen into a centered "start" box, triggering one presentation of the stimulus. The cursor then reset to the bottom of the screen, the start box reappeared, and a second cursor movement produced another stimulus presentation. After a 1-s delay, four different photographs appeared on the screen (for examples see Fig. 2). One of these items was the correct match to the audio stimulus, and the others were foils chosen randomly by the controlling computer program. Visual items were positioned randomly in four of six possible locations on each trial. Photograph foils were those of words used in the same session, thereby reducing the chance that Panzee could rule out items corresponding to words she was not hearing (Beran and Washburn 2002).

Panzee's task was to use the joystick to move the cursor from the middle of the screen to the photograph corresponding to the stimulus word (see Fig. 3). When Panzee heard a word in natural form and made a correct choice, she heard an ascending ("correct") tone and received a food reward. When she made an incorrect choice on a natural word trial, she heard a buzzer-like ("incorrect") sound and did not receive a reward. Neither feedback sounds nor food rewards

Fig. 3 Panzee working on a computer task, hearing words and choosing corresponding photos

were provided on trials with synthetic stimuli. For correct, natural trials Panzee was rewarded with highly valued food, including pieces of cherries, raspberries, raisins, or Chex Mix®. This reward regimen kept Panzee motivated.

Human procedure

A word-recall method was used whereby participants had to transcribe the sounds they heard. Participants were familiarized with what the synthetic speech sounded like by listening to a recording of the words "one" through "ten" and then "ten" through "one" in the particular test form. They were instructed to inform the experimenter as soon as they were able to identify these sounds as speech, and then they proceeded to the experimental phase.

In the experimental phase, participants heard each stimulus word twice on each trial with a 1200-ms delay between presentations. They then had eight seconds in which to transcribe that word. Stimuli were presented in two different randomized blocks, with block order counterbalanced across individuals. One block consisted of Group A words in natural form and Group B words in the synthetic forms, and the other block included Group B words in natural form and Group A words in the synthetic forms. Within a block, natural words were presented for two trials, and synthetic words were presented for one trial in each of the synthetic test forms. Each test session, therefore, included a total of 72 trials.

Data analysis

Panzee's mean percentage-correct performance in orientation versus test sessions with natural words was compared

Fig. 2 Samples of the photographs used in Panzee's spoken-word recognition task

using an unpaired *t*-test for a possible learning effect. The mean percentage-correct performance for each synthetic word form within and across the six test sessions was compared to the chance-rate performance of 25% using binomial tests. Pearson's chi-squared tests with a Bonferroni correction were conducted to compare Panzee's performance across the various sine-wave versions.

Humans' percentage-correct performance was computed for each synthetic word form, with a mean performance for the 12 participants then examined separately for natural and synthetic versions. An ANOVA was used to test for an overall effect of synthetic word forms, and Tukey post-hoc comparisons were used for subsequent pair-wise comparisons among them. Finally, an independent *t*-test was conducted to test for possible effects of block-presentation order.

Experiment 1

Since 1981, sine-wave speech has been investigated for the purpose of understanding spoken language processing (Lewis and Carrell 2007; Remez et al. 1981; Rosner et al. 2003). In this synthesis form, words or sentences are produced from three sine waves that track the first three formants of the natural speech signal (see Fig. 1b). Sine-wave speech is extremely unnatural-sounding and is considered to preserve key phonetic properties only in an abstract form (Remez et al. 2011). In their experiments, Remez and colleagues (1981) presented a sine-wave sentence to human listeners. When the participants were not told that these sounds could be understood as speech, they described them as "science-fiction sounds" or "whistles." When they were told that they would be hearing sentences produced by a computer, however, the listeners were typically able to identify a substantial number of the syllables and words in the sentence. The researchers concluded that perception of sine-wave speech was evidence for a "speech mode of perception," and that listeners expecting to hear a language-like stimulus tuned into this mode. Even in the absence of traditional acoustic cues, listeners were able to perceive phonetic content in the sine-wave signal.

In addition, their experiments revealed that sine-wave speech becomes more difficult to recognize when either the first (SW1) or second sine wave (SW2) of the three sine waves is removed (Remez et al. 1981). It was hypothesized that Panzee would also show evidence of relying disproportionately on these cues. Experiment 1 compared her performance to that of human participants when hearing four critically different versions of sine-wave words that included varying combinations of individual tones.

Methods

Stimuli

The 24-word set contained 9 two-syllable words, 13 three-syllable words, 1 four-syllable word, and 1 five-syllable word. An additional 12 words were used during an initial, "orientation" phase that included both natural and SW123 versions (see Table 1 for a complete list of orientation and experimental words). To produce the sine-wave stimuli in the three incomplete forms either SW1, SW2, or SW3 were removed from the previously constructed, processed SW123 versions (Heimbauer et al. 2011). Individual sine-waves were removed using Hanning-window, band-pass filtering.

Chimpanzee procedure

In all sessions, words were presented in four randomized blocks, for a total of 96 trials. Initial sessions assessed Panzee's performance when hearing the 24 test words in natural form to ensure normative performance. Criterion performance to progress to the orientation phases was set for at least 70% correct for three consecutive sessions.

During orientation, Panzee completed both natural and sine-wave sessions with non-test ("Orientation") words. First, she heard eight blocks of 12 natural Orientation words for two sessions, averaging 72% correct. Then she heard Orientation words in natural and sine-wave form for two sessions. In the first session, she heard six orientation words in natural form and the other six in SW123 form for eight blocks. In the second session, the six words Panzee previously heard as natural words were presented in SW123 form, and vice versa. After the sine-wave orientation phase, Panzee participated in one additional session with the 24 natural test words presented in four randomized blocks to refresh her on the test word set, performing at 80% correct.

In the testing phase, Panzee completed one session with the 12 (Group A) words in natural form and the remaining 12 (Group B) words in SW123, SW12, SW13, and SW23 forms. In a second session on a different day, she heard the Group B words in natural form and Group A words in the four sine-wave forms. Trials were randomized within blocks in these sessions, with Panzee hearing natural words four times each and sine-wave words once in each form. She participated in these two types of sessions three times each, in an alternating order, resulting in a total of 12 trials for each word in natural form and 3 trials for each word in every sine-wave form.

Human procedure

The word-recall and transcription method as described in the General Methods section was used in this experiment.

Table 1 Orientation and test word groups. Test words used in Experiment 1 (Group A and Group B) and Experiment 2 (Group C and Group D), as well as orientation words (O)

Word	Experiment 1	Experiment 2
Apple	O	O
Apricot	A	D
Balloon	O	O
Banana	A	
Blueberries	B	D
Bubbles	A	C
Carrot	O	
Celery		D
Cereal	O	O
Coffee	O	
Colony Room	B	D
Gorilla	B	C
Hotdog	O	
Jello	O	
Kiwi	O	
Kool-Aid	O	
Lemonade	B	D
Lettuce		C
Lookout	A	O
M&M	B	
Melon	A	
Mushroom trail	B	O
Noodles		C
Observation room	B	D
Orange	A	C
Orange drink	A	O
Orange juice	B	D
Peaches	O	C
Pine needle	B	O
Plastic bag	B	D
Popsicle	O	D
Potato	B	D
Raisin	O	C
Sparkler	A	C
Strawberries		O
Sugarcane	B	D
Surprise	A	C
Sweet potato		O
Tickle	A	C
Tomato		D
Toothpaste		C
TV		C
Vitamins		O
Water	A	O
Yogurt	A	O

During orientation, participants listened to a recording of the words "one" through "ten," and then "ten" through "one" in SW123 form. After participants informed the experimenter that they were able to identify these sounds as speech, they were presented with natural and manipulated word stimuli and had eight seconds in which to transcribe that word.

Results

As illustrated in Fig. 4a Panzee's mean performance over the three natural word orientation sessions was 73.3%, which was statistically above chance level ($p < 0.001$). Correct natural-word trials in the six test sessions ranged from 81.3% to 93.8%, averaging 87.2% overall, which also was significantly above chance ($p < 0.001$). An unpaired, two-tailed *t*-test revealed that Panzee's performance with natural words was significantly higher in test sessions than in orientation sessions, $t(7) = 3.95$, $p < 0.01$. Overall, correct performance for all sine-wave words was statistically above chance level (SW123 and SW12 forms, $p < 0.001$; SW23 and SW13, $p < 0.05$). As illustrated in Fig. 4b, Panzee was 36.1% correct for SW23 and SW13 words, and 58.3% correct for SW123 and SW12 words. A chi-squared test examining SW123, SW23, and SW13 data, with a Bonferroni corrected alpha value of 0.025, revealed that correct performance for SW123 words was significantly higher than for SW23 ($p = 0.006$) and SW13 versions ($p = 0.004$).

Mean transcription performance of natural words by humans was 100% correct, as shown in Fig. 4a. Mean percentage-correct values for SW123, SW12, SW23, and SW13 word forms were 42.7%, 35%, 30.6%, and 27.4%, respectively. A Kolmogorov–Smirnov test for normality validated use of ANOVA and results revealed a statistically significant difference among the various outcomes, $F(3,44) = 6.00$, $p = 0.002$. Tukey post-hoc comparisons showed that outcomes were significantly higher for SW123 stimuli than for SW13 versions, $p = 0.001$, as illustrated in Fig. 4c. No other differences were found. Independent, two-tailed, *t*-test results revealed an effect of presentation order. Five participants transcribed Group B words first and performed significantly better on Group A sine-wave words than did participants hearing Group A words first, $t(46) = 2.28$, $p = 0.027$. Similarly, the seven participants transcribing Group A words first performed significantly better on Group B sine-wave words than those hearing Group B words first, $t(29.6) = 9.33$, $p < 0.001$.

Examining individual performance, six participants performed similarly to Panzee, either overall or in one of the blocks. Specifically, for these participants, the performance was the same for SW123 and SW12, or for SW23 and SW13 word forms. These six participants also performed notably better on SW123 and SW12 words,

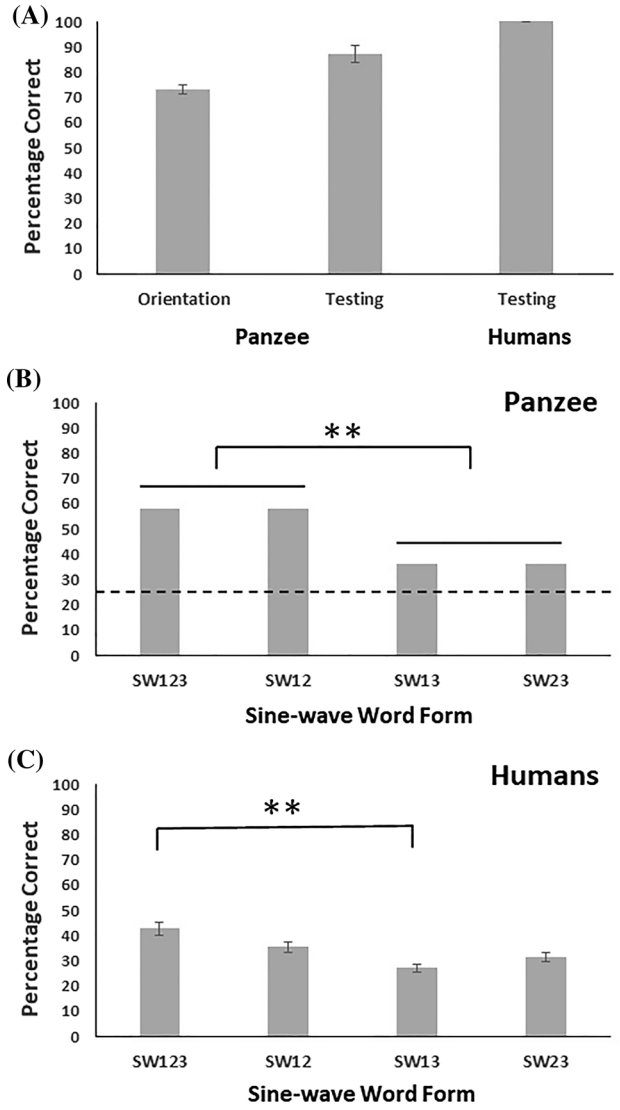

Fig. 4 Experiment 1 chimpanzee and human word recognition. **a** Mean performance with natural words by Panzee and the human participants, with applicable 95% confidence intervals. **b** Panzee's sine-wave word performance, with chance-level accuracy shown by the dashed line. **c** Mean human sine-wave word performance, with 95% confidence intervals

than on SW23 and SW13 forms. Although we only found a correlation between humans' and Panzee's performance on SW123 words, $r(22) = 0.47$, $p = 0.02$, humans and Panzee never recognized the words "banana," "bubbles," "orange drink," and "pine needle" in some of the SW forms. In addition, Panzee and the humans performed poorly (between 0 and 33%) with the same 12 words (50%) in SW13 form, and with the same 14 words (58.3%) in SW23 form.

Discussion

Panzee demonstrated consistent natural word-recognition performance, showing similar outcomes in orientation and test sessions as in earlier annual testing and synthetic-speech experiments (Heimbauer et al. 2011; Beran and Heimbauer 2015). Her recognition of SW123 words was also similar to performance in previous sine-wave testing (Heimbauer et al. 2011). In the current experiment, Panzee identified more words in SW123 and SW12 form than in SW13 and SW23 form. Humans performed similarly with a decrease in performance across conditions from SW123 to SW13, although only the difference between SW123 (with both SW1 and SW2 present) and SW13 (missing SW2) performance was statistically significant. Despite the difference in methodology between Panzee and the humans, 6 of the 12 human participants did perform similarly to Panzee, either overall or in one of the two test-word blocks. In other words, these participants performed exactly the same with SW123 and SW12 forms, or with SW23 and SW13 forms, and were better at identifying SW123 and SW12 words than SW23 and SW13 words in those instances.

Unexpectedly, Panzee's performance on SW123 words was 58% correct, which was higher than the mean human outcome of 43% correct (although, again, the response modality was different for the two species). Panzee's higher accuracy may be due to the fact that although sine-wave words can be quite challenging even to humans (Heimbauer et al. 2011), she was very familiar with her word set and had heard them in SW123 form in earlier experiments. Additionally, we acknowledge that closed-set testing procedures do not necessarily present the same challenges as open-set testing, as demonstrated by Sommers et al. (1997). However, in an attempt to compensate for this possible advantage, as noted in the *General Methods Human procedure* section, human participants in the current experiments were exposed to and tested with a block of Group A natural words and Group B test words, and then with a block of Group B natural words and Group A test words (presentation of blocks counter-balanced across participants). This means that they were hearing half of the natural words before being tested with them in SW (or NV) form. Also, as noted in the *Human procedure* section, we familiarized our human participants with sine-wave speech in SW123 form before they started the experiment (in Experiment 2 we familiarized them with a noise-vocoded speech in NB7 form first).

Sommers et al. (1997) also noted that closed-set testing protocols are important when examining discrimination of phonetic features in speech perception; and both sine-wave and noise-vocoded speech contain and preserve some

phonetic properties, albeit in abstract form (Dorman et al. 2002; Remez et al. 2011; Sawusch 2005). Therefore, we would argue that the more important result of our experiment is that both species showed a statistically significant performance difference between complete sine-wave words and those containing the F1 and F2, the first two formants (SW123 and SW12, respectively), and the same words when missing the tone analog to either F1 or F2 (SW23 and SW13, respectively).

Panzee's demonstrated ability to interpret sine-waves, possibly as cues to phonetic content, suggests that she was drawing on implicit knowledge of speech acoustics and corresponding phonetics (Davis and Johnsrude 2007; Mann and Liberman 1983; Newman 2006; Whalen and Liberman 1987). This outcome is indicative of some form of cognitive top-down processing of speech sounds, and the possibility that Panzee is responding to the same cues in sine-wave speech that humans respond to, with the further implication that she is attending to the same features as humans in natural speech as well. This conclusion is based on the findings that Panzee was most successful in identifying sine-wave speech that included information concerning both F1 and F2, the most important formants in human perception of natural speech (Drullman 2006; Remez and Rubin 1990).

Experiment 2

The second synthetic speech form presented to Panzee and human participants was "noise-vocoded" speech, which is synthesized from noise bands (see Fig. 1c). To create noise-vocoded speech, the natural signal is divided into a number of frequency bands using individual band-pass filters. The intensity pattern, or amplitude envelope, of each band is extracted over the length of that signal. Resulting envelopes are then used to modulate corresponding, frequency-limited bands of white noise. The result is a series of amplitude-modulated, noise waveforms that when summed potentially becomes recognizable as harsh, but comprehensible speech (Davis et al. 2005; Shannon et al. 1995).

Perception of noise-vocoded speech is of particular interest because it is a simulation of the input produced by a cochlear implant—a surgically implanted, electronic device for the hearing-impaired (Dorman et al. 2002). However, noise-vocoded speech is also useful in investigating speech perception in normally hearing individuals, as it preserves the amplitude and temporal information of the original utterance while omitting most spectral detail (Shannon et al. 1995). Even in the absence of F0 and formants noise-vocoded speech can carry a surprising amount of information regarding phonemes (Dorman et al. 2002; Sawusch 2005). One critical factor is the number of noise bands used in the synthesis process. Listeners cannot reliably recognize noise-vocoded speech created with only two noise bands. However, recognition becomes much more consistent if three or four bands are present (Shannon et al. 1995). When ten or more noise bands are used, noise-vocoded speech is readily intelligible even to naïve listeners (Davis et al. 2005). Individuals hearing speech in this synthesis form typically show improvement with practice. For example, Davis et al. (2005) reported that the identification of noise-vocoded words in sentences increased from less than 10% to 70% correct within just a few minutes.

Previous research has shown that humans find it easier to recognize sentences produced with four or more noise bands (Shannon et al. 1995), attributed to the fact that increased numbers of noise bands enhance the amplitude and frequency modulation information represented (see also Davis et al. 2005). Based on Panzee's previous performance with words in noise-vocoded form (Heimbauer et al. 2011), it was again hypothesized that she would show similar performance to humans as a function of the number of noise bands used in synthesis.

Methods

Stimuli

Stimuli consisted of 24 previously recorded and processed words, which were those that Panzee had best identified in noise-vocoded form in earlier testing (Heimbauer et al. 2011). The word list consisted of 11 two-syllable words, 11 three-syllable words, 1 four-syllable word, and 1 five-syllable word (see Table 1). Fifteen of the words also were used in Experiment 1. Noise-vocoded test versions of these words varied from two to five noise bands with an additional 12 words synthesized using seven noise bands (NB7) for orientation purposes. To produce the various noise-band test stimuli (NB2, NB3, NB4, and NB5), the natural speech signal was divided into 2, 3, 4, and 5 frequency bands using a band-pass filter with lower- and upper-cutoff frequencies (see Table 2) calculated using the "Greenwood Function" (Souza and Rosen 2009). This function calculates frequency

Table 2 Lower-to-upper cutoff frequencies for noise-band stimuli in Experiment 2

Bands	Frequencies (Hz)				
2	100–1005	1005–5000			
3	100–548	548–1755	1755–5000		
4	100–392	392–1005	1005–2294	2294–5000	
5	100–315	315–705	705–1410	1410–2687	2687–5000

ranges corresponding to equal distances along the basilar membrane of the cochlea and can be applied to both humans and other mammals, including nonhuman primates (Greenwood 1961, 1990). The approach was used to ensure the orderly selection of frequency-cutoff values as they relate to hearing. The amplitude envelope of each band was then extracted and used to modulate a corresponding white-noise band. The resulting amplitude-modulated noise waveforms were then summed.

Chimpanzee procedure

The testing procedure and reward regimen were the same as those used in Experiment 1, and all sessions consisted of four randomized blocks for a total of 96 trials. However, the orientation phase was somewhat different (and briefer) than in Experiment 1. Now during orientation, Panzee first completed one session with 12 non-test, orientation words in natural form, a second session with six of these in natural and six in NB7 form, and a third session with these words in the converse forms. In the final orientation phase, Panzee completed one more session with the 24 natural test words and then three sessions of these words in natural and NB7 forms. In each of these latter sessions, a different eight words were in NB7 form and the remaining 16 were natural versions.

In test sessions, we added an additional programming contingency. Here, words of the same type, meaning natural or noise-vocoded, were not presented more than three times in a row. We made this adjustment to avoid the frustration that could possibly result from Panzee hearing a series of challenging noise-vocoded words consecutively. In the first test session, Panzee heard 12 words (Group C) in natural form and the remaining 12 words (Group D) in NB2, NB3, NB4, and NB5 versions. In the second test session, on a different day, she heard Group D words in natural form and Group C words in the four NB versions (see Table 1). Within a session, there were four trials with each natural word and one trial with each of the words in every NB form. Panzee participated in these two sessions types three times, each in alternation order, resulting in a total of 12 trials for each of the 24 natural words and 3 trials for each of the NB word forms.

Human procedure

The word-recall and transcription method as described in the General Methods section was used in this experiment. During orientation, participants listened to a recording of the words "one" through "ten," and then "ten" through "one" in NB7 form.

Results

As shown in Fig. 5a Panzee's natural word-recognition performance in orientation sessions ranged from 77.2% to 83.3%, with an overall mean of 80.6%, which was statistically above chance level, $p < 0.001$. Percentage-correct on natural words in the six test sessions ranged from 77.1% to 87.5%, with an overall mean of 82.8% correct, which was also significantly above chance level ($p < 0.001$). An unpaired, two-tailed t-test revealed that Panzee's

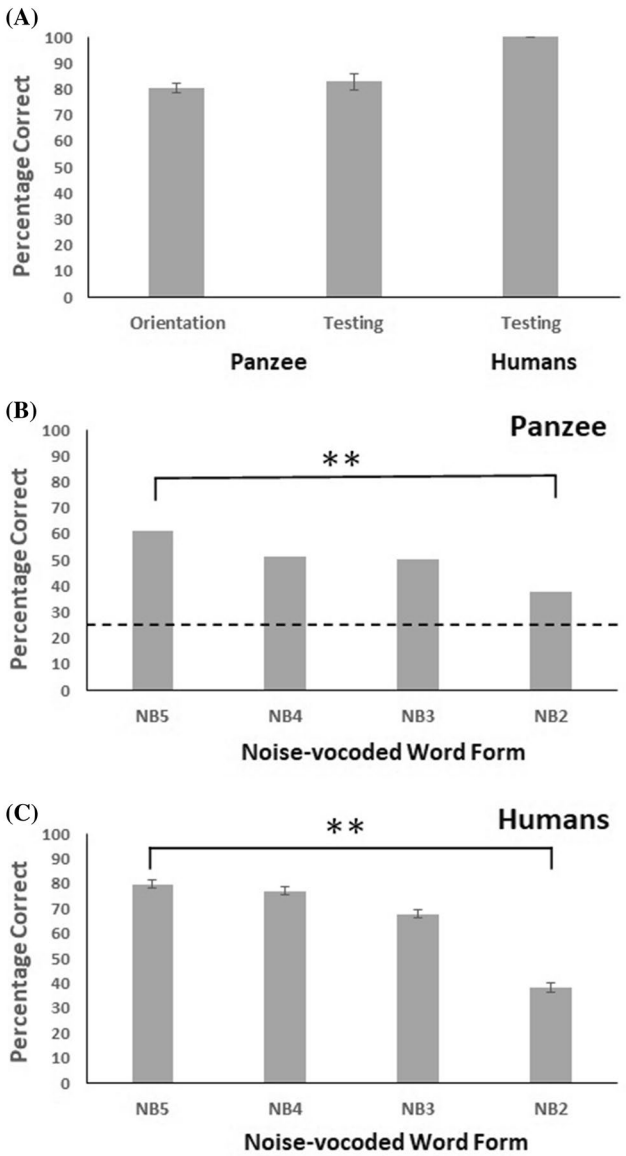

Fig. 5 Experiment 2 chimpanzee and human word recognition. **a** Mean performance with natural words by Panzee and the human participants, with applicable 95% confidence intervals. **b** Panzee's noise-vocoded word performance, with chance-level accuracy shown by the dashed line. **c** Mean human performance for noise-vocoded words, with 95% confidence intervals

natural-word performance was not statistically different between these two session types, $t(7) = 0.70$, *ns*.

Panzee's percentage correct for NB5, NB4, and NB3 word forms ranged from 61.1% to 50% (see Fig. 5b), and overall was significantly above chance ($p < 0.001$). Her NB2 word performance was lower at 37.5% correct, and not significantly different from chance. A one-tailed chi-squared test, with a Bonferroni adjusted alpha value of 0.017, showed that Panzee's recognition of NB5 words was significantly higher than NB2 versions ($p = 0.002$), but not higher than either NB4 or NB3 forms.

Human word transcription of natural words was 100% correct (see Fig. 5a). Mean percentage-correct values for NB2, NB3, NB4, and NB5 forms were 79.9%, 77.0%, 67.7%, and 38.2%, respectively (see Fig. 5c). After a Kolmogorov–Smirnov test showed the data to be normally distributed, ANOVA revealed an overall effect across these noise-vocoded word forms, $F(3, 44) = 24.0$, $p < 0.001$. Tukey post-hoc comparisons revealed a significant difference between performance with NB5 and NB2 forms ($p < 0.001$), but no other condition effects.

Examining the performances of individual participants revealed that four human participants performed much as Panzee did. They showed the best performance with NB5 words, worst for NB2 forms, and virtually identical outcomes for NB4 and NB3 words. Panzee never recognized the words "celery," "noodles," and "raisin" in NB2 form, and 11 of 12 human participants completely failed with these items as well. Additionally, Panzee and the human participants demonstrated their best performance (67% to 100% correct) with 41.7% of the same words (10 of the 24) in NB5 form, 37.5% of the same words (9 of the 24) in NB4 form, and 41.7% of the same words (10 of the 24) in NB3 form – and most poorly (0% to 33% correct) with 33.3% of the same words (8 of 24) in NB2 form.

Discussion

As in earlier testing (Heimbauer et al. 2011), Panzee again reliably identified words in noise-vocoded form. However, her performance was significantly better for words in NB5 form than in corresponding NB2 versions, a pattern that was similar to human performance in comparable earlier studies (Shannon et al. 1995). As expected, increasing numbers of noise bands was associated with higher word-identification performance for Panzee with her performance on NB4 and NB5 forms similar to earlier testing with NB7 stimuli (Heimbauer et al. 2011). The results confirm that noise-vocoded speech based on as few as four noise bands is reliably comprehensible for this chimpanzee as is the case for human listeners in this and in previous experiments (Souza and Rosen 2009; Shannon et al. 1995). Panzee's similar performance with words in NB4 and NB3 may be an artifact of her small word set as discussed earlier. However, as hypothesized, Panzee's performance with noise-vocoded words does show evidence of successful word identification in spite of the absence of basic speech features, such as F0 and formant information. We propose that this language-trained chimpanzee was able to take advantage of whatever spectro-temporal cues remained. Results are also again indicative of top-down processing, with Panzee making use of her previous knowledge of speech acoustics in interpreting these fundamentally altered, synthetic versions.

General discussion

The current experiments investigated the possibility that Panzee uses the same information as humans to identify synthetic speech in sine-wave and noise-vocoded forms (Heimbauer et al. 2011). In sine-wave speech perception, humans perform significantly better when both SW1 and SW2 are present (Remez et al. 1981) – the components that are modeled on formants F1 and F2 in the natural speech signal. Panzee's human-like performance with sine-wave speech indicates that she may also be attending more to these particular tones in synthetic speech, with implications for sensitivity to the corresponding formants in natural speech. Similar results occurred with noise-vocoded speech, showing that Panzee performed best with stimuli that included four or five noise bands, and less well with three and two bands. These outcomes are comparable to earlier findings with noise-vocoded speech, showing differences in relative intelligibility of these synthesis forms by humans. Although it is difficult to specify exactly which acoustic cues are critical in these types of synthesized speech or what both have in common as discussed by Remez et al. (1994), it is clear that amplitude and frequency modulation over time is important, both in sine-wave and noise-vocoded speech (Remez et al. 1981; Shannon et al. 1995).

Additionally, Panzee's abilities provide evidence pertaining to top-down processing in speech perception. When humans listen to the speech, including in the difficult experimental situations we presented in our experiments, they take the knowledge they have of speech sounds into account to identify the speech. Bottom-up processing involves the incoming acoustic signal, but then top-down processing allows for the processing of the signal based on the individual's prior knowledge of the phonetic content as it relates to the word and its meaning. Each of the tasks presented to Panzee required top-down processing, which is, as noted above, considered to be the case in the context of human speech perception (Davis et al. 2005; Davis and Johnsrude 2007; Hillenbrand et al. 2011). An innate, speech-perception module would be an extreme form of top-down processing,

although this approach then downplays the role of experience with speech. Panzee's abilities argue against such a module and in favor of a strong role of experience. As noted earlier, Panzee's exposure to and experience hearing speech sounds began in infancy, as is the case with humans. Based on her performance in the current experiments, it is more likely that the critical factor in top-down processing is the vast amount of passive experience that human infants have hearing speech from birth, rather than a speech module. For instance, experience hearing speech allows infants to learn what speech sounds are being used, how differences among sounds may or may not be significant to categorizing them, and the meanings that sound combinations convey (Marcus et al. 1999; Saffran et al. 1996; Werker and Desjardins 1995). Whereas it is evident that humans can be much more successful in these tasks than Panzee was, it may be because their powerful language abilities are a result of human specializations for efficiency in processing language rather an innate ability to process the specific sounds and their acoustics.

As noted earlier, we tested Panzee using her familiar four-choice testing procedure, which did not present the same level of difficulty as the testing procedure used with the human participants because Panzee had a closed set of possible options from which to choose. We acknowledge that if we tested the humans with the closed-set method used with Panzee many participants would have reached ceiling performance due to their extensive experience with and use of the English language. This would have prevented us from measuring and comparing any differences in performance in response to the different types of sine-wave and noise-vocoded stimuli presented (i.e., different combinations of sine waves and different numbers of noise bands, respectively). Employing the open-set testing method with the human participants allowed us to make the necessary comparisons while controlling somewhat for any possible advantages that humans have due to their extensive language experience.

Despite these methodological differences, it is evident from Panzee's performance in these experiments that a language-trained chimpanzee can provide for interesting discussion regarding the uniqueness of human speech perception capabilities, in addition to informing about the likely speech perception and cognitive processing capabilities of an ape-human common ancestor. We do not want to ignore the impact that different methods and types of stimuli (i.e., identification and confidence ratings; and use of vowels, consonants, words, and sentences, respectively) may have played in performance, and so we cannot make strong conclusions directly comparing Panzee to humans. That said, the results of the current experiments demonstrate the possibility that Panzee may be drawing on the same processing strategies as humans to identify these words (or as in open-set testing), and she may be relying on the same cues as humans due to her rearing history as it relates to her experience with speech. We would expect that chimpanzees without this experience and the ability to understand spoken words would not demonstrate these same abilities and that early life experience thus played a crucial role in the emergence of this capability.

Acknowledgements We thank the care staff at Georgia State University for their care of Panzee throughout this experiment and the College of Arts and Sciences at Georgia State University for financial support for this project.

Funding Work was supported by the National Institute of Child Health and Human Development, National Science Foundation, GSU's RCALL and Brains & Behavior programs, the Center for Behavioral Neuroscience under the Science and Technology Centers Program of the National Science Foundation under agreement IBN-9876754, and ERC grant agreement AdG249516. LAH was funded by an RCALL Fellowship and a Duane M. Rumbaugh Fellowship.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no competing interests and no conflict of interest.

Availability of data and material The data are available in the Open Science Framework.

References

Beran MJ (2010) Use of exclusion by a chimpanzee (*Pan troglodytes*) during speech perception and auditory-visual matching-to-sample. Behav Process 83:287–291

Beran MJ, Heimbauer LA (2015) A longitudinal assessment of vocabulary retention in symbol-competent chimpanzees (*Pan troglodytes*). PLoS ONE 10:e0118408. https://doi.org/10.1371/journal.pone.0118408

Beran MJ, Washburn DA (2002) Chimpanzee responding during matching to sample: Control by exclusion. J Exp Anal Behav 78:497–508

Beran MJ, Pate JL, Washburn DA, Rumbaugh DM (2004) Sequential responding and planning in chimpanzees (*Pan troglodytes*) and rhesus macaques (*Macaca mulatta*). J Exp Psychol Anim Behav Process 30:203–212

Boersma P, Weenink D (2008) Praat: doing phonetics by computer [Computer program]. Version 5.1.11. http://www.praat.org/. Retrieved 1 Sept 2008

Brakke KE, Savage-Rumbaugh ES (1995a) The development of language skills in bonobo and chimpanzee–I. Comprehens Lang Commun 15:121–148

Brakke KE, Savage-Rumbaugh ES (1995b) The development of language skills Pan-II. Prod Lang Commun 16:361–380

Davis MH, Johnsrude IS (2007) Hearing speech sounds: Top-down influences on the interface between audition and speech perception. Hear Res 229:132–147

Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, McGettigan C (2005) Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. J Exp Psychol Gen 134:222–241

Dorman MF, Loizou PC, Spahr AJ, Maloff E (2002) A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants. J Speech Lang Hear R 45:783–788

Drullman R (2006) The significance of temporal modulation frequencies for speech intelligi-bility. In: Greenberg S, Ainsworth WA (eds) Listening to speech: An auditory perspective. Erlbaum, NJ, pp 43–67

Greenwood DD (1961) Critical bandwidth and the frequency coordinates of the basilar membrane. J Acoust Soc Am 33:1344–1356

Greenwood DD (1990) A cochlear frequency-position function for several species – 29 years later. J Acoust Soc Am 87:2592–2605

Heimbauer LA, Beran MJ, Owren MJ (2011) A chimpanzee recognizes synthetic speech with significantly reduced acoustic cues to phonetic content. Curr Biol 21:1210–1214

Heimbauer LA, Beran MJ, Owren MJ (2018) A chimpanzee's (Pan troglodytes) perception of variations in speech: Identification of familiar words when whispered and when spoken by a variety of talkers. Int J Psychol 31:1–16

Hillenbrand JM, Clark MJ, Baer CA (2011) Perception of sinewave vowels. J Acoust Soc Am 129:3991–4000

Kluender KR, Diehl RL, Killeen PR (1987) Japanese quail can learn phonetic categories. Science 237:1195–1197

Kuhl PK (1988) Auditory perception and the evolution of speech. Hum Evol 3:19–43

Kuhl PK, Miller JD (1975) Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science 190:69–72

Kuhl PK, Padden DM (1982) Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. Percept Psychophys 35:542–550

Kuhl PK, Padden DM (1983) Enhanced discriminability at the phonetic boundaries for the place feature in macaques. J Acoust Soc Am 73:1003–1010

Lenneberg EH (1967) Biological foundations of language. Wiley, NY

Lewis DE, Carrell TD (2007) The effect of amplitude modulation on intelligibility of time-varying sinusoidal speech in children and adults. Percept Psychophys 69:1140–1151

Liberman AM (1982) On finding that speech is special. Am Psychol 37:148–167

Mann VA, Liberman AM (1983) Some differences between phonetic and auditory modes of perception. Cognition 14:211–235

Marcus G, Vijayan S, Rao S, Vishton PM (1999) Rule learning by seven-month-old infants. Science 283:77–80

Newman RS (2006) Perceptual restoration in toddlers. Percept Psychophys 68:625–642

Nygaard LC, Pisoni DB (1998) Talker specific learning in speech perception. Percept Psychophys 60:355–376

Nygaard LC, Sommers MS, Pisoni DB (1994) Speech perception as a talker-contingent process. Psychol Sci 5:42–46

Owren MJ (2010) GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software. Behav Res Methods 40:822–829

Pisoni DB (1995). Some thought on "normalization" in speech perception. Research on Spoken Language Processing, Progress Report No. 20, Indiana University 3–29.

Remez RE (2005) Perceptual organization of speech. In: Pisoni DB, Remez RE (eds) The handbook of speech perception. Wiley-Blackwell, Oxford, pp 28–50

Remez RE, Rubin PE (1990) On the perception of speech from time-varying acoustic information: Contributions of amplitude variation. Percept Psychophy 48:313–325

Remez RE, Rubin PE, Pisoni DB, Carrell TD (1981) Speech perception without traditional speech cues. Science 212:947–949

Remez RE, Rubin PE, Berns SM, Lang PJS, JM, (1994) On the perceptional organization of speech. Psychol Rev 101:129–156

Remez RE, Dubowski KR, Broder RS, Davids ML, Grossman YS, Moskalenko M, Pardo JS, Hasbun SM (2011) Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. J Exp Psychol Human 37:968–977

Remez RE, Thomas EF, Dubowski KR, Koinis SM, Porter NAC, Paddu NU, Moskalenko M, Grossman YS (2013) Modulation sensitivity in the perceptual organization of speech. Atten Percept 75:1353–1358

Rosner BS, Talcott JB, Witton C, Hogg JD, Richardson AJ, Hansen PC, Stein JF (2003) The perception of "Sine-Wave Speech" by adults with developmental dyslexia. J Speech Lang Hear Res 46(1):68–79

Rumbaugh DM, Washburn DA (2003) Intelligence of apes and other rational beings. Yale University, CT

Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8 month-old infants. Science 274:1926–1928

Sawusch JR (2005) Acoustic analysis and synthesis of speech. In: Pisoni DB, Remez RE (eds) The handbook of speech perception. Blackwell, Oxford, pp 7–27

Shannon RV, Zeng F, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304

Sommers MS, Kirk KI, Pisoni DB (1997) Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I. The effects of response format. Ear Hear 18:89–99

Souza P, Rosen S (2009) Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. J Acoust Soc Am 126:792–805

Trout JD (2001) The biological basis of speech: What to infer from talking to the animals. Psychol Rev 108:523–549

Werker JF, Desjardins RN (1995) Listening to speech in the first year of life: Experiential influences on phoneme perception. Curr Dir Psychol Sci 4:76–81

Whalen DH, Liberman AM (1987) Speech perception takes precedence over nonspeech perception. Science 237:169–171

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.