INTERNATIONAL JOURNAL OF COMPARATIVE PSYCHOLOGY

2018, 31
David Washburn
Special Issue Editor
Peer-reviewed

A Chimpanzee's (*Pan troglodytes*) Perception of Variations in Speech: Identification of Familiar Words When Whispered and When Spoken by a Variety of Talkers

Lisa A. Heimbauer[1], Michael J. Beran[2] , and Michael J. Owren[3]

[1]The State University of New York at Delhi
[2]Georgia State University
[3]Georgia State University

When humans perceive speech, they process the acoustic properties of the sounds. The acoustics of a specific word can be different depending on who produces it and how they produce it. For example, a whispered word has different acoustic properties than a word spoken in a more natural manner; basically, the acoustics are "noisier." A word will also sound differently depending on who speaks it, due to the different physical and physiological characteristics of the talker. In this instance, humans routinely normalize speech to retrieve the lexical meaning by solving what is termed the "lack of invariance" problem. We investigated these speech perception phenomena in a language-trained chimpanzee (*Pan troglodytes*) named Panzee to ascertain if more generalized auditory capabilities, as opposed to specialized human cognitive processes, were adequate to accomplish these perceptual tasks. In Experiment 1, we compared the chimpanzee's performance when identifying words she was familiar with in natural versus whispered form. In Experiment 2, we investigated Panzee's ability to solve the "lack of invariance" problem when familiar words were spoken by a variety of talkers (familiar and unfamiliar male and female adults and children). The results of Experiment 1 demonstrated that there was no difference in her recognition for the two word types. The results of Experiment 2 revealed no significant difference in Panzee's performance across all talker types. Her overall performance suggests that more generalized capabilities are sufficient for solving for uncertainty when processing the acoustics of speech and instead favor a strong role of early experience.

Keywords: chimpanzee, speech perception, auditory perception, whispers, talker variation

Variety of Talkers

Speech perception is the ability to hear and recognize the acoustics of spoken language. It involves many levels of processing—from the auditory input to the comprehension of lexical meaning (Carroll, 2008). Listeners accomplish a broad array of auditory perceptual tasks both when learning and after having mastered language, many of which seem effortless. Many language skills are learned implicitly, without instruction, relatively passively, and with minimal conscious attention; and some of these skills emerge as the learner builds on previous knowledge and predictive relationships that help solve novel challenges (Gomez, 2002; Gomez & Gerken, 1999; Marcus, Vijayan, Bandi-Rao, & Vishton, 1999; Saffran, Aslin, & Newport, 1996). In the words of Rumbaugh (2002), demonstration of these types of behaviors would be considered *emergents* in the sense that they occur without explicit training and as a result of the accumulation of varied experiences over time. They are, without question, skills that reflect the communicative complexity and sophistication of our species.

Please send correspondence to Lisa HeimBauer, The State University of New York at Delhi. (Email: lisa.heimbauer@gmail.com)

Insights into the generality of the auditory and cognitive processes involved in speech perception for a nonhuman primate are fundamental to the discussion regarding the evolution of the associated capabilities in humans. Research with great apes has argued against the view termed "Speech is Special," whereby it is proposed that humans possess a specialized cognitive module for speech perception (Mann & Liberman, 1983) that is absent in all other species. That research instead supports the *Auditory Hypothesis* view that suggests spoken-language evolution took advantage of existing auditory-system capabilities that were present in ancestors of modern humans and potentially also present in species closely related to humans (Kuhl, 1988).

Specifically, research with language-trained apes (*Pan paniscus* and *Pan troglodytes*) has revealed that several individuals readily comprehend human speech (Beran & Heimbauer, 2015; Williams, Brakke, & Savage-Rumbaugh, 1997), with one individual in particular (a female chimpanzee named Panzee) doing so reliably throughout her lifetime, and even when speech was altered and missing traditional acoustic cues (Heimbauer, Beran, & Owren, 2011). Panzee was raised from infancy in a speech-rich environment, which provided her the opportunity to learn about language in the communicative context that humans experience – in this case, the English language (Brakke & Savage-Rumbaugh, 1995, 1996).

In light of Panzee's demonstrated speech comprehension abilities, several experiments were conducted to extend investigations of the perceptual processes underlying her perception of speech. More specifically, these experiments assessed her perceptual capabilities in comparison to some of the natural speech perception phenomena that occur in humans. The first experiment investigated whether Panzee could recognize speech when words were whispered, a form of speech that is acoustically very different from naturally spoken speech. The second experiment was designed to determine if she could solve what is termed the "lack of invariance" problem and normalize speech across different talkers. In particular, talkers of different sexes and ages may show differences in fundamental frequency (F0), formant frequencies, pitch, phoneme duration, and spectral and temporal variability (see Gerosa, Giuliani, & Brugnara, 2007; Gerosa, Lee, Giuliani, & Narayanan, 2006; Lee, Potamianos, & Narayanan, 1999; Pisoni, 1995).

Whispered speech. Acoustically, one way to describe parts of speech is as "voiced" and "unvoiced." Thomson, Boland, Wu, Epps, and Smithers (2003) described the voiced portions of speech as quasiperiodic energy and the unvoiced portions as noise-like energy, with voiced speech produced by vocal-fold vibrations, and unvoiced speech produced by turbulence in vocal-tract airflow. In the former case, the vocal folds vibrate in response to air pressure from the lungs, with filtering of this signal based on the resonances created by various articulatory positions of the jaw, tongue, and lips. These resonances, known as formants, emphasize the harmonics in this voiced energy and determine the phonetic quality of the sound. Similar effects occur in unvoiced speech, except that, in this case, the acoustical signal is noisy energy created by constricting the vocal tract at any of several points along its length to create turbulence and random, noisy energy. Formant effects also occur in unvoiced segments, which can reflect both the articulatory position associated with that particular sound and the resonances of adjacent, voiced sounds.

In normal speech, the majority of vowels are voiced, and some consonants are unvoiced; but in whispered speech, there are no vibrations of the vocal cords at all (Ito, Takeda, & Itakura, 2005). Because exhalation is the only source of excitation in whispered speech, the acoustic characteristics are different than those in normal speech (Fujimura & Lindqvist, 1971), specifically consisting of "noisier" energy (see Figure 1). A study on the acoustic analysis of vowel sounds (Konno, Toyama, Shimbo, & Murata, 1996) showed that the formant frequencies for a vowel in whispered speech shift to higher frequencies compared to normal speech. However, voiced consonants in whispered speech have lower energy at low frequencies up to 1.5 kHz

and their spectral flatness is greater compared to the normal speech (Ito et al., 2005), with the distinction between consonants and vowels reduced in whispered speech (Irino, Aoki, Kawahara, & Patterson, 2012). Whispered speech takes more effort to produce than normal speech (Schwartz, 1968), and is typically 15-20 decibels softer than the voiced components of speech (Traunmüller & Eriksson, 2000). Additionally, whispered speech is missing an important feature of everyday speech, namely, voice pitch (Irino et al., 2012). Due to these various differences, whispered speech signals can be more difficult to process and recognize than normal speech (Morris & Clements, 2002; Wenndt, Cupples, & Floyd, 2002).

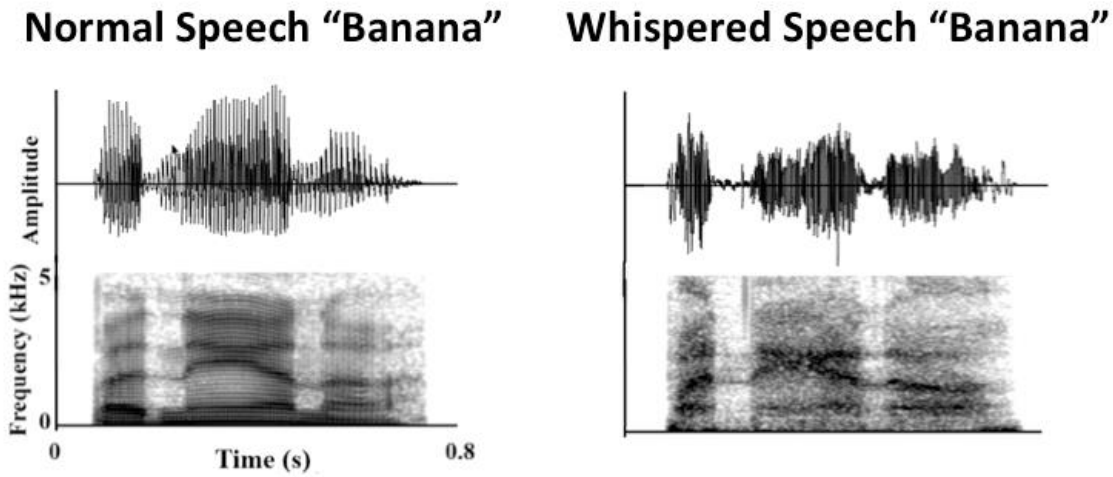

Figure 1. **Waveforms (top) and narrowband spectrograms (bottom) of the word banana in different acoustic forms: a) normal speech, and b) whispered speech.**

The lack-of-invariance problem. In addition to the questions regarding how specific acoustic elements contribute to lexical information, there are more general perceptual problems that arise during speech processing. The lack-of-invariance problem reflects the high variability in speech acoustics that results from the different physical and physiological properties of individual talkers (Pisoni, 1995). Because of these differences in talker acoustics, speech must be normalized in order for listeners to perceive the common lexical identities of individual words. Listeners routinely normalize speech from both familiar and unfamiliar talkers, despite differences such as age and sex classes and language backgrounds. Acoustically, this variation affects a variety of features, such as F0 range, formant frequencies, speaking rate, and acoustical patterning for a given phoneme (for a review see Benzeghiba et al., 2007).

Not only is there a difference between male and female voices, in that male voices have a lower F0, but children's speech is also very different from adult's speech. Children's speech is typically characterized by higher pitch and formant frequencies, especially for vowels (Gerosa et al., 2007; Lee et al., 1999). In addition, children under the age of seven typically have longer phoneme duration and larger spectral and temporal variability in consonants and consonant-vowel transitions than older children and adults (Gerosa et al., 2006). It is these characteristics that can often make children's speech more difficult to understand than adult speech.

Talker normalization emerges early in human ontogeny, as shown by the finding that infants are more sensitive to talker variation at seven and a half months of age than at ten and a half months (Houston & Jusczyk, 2000). However, the process by which talker normalization occurs is not well understood, and different models have been proposed (Creel & Tumlin, 2009). Some researchers believe that the process involves the listener stripping away individual talker information to extract phoneme content in abstract form. Others instead propose that generalizing across talkers is based on learning and implicitly remembering a large number of instances of speech sounds from many different individual talkers (e.g., Creel & Tumlin, 2009; Sumner, 2011). The latter view is supported by the fact that learning talker-specific characteristics can improve linguistic processing (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Pisoni, 1995).

Current Experiments

Because the chimpanzee Panzee had been listening to and responding to speech since she was born, we hypothesized that she should be able to process the variable acoustic properties of speech, such as those in whispered speech. Therefore, we designed an experiment to see if she would recognize English words she knew in normal form when they were whispered, a word form acoustically different from normally spoken speech (Ito et al., 2005). Additionally, we investigated Panzee's ability to recognize speech from a variety of talkers. Although there was indirect evidence that apes normalize speech across talkers – for example, Panzee (and several language-trained bonobos) interacted with at least a dozen humans on an everyday basis, and reacted appropriately to speech commands and requests (M. J. Beran, personal communication, January 2010) – we designed an experiment to test her talker-normalization capabilities more systematically, using recordings from a large number of individuals, both familiar and unfamiliar, and including male and female talkers of different ages.

Method

Subject

The subject was the female chimpanzee Panzee, who was 23 years old when the current experiments began. Panzee was taught to communicate using visuographic symbols called lexigrams (Rumbaugh, 1977), and she used these symbols and associated photographs to communicate in everyday situations (Beran, Savage-Rumbaugh, Brakke, Kelley, & Rumbaugh, 1998). Over a ten-year period of word comprehension testing, Panzee showed reliable recognition of approximately 130 spoken English words during computerized assessments of her vocabulary (Beran & Heimbauer, 2015).

Panzee was socially housed with three conspecifics at the Language Research Center at Georgia State University. She 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 could choose not to participate or to stop responding at any point during a session. Panzee used her language-like, lexigram-based communication system to request items throughout the day and often during experimental situations.

Materials

In addition to language-comprehension testing using lexigrams and photographs, Panzee also had experience with numerous computer-based protocols (see Rumbaugh & Washburn, 2003). In the two experiments in this study, she participated in three to four 20- to 30-min sessions per week, and worked for favored 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 either indoors or outdoors, with the option of moving between those areas at will.

Apparatus. Computer programs used to test Panzee were written in Visual Basic Version 6.0 (Microsoft Corp., Redmond, WA) and presented 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. Panzee registered her choices using a customized Gravis 42111 Gamepad Pro video-gaming joystick (Kensington Technology Group, San Francisco, CA).

Audio-recording of human speech was conducted with a Shure PG14/PG30-K7 head-worn wireless microphone system (Shure Inc., Niles, IL) and either a Realistic 32-12008 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 & Weenink, 2008), and custom-written scripts (Owren, 2010).

Stimuli. Spoken stimuli were chosen from the list of approximately 130 English words that Panzee had consistently identified in a decade of annual word-comprehension testing (Beran & Heimbauer, 2015). 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. Individual words were isolated by cropping corresponding segments at zero crossings, with 100 ms of silence 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.

Procedure

Panzee was tested using the procedure employed for annual word-comprehension testing (Beran & Heimbauer, 2015; Beran et al., 1998). 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 auditory 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 presentation of the same stimulus. After a 1-s delay, four different lexigrams (see Figure 2) appeared on the screen. One of these items was the correct match to the audio stimulus, and the others were lexigram foils chosen randomly by the controlling computer program from the word list used in each experiment. Visual items were positioned randomly in four of six possible locations on the computer screen—three on the left side of the screen and three on the right. Incorrect stimuli were 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 & Washburn, 2002).

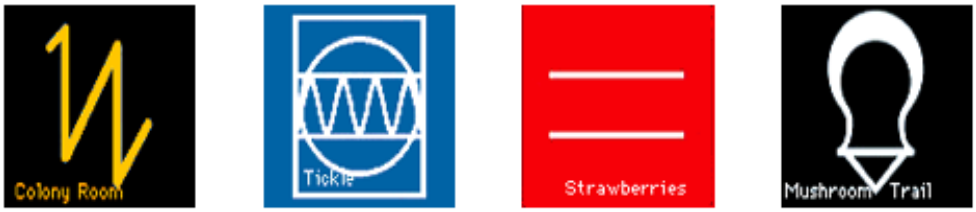

Figure 2. **Samples of lexigrams used in Panzee's spoken-word recognition task.**

Panzee's task was to use the joystick to move the cursor from the middle of the screen to the lexigram corresponding to the stimulus word. Panzee was rewarded with highly valued food, including pieces of cherries, grapes, blueberries, peaches, raspberries, strawberries, mixed fruit, or Chex Mix®. Panzee received auditory feedback on every trial (a melodic tone for correct answers and a buzz tone for incorrect answers) and was rewarded for all correct responses. This reward schedule kept her highly motivated and could be used because each trial was unique.

Experiment 1

This experiment investigated Panzee's ability to recognize whispered speech, despite the acoustic variations that exist in whispered speech as compared to natural speech (see Ito et al., 2005). Although caregivers and researchers did not typically whisper to Panzee, we could not say if she had any experience identifying speech in this form in normal daily interactions. We designed this experiment to ascertain if she would be able to reconcile the noisier acoustics of whispered words to her experience with the same words in natural speech.

Method

Stimuli. Test stimuli consisted of 1 five-syllable, 2 four-syllable, and 21 three-syllable words (see Table 1). The 24 words were chosen from the list of 130 words that Panzee routinely performed best on when identifying them in English. These words were recorded when spoken by a male researcher (MJB) in both a natural manner and when whispered.

Panzee was tested over three sessions, each of which included 96 randomized trials (each word in a session was presented four times). In each of the three sessions, a different group of 8 of the 24 words (Groups A, B, and C; see Table 1) were heard in whispered form with the remaining 16 words spoken naturally.

Table 1

Test Word Groups. Test Words used in Experiments 1 (Groups A, B, and C) and 2 (Groups D and E)

Words	Experiment 1	Experiment 2
Apple		E
Apricot	A	D
Balloon		E
Banana	C	D
Blueberries	B	D
Bubbles		E
Carrot		D
Celery	B	E
Cereal	A	E
Clover		E
Coffee		D
ColonyRoom	C	D
Gorilla	A	
Honeysuckle	B	D
Hotdog		E
Lemonade	A	
M&M	B	
Melon		D
MushroomTrail	C	E
Noodles		D
ObservationRoom	A	
OrangeDrink	B	E
OrangeJuice	C	D
Peaches		E
Pineapple	B	D
Pineneedle	A	E
PlasticBag	B	
Popsicle	C	
Potato	C	D
Strawberries	A	E
SweetPotato	C	E
Tickle		D
Tomato	A	E
Toothpaste		D
TV		E
Vitamins	C	D
Water		E
Yogurt		D

Data analysis. One-sample *t*-tests were conducted to assess if performance on the two types of spoken words was above the chance-rate of 25%, in addition to ascertaining whether Panzee performed above chance-rate performance for all 24 words as a set when presented for the first time in whispered form. A two-tailed, paired *t*-test was conducted to analyze results for a difference in performance between the whispered and natural words.

Results

As shown in Figure 3, the percentage of correct lexigram choices for the 24 whispered words was 82.3%, and performance with these words was statistically better than chance, $t(23) = 13.92$, $p < 0.01$. The percentage of correct choices for the 24 words in natural form was 83.6%, also statistically above chance, $t(23) = 25.82$, $p < 0.01$. There was no significant difference on performance between the two types of spoken words, $t(23) = 0.30$, $p = 0.76$), nor was there a correlation in performance between specific words in the two forms, $r(22) = 0.14$, $p = 0.52$ (two-tailed). In addition, the first time Panzee heard each of the words in the whispered version, a one-sample *t*-test showed her performance was statistically above chance, $t(23) = 6.40$, $p < 0.01$, as she correctly identified 19 of the 24 words (79%).

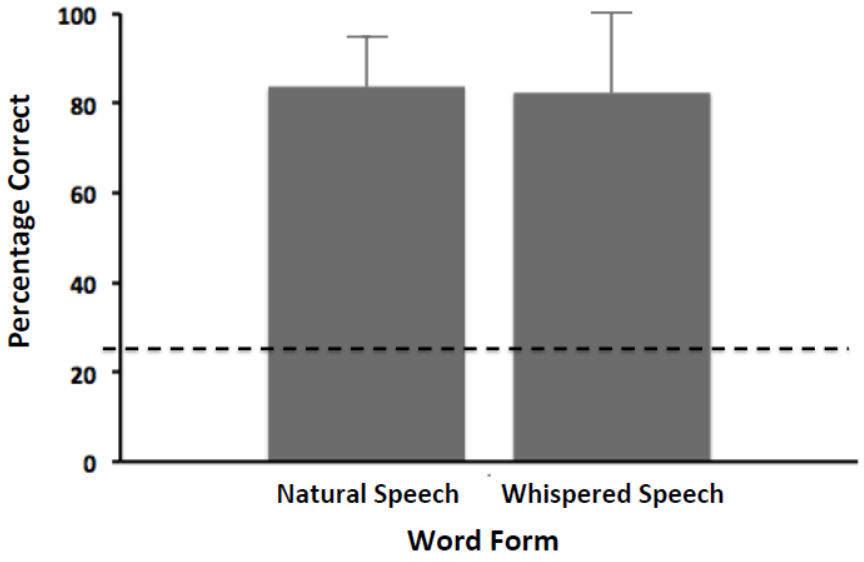

Figure 3. **Percentage of correct choices for the 24 words when in natural and whispered form.** The error bars represent standard deviations, and the dashed line signifies chance-rate performance of 25%.

Discussion

Panzee's performance was similar regardless of whether words were in natural or whispered form. She also performed above chance when hearing the whispered words for the first time, at a rate comparable to

testing with natural words. These results confirm that she could understand whispered speech despite its noisier form caused by the acoustic differences from the natural words with which she had experience.

Because Panzee rarely heard whispered speech, it is likely that she accomplished the task by using underlying perceptual top-down processing capabilities, whereby a listener takes advantage of previously learned acoustic and phonetic information (Davis & Johnsrude, 2007; Mann & Liberman, 1983; Newman, 2006; Whalen & Liberman, 1987). Top-down processing is likely critical in normative speech perception as well as in difficult listening situations, with processing of synthetic words and sentences becoming useful in understanding how acoustic input can contribute to recognition of speech at various levels of organization (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Hillenbrand, Clark, & Baer, 2011; Remez et al., 2009; Saberi & Perrott, 1999).

Experiment 2

This experiment investigated Panzee's ability to understand speech from a variety of talkers. This capability is important because, as mentioned earlier, the acoustic characteristics of speech can vary significantly among both individual talkers and classes of talkers (Pisoni, 1995; Remez, 2005), causing the lack-of-invariance problem. As Panzee routinely interacted with and responded to different humans who spoke to her, we anticipated that she would correctly perceive familiar words regardless of the talker, demonstrating normalization across talkers. Earlier speech perception work with Panzee (Beran & Heimbauer, 2015; Heimbauer et al., 2011) involved the speech of only one talker, a familiar male researcher (MJB). Hence, there had been no specific information regarding her potential talker-normalization abilities. It could be that Panzee understood other caregivers and researchers because she simply had become accustomed to the speech of these particular known individuals over time without being able to normalize to novel talkers. Therefore, to investigate this hypothesis, Panzee was tested for recognition abilities with the speech of a variety of male and female familiar adults, unfamiliar adults, and unfamiliar children. Panzee had been exposed to children's voices much less frequently, and only very rarely later in her life, whether in experimental or in informal interactions.

Method

Audio recording and stimuli. Test stimuli consisted of 15 two-syllable, 14 three-syllable, and 3 four-syllable words (see Table 1). All talkers were recorded speaking 48 words, but only 32 words were used in the experiment. Some of the words were difficult for the children to pronounce, and they did not always speak clearly. The 32 words were chosen on the basis of finding the best recordings from each talker. Stimuli included speech from a diverse set of talkers, both familiar and unfamiliar people to Panzee, including variation in biological sex, age, and dialect background. In total, there were 31 different native-English speakers including 21 adults and 10 children. These talkers included a familiar male researcher (MJB), five additional familiar adult males (FAM), five unfamiliar adult males (UAM), five familiar adult females (FAF), five unfamiliar adult females (UAF), five unfamiliar male children (UCM), and five unfamiliar female children (UCF). The age range of adult talkers was 20 to 72 years old, and the age range for children was 4 to 7 years old.

MJB and the 20 additional adults were recorded as they read the individual words from index cards. The 10 children were recorded as they named photographs appearing individually in a Microsoft PowerPoint

presentation. If a child could not name a photograph, they were told the word explicitly. Talkers were from a wide range of areas within the United States, with a variety of regional dialect backgrounds. MJB was born in Ohio, and had also lived in Alabama and Georgia. The other 10 familiar talkers were born in six different states and in Germany and had lived in a total of 14 other states and Washington, DC. The northern-most of these states was Michigan, the southern-most was Florida, the eastern-most was New York, and the western-most was California and Oregon. Some talkers had also lived in Germany, Japan, Nepal, Switzerland, and Taiwan. The 10 unfamiliar talkers were born in six different states and in Puerto Rico and had lived in a total of nine other states. The northern-most of these was New York, the southern-most was Louisiana, the eastern-most was New Jersey, and the western-most was California and Hawaii. One of these talkers had also lived in Germany. All of the children had been born and raised exclusively in either Georgia or New York.

Design and procedure. Panzee was tested for a total of 14 sessions, each of which included 80 randomized trials. In the first session, she heard 16 test words (Group D; see Table 1) five times each, spoken by MJB. In the next six sessions, she heard Group D words spoken by the five talkers within each of the specific talker-type groups. In these sessions, Panzee heard the 16 words spoken once by each for the talkers. The session order of talker types for testing was FAM, UAM, FAF, UAF, UCM, and then UCF. In the eighth session, Panzee heard the remaining 16 test words (Group E; see Table 1) five times each, spoken by MJB, which was followed by six sessions with Group E words—one session for each of the six talker-type groups. Testing order differed relative to the earlier sessions and was UAM, FAF, UCF, FAM, UCM, and then UAF.

Data analysis. Panzee's percentage-correct performance was computed for each of the seven different talker types, and one-tailed binomial tests were conducted to compare performance to a chance rate of 25% for each talker type separately. A Kruskal-Wallis test was used to test for an overall effect among word forms. Performance with these talkers was compared to recognition of speech from the familiar male researcher (MJB), who is arguably one of the talkers with which Panzee was most familiar, and across all talkers.

Results

As shown in Figure 4, Panzee's mean performance was calculated for each talker and averaged for the two sessions for each different talker type, ranging from 75.6% (MJB, UCF) to 81.3% (FAM). Word recognition was significantly above chance level for all talker types, $p < 0.01$. A Kolmogorov-Smirnov test showed that the data were not normally distributed, and a Kruskal-Wallis test was conducted using combinations of talker types. Because performances for all talker-types were similar to previous annual test performance levels (Beran & Heimbauer, 2015), some talker-types were combined to analyze performance comparisons. These analysis categories were "All Familiar Adults" (FAM and FAF), "All Unfamiliar Adults" (UAM and UAF), "All Adult Males" (FAM and UAM), "All Adult Females" (FAF and UAF), and "All Children" (UCM and UCF). The rationale for combining data from boys and girls was that, prior to puberty, vocal tracts and vocal folds of girls and boys are very similar (Simpson, 2009). These analyses revealed no overall difference in performance among the collapsed categories, $X^2(4) = 0.58$, $p = 0.96$.

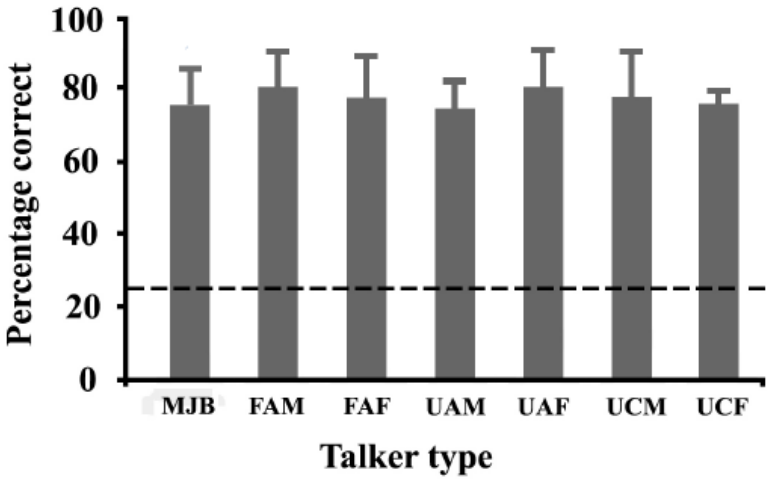

Figure 4. **Experiment 2 chimpanzee word-recognition performance across talkers.** Talker-type groups were as follows: MJB is the familiar male researcher; FAM, other familiar adult males; FAF, familiar adult females; UAM, unfamiliar adult males; UAF, unfamiliar adult females; UCM, unfamiliar male children; and, UCF, unfamiliar female children. The error bars represent the standard deviations, and the dashed line signifies chance-rate performance of 25%.

Discussion

Panzee readily recognized words not only from a variety of familiar adult males and females, but also from unfamiliar adults and children of both sexes, demonstrating her ability to solve the lack-of-invariance problem. Similar to humans, she was able to accommodate the acoustic variability found in speech from different talkers, and she appeared to do so rather effortlessly.

Because speech acoustics are highly variable over a variety of characteristics, listeners have to solve the lack-of-invariance problem almost every time they hear spoken language. Although the process by which this occurs remains unclear (Creel & Tumlin, 2009), speech experience likely plays a major role for top-down processing. Human infants demonstrate at least some talker normalization ability by the age of ten and a half months (Houston & Jusczyk, 2000), and human adults, having a vast amount of speech experience, routinely normalize across a wide range of talkers (Benzeghiba et al., 2007). Although some nonhumans have shown an ability to discriminate and categorize speech sounds (e.g., Dooling & Brown, 1990; Kuhl & Miller, 1975; Loebach and Wickesberg, 2006; Ohms, Gill, Van Heijningen, Becker, & ten Cate, 2010), talker normalization had not previously been systematically investigated in nonhumans.

In the current experiment, Panzee was tested thoroughly, demonstrating the ability to normalize across a range of talkers producing her familiar words. Not surprisingly, Panzee recognized these words when spoken by the familiar researcher, MJB, at a rate similar to that found in many previous tests with his voice. However, she also showed essentially the same performance when hearing the voices of familiar males and females, unfamiliar males and females, and unfamiliar boys and girls. In addition to age- and sex-related variation, these

talkers came from a variety of regional dialect backgrounds. The adults had lived in a total of 26 states, Washington, DC, Puerto Rico, and five other countries. Although the children were only from the northern state of New York and the southern state of Georgia, their voices were the least familiar to Panzee, as well as being very different from the adult voices (Gerosa et al., 2007; Lee et al., 1999).

As stated earlier, one possible interpretation of Panzee's performance with familiar talkers is that she had previously learned the features of each person's voice individually, thereby knowing from experience what each of these words sounded like when spoken by these particular individuals. However, that explanation cannot account for her ability to recognize the unfamiliar adults or children. A more likely explanation is that, as hypothesized, Panzee was showing human-like, talker-normalization abilities.

General Discussion

Panzee's success in identifying whispered speech in Experiment 1 revealed that she was able to identify speech even when presented in a noisier form than that of normal speech. In fact, the results of Experiment 1 were the impetus for a later experiment which demonstrated that she was able to recognize speech in synthetic forms missing traditional acoustic cues to phonetic content (i.e., noise-vocoded and sine-wave speech; Heimbauer et al., 2011).

In Experiment 2, Panzee showed that she could easily identify speech in other variable forms as words spoken by a variety of talkers including both familiar and unfamiliar individuals, adults and children, and in a variety of dialects. These results provided evidence that she was able to solve the lack-of-invariance problem created by this variability across talkers. Despite the high acoustic variability of human voices (Evans & Iverson, 2004; Hillenbrand, Getty, Clark, & Wheeler, 1995; Pisoni, 1995; Remez, 2005), Panzee was able to recognize words from all talker-types equally well, including the novel conditions of children's voices. Panzee's performance was similar to her historic levels in every case, with no significant performance difference between any of the talker groups.

Panzee's success in these experiments is most likely due to underlying perceptual processing abilities; in this case, due to particularly well-developed top-down processing—the general cognitive strategy of using pre-existing knowledge to interpret inherently ambiguous sensory input (David & Johnsrude, 2007; Plomp, 2001). In the case of whispered speech, despite the fact that it is produced without vocal cord vibrations (Ito et al., 2005) changing the vocal characteristics of both vowels and consonants, Panzee showed no difference in performance with whispered versus natural words, even when considering only first trials of the 24 words. Hence, Panzee was again demonstrating top-down interpretation of impoverished speech input (Heimbauer et al., 2011) in the same way that humans have been shown to do with altered and synthetic speech (Davis et al., 2005; Davis & Johnsrude, 2007; Hillenbrand et al., 2011). These results provide further evidence for the Auditory Hypothesis and the view that early hominins took advantage of pre-existing auditory capabilities when spoken language evolved.

The results of the current experiments also favor a strong role of experience when solving speech perception uncertainty caused by variations in speech. Based on Panzee's performance, it is likely that a critical factor in human top-down processing is the vast amount of passive experience that human infants have hearing speech from birth on, rather than human, species-specific, speech-processing capabilities. 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 & Desjardins, 1995). At present, there are only a few remaining apes that have had this type of immersive early experience with human language in auditory form. This is unfortunate, because Panzee's successes here and in previous research suggest that apes are excellent participants in studies examining the role of early experience on language acquisition, speech perception, and higher order cognition.

Panzee's speech-perception abilities also represent an example of emergents (Rumbaugh, 2002; Rumbaugh, King, Beran, Washburn, & Gould, 2007), defined for both humans and animals as important components of learning and cognition as new behaviors with antecedents in previously gained knowledge or experience (Rumbaugh & Washburn, 2003). They differ from behaviors learned through operant or classical conditioning, are considered common in some species of nonhuman animals, and are argued to provide the potential basis for new and innovative actions. Emergents may be necessary for adaptive and behaviorally flexible species to meet new challenges in complex environments and can be expressed in different situations at the first opportunity. The speech-processing abilities that Panzee has demonstrated in the current experiments are emergents in this sense, with her extensive experience with natural speech providing the basis for solving a variety of perceptual problems—including new ones.

Acknowledgments

We thank David Washburn for inviting us to submit this paper as part of the memorial issue for Duane M. Rumbaugh. Professor Rumbaugh was a mentor to two of us (Michael J. Beran and Lisa A. Heimbauer), and we both had the great honor to serve as the Duane M. Rumbaugh Fellow at Georgia State University early in our careers. Duane was a wonderful mentor and an enthusiastic supporter of the research we each have done in our careers. He will be greatly missed, but his legacy will continue. We also thank Mike Hart, John Kelley, Sarah Hunsberger, and Dan Rice for expert animal care.

References

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., … Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*, 763–786.

Beran, M. J., & Heimbauer, L. A. (2015). A longitudinal assessment of vocabulary retention in symbol-competent chimpanzees (*Pan troglodytes*). *PLOS ONE.* doi:10.1371/journal.pone.0118408

Beran, M. J., & Washburn, D. A. (2002). Chimpanzee responding during matching to sample: Control by exclusion. *Journal of Experimental Analysis of Behavior, 78*, 497–508.

Beran, M. J., Savage-Rumbaugh, E. S., Brakke, K. E., Kelley, J. W., & Rumbaugh D. M. (1998). Symbol comprehension and learning: A "vocabulary" test of three chimpanzees. *Evolution Communication, 2*, 171–188.

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

Brakke, K. E., & Savage-Rumbaugh, E. S. (1995). The development of language skills in bonobo and chimpanzee—I. Comprehension. *Language & Communication, 15*, 121–148.

Brakke, K. E., & Savage-Rumbaugh, E. S. (1996). The development of language skills in Pan—II. Production. *Language & Communication, 16*, 361–380.

Carroll, D. W. (2008). *Psychology of language* (5th ed.). Belmont, CA: Wadsworth.

Creel, S. C., & Tumlin, M. A. (2009). Talker information is not normalized in fluent speech: Evidence from on-line processing of spoken words. *31st Annual Meeting of the Cognitive Science Society*, Amsterdam, 845–850.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research, 229*, 132–147

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*, 222–241.

Dooling, R. J., & Brown, S. D. (1990). Speech perception by budgerigars (*Melopsittacus undulatus*): Spoken vowels. *Perception and Psychophysics, 47*, 568–574.

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *Journal of the Acoustical Society of America, 115*, 3523–3561.

Fujimura, O., & Lindqvist, J. (1971). Sweep-tone measurements of vocal-tract characteristics. *Journal of the Acoustical Society of America, 49*, 541–558.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431–436.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135.

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication, 49*, 847–860.

Gerosa, M., Lee, S., Giuliani, D., & Narayanan, S. (2006) Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP06), France, 1*, 393–396.

Heimbauer, L. A., Beran, M. J., & Owren, M. J. (2011). A chimpanzee recognizes synthetic speech with significantly reduced acoustic cues to phonetic content. *Current Biology, 21*, 1210–1214.

Hillenbrand, J. M., Clark, M. J., & Baer, C. A. (2011) Perception of sinewave vowels. *Journal of the Acoustical Society of America, 129*, 3991–4000.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099–3111.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1570–1582.

Irino, T., Aoki, Y., Kawahara, H., & Patterson, R. D. (2012). Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency discrimination. *Speech Communication, 54*, 998–1013.

Ito, T., Takeda, K., & Itakura, F. (2005). Analysis and recognition of whispered speech. *Speech Communication, 45*, 139–152.

Konno, H., Toyama, J., Shimbo, M., & Murata, K. (1996). The effect of formant frequency and spectral tilt of unvoiced vowels on their perceived pitch and phonemic quality. *IEICE Technical Report, SP95-140, March*, pp. 39–45.

Kuhl, P. K. (1988). Auditory perception and the evolution of speech. *Human Evolution, 3*, 19–43.

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190*, 69–72.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America, 105*, 1455–1468.

Loebach, J. L., & Wickesberg, R. E. (2006). The representation of noise vocoded speech in the auditory nerve of the chinchilla: Physiological correlates of the perception of spectrally reduced speech. *Hearing Research, 213*, 130–144.

Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition, 14*, 211–235.

Marcus, G., Vijayan, S., Bandi-Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283*, 77–80.

Morris, R., & Clements, M. (2002). Estimation of speech spectra from whispers. In: *Proceedings of International Conference on Acoustic Speech and Signal Processing (ICASSP), 7–11 May 2002*, Orlando, p. IV-4159

Newman, R. S. (2006). Perceptual restoration in toddlers. *Perception and Psychophysics, 68*, 625–642.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60*, 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.
Ohms, V. R., Gill, A., Van Heijningen, C. A. A., Becker, G. J. L., & ten Cate, C. (2010). Zebra finches exhibit speaker-independent phonetic perception of human speech. *Philosophical Transactions of the Royal Society B: Biological Sciences, 277*, 1003–1009.
Owren, M. J. (2010). GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software. *Behavior Research Methods, 40*, 822–829.
Pisoni, D. B. (1995). Some thoughts on "normalization" in speech perception. *Research on Spoken Language Processing, Progress Report No. 20*, Indiana University, 3–29.
Plomp, R. (2001). *The intelligent ear*. Mahwah, NJ: Erlbaum Associates.
Remez, R. E. (2005). The perceptual organization of speech. In D. B. Pisoni and R. E. Remez (Eds.), *Handbook of speech perception* (pp. 9-26). Oxford, England: Wiley-Blackwell.
Remez, R. E., Dubowski, K. R., Broder, R. S., Davids, M. L., Grossman, Y. S., Moskalenko, M., …Hasbun, S. M. (2009). Auditory-phonetic projection and lexical structure in the recognition of sine-wave words. *Journal of Experimental Psychology: Human Perception and Performance, 125*, 2656.
Rumbaugh, D. M. (1977). *Language learning by a chimpanzee: The LANA Project*. New York: Academic Press.
Rumbaugh, D.M. (2002). Emergents and rational behavior. *Eye on Psi Chi, 6*, 8–14.
Rumbaugh, D. M., King, J. E., Beran, M. J., Washburn, D. A., & Gould, K. L. (2007). A salience theory of learning and behavior: With perspectives on neurobiology and cognition. *International Journal of Primatology, 28*, 973–996.
Rumbaugh, D. M., & Washburn, D. A. (2003). *Intelligence of apes and other rational beings*. New Haven, CT: Yale University.
Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature, 398*, 760.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8 month-old infants. *Science, 274*, 1926–1928.
Schwartz, M. F. (1968). Air consumption, per syllable, in oral and whispered speech. *Journal of the Acoustical Society of America, 43*, 1448–1449.
Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and linguistics compass, 3*, 621–640.
Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition, 119*, 131–136.
Thomson, M., Boland, S., Wu, M., Epps, J., & Smithers, M. (2003). Decomposition of speech into voiced and unvoiced components based on a state-space signal model. *Acoustics, Speech, and Signal Processing Proceedings, 3*, I160–I163.
Traunmüller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America, 107*, 3438–3451.
Wenndt, S. J., Cupples, E. J., & Floyd, R. M. (2002). A study on the classification of whispered and normally phonated speech. In: *Proceedings of International Conference on Spoken Language Processing (ICSLP), 16–20 September 2002*, Denver, pp. 649–652.
Werker, J. F., & Desjardins, R. N. (1995). Listening to speech in the first year of life: Experiential influences on phoneme perception. *Current Directions in Psychological Science, 4*, 76-81.
Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science, 237*, 169–171.
Williams, S. L., Brakke, K., & Savage-Rumbaugh, E. S. (1997). Comprehension skills of language-competent and nonlanguage competent chimpanzees. *Language & Communication, 17*, 301–317.

Financial conflict of interest: This research was supported by the National Institute of Child Health and Human Development (HD-060563), 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 AdG 249516. Lisa A. Heimbauer also was supported by an RCALL Fellowship.
Conflict of interest: No stated conflicts.

Submitted: June 20th, 2018
Resubmitted: August 14th, 2018
Accepted: September 21st, 2018