8-1-2012

Investigating Speech Perception in Evolutionary Perspective: Comparisons of Chimpanzee (Pan troglodytes) and Human Capabilities

Lisa A. Heimbauer

Georgia State University

Follow this and additional works at: http://scholarworks.gsu.edu/psych_diss

Recommended Citation
http://scholarworks.gsu.edu/psych_diss/106

This Dissertation is brought to you for free and open access by the Department of Psychology at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Psychology Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
INVESTIGATING SPEECH PERCEPTION IN EVOLUTIONARY PERSPECTIVE:

COMPARISONS OF CHIMPANZEE (PAN TROGLODYTES) AND HUMAN CAPABILITIES

by

LISA A. HEIMBAUER

Under the Direction of Michael J. Owren

ABSTRACT

There has been much discussion regarding whether the capability to perceive speech is uniquely human. The “Speech is Special” (SiS) view proposes that humans possess a specialized cognitive module for speech perception (Mann & Liberman, 1983). In contrast, the “Auditory Hypothesis” (Kuhl, 1988) suggests spoken-language evolution took advantage of existing auditory-system capabilities. In support of the Auditory Hypothesis, there is evidence that Panzee, a language-trained chimpanzee (Pan troglodytes), perceives speech in synthetic “sine-wave” and “noise-vocoded” forms (Heimbauer, Beran, & Owren, 2011). Human comprehension of these altered forms of speech has been cited as evidence for specialized cognitive capabilities (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGgettigan, 2005).
In light of Panzee’s demonstrated abilities, three experiments extended these investigations of the cognitive processes underlying her speech perception. The first experiment investigated the acoustic cues that Panzee and humans use when identifying sine-wave and noise-vocoded speech. The second experiment examined Panzee’s ability to perceive “time-reversed” speech, in which individual segments of the waveform are reversed in time. Humans are able to perceive such speech if these segments do not much exceed average phoneme length. Finally, the third experiment tested Panzee’s ability to generalize across both familiar and novel talkers, a perceptually challenging task that humans seem to perform effortlessly.

Panzee’s performance was similar to that of humans in all experiments. In Experiment 1, results demonstrated that Panzee likely attends to the same “spectro-temporal” cues in sine-wave and noise-vocoded speech that humans are sensitive to. In Experiment 2, Panzee showed a similar intelligibility pattern as a function of reversal-window length as found in human listeners. In Experiment 3, 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. Overall, results suggest that a combination of general auditory processing and sufficient exposure to meaningful spoken language is sufficient to account for speech-perception evidence previously proposed to require specialized, uniquely human mechanisms. These findings in turn suggest that speech-perception capabilities were already present in latent form in the common evolutionary ancestors of modern chimpanzees and humans.

INDEX WORDS: Speech perception, Evolution, Language-trained chimpanzee, Synthetic speech
INVESTIGATING SPEECH PERCEPTION IN EVOLUTIONARY PERSPECTIVE:

COMPARISONS OF CHIMPANZEE (PAN TROGLODYTES) AND HUMAN CAPABILITIES

by

LISA A. HEIMBAUER

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2012
INVESTIGATING SPEECH PERCEPTION IN EVOLUTIONARY PERSPECTIVE:

COMPARISONS OF CHIMPANZEE (PAN TROGLODYTES) AND HUMAN CAPABILITIES

by

LISA A. HEIMBAUER

Committee Chair: Michael J. Owren

Committee: Rose A. Sevcik

Gwen Frishkoff

Michael J. Beran

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2012
DEDICATION

I dedicate this dissertation to my family, especially my loving husband Gary who has supported me unconditionally and unselfishly since I decided to begin my education. I am also grateful to my parents, Albert and Marie, for instilling in me a love of education, and my brother Bill and daughter-in-law Tara who were always there when I needed someone to listen. To my children, Randy, Stephanie, Melanie, and Gary, and my grandchildren, Holly, Samantha, Thomas, John, and Leah, I especially hope my journey inspires you to always follow your dreams—wherever they lead you.
ACKNOWLEDGMENTS

I am truly indebted and grateful to my advisor, Michael Owren, for the support and guidance he provided throughout the writing of my dissertation. Additionally, his rigorous scholarship and theoretical beliefs on the evolution of communication have been inspiring and have helped to shape my program of research during my graduate school education. These pages also reflect the relationships with other generous and inspiring individuals who have influenced me, served on my committees, and provided support and friendship during my time at Georgia State University and the Language Research Center, especially Mike Beran.

In addition, I am obliged to others who have supported me by contributing to my education, especially dissertation committee members Rose Sevcik and Gwen Frishkoff, and faculty members Sarah Brosnan and Seyda Özçaliskan. I would also like to thank the members of the Research on the Challenges of Acquiring Language and Literacy initiative and the Duane M. Rumbaugh Fellowship committee for the generous fellowship support. Finally, I am grateful to the graduate students who have provided intellectual and moral support, and more importantly, friendship.
TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................... v

LIST OF TABLES ..................................................................................................................... x

LIST OF FIGURES .................................................................................................................... xi

1. INTRODUCTION .................................................................................................................. 1
   1.1 Human speech ................................................................................................................... 3
      1.1.1 *Natural speech-perception phenomena* ................................................................. 6
      1.1.2 *Synthetic speech perception* ................................................................................... 7
   1.2 The SiS argument ............................................................................................................. 11
   1.3 The Auditory Hypothesis ............................................................................................... 13
      1.3.1 *Mammalian hearing* ............................................................................................... 14
      1.3.2 *Speech perception in non-primates* ....................................................................... 17
      1.3.3 *Speech perception in nonhuman primates* ............................................................ 19
      1.3.4 *Evidence of top-down processing when perceiving speech* ............................ 22
   1.4 Speech perception in a language-trained chimpanzee ..................................................... 24
      1.4.1 *Previous experiments with Panzee* ....................................................................... 24

2. CURRENT EXPERIMENTS ................................................................................................... 28
   2.1 General Methods .............................................................................................................. 30
      2.1.1 *Subject* ................................................................................................................... 30
      2.1.2 *Participants* .......................................................................................................... 30
      2.1.3 *Apparatus* ........................................................................................................... 31
      2.1.4 *Stimuli* ................................................................................................................ 31
2.1.5 Chimpanzee procedure ................................................................. 32
2.1.6 Human procedures .................................................................... 33
2.1.7 Data analysis ............................................................................. 34

3. EXPERIMENT 1 .................................................................................. 34

3.1 Experiment 1a ................................................................................. 35

3.1.1 Subject ........................................................................................ 36
3.1.2 Participants .................................................................................. 36
3.1.3 Stimuli .......................................................................................... 36
3.1.4 Chimpanzee procedure ................................................................. 38
3.1.5 Human procedure ....................................................................... 39
3.1.6 Data analysis ................................................................................. 40
3.1.7 Results ........................................................................................ 41
3.1.8 Discussion .................................................................................... 43

3.2 Experiment 1b ................................................................................. 45

3.2.1 Subject ........................................................................................ 46
3.2.2 Participants .................................................................................. 46
3.2.3 Stimuli .......................................................................................... 46
3.2.4 Chimpanzee procedure ................................................................. 48
3.2.5 Human procedure ....................................................................... 49
3.2.6 Data analysis ................................................................................. 49
3.2.7 Results ........................................................................................ 49
3.2.8 Discussion .................................................................................... 51
3.3 General Discussion ........................................................................................................ 52

4. EXPERIMENT 2 .............................................................................................................. 53
   4.1 Subject ......................................................................................................................... 54
   4.2 Participants .................................................................................................................. 54
   4.3 Stimuli ......................................................................................................................... 55
   4.4 Chimpanzee procedure ............................................................................................. 55
   4.5 Human procedure ...................................................................................................... 56
   4.6 Data analysis ............................................................................................................... 56
   4.7 Results ......................................................................................................................... 58
   4.8 Discussion ................................................................................................................... 60

5. EXPERIMENT 3 ............................................................................................................... 62
   5.1 Subject ......................................................................................................................... 65
   5.2 Participants .................................................................................................................. 65
   5.3 Audio-recording and stimuli ...................................................................................... 65
   5.4 Chimpanzee procedure ............................................................................................. 67
   5.5 Data analysis ............................................................................................................... 67
   5.6 Results ......................................................................................................................... 68
   5.7 Discussion ................................................................................................................... 69

6. GENERAL DISCUSSION ................................................................................................. 70
   6.1 Current results ............................................................................................................. 71
   6.2 Implications of experimental results ......................................................................... 72
       6.2.1 The SiS view versus the Auditory Hypothesis ....................................................... 72
6.2.2  Top-down processing and speech perception experience ...................... 74

6.3  Cognitive processing and language ................................................................................................................................. 75

6.4  Future directions ................................................................................................................................................................. 78

REFERENCES ............................................................................................................................................................................. 80
LIST OF TABLES

Table 1. Orientation and test word groups................................................................. 37

Table 2. Lower-to-upper noise-band cutoff frequencies.................................................. 47
LIST OF FIGURES

Figure 1. Spectrographic word examples................................................................. 5

Figure 2. Panzee’s synthetic speech word-recognition performance.......................... 27

Figure 3. Samples of photographs used in Panzee’s spoken-word identification........... 32

Figure 4. Panzee working on a computer task.......................................................... 33

Figure 5. Intelligibility of sine-wave speech to humans............................................. 36

Figure 6. Experiment 1a chimpanzee and human word recognition.......................... 42

Figure 7. Intelligibility of noise-vocoded speech to humans....................................... 46

Figure 8. Experiment 1b chimpanzee and human word recognition.......................... 50

Figure 9. Intelligibility of time-reversed speech to humans....................................... 54

Figure 10. Experiment 2 chimpanzee and human word recognition......................... 59

Figure 11. Samples of lexigrams used in Panzee’s spoken-word identification task........ 67

Figure 12. Experiment 3 chimpanzee word-recognition performance across talkers........ 68
1. INTRODUCTION

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. Lenneberg (1967) claimed that both speech production and speech perception are uniquely human adaptations; a view later termed “Speech is Special” (SiS) by Liberman (1982). In contrast, studies with nonhumans have revealed that some animals are able to discriminate and categorize phonemes—the smallest unit of speech sounds—much as humans do (Kluender, Diehl, & Killeen, 1987; Kuhl & Miller, 1975; Kuhl & Padden, 1982, 1983). Therefore, it may be that auditory processing in humans and nonhumans are fundamentally similar, as proposed by Kuhl’s (1988) “Auditory Hypothesis.” 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.

Numerous experiments have investigated both 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 (Remez, 2005; Trout, 2001). These abilities have been difficult to show in nonhumans. Animals typically do not understand word meaning and are tested with the very brief, meaningless phoneme components of speech instead. More recently, however, experiments with a language-trained chimpanzee (Pan troglodytes) named Panzee have demonstrated the ability to recognize synthetic speech in some of the highly reduced forms humans have been tested with (Heimbauer, Beran, & Owren, 2011). These
outcomes support the Auditory Hypothesis, suggesting that comprehension of meaningful speech does not require perception specializations. Instead, listeners can apply general auditory processes shared with pre-hominin ancestors.

After discussing general aspects of human speech perception and production, key findings from experiments investigating SiS and the Auditory Hypothesis will be reviewed. Three new experiments conducted with Panzee, which further investigated the speech-perception capabilities of this chimpanzee, will then be presented. The first of these experiments extended an earlier finding which demonstrated that Panzee was able to recognize spoken words presented in “sine-wave” and “noise-vocoded” forms (Heimbauer et al., 2011). Both are synthetic versions of speech that reduce normal speech acoustics to small sets of sine waves and noise bands, respectively. It was hypothesized based on Panzee’s previous results that she would rely on the same acoustic cues as humans when hearing these speech forms (Remez, Rubin, Pisoni, & Carrell, 1981; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Alternatively, Panzee could rely on a more holistic approach, for instance matching general characteristics of synthetic words to corresponding natural versions. To test this hypothesis, both Panzee and human participants were tested with varying versions of sine-wave and noise-vocoded speech, using stimuli similar to those used in previous human testing (Remez et al., 1981; Shannon et al., 1995).

In the second experiment, Panzee and humans heard words presented in “time-reversed” form. This manipulation involves separating the speech signal into fixed time-length segments and then reversing these “windows” throughout the signal. Humans find speech in this form highly intelligible for reversal windows up to approximately 100 milliseconds in length
(Saberi & Perrott, 1999), with intelligibility decreasing as window length increases. Here, the relationship between phoneme- and window-length is proposed to be the critical feature. Specifically, if the reversal window is smaller than the average duration of a phoneme, the manipulated speech remains comprehensible. However, if window length exceeds phoneme duration, then the speech becomes critically distorted. Based on Panzee’s robust speech-perception abilities, it was again hypothesized that Panzee would perform similarly to human listeners, thereby providing evidence of attending to the phoneme content of speech. Alternatively, if Panzee’s speech recognition relies on a different strategy than humans use—for instance, a more holistic approach—she would not show the same window-length-dependent intelligibility function.

The third experiment investigated Panzee’s ability to understand speech from a variety of talkers. This capability is important because the acoustic characteristics of speech can vary significantly among both individual talkers and classes of talkers (Pisoni, 1995; Remez, 2005). As Panzee routinely interacts with, and responds to, different individuals, it was hypothesized that she would readily perceive familiar words regardless of the talker. Alternatively, she may simply have become accustomed to the speech of particular, known individuals without being able to normalize to novel talkers. 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 children.

1.1 Human speech

When perceiving natural speech, listeners are faced with many perceptual challenges. The most important is to combine and categorize the diverse acoustic elements making up the
speech stream (Remez, 2005), thereby mapping signal acoustics onto their linguistic correlates (Pisoni, 1995). Words are produced differently each time they are uttered, and minimal differences between words often matter. Researchers, therefore, have focused on how listeners can recognize specific speech sounds in the face of potentially large acoustic differences and then use this information to understand the phonetic content of spoken language (Appelbaum, 1996).

The acoustic characteristics of speech are illustrated in spectrographic form in Figure 1a. A spectrogram is a visual depiction of the time-varying properties of sound, with time displayed on the horizontal axis, frequency on the vertical axis, and amplitude shown as the darkness of shading at any particular point (Olive, Greenwood, & Coleman, 1993). One important characteristic is the fundamental frequency (F0), which corresponds to the basic rate of vibration of the vocal folds in the larynx. F0 is typically the lowest prominent frequency visible in a speech spectrogram. Regular vibration also produces energy at integer multiples of F0, referred to as “harmonics” (H2 and H3 are labeled in the figure). Energy from the larynx subsequently passes through the pharynx, oral cavity, and nasal cavity, whose resonances act to filter this energy. 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. The lowest three bands, often also called formants, are also labeled in the figure (F1, F2, and F3).
Figure 1. Spectrographic word examples. The spectrograms were created using a sampling rate of 44100 Hz and 0.03 s Gaussian analysis window. a) The natural word “tickle,” showing its fundamental frequency (F0), next higher harmonics (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). d) The word “tickle” in time-reversed form, with a 50 ms time-reversal (TR) window noted.
The perceptual challenge of recognizing acoustically variable speech is formally known as the “lack of invariance” problem. Speech acoustics can be highly variable, with individuals showing systematic differences in speech rate, F0s, and formant values. The latter are even more different among classes of talkers, for instance males versus females, and adults versus children. In addition to varying due to physical and physiological properties, speech acoustics can also differ due to factors such as talker emotional state or regional dialect (Pisoni, 1995). The result is that there is no simple mapping between acoustic structure and phonetic units, meaning that listeners have to categorize a talker’s phonemes in the absence of invariant cuing. Humans nonetheless routinely recognize speech from both familiar and unfamiliar talkers, an ability referred to as “talker normalization.”

### 1.1.1 Natural speech perception phenomena

Listeners accomplish a broad array of auditory perceptual tasks both when learning and after having mastered language, many of which seem effortless. The majority of language skills are learned implicitly, without instruction, relatively passively, and with minimal conscious attention. Infants as young as eight months old can segment and organize phonemes and words from a continuous, acoustic speech stream (Marcus, Vijayan, Rao, & Vishton, 1999; Saffran, Aslin, & Newport, 1996). Categorization of individual phonemes has been demonstrated at even younger ages (Werker & Desjardins, 1995), which is likely important in learning how different language elements combine. Experiments by Werker and Desjardins (1995) have revealed that at 6 to 8 months of age, infants discriminated phonemes across a variety of languages, but by 10 to 12 months of age had become tuned to the phonemes of the language spoken by their caregivers.
In addition to implicit perception of natural speech, there are automatic perceptual phenomena that occur when humans hear speech in altered or distorted forms. One notable example is the ability to identify words containing a phoneme or a syllable that has been replaced by white noise. This capability has been termed "phonemic restoration" by Warren (1970) and is so deeply rooted cognitively that participants can not only identify the words, but also report hearing them in their entirety without perceptible gaps (Warren & Obusek, 1971). One interpretation of this phenomenon is that the missing segment is re-created in the brain, even in the absence of that sound (Kashino, 2006). This ability to “hear” the missing segment presumably enables listeners to fill in the gaps that routinely occur when speech is heard in noisy, everyday situations. Another phenomenon of interest has been termed “duplex perception” (Rand, 1974). Duplex perception occurs when a short sine wave, sounding like a chirp, is presented to one ear and is perceptually integrated into an otherwise incomplete phoneme being played to the other ear (Whalen & Liberman, 1987). Davis and Johnsrude (2007) suggested that duplex perception illustrates that human listeners actively attempt to organize sound into perceptible speech whenever possible. In fact, this tendency to organize likely extends to any form of distorted speech.

1.1.2 Synthetic speech perception. 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. Here, three forms of altered speech are of particular interest. Two of these, sine-wave and noise-vocoded speech, lack many of the acoustic features traditionally considered crucial to speech perception, including F0 and formants (see Figures 1b and 1c). The third form, time-reversed speech, alters moment-to-
moment temporal patterning in the signal (see Figure 1d). In each case, human listeners rely on their extensive experience with spoken language to make sense of the input. Being able to do so is considered a form of “top-down processing,” 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).

Since 1981, sine-wave speech has been investigated for the purpose of understanding spoken language processing (Lewis & Carrell, 2007; Remez et al., 2009; 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 Figure 1b). Sine-wave speech is extremely unnatural-sounding and is considered to preserve key phonetic properties only in an abstract form (Remez et al., 2009). 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.

Another altered form of interest is “noise-vocoded” speech, which is synthesized from noise bands (see Figure 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, Loizou, Spahr, & Maloff, 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
identification of noise-vocoded words in sentences increased from less than 10% to 70% correct within just a few minutes.

The most recently developed synthesis form of interest is time-reversed speech (Barkat, Meunier, & Magrin-Chagnolleau, 2002; Purnell & Raimy, 2008; Saberi & Perrott, 1999). As shown in Figure 1d, such sounds contain segments of equal length that have been reversed in time, typically on a millisecond (ms) scale. This manipulation preserves the amplitude of each frequency component at every point in time, but reverses the pattern of energy changes within each window. The resulting disruption of the amplitude envelope could make the signal unintelligible, but listeners are reliably able to recognize speech content at window lengths up to 100 ms (Saberi & Perrott, 1999). Window lengths exceeding 100 ms produce partial intelligibility, with the 50% “threshold” point for intelligibility occurring at 130 ms or more. The interpretation is that individual phonetic segments (“phones”) in speech range from approximately 50 to 100 ms (Crystal & House, 1988). In other words, reversal-windows up to 100-ms long leave many individual phonemes undisturbed. However, longer windows break up more and more individual phonemes, thereby making the speech unintelligible. Barkat et al. (2002) confirmed this view of phoneme-length perception in finding that French-speaking participants hearing French sentences showed a different 50% intelligibility threshold for time-reversed speech. Although these listeners also demonstrated decreased speech recognition as window-lengths increased, intelligibility fell more slowly. The 50% threshold was reached at a window approximately 20 ms longer than for English-speaking listeners hearing English sentences, likely reflecting a longer mean phoneme duration in French compared to English.
Overall, results of synthetic-speech perception experiments provide evidence that sine-wave, noise-vocoded, and time-reversed versions can be intelligible. It is not clear, however, what exactly these three synthesis forms have in common to make each recognizable to listeners. For synthetic speech, Remez et al. (1994) characterize the critical cues for sine-wave and noise-vocoded speech as “spectro-temporal” patterning, while also noting that the two versions do not show obvious commonalities. This interpretation is arguably supported by findings from time-reversed speech experiments, which show that preserving spectro-temporal information within phonemes is a critical factor (see also Drullman, 2006; Remez, 2005).

1.2 The SiS argument

In 1967, Liberman and colleagues proposed that in order to perceive speech, humans must draw on their implicit knowledge of how phonemes are articulated. They hypothesized that spoken words are perceived by identifying associated vocal tract gestures, rather than by identifying sound patterns of speech itself. This hypothesis was termed the “Motor Theory of Speech Perception,” and implies that only producers of speech can also perceive speech. Therefore, both speech production and speech perception would be “special” to humans. Later, Fodor (1983) revived the historical concept of mental modularity from Gall’s phrenology, in which individual mental faculties are associated with domain-specific areas of the brain. Fodor suggested that these specialized cognitive modules operate individually on domain-specific input. This theory “upped the ante” for uniquely human speech perception, with Liberman and colleagues (Liberman & Mattingly, 1989; Mann & Liberman, 1983; Whalen & Liberman, 1987) then claiming that humans have a specialized, cognitive module responsible for speech perception.
In support of this SiS argument, a variety of human speech perception phenomena have been proposed as evidence for a speech mode, a Fodorian phonetic module, or both. For example, both phonemic restoration and duplex perception have been cited as evidence for a uniquely human, speech module (Liberman & Mattingly, 1989; Mann & Liberman, 1983; Whalen & Liberman, 1987). Findings pertaining to neural mechanisms of speech perception have also been interpreted in this way. For instance, most humans show a “right-ear advantage” (REA) when processing the phonetic elements of speech, meaning that input presented to the right ear is recognized more quickly and accurately than that presented to the left ear. The REA is attributed to phonetically based language processing being primarily lateralized to the left hemisphere in approximately 95% of the population (Hughdahl, 2004; Kimura, 1961; Studdert-Kennedy & Shankweiler, 1970).

Historically, dichotic listening experiments have been used to investigate language-related lateralization effects (for a review, see Hugdahl & Davidson, 2004). In these experiments, different phonetic stimuli are presented simultaneously to the two ears, and participants are instructed to report what is heard in one ear or the other. Studdert-Kennedy and Shankweiler (1970) used dichotic listening to show that both hemispheres are involved in processing purely auditory parameters of a speech signal. However, they found that left lateralization occurs when specifically linguistic features are attended to. They concluded that these findings provide evidence of a specialized, linguistic device located in the left hemisphere—thereby supporting the SiS view. Cutting (1974) investigated the REA associated with syllable and phoneme perception, using stylized, synthetic vowel and consonant-vowel syllables and sine-wave facsimiles that were not perceived as speech sounds. He concluded that there
might be two processing mechanisms in the left hemisphere—one dedicated to speech per se, and the other responding to rapid frequency changes in any auditory input, which processes formant movement in speech as well as frequency modulation in other sounds.

It should be noted that some animals, including nonhuman primates, have also demonstrated asymmetries for perception of vocalizations (for reviews see Corballis, 2009; Taglialatela, 2007). Japanese macaques (Macaca fuscata) have shown an REA and left-hemisphere advantage for species-specific vocalizations (Petersen, Beecher, Zoloth, Moody, & Stebbins, 1978; Petersen et al., 1984), as have mice (Ehert, 1987). Magnetic resonance imaging studies of gorilla (Gorilla gorilla gorilla), orangutan (Pongo pygmaeus), chimpanzee (Pan troglodytes), and bonobo (Pan paniscus) brains have revealed a larger planum temporale in the left hemisphere, paralleling an asymmetry in the corresponding area of the human brain (Gannon, Holloway, Broadfield, & Braun, 1998; Hopkins, Marino, Rilling, & MacGregor, 1998). In humans, the planum temporale is known to be involved in both production and perception of spoken language (discussed in more detail below).

1.3 The Auditory Hypothesis

Despite a variety of evidence in support of the SiS hypothesis, the claim that only speech producers can also perceive speech is problematic, because only humans have the vocal-tract morphology capable of producing speech (deBoer, 2006; Fitch, 2000). The descended larynx, in conjunction with the hyoid bone and tongue root, enables production of diverse formant patterns by humans (Lieberman, 1968). Therefore, the uniqueness of human anatomy makes it difficult to refute SiS claims. In addition, humans who are mute can still routinely perceive speech (Studdert-Kennedy, 1980), and some animals have demonstrated at least the ability to
discriminate and categorize speech sounds much as humans do (Kluender, Diehl, & Killeen, 1987; Kuhl & Miller, 1975; Kuhl & Padden, 1982, 1983). Trout (2001) has proposed that refuting the SiS view requires demonstrating a common cognitive or biological substrate for speech perception in humans and nonhuman animals. In his view, such a demonstration is needed to support the argument that although the human ability to produce speech is unique among mammals, the same may not be true for speech perception. However, to date, neither a common mechanism nor common functional organization for speech perception has been shown across species. It may be, instead, that human speech-production capabilities are more specifically adapted to language than the auditory system is.

Not surprisingly, numerous auditory-perception studies with animals have been conducted to test claims of human specializations for speech. At least some evidence suggests that perception of speech is rooted in the same general mechanisms of audition and perception evolved to handle other classes of environmental sounds (Diehl, Lotto, & Holt, 2004). The next section, therefore, reviews some basics of mammalian hearing, followed by speech-perception evidence from research with animals, including both non-primates and nonhuman primates.

1.3.1 Mammalian hearing. The auditory-perception capabilities of humans and other mammals may very well have evolved similarly, as the general morphology and functions of the auditory system have been similar in all mammals since the middle ear was transformed from its reptilian form (Allin, 1975; Fleagle, 1999; Stebbins, 1983). In mammals, the basic function of the ear is amplification and transmission of sound, in addition to sensory transduction. The outer ear consists of the pinna, used for gathering and localizing sound. Sound waves are then funneled along the ear canal to the tympanic membrane (ear drum). The external ear filters
sound somewhat, passing energy in 2 to 5 kHz frequency best—a range that is important for speech perception (Breedlove, Watson, & Rosenzweig, 2007). Sound pressure waves cause the tympanic membrane to vibrate, which in turn moves the three small bones (ossicles) of the middle ear. The last bone in the chain—the stapes—transmits this energy to the oval window, a small membrane at the base of the cochlea of the inner ear. Here, energy from the original cyclical air pressure changes to waves in fluid (Kalat, 2009; Pickles, 1988). Finally, transduction occurs when these cochlear fluid waves cause neural firing in the eighth cranial nerve, which is transmitted to the auditory cortex (for further details, see Hackney, 2006).

The auditory system is similar in all mammals, functioning as a sound localizer and showing sensitivity across a range of sound frequencies. However, that range varies with head size (and associated inter-aural distance), as sound localization requires species with smaller heads to be sensitive to higher-frequency energy (Gans, 1992; Heffner, 2004; Masterton, Heffner, & Ravizza, 1969). Smaller mammals thus typically hear well above the nominal 20-kHz limit of human hearing. The earliest mammals were likely similar (Rosowski, 1992), as are most nonhuman primates (Beecher, 1974; Jackson, Heffner, & Heffner, 1999; Masterton et al., 1969; Owren, Hopp, Sinnott, & Petersen, 1988; Stebbins, 1973; Stebbins & Moody, 1994). However, frequency sensitivity has also been subject to change over the course of primate evolution, for instance due to habitat differences as well as body-size effects (Owren et al., 1988).

One such change was that larger-bodied monkeys and apes acquired better low-frequency hearing. This trend has resulted in both chimpanzees and humans showing well-developed hearing both at low (below approximately 1 kHz) and mid-range frequencies (1 to 8 kHz). Stebbins (1973) has proposed that these hearing changes may have been related to
evolutionary pressures for more intricate, intra-specific vocal communication systems. This view is consistent with human speech perception potentially being grounded in the more general auditory processing capabilities of larger mammals than being species-specific. Stebbins (1983) has also shown that hearing in chimpanzees is more similar to that of humans than to Old and New World monkeys—who can hear frequencies as high as 40 to 45 kHz. His review included several auditory parameters—high- and low-frequency sensitivity, lowest frequency threshold, best frequency, and area of the audible field—for five mammalian species, including a non-primate (tree shrew), a prosimian (bush baby), and three anthropoid primates (macaque, chimpanzee, and human). Because low- and mid-range frequency hearing is common to chimpanzees and humans, overall human auditory sensitivity may facilitate communication without being a specialization for speech.

In contrast to the chimpanzee hearing capabilities reviewed by Stebbins (1983), Kojima (1990) reported that two chimpanzees were notably less sensitive than humans to frequencies between 2 and 4 kHz. Although Kojima found external ear resonances to be approximately the same for both species, the chimpanzee sensitivity pattern showed a pronounced decrease in mid-range sensitivity-- more similar to patterns found in Old and New World monkeys (Beecher, 1974; Stebbins, 1973) than in humans. However, 2 to 4 kHz is also the range of highest energy in chimpanzee screams (see Figure 1b in Riede, Arcadi, & Owren, 2007)—which are frequently very loud. It may, instead, be that the decreased mid-range sensitivity of Kojima’s two captive subjects reflected hearing loss due to repeated exposure to high-amplitude screaming by conspecifics in the animals’ confined housing spaces. If so,
chimpanzees without any hearing impairment should be able to process critical frequencies of speech as well as humans.

1.3.2 Speech perception in non-primates. Research attempting to find human-nonhuman animal commonalities in speech perception mainly focused on perception of rudimentary elements of spoken language. In support of the Auditory Hypothesis, experiments with chinchillas by Kuhl and Miller (1975) and Loebach and Wickesberg (2006) have revealed that these animals perceive and discriminate at least some individual phonemic features of speech. In their seminal study, Kuhl and Miller (1975) demonstrated that chinchillas discriminated between consonant-vowel syllables differing in voice-onset time. This feature is the length of time between initial consonant articulation movements and onset of vocal-fold vibration. Specifically, the chinchillas were trained to respond differently to a variety of initial /t/ and /d/ consonant-vowel syllables produced by eight talkers. The animals also correctly classified novel instances of initial /t/ and /d/ syllables, including syllables produced by new four talkers, produced in new vowel contexts, and computer-synthesized /ta/ and /da/ syllables.

In a neurological study, Loebach and Wickesberg (2006) demonstrated that there might be a common physiological substrate in the peripheral auditory system of chinchillas and humans involved in recognition of speech cues. The animals showed auditory-nerve responses when hearing syllables in natural and noise-vocoded form that resembled those shown by humans hearing the same sounds. Loebach and Wickesberg presented four syllables produced by male talkers in both natural form and resynthesized form using one, two, three, or four noise bands. Despite the different spectral profiles of natural and noise-vocoded speech, chinchillas responded similarly to humans to the cues to consonant identity. In humans,
common spectro-temporal features of natural and noise-vocoded speech provide these cues for speech recognition, and perception is enhanced as the number of noise bands used in noise-vocoding synthesis increases (Shannon et al., 1995). Finding parallel performance in chinchillas suggests that these animals respond to the same spectro-temporal cues in noise-vocoded and natural speech that humans respond to.

Birds are not mammals, and therefore are not close evolutionary relatives of humans, but avian communication abilities should also be mentioned. Although the class Aves originated approximately 50 million years ago (Sibley & Alquist, 1990), research with birds also provides evidence that humans may be utilizing general mechanisms in speech processing. Vocal communication in birds shows some parallels to human abilities. For example, many songbird species use complex vocalizations and similar underlying developmental and mechanistic processes may be involved (for a review, see Beckers, 2011, and Doupe & Kuhl, 1999).

Psychoacoustic studies have shown that at least some birds demonstrate some of the perceptual phenomena proposed to be special to humans hearing speech—specifically, phoneme discrimination and categorization, compensation for coarticulation, and an ability to solve the lack of invariance problem. As one example, both songbirds and non-songbirds have demonstrated the ability to discriminate phoneme variations of the vowel /a/ (Hienz, Sachs, & Sinnott, 1981). Japanese quail (Coturnix coturnix) can also discriminate and categorize initial-consonant syllables (Kluender et al., 1987), while budgerigars (Melopsittacus undulates) discriminate vowel categories and are more sensitive to phonemic vowel distinctions than to talker-related vowel variation (Dooling and Brown, 1990). More recently, Ohms, Gill, Van Heijningen, Beckers, and ten Cate (2010) found that zebra finches (Taeniopygia guttata) could discriminate and
categorize monosyllabic words that differ in vowels, and can generalize this ability to unfamiliar male and female talkers. Results such as these suggest that some birds have an ability to normalize specific components of speech across talkers. Budgerigars and zebra finches have even demonstrated the ability to discriminate both full-formant and sine-wave versions of /ra-la/, revealing some similarities to human discrimination of speech and speech-like sounds (Best et al., 1989; Dooling et al., 1995). Clearly, birds share at least some rudimentary speech-perception capabilities with humans.

1.3.3 Speech perception in nonhuman primates. Nonhuman primates, closer evolutionary relatives of humans than of other mammals and birds, also demonstrate auditory processing capabilities that support the Auditory Hypothesis. In two studies, Kuhl and Padden (1982, 1983) tested rhesus macaques (M. mulatta), an Old World monkey, and human infants to compare discrimination of voiced and voiceless phonemes. The difference between these two is that voiced sounds include vocal-fold vibration and voiceless sounds do not. Monkeys were first trained to categorize these phoneme types in a “same-different” procedure, and after were tested with unfamiliar syllable pairs on a stimulus continuum ranging from voiced to voiceless (e.g., /ba-pa/). In the second experiment, the subjects heard syllable pairs differing in “place of articulation” (e.g., /b/-/d/). This feature can refer to tongue placement during speech production. For example, when /b/ is produced, the lips come together and the tongue is held away from the teeth. However, when /d/ is produced, the lips are separated and the tongue touches the ridge above the top of the teeth. Both experiments revealed that the macaques divided the stimulus continuum at the same physical points as humans, meaning they showed similar boundary points in categorization based on both voicing and place.
More recently, it has been demonstrated that Japanese macaques can perceive the articulation events of speech, although their performance better resembles that of human infants than of adults. Sinnott and Gilmore (2004) investigated perception of place-of-articulation information in natural speech by monkeys and adult humans, presenting consonant-vowel tokens consisting of /b/ or /d/ combined with /i/, /e/, /a/, or /u/. Tongue placement is different across these sounds, with /a/ and /u/ described as “back” vowels, and /i/ and /e/ as “front” vowels. In front vowels the tongue is positioned as far forward as possible without creating a constriction that would produce a consonant sound. In back vowels the tongue is positioned as far back as possible without creating a constriction.

Sinnott and Gilmore (2004) used a two-choice identification task, whereby the monkeys and human participants had to actively classify /b/ versus /d/ consonant-vowel stimuli by moving a lever. Humans performed well with all stimuli, while the monkeys performed better with tokens based on the back vowels /a/ and /u/ than with front vowels /i/ and /e/. An earlier study had shown that three- to four-month-old human infants also classify back vowels more easily (Eimas, 1999), evidently learning how to reliably differentiate front vowels over time. Sinnott and Gilmore therefore concluded that the monkeys’ performance reflected basic auditory-system processing, as is also found in preverbal infants before critical speech-related learning occurs.

There is also evidence that cotton-top tamarins (Saguinus oedipus), a New World monkey, perform similarly to one-month-old infants when discriminating Dutch versus Japanese sentences (Ramus, Hauser, Miller, Morris, & Mehler, 2000). Although testing methods were quite different for the two species, they found that both were able to distinguish these
languages. In addition, both species did so in the face of at least modest talker-related variability, as stimuli from a total of four speakers of each language were used. Neither monkeys nor infants were able to discriminate the languages when sentences were played backwards. The researchers concluded that the infants were likely using innate, generalized auditory processing shared across nonhuman primates, if not all mammals.

As mentioned earlier, both behavioral and neuroanatomical studies have revealed that primates other than humans show hemispheric lateralization effects (Taglialetela, 2007). Specifically, several macaque species exhibit an REA for communicatively relevant vocalizations. In chimpanzees, brain-imaging studies have revealed parallels to Wernicke’s and Broca’s areas in humans—both of which are considered critical in language functions. Wernicke’s area is located in the planum temporale of the temporal lobe and is involved in human language perception and comprehension. This region can be as much as five times larger in the left hemisphere than in the right. In chimpanzees, this brain structure was significantly larger in the left hemisphere for 94% of MRI scans examined by Gannon and colleagues (1998). Broca’s area is located in the inferior frontal gyrus, and is critical in speech production. Similarly, this area is activated in chimpanzees when these animals vocalize (Taglialetela, Russell, Schaeffer, & Hopkins, 2008).

Only a limited number of studies have tested perceptual discrimination of speech sounds in apes. In one such experiment, Kojima et al. (1989) presented two chimpanzees with consonant-vowel syllables, using synthetic /ga/-/ka/ and /ba/-/da/ continua to examine perception of voicing and place-of-articulation contrasts, respectively. Using response times as an index of perceived similarity, chimpanzees demonstrated better discrimination of syllables
when phonetic contrasts were based on the same features of voicing and articulation that humans attend to. As in the earlier-mentioned Sinnott and Gilmore (2004) experiment with macaques, chimpanzees did not perform as well as humans, likely reflecting the importance of early experience in human speech perceptual development. As noted earlier, one unique aspect of human language acquisition is the importance of learning and environmental input.

1.3.4 Evidence of top-down processing when perceiving speech. Speech perception studies with nonhuman primates appear to demonstrate the operation of general auditory processing—in support of the Auditory Hypothesis. As in studies with non-primates, however, this work has not provided evidence of the higher-level, top-down-processing capabilities claimed to be evidence of a uniquely human, cognitive module (Liberman & Mattingly, 1989; Mann & Liberman, 1983; Whalen & Liberman, 1987). Such processing is potentially difficult to demonstrate with the rudimentary types of stimuli used in these studies, with top-down effects emerging more clearly in humans at the level of words and sentences. In other words, distinguishing SiS theory and the Auditory Hypothesis also requires examining higher-order processing of meaningful speech in nonhumans, rather than for meaningful speech segments.

In fact, acoustic cues to phonemes and syllables may not be processed in the same way as lexical components of language. For example, the former could be processed as non-speech sounds are, using “bottom-up” perception that is not strongly guided by higher-level knowledge of meaningful language. Because chimpanzees and humans are close evolutionary relatives—divergence from a common ancestor occurred 5- to 8-million years ago (Wood, 1996)—investigating word recognition by these apes could provide compelling evidence for the SiS versus Auditory Hypothesis debate. As Stebbins (1983) noted, chimpanzees may be the species
to shed light on whether the ability to perceive meaningful speech was present in latent form in hominins before the evolution of mechanisms to produce speech.

Revising the Motor Theory of Speech Perception, Galantucci, Fowler, and Turvey (2006) recently suggested that the only compelling evidence for neural hardware specialized for speech would be discovering a dedicated circuit active “if and only if” speech is perceived or produced. Galantucci et al. further argue that speech cannot be understood in isolation, that production and perception components necessarily work together, and that spoken language is critically embedded in a communicative context. These are restrictive arguments that again tend to inherently rule out the possibility of animal experiments. As mentioned earlier, given that nonhuman primates do not speak, it is impossible to test them for a dedicated circuit involved in both speech production and perception.

Another point is that very few nonhumans are raised in a speech-rich environment, meaning that animals typically have no opportunity to experience the input that is crucial to human language development. A very few apes, however, have been raised in a manner similar to humans. Through a combination of language exposure and enculturation by humans, these animals acquired both meaningful communicative and word-recognition abilities (Beran, Savage-Rumbaugh, Brakke, Kelley, & Rumbaugh, 1998; Brakke & Savage-Rumbaugh, 1995; Rumbaugh & Savage-Rumbaugh, 1996; Savage-Rumbaugh, Murphy, Sevcik, & Brakke, 1993). While approaching the impossible criteria proposed by Galantucci et al. (2005; see also Trout, 2001), language-capable apes also have the potential to be tested with meaningful speech.
1.4 Speech perception in a language-trained chimpanzee

Language-trained apes arguably present a convincing, and possibly unique, opportunity for investigating speech perception in nonhumans and settling the SiS versus Auditory Hypothesis debate. Specifically, if speech perception utilizes a specialized, human cognitive module, then a language-trained ape should not be able identify or understand speech presented in the altered, synthetic forms proposed to require this specialization. Furthermore, such an animal should arguably not exhibit talker normalization for meaningful speech, given the associated lack of invariance problem. However, previous research with one particular language-trained ape suggests otherwise, in support of the Auditory Hypothesis.

1.4.1 Previous experiments with Panzee. To examine whether apes and humans can show fundamental similarities in speech processing, a recent series of experiments assessed the perceptual capabilities of a chimpanzee named Panzee. This animal is an adult female, housed at the Language Research Center (LRC) at Georgia State University (GSU). She was raised routinely hearing speech from the age of eight days old, showing reliable recognition of approximately 130 spoken English words. Panzee was also taught corresponding visuo-graphic symbols, called lexigrams, and can use both these symbols and associated photographs to communicate in everyday and experimental situations. When Panzee hears a familiar English word, she is reliably able to choose the correct, corresponding lexigram or photograph from among multiple alternatives (Beran et al., 1998). To date, Panzee has been tested with natural, whispered, and several forms of synthetic speech (Heimbauer et al., 2011; Heimbauer, unpublished data).
To test Panzee, 48 two- to five-syllable words were chosen from among her familiar English words. In all the experiments, Panzee first heard a word in natural or synthetic form. Her task was then to use a joystick to select the one lexigram or photograph corresponding to that word from among four such items appearing on a computer screen. Panzee received no feedback for correct or incorrect choices during test sessions, ruling out the possibility of learning how to respond to altered versions of the words. However, she did receive a reward every three or four trials on a randomized, noncontingent basis. Sixteen different natural and eight different test words were presented in each experimental session, with sessions including four blocks of these words in randomized order, for a total of 96 trials. Typically, a session lasted 20 to 30 minutes.

Annual testing over a 10-year period ending in 2008 has demonstrated that Panzee’s session performance for natural words is consistently between 75% and 85% correct (M. J. Beran, personal communication, January 2010). When tested with words in whispered form, she performed similarly at or above 75%. Subsequent testing with words in four different synthetic forms produced the same results (Heimbauer, unpublished data). One form simply reproduced the original utterance as closely as possible, but another form was less complete. This less complete, “voiced-only” form included only tonal elements of the original, meaning that any noise-based components were removed. Perception of voiced-only speech thus involves top-down processing, as unvoiced, noisy acoustics are important contributors to many English phonemes. Despite the missing sounds, Panzee showed no difference in performance with voiced-only versus natural words, even when considering only “first trials”--the 48
instances in which she heard a given word in synthetic form for the first time (Heimbauer, unpublished data).

The two remaining synthesis forms were selected to be directly relevant to the SiS versus Auditory Hypothesis debate, with outcomes shown in Figure 2. One was noise-vocoded speech, synthesized using seven noise-bands. This form can be challenging to humans, but is relatively comprehensible to most listeners (Davis & Johnsrude, 2007; Davis et al., 2005; Hervais-Adelman et al., 2008; Shannon et al., 1995). Despite this potential challenge, Panzee’s performance was statistically well above chance level both on first trials and overall. Her performance with these versions was statistically below the outcomes for natural words presented in the same session, but she showed no evidence of having learned how to respond to them (Heimbauer et al., 2011). The last synthetic form tested was sine-wave speech, described as “science-fiction sounds” and “whistles” by naïve human listeners (Remez et al., 1981) and “impossibly unspeechlike” by prominent speech researchers (Remez et al., 1994). Again, Panzee’s performance was above chance on first trials and overall. While she was less accurate identifying sine-wave words than natural words, human performance was similar to this in a transcription task—even with humans receiving orientation to sine-wave speech and hearing all the natural-word stimuli ahead of time (Heimbauer et al., 2011).

The results of these noise-vocoded and sine-wave speech experiments provided the first evidence of human-like performance by a nonhuman responding to meaningful, but incomplete and perceptually difficult synthetic speech. Top-down processing abilities are necessary to identify speech in both forms, with both Panzee and human listeners needing to access previous knowledge about natural speech in order to identify altered versions of the words. Given
that some researchers have linked perception of both noise-vocoded and sine-wave speech to a proposed phonetic module (e.g., Trout, 2001; Whalen & Liberman, 1987), these results directly contradict SiS arguments while supporting the Auditory Hypothesis. However, similarities in recognition performance do not necessarily imply similarities in underlying processing, and it remains unclear if Panzee was using the same perceptual strategies as humans to perform this task.

**Figure 2.** Panzee’s synthetic-speech, word-recognition performance. The figure is reproduced from Heimbauer, Beran, & Owren, 2011, and includes means and standard errors of percentage-correct performance for 48 words heard in natural, noise-vocoded, and sine-wave forms. First trials represent the 48 first instances of the chimpanzee hearing a word in given synthetic form. The first set of sine-wave results shows performance with noncontingent, intermittent reward delivery and no response feedback. The second set shows performance with contingent reward received on natural trials but with no reward or response feedback on sine-wave trials. The dashed line indicates the chance-performance rate of 25% correct. All comparisons to chance performance were statistically significant at \( p < 0.008 \) and are marked by a pair of asterisks.
2. CURRENT EXPERIMENTS

In the debate about the existence of, or necessity for, human specializations for speech perception, Panzee’s natural and synthetic speech word-recognition abilities provide evidence for the Auditory Hypothesis view. Her performance with noise-vocoded and sine-wave words in particular suggests that human speech perception is grounded in generalized auditory capabilities and extensive experience with speech rather than specialized processing mechanisms. These studies do not, however, allow an unequivocal conclusion that Panzee’s speech processing is fundamentally similar to human perception. For example, it is possible that Panzee is able to recognize her relatively small number of familiar words through more holistic judgments, either based on duration or overall aural impressions (see Heimbauer et al., 2011). To test the SiS view versus the Auditory Hypothesis more definitively, it would require specific evidence about detailed aspects of Panzee’s processing strategies.

Therefore, the current experiments were designed to extend previous work on Panzee’s speech perception, specifically investigating whether she relies on similar auditory perceptual mechanisms as humans. Experiment 1 examined the acoustic cues Panzee may be attending to when hearing speech in sine-wave (Experiment 1a) and noise-vocoded (Experiment 1b) forms. For humans, 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). Hypothesizing that Panzee would also show evidence of relying disproportionately on these cues, Experiment 1a compared her performance to that of human participants when hearing four critically different versions of sine-wave words. In Experiment 1b, Panzee and human participants were tested with noise-vocoded words produced from varying numbers of noise
bands. 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 (Davis et al., 2005; Shannon et al., 1995). Based on her previous performance with words in noise-vocoded form, it was again hypothesized that Panzee would show similar performance to humans as a function of the number of noise bands used in synthesis.

Experiment 2 focused on time-reversed speech. Here, previous research with humans has revealed that phonemes are cued over segments of roughly 50 to 100 ms (Crystal & House, 1988), with shorter windows in time-reversed form leaving intelligibility largely unaffected, but longer windows having a significant detrimental impact (Saberi & Perrott, 1999). The second experiment, therefore, was designed to investigate whether Panzee’s perception also relies on cuing over this time frame. In this study, she heard test words with reversal windows that ranged from 25 ms to 200 ms in eight 25-ms increments, and human participants were tested for comparison purposes. The prediction tested was that, like humans, Panzee would perform best with window lengths less than the average duration of English phonemes, with intelligibility decreasing to 50% threshold level for a window-length of approximately 130 ms—as found for human performance by Saberi and Perrott (1999).

Experiment 3 investigated the lack of invariance problem. While humans seem to effortlessly accommodate the acoustic variability found in speech from different talkers, evidence of corresponding talker normalization capabilities in nonhumans is suggestive, but limited. As Panzee has heard and responded to speech from a variety of talkers through her lifetime, this last experiment was designed to test her talker-normalization capabilities more
systematically. Stimuli included speech from a diverse set of talkers, including variation in biological sex, age, and dialect background. In addition, some of the individual talkers were familiar to Panzee and others were unfamiliar.

2.1 General Methods

2.1.1 Subject. The subject was the female chimpanzee Panzee, who was 25 years old when the current experiments began. This animal is socially housed with three conspecifics at the LRC at GSU. Panzee has daily access to indoor and outdoor areas, unlimited access to water, and is fed fruits and vegetables three times a day. She participates in testing on a voluntary basis and may choose not to participate or to stop responding during a session. Panzee uses a language-like, 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, this animal also has experience with numerous, computer-based protocols (Rumbaugh & Washburn, 2003). In the three experiments, she participated in three to four, 20- to 30-minute 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.

2.1.2 Participants. Human participants were undergraduates, aged 18 to 55 years old, and were recruited via the GSU on-line, experiment participation system. Each participant was tested in only one of the applicable experiments—either in Experiment 1a, 1b, or 2. Only participants without reported hearing problems and that were native English-speakers were included in analyses.
2.1.3 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 32-12008 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, 2008), and custom-written scripts (Owren, 2010).

2.1.4 Stimuli. Spoken stimuli were chosen from a list of approximately 130 words that Panzee has consistently identified in a decade of annual word-comprehension testing. Natural word stimuli were recorded at 44100 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 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.

2.1.5 Chimpanzee procedure. Panzee was tested using the general procedure employed for annual word-comprehension testing. 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-sec delay, four different photographs (Experiments 1 and 2; see Figure 3) or lexigrams (Experiment 3; see Figure 11) appeared on the screen. One of these items was the correct match to the audio stimulus, and the others were foils chosen randomly by the controlling computer program. As illustrated in Figure 4, visual items were positioned randomly in four of six possible locations—three on the left side of the screen and three on the right. 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 & Washburn, 2002).

![Figure 3. Samples of the photographs used in Panzee’s spoken-word recognition task.](image)
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 Figure 4). In Experiments 1a and 1b both natural and altered words were presented in randomized order within each trial block. In Experiment 2, only altered words were presented; and in Experiment 3, only natural-word stimuli were used. Panzee was rewarded with highly valued food, including pieces of cherries, grapes, blueberries, peaches, raspberries, strawberries, mixed fruit, or Chex Mix®. The reward regimen was specific to each experiment, and is noted in the individual procedure sections.

Figure 4. Panzee working on a computer task. She was hearing words and choosing corresponding photos.

2.1.6 Human procedures. Human testing varied somewhat by experiment and is described in the individual procedure sections. Common elements included that stimuli were presented in randomized blocks, that the stimulus was heard twice on each trial 1200-ms apart, and that listeners had eight seconds in which to transcribe that word.
2.1.7 Data analysis. Statistical testing varied by experiment and is discussed separately in each case.

3. EXPERIMENT 1

Although Panzee has demonstrated the ability to identify sine-wave and noise-vocoded speech (Heimbauer et al., 2011), the cognitive mechanisms that she is employing are unknown. Thus, two studies investigated the acoustic cues she may be attending to when identifying these altered forms of speech. It was hypothesized that Panzee would show evidence of using the same information humans are proposed to use, namely the spectro-temporal cues produced by amplitude and frequency modulations over time (Remez et al., 1981; Remez et al., 1994; Shannon et al., 1995). Alternatively, she might utilize more holistic cues, such as word length or general sound impressions. Analyses conducted by Heimbauer et al. (2011) argued against this possibility. In these 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.

Here, Panzee was tested with sine-wave words that included varying combinations of individual tones (Experiment 1a), and noise-vocoded speech produced from varying numbers of noise bands (Experiment 1b). As a result, these different synthetic words included differing degrees of time-varying amplitude and frequency cuing of a kind previously shown to systematically affect human performance. The rationale was that if performance by both 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. Although difficult
to specify precisely, this information has been characterized as critical spectro-temporal pattering in each natural word that is preserved in synthetic versions.

3.1 Experiment 1a

A seminal study by Remez and colleagues (1981) explored human perception and identification of sine-wave speech with the objective of investigating the role of time-varying properties in speech perception. They found that sine-wave speech, despite lacking the traditional acoustic information—such as F0 and formants—could be intelligible. In addition, as shown in Figure 5, their listeners were more successful in identifying sentence components when SW1 and SW2 were both present (SW123 and SW12 forms) than when either one was absent. SW1 and SW2 model the corresponding amplitude and frequency modulation patterns of natural-speech formants F1 and F2, respectively. That outcome was expected, as it is the lowest two formants that typically most clearly cue vowel identity, in addition to providing articulation information for adjacent consonants (Drullman, 2006; Ladefoged, 2001). Thus, Remez et al.’s results demonstrate that corresponding tone analogs to F1 and F2 contribute disproportionately to sine-wave speech identification as well.

To ascertain if Panzee also differentially uses these components of sine-wave speech, both she and human participants were presented with 24 words synthesized in four of the same forms used by Remez and colleagues. All three sine waves were present in one version (SW123), while one of the three was removed in the others. It was hypothesized that if Panzee identifies sine-wave speech using similar acoustic cues as humans, her performance when particular sine waves were missing would be similar to humans. Specifically, she was predicted
to perform best with words that included both SW1 and SW2, (SW123, SW12) and less well when either of these components was missing (SW13, SW23).

Figure 5. Intelligibility of sine-wave speech to humans. The figure is reproduced from Remez et al., 1981, and shows syllable-transcription results of sine-wave sentences in seven different forms.

3.1.1 Subject. The subject was the chimpanzee Panzee.

3.1.2 Participants. There were 12 human participants (eight females).

3.1.3 Stimuli. For Experiment 1a, natural word stimuli were recorded spoken by an adult male researcher (MJB), who is very familiar to Panzee and who conducted her annual, word-comprehension testing over a 10-year period. Stimuli consisted of natural versions and sine-wave versions of 24 spoken words that Panzee had previously successfully identified in sine-wave form (Heimbauer et al., 2011). 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 from the larger word list that Panzee can routinely identify 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. Individual sine-waves were removed using Hanning-window, band-pass filtering.

Table 1.

Orientation and test word groups. Test words used in Experiments 1a (A and B), 1b (C and D), 2 (*), and 3 (E and F), as well as orientation words used in Experiments 1a and 1b (O).

<table>
<thead>
<tr>
<th>Words</th>
<th>Experiment 1a</th>
<th>Experiment 1b</th>
<th>Experiment 2</th>
<th>Experiment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple</td>
<td>O</td>
<td>O</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Apricot</td>
<td>A</td>
<td>D</td>
<td>*</td>
<td>E</td>
</tr>
<tr>
<td>Balloon</td>
<td>O</td>
<td>O</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Banana</td>
<td>A</td>
<td></td>
<td>*</td>
<td>E</td>
</tr>
<tr>
<td>Blueberries</td>
<td>B</td>
<td>D</td>
<td>*</td>
<td>E</td>
</tr>
<tr>
<td>Bubbles</td>
<td>A</td>
<td>C</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Carrot</td>
<td>O</td>
<td></td>
<td></td>
<td>E</td>
</tr>
<tr>
<td>Celery</td>
<td></td>
<td>D</td>
<td>*</td>
<td>F</td>
</tr>
<tr>
<td>Cereal</td>
<td>O</td>
<td>O</td>
<td>*</td>
<td>F</td>
</tr>
<tr>
<td>Clover</td>
<td></td>
<td></td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Coffee</td>
<td>O</td>
<td></td>
<td></td>
<td>E</td>
</tr>
<tr>
<td>ColonyRoom</td>
<td>B</td>
<td>D</td>
<td></td>
<td>E</td>
</tr>
<tr>
<td>Gorilla</td>
<td>B</td>
<td>C</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Honeysuckle</td>
<td></td>
<td></td>
<td></td>
<td>E</td>
</tr>
<tr>
<td>Hotdog</td>
<td>O</td>
<td></td>
<td></td>
<td>F</td>
</tr>
<tr>
<td>Jello</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kiwi</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Koolaid</td>
<td>O</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lemonade</td>
<td>B</td>
<td>D</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Lettuce</td>
<td></td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lookout</td>
<td>A</td>
<td>O</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M&amp;M</td>
<td>B</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Melon</td>
<td>A</td>
<td></td>
<td></td>
<td>E</td>
</tr>
<tr>
<td>MushroomTrail</td>
<td>B</td>
<td>O</td>
<td>*</td>
<td>F</td>
</tr>
<tr>
<td>Noodles</td>
<td>C</td>
<td></td>
<td></td>
<td>E</td>
</tr>
</tbody>
</table>
### 3.1.4 Chimpanzee procedure.

In all sessions, words were presented in four randomized blocks, for a total of 96 trials. On each trial, Panzee chose from among four photographs that all corresponded to words being used in that particular session. 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. In sessions when Panzee heard both natural and synthetic words, this feedback helped keep her motivated. Neither feedback sounds nor food rewards were provided on trials with synthetic stimuli.
Initial sessions assessed Panzee’s performance when hearing the 24 test words in natural form. To progress, Panzee was required to perform at or above 70% correct with natural test words for three consecutive sessions (chance was 25%). Panzee then completed both natural and sine-wave sessions with orientation words. Here, she heard eight blocks of 12 natural orientation words for two sessions, averaging 72% correct. Then she heard these words in both natural and sine-wave form for two sessions. In the first, 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 natural test words to refresh her on the test word set, and performed 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.

3.1.5 Human procedure. Pilot experimentation demonstrated that humans were at ceiling performance for all word forms when stimuli were presented using the word recognition method employed with Panzee. Therefore, instead of having participants choose from four photographs to identify the stimuli, a word-recall method was used whereby participants had
to transcribe the sounds they heard. First, however, participants were familiarized with the word set by exposure to the 24 test words presented on a PowerPoint presentation. Photographs of test-word objects were shown, one at a time, while the corresponding, naturally recorded word was heard. Then, the participant was asked to write down the name of each photograph as it was presented without word labels or sounds. Participants were also familiarized with sine-wave speech by listening to a recording of the words “one” through “ten” and then “ten” through “one” in SW123 form. They were instructed to inform the experimenter as soon as they were able to identify these sounds as speech.

In the test session, participants heard the stimuli 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 four sine-wave forms, and the other block included Group B words in natural form and Group A words in the four sine-wave forms. Within a block, natural words were presented for two trials, and sine wave words were presented for one trial each in every sine-wave form. These sessions included a total of 72 trials.

**3.1.6 Data analysis.** Panzee’s mean percentage-correct performance in orientation versus test sessions with natural words was compared using an unpaired t-test for a possible learning effect. Mean percentage-correct performance for each sine-wave word form within and across the six test sessions was compared to 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. Human percentage-correct performance was computed for each word form, with mean performance for the 12 participants then examined separately for natural and sine-wave versions. ANOVA was used to
test for an overall effect of sine-wave 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.

**3.1.7 Results.** Panzee’s mean performance over the three natural word orientation sessions was 73.3% ($SD = 1.58$), 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% ($SD = 4.06$) overall, which also was significantly above chance ($p < 0.001$). An unpaired, 2-tailed $t$-test revealed that Panzee’s performance with natural words was significantly higher in test sessions than in orientations sessions, $t(7) = 3.95, p < 0.01$, as shown in Figure 6a. 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 Figure 6b, Panzee was 36% correct for SW23 and SW13 words, and 59% correct for SW123 and SW12 words. A chi-squared test with a Bonferroni corrected alpha value of 0.025 revealed that correct performance for SW123 and SW12 words was significantly greater than for SW23 ($p = 0.006$) and SW13 versions ($p = 0.004$).
Figure 6. Experiment 1a chimpanzee and human word recognition. a) Mean performance with natural words by Panzee and the human participants, with applicable standard deviations.

b) Panzee’s sine-wave word performance, with chance-level accuracy shown by the dashed line.

c) Mean human sine-wave word performance, with standard error bars.
Mean transcription performance of natural words by humans was 99.8% correct (SEM = 0.06), as shown in Figure 6a. Mean percentage-correct values for SW123, SW12, SW23, and SW13 word forms were 43%, 35%, 31%, and 27%, respectively. A Kolmogorov-Smirnov test for normality validated use of ANOVA, and results revealed a statistically significant difference among the various outcomes, $F(3, 44) = 6.00, p = 0.002$. Furthermore, Tukey post-hoc comparisons showed that outcomes were significantly higher for SW123 stimuli than for SW13 versions, $p = 0.001$, as illustrated in Figure 6c. No other differences were found. Independent, 2-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. For these participants, 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, than on SW23 and SW13 forms. Finally, both the humans and Panzee never recognized the words “banana,” “bubbles,” “orangedrink” and “pineneedle” in several of the SW forms.

3.1.8 Discussion. Panzee demonstrated consistent natural word-recognition performance, showing similar outcomes in both orientation and test sessions as in earlier annual testing synthetic-speech experiments (Heimbauer et al., 2011). Her recognition of SW123 words was also similar to performance in previous sine-wave testing (Heimbauer et al., 2011).
Panzee identified more words in SW123 and SW12 form than in SW13 or SW23 form. Human performed similarly, although only the difference between SW123 (with both SW1 and SW2 present) and SW13 (missing SW2) performance was statistically significant, while the SW123 and SW23 performance difference was not. However, 6 of the 12 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 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 mean human outcome of 43% correct. 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. Although human participants were exposed to and tested with the natural words before hearing the sine-wave forms, they were less familiar with them than was the chimpanzee.

The more important result is that both species showed a statistically significant performance difference between complete sine-wave words (SW123) and the same words when missing the tone analog to F2 (SW2). This result is consistent with the hypothesis that Panzee responds 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 & Rubin, 1990). Both Panzee and humans demonstrated an ability to interpret sine-waves as cues to phonetic content, also suggesting that both were drawing on implicit knowledge of speech acoustics and corresponding phonetics (Davis & Johnsrude, 2007; Mann & Liberman, 1983; Newman, 2006; Whalen & Liberman, 1987). Taken together, these outcomes are indicative of cognitive top-down processing.

3.2 Experiment 1b

Panzee’s ability to identify words in noise-vocoded form also provided an opportunity to examine the cues she is sensitive to in synthetic speech, with corresponding implications for natural speech processing. Although her previous performance has suggested that general auditory processing capabilities may be sufficient for human-like speech perception (Heimbauer et al., 2011), more detailed testing with noise-vocoded words could further strengthen this conclusion. Hence, the purpose of the next experiment was to compare Panzee’s performance with noise-vocoded words with varying degrees of spectro-temporal information to that of humans.

In 1995, Shannon and colleagues found that as the number of bands used to synthesize noise-vocoded phonemes and sentences increased, participants showed corresponding improvements in identification accuracy (see Figure 7). With trained listeners, four noise bands are often sufficient for speech recognition, while at least ten noise bands are necessary with untrained participants (Davis et al., 1995; Shannon et al., 1995). Previously, Panzee demonstrated recognition of familiar words in noise-vocoded form synthesized with seven noise bands (Heimbauer et al., 2011). Therefore, this experiment assessed her word-recognition ability as a
function of the number of noise bands used to produce the stimuli. It was hypothesized that if Panzee uses the same available cues as humans, she would show a similar pattern of performance across those forms. Specifically, her performance was predicted to increase linearly with increasing numbers of noise bands.

Figure 7. Intelligibility of noise-vocoded speech to humans. Percentage-correct performance for eight human listeners identifying consonants, vowels, and sentences as a function of noise-band number in noise-vocoded speech tested by Shannon et al. (1995). The dashed line denotes chance-level accuracy.

3.2.1 Subject. Panzee again was the subject.

3.2.2 Participants. There were 12 human participants (eight females).

3.2.3 Stimuli. Stimuli consisted of 24 previously recorded and processed natural words (see Table 1), which were those that Panzee had best identified in noise-vocoded form in earlier testing (Heimbauer et al., 2011). Noise-vocoded versions varied from two to five noise bands,
and were synthesized using lower- and upper-cutoff frequencies (see Table 2) calculated using the “Greenwood function” (Souza & Rosen, 2009). This function calculates frequency 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; Greenwood, 1990). The approach was used to ensure orderly selection of frequency-cutoff values as they relate to hearing.

Test stimuli consisted of 11 two-syllable words, 11 three-syllable words, 1 four-syllable word, and 1 five-syllable word. All words were chosen from a list of those that Panzee previously successfully identified in noise-vocoded form. Fifteen of these words were also used in Experiment 1a. Twelve additional words, in natural and a previously synthesized form using seven noise bands (NB7), were used during an orientation phase (see Table 1). To produce the various noise-band test stimuli (NB2, NB3, NB4, and NB5), the natural speech signal was divided into 2, 3, 4, or 5 frequency bands using a band-pass filter. 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.

Table 2.

Lower-to-upper cutoff frequencies for noise-band stimuli in Experiment 1b.

<table>
<thead>
<tr>
<th>Bands</th>
<th>Frequency (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>100-1005</td>
</tr>
<tr>
<td>3</td>
<td>100-548</td>
</tr>
<tr>
<td>4</td>
<td>100-392</td>
</tr>
<tr>
<td>5</td>
<td>100-315</td>
</tr>
</tbody>
</table>
3.2.4 Chimpanzee procedure. The testing procedure and reward regimen were the same as those used in Experiment 1a, and all sessions consisted of 96 trials. Panzee first completed three natural word sessions to prepare her for the test sessions, and to ensure normative performance with natural words. Criterion performance to progress to the orientation phases was set at 75% in three consecutive sessions. During orientation, Panzee completed one session with 12 non-test 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.

An additional programming contingency was added in test sessions. Here, words of the same type, meaning natural or noise-vocoded, were not presented more than three times in a row. This adjustment was made to avoid 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 a 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 session types three times each in alternating 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.
**3.2.5 Human procedure.** Several human listeners were first tested in pilot sessions using the same orientation and test procedures as in Experiment 1a. However, as these participants demonstrated high accuracy with test words in all NB forms, the orientation procedure was changed. Experimental participants were instead familiarized with noise-vocoded speech only by listening to a recording of the words “one” through “ten” and then “ten” through “one” in NB7 form. Following this simple orientation, they heard and transcribed one block of the 24 natural words in randomized order. Lastly they heard and transcribed a randomized test block of the same words in each of the four noise-band forms, for a total of 96 trials.

**3.2.6 Data analysis.** Data for both Panzee and the human participants was analyzed as in Experiment 1a.

**3.2.7 Results.** Panzee’s natural word-recognition performance in orientation sessions ranged from 77.2% to 83.3%, with an overall mean of 80.6 % (SD = 3.11), which was statistically above chance level, $p < 0.001$ (see Figure 8a). Percentage-correct on natural words in the six test sessions ranged from 77.1% to 87.5%, with an overall mean of 82.8% (SD = 3.8) correct, which was also significantly above chance level ($p < 0.001$). An unpaired, 2-tailed t-test revealed that Panzee’s 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% to 50% (see Figure 8b), and overall was significantly above chance ($p < 0.001$). Her NB2 word performance was lower at 38% 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.

**Figure 8.** Experiment 1b chimpanzee and human word recognition. a) Mean performance with natural words by Panzee and the human participants. 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 standard error bars.
Human word transcription of natural words was 100% correct (see Figure 8a). Mean percentage-correct values for NB2, NB3, NB4, and NB5 forms were 80%, 78%, 68%, and 38%, respectively (see Figure 8c). 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 performed much as Panzee did. In other words, 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.

3.2.8 Discussion. As in earlier testing, Panzee again demonstrated the ability to reliably identify words in noise-vocoded form. However, her performance was significantly better for words in NB5 form than in corresponding NB2 versions. Humans performed similarly, both in the current work and in comparable earlier studies (Shannon et al., 1995). As expected, increasing numbers of noise bands was associated with higher word-identification performance for both Panzee and the humans, with both species performing as well with NB4 and NB5 forms as in earlier testing with NB7 stimuli (Heimbauer et al., 2011). The results again confirm that noise-vocoded speech based on as few as four noise bands is reliably comprehensible (Souza & Rosen, 2009; Shannon et al., 1995)—in this case for a chimpanzee as well. While humans recognized more NB5, NB4, and NB3 words than Panzee in the current experiment, they
performed no better than she did with NB2 stimuli. Only one of twelve humans was able to identify the three words that were unintelligible to Panzee in this form.

As hypothesized, Panzee’s performance with noise-vocoded words showed evidence of sensitivity to the same cues as human listeners. In both cases, perception was successful in spite of the absence of basic speech features, such as F0 and formant information. Whatever the spectro-temporal cues that remained, this language-trained chimpanzee was able to take advantage of them. As with sine-wave speech, the outcomes are inconsistent with the SiS perspective and instead support the Auditory Hypothesis. Results are also again indicative of top-down processing, with both species evidently making use of previous knowledge of speech acoustics and phonetic categories in interpreting these fundamentally altered, synthetic versions.

3.3 General Discussion

Experiment 1 was designed to investigate the acoustic cues that Panzee may be utilizing when listening to sine-wave and noise-vocoded speech, comparing her outcomes to analogous human performance. As hypothesized, Panzee showed evidence of sensitivity to the same cues in both synthetic forms as humans—stimulus patterning argued to reflect spectro-temporal properties of the original, natural speech signal (Remez et al., 1994). For both sine-wave and noise-vocoded speech recognition, Panzee and current human participants performed better when the synthetic word forms included the attributes shown to facilitate human performance in comparable previous work. Additionally, Panzee’s performance in both Experiments 1a and 1b demonstrated that she can reliably understand speech that has been characterized as missing traditional acoustic cues to phonetic content (Remez et al., 1994). This supports the
view that human specializations for perception are not necessary for speech perception (Kuhl, 1988), and that Panzee is demonstrating top-down interpretation of this impoverished speech input in the same way as humans (Davis et al., 2005; Davis & Johnsrude, 2007; Hillenbrand et al., 2011).

In both experiments, a strategy of matching synthetic versions to holistic properties of known words would make identification difficult or impossible. For example, performance by matching on overall duration would not likely be dramatically affected across the various forms, as gross temporal properties were captured across individual sine-wave components and noise-bands. In addition, both sine-wave and noise-vocoded versions are dramatically altered relative to natural speech, and overall “auditory impressions” are markedly different among the two. The reasonable conclusion is that both Panzee and human listeners were able to perceive the synthetic versions as perceptible speech in spite of their unusual acoustics.

4. EXPERIMENT 2

The manipulations used to investigate acoustic cues to phonetic content in Experiment 1 are, of course, not the only way to approach basic speech-perception problems. One alternative is to examine phonemes as individual segments within the speech signal, which is the rationale behind recent work with time-reversed speech. As discussed earlier, creating this speech form involves reversing short, fixed portions of the signal. Although temporal properties of the waveform are markedly changed by this time-reversal, they have little effect on speech intelligibility as long as window-length is less than 100 ms (Barkat et al., 2002; Saberi & Perrott, 1999)—meaning no longer than the approximate duration of phonetic segments (Crystal & House, 1988). In fact, Saberi and Perrott found that intelligibility of time-reversed speech for English
listeners decreased to 50% threshold level when window-lengths were 130 ms or more (see Figure 9). The goal of Experiment 2, therefore, was to assess Panzee’s ability to recognize time-reversed speech, testing whether she showed similar processing over phoneme-length window lengths. It was hypothesized that Panzee does perceive speech based on phonemic segments, and would demonstrate human-like performance when tested with time-reversed speech. If so, the result would constitute evidence that speech perception mechanisms are based on generalized mechanisms and provide support for the Auditory Hypothesis.

Figure 9. Intelligibility of time-reversed speech to humans. The figure is reproduced from Saberi and Perrott (1999), and shows subjective intelligibility ratings of time-reversed sentences by seven participants. Note that 50% intelligibility occurs at a window-length of approximately 130 ms.

4.1 Subject

The subject was the chimpanzee Panzee.

4.2 Participants

There were 12 human participants (8 females).
4.3 Stimuli

Twenty of Panzee’s previously recorded and processed, three-syllable, natural words were chosen (see Table 1). Three-syllable words were used exclusively to maximize both word length and uniformity among the stimuli, as well as being a control for syllable count as a possible cue to word identification. Each of these words was manipulated by applying time-reversal windows of varying lengths, starting at the beginning of the file and continuing through contiguous sections to the end. Words were reproduced in eight forms, using window lengths of 25 ms (TR25), 50 ms (TR50), 75 ms (TR75), 100 ms (TR100), 125 ms (TR125), 150 ms (TR150), 175 ms (TR175), and 200 ms (TR200). The final time-reversed segment in each sound file was almost always smaller than the nominal window length, a factor that will be discussed later in the Results section.

4.4 Chimpanzee procedure

The computer testing procedure was the same as in Experiments 1a and 1b; however, test sessions now consisted of 80 trials. Panzee received no auditory feedback on individual trials and was rewarded after every three to four trials, independent of performance. This non-contingent reward regimen was used to avoid the possibility of learning effects across the various time-reversed forms. Panzee did not receive any orientation for time-reversed stimuli. Instead, she simply participated in sessions hearing the 20 words in only natural form before testing began. Each of these sessions consisted of four, randomized word blocks, and she was required to perform at a level of at least 70% correct over three consecutive sessions. She reached this criterion after six sessions.
During the first test session, Panzee heard the 20 words one time each in four time-reversed forms—TR50, TR100, TR150, and TR200. In the second test session, on a different day, she heard the same 20 words one time each in the other four forms—TR25, TR75, TR125, and TR175 (see Table 1). Stimuli were randomized within blocks, and Panzee completed both types of sessions four times each, in alternating order. Overall, testing included 4 trials for each of the 20 words in every TR form.

4.5 Human procedure

Two sets of 12 participants were tested. “Group 1” was tested with the same 20 test words used with Panzee in forms TR25, TR75, TR125, and TR175, for a total of 80 randomized trials heard in a single session. “Group 2” was tested similarly, but with the 20 words in forms TR50, TR100, TR150, and TR200.

4.6 Data analysis

Panzee’s percentage-correct performance was computed for each of the eight word forms, both within and across six test sessions, which included three trials for each word in each of the eight time-reversed forms. Two sessions were excluded from data analysis, because Panzee was consistently distracted and disinterested. During one of these sessions, she constantly moved away from the test screen to look out a window at cars coming and going. During another session, she repeatedly moved away from the test screen to ask for different food rewards. Although she eventually completed these two sessions, her performance was less than 60% correct with even the easiest (most natural-sounding) time-reversed words (e.g., TR25 and TR50).
Binomial tests were conducted to compare performance to a chance rate of 25% for each version separately. A Kruskal-Wallis test was used to test for an overall effect among word forms, with Mann-Whitney U tests applied in post-hoc, pair-wise comparisons. The relationship between performance and the time-reversed value form was modeled using linear regression.

In addition to this statistical testing, two threshold values were computed for word intelligibility as a function of window-length. These values were based on the rationale that the threshold represents the halfway point between no perception and perfect perception. While the low boundary could be taken as chance-rate performance at 25% correct, the high boundary was not as clear-cut. A “high” threshold was, therefore, set as the midpoint between 25% and 100% (62.5%), and a “low” threshold was set as relative to 80% (52.5%), and both values were at above chance-rate performance ($p < 0.001$). The latter value represents Panzee’s overall, long-term performance level, based on a historic range of 75% to 85% in natural word recognition (M. J. Beran, personal communication, January 2010), as well as her performance in Experiment 1. Panzee’s 50% threshold performance using these two threshold values was then interpolated based on intelligibility rates at the closest tested reversal-windows.

Human mean percentage-correct performance for each word form was computed for Group 1 and Group 2 separately. ANOVA was used to test for an overall effect with individual word-form results compared using Tukey post-hoc tests. Linear-regression analysis was also applied; and for regression purposes, Group 1 and Group 2 participants were treated as a single sample.
4.7 Results

As shown in Figure 10, Panzee’s correct word recognition for the six sessions ranged from 49% to 63%, with a mean of 56% over all sessions, and was statistically above chance level, \( p < 0.001 \). Performance was also statistically above chance level for all individual word forms. Regression analysis revealed that window length predicted percentage correct, \( \beta = -0.17, p < 0.01 \), and a significant amount of the variance, \( R^2 = .80, F(1,6) = 23.7, p < 0.01 \).

A Kolmogorov-Smirnov test showed that the data were not normally distributed. A Kruskal-Wallis test was conducted, and revealed an overall performance difference among the TR word forms, \( X^2(7) = 16.7, p = 0.019 \). Because Panzee’s performance was less than 60% correct for TR125, TR150, TR175, and TR200 words, a mean comparison Mann Whitney U test was conducted between TR125 and TR175 to test for a potential difference in intelligibility performance, similar to that reported for humans by Saberi and Perrott (1999). Results did, in fact, reveal a significant difference, \( p < 0.05 \). Panzee’s high and low 50%-intelligibility threshold points were 92.5 ms and 141.0 ms, respectively.
Figure 10. Experiment 2 chimpanzee and human word recognition. The top figure shows Panzee’s time-reversed word performance, corresponding regression line, and chance-level accuracy. Both “high” and “low” thresholds are also noted (see text for further details). The bottom figure shows mean time-reversed word performance for humans, with standard error bars and the regression line. The white line on the TR200 bars represents percentage-correct values for Panzee and human participants when words with final window-lengths of 130 ms or less were excluded from analysis. Asterisks denote statistical significance: * = p < 0.05, ** = p < 0.001.

Mean percentage-correct was calculated for humans for each TR value, and ranged from almost 100% correct for TR25 to 23% for TR200 (see Figure 10). Regression analysis revealed that reversal window-length was a strong statistical predictor of percentage correct, $\beta = -0.45,$
\( p < 0.001 \), and accounted for almost all of the associated variance, \( R^2 = .97, F(1,6) = 200.3, \)
\( p < 0.001 \) (see Figure 10). After a Kolmogorov-Smirnov test confirmed normality, ANOVA of
Group 1 data showed an overall effect among TR25, TR75, TR125, and TR175, \( F(3, 44) = 62.3, \)
\( p < 0.001 \). Tukey post-hoc comparison tests revealed statistically significant differences
between TR25 and TR75, and TR75 and TR125 (both, \( p < 0.001 \)). However, there was no
performance between TR125 and TR175 forms. After again confirming normality, an ANOVA
showed an overall effect among TR50, TR100, TR150, and TR200 Group 2 performance,
\( F(3, 44) = 61.2, p < 0.001 \). Tukey post-hoc comparisons revealed statistical differences between
TR50 and TR100, and TR100 and TR150 (both, \( p < 0.001 \)), but not between TR150 and TR200.
The intelligibility threshold value calculated for humans was 121.6 ms.

4.8 Discussion

In this experiment, Panzee again demonstrated proficiency with words in natural form
during orientation sessions. More importantly, her performance revealed her ability to identify
words in time-reversed form. Time-reversed speech has been described as “the most drastic
form of time scale distortion” (Licklider & Miller, 1960). Despite the dramatic changes involved,
Panzee, like human participants, recognized speech in many of these forms. In addition, the
decline in recognition performance, expected as the reversal manipulation comes to affect
adjacent phonemes, occurred similarly in Panzee and the humans.

Both Panzee’s and the humans’ performance revealed that time-reversal window length
significantly predicted percentage correct outcomes. As in Saberi and Perrott’s (1999) work
with sentences, word intelligibility was below 50% at reversal-window lengths of approximately
130 ms for both Panzee and human participants. Using a maximum possible performance of
100%, her 50% intelligibility threshold was just above 90 ms. Setting that level at 80%, her threshold was at 141 ms. Both values are similar to the current mean threshold value of 122 ms for humans, as well as Saberi and Perrott’s (1999) outcome of approximately 130 ms. Although Group 1 participants did not demonstrate a significant decrease in performance between TR125 and TR175, they did show a difference between TR75 and TR125 (81% and 48%, respectively). Group 2 participants showed a significant decrease between TR100 and TR150 (63.3% and 33.8%, respectively).

Although Panzee’s pattern of decreasing performance from TR25 to TR200 was somewhat different than mean human performance, one participant showed exactly the same pattern as Panzee for TR25, TR75, TR125, and TR175 word forms. This participant performed at 100% on TR25 and TR75 words, and performance for TR125 (50%) was less than at TR75 (15%). Furthermore, both Panzee and humans found it difficult to identify “tomato” and “potato” for TR lengths of 125 or more, with Panzee only doing so in one instance (“potato,” in TR200 form). Humans showed a total of only 29% correct trials for these two words in the four longer-window forms.

Unexpectedly, Panzee demonstrated better performance on TR200 words than with either TR150 or TR175 forms. While earlier work has only used sentence stimuli (Saberi & Perrott, 1999), this work used word stimuli, which are shorter in duration. It may be that the difference between using sentence- and word-length stimuli had a minor effect on the current results. For 12 of the 20 words the final reversal window encompassed only 130 ms or less when the words were in TR200 form. This final segment could, therefore, have provided a stronger indication of word identity than intended for both Panzee and the human listeners. In
all, 18 of Panzee’s 28 (64.3%) correct responses to TR200 words were for these 12 items. In humans, the 12 words accounted for 38.9% of the correct TR200 trials. Excluding these words from analysis produced 13.9% and 13.8% correct for Panzee and the humans, respectively. These outcomes are lower than for any other window length (see Figure 10).

If the hypothesis that time-reversed speech “works” because reversal within a phoneme-length window does not affect processing, then Panzee is evidently listening to phoneme-length segments. If phonemes are typically considered to be 50- to 100-ms long (Crystal & House, 1988), then Panzee did well at those window lengths. It was not clear exactly how to set Panzee’s 50% intelligibility threshold. However, the high and low points that were chosen produced values similar to human threshold. Using a very different rationale than was used for producing sine-wave and noise-vocoded speech, Panzee’s performance suggests that she segments speech based on the same phoneme-based organization as humans, and may also hear words as a sequence of phonemes. The results indicate that detailed auditory analysis of the short-term acoustic spectrum is not essential. Rather, the amplitude envelope is more likely important, and short time-reversal windows do not appear to significantly impair speech perception of the phoneme information within the window. However, performance declines when windows are long enough to cause overlap of phoneme-length information. This decrease reveals that for both species, extensive alteration of the spectro-temporal cuing can critically affect performance.

5. EXPERIMENT 3

In addition to the questions regarding how specific acoustic elements contribute to lexical information, there are more general perceptual problems that arise. The lack of
invariance problem, as described earlier, reflects the high variability in speech acoustics that results from the different physical and physiological properties of individual talkers (Pisoni, 1997). 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 speech. Children’s speech is typically characterized by higher pitch and formant frequencies, especially for vowels (Gerosa, Giuliani, & Brugnara, 2007; Lee, Potamianos, & Narayanan, 1999). In addition, children under the age of seven typically have longer phone duration and larger spectral and temporal variability in consonants and consonant-vowel transitions than older children and adults (Gerosa, Lee, Giuliani, & Narayanan, 2006). It is these characteristics that can often make child speech more difficult to understand than adult speech.

Talker normalization emerges early in human ontogeny, as shown by 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 is one whereby the listener strips 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 (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).

Although there are different views regarding how talker normalization occurs, this phenomenon has not been strongly linked to either SiS or Auditory Hypothesis positions. On the one hand, the difficulty of talker normalization could be interpreted as evidence for the SiS argument. On the other hand, some evidence favors the Auditory Hypothesis. For example, some non-primates, such as chinchillas and birds, have demonstrated the ability to categorize speech sounds and monosyllabic words and then generalize performance to at least a small number of unfamiliar talkers (Kuhl & Miller, 1975; Ohms et al., 2010). Budgerigars have demonstrated that they are less sensitive to vowels from different talkers than to vowels between categories (Dooling & Brown, 1990). There is also indirect evidence that apes normalize speech across talkers. For example, Panzee and several language-trained bonobos interact with at least a dozen humans on an everyday basis, and react appropriately to speech commands and requests (M. J. Beran, personal communication, January 2010).

However, the empirical nonhuman normalization studies have typically involved only a few talkers and short, often rudimentary sounds. Thus, Experiment 3 was designed to address talker normalization more systematically in Panzee, providing an opportunity both to test her with speech from a large number of individuals, and to present more complex, lexically meaningful stimuli. As before, the hypothesis was that she would be similar to humans in
demonstrating normalization, because she has heard speech from many different talkers through her life and currently responds to English spoken by various individuals. However, both earlier work (Heimbauer et al., 2011) and Experiments 1 and 2 involved the speech of only talker, MJB. Specific information regarding her potential talker normalization abilities, therefore, is not available. Experiment 3 included both familiar and unfamiliar adult talkers, with a variety of dialect backgrounds. Words from young children were also presented. Panzee has been exposed to children’s voices much less frequently, especially later in her life, either in experiments or in informal interactions. Performance with these talkers was compared to recognition of speech from talker MJB, who is arguably one of the talkers Panzee is most familiar with.

5.1 Subject

The subject was the chimpanzee Panzee.

5.2 Participants

No human listeners were tested. However, audio-recording included a total of 31 different native-English speakers including 21 adults and 10 children. These talkers included the familiar talker MJB, 5 familiar males, 5 unfamiliar males, 5 familiar females, 5 unfamiliar females, and 5 boys and 5 girls (all unfamiliar). The age range of adult talkers was 20 to 72 years old, and the age range for children was 4 to 7 years old.

5.3 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 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 all 31 talkers. MJB and 20 additional native English-speaking adults were recorded reading the individual words from index cards. Ten native English-speaking 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. The 30 new talkers were grouped as familiar adult males (FAM), familiar adult females (FAF), unfamiliar adult males (UAM), unfamiliar adult females (UAF), unfamiliar male children (UCM), and unfamiliar female children (UCF). MJB was re-recorded for this experiment using the same equipment that was used to record the other 30 talkers.

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 were 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 were 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.
5.4 Chimpanzee procedure

Panzee was tested for a total of 14 sessions, each of which included 80 trials. In the first session, she heard 16 test words (Group E) spoken by MJB. In the next six sessions she heard Group E words spoken by the five talkers within each of the specific talker-type groups. The session order of talker types for testing was: FAM, UAM; FAF; UAF; UCM; UCF. In the eighth session, Panzee heard the remaining 16 test words (Group F) spoken by MJB in a session, which was followed by six sessions with Group F 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; UAF. In this experiment, Panzee chose from four lexigrams (see Figure 11), instead of four photographs. This change was made to be able to eventually compare resulting data to an earlier, unpublished experiment that also used lexigrams. Panzee received auditory feedback on every trial, and was rewarded for all correct responses. This reward schedule kept her highly motivated, and could be used because each trial was unique.

![Figure 11](image)

Figure 11. Samples of lexigrams used in Panzee’s spoken-word recognition task.

5.5 Data analysis

Panzee’s data were analyzed as in Experiment 2.
5.6 Results

As shown in Figure 12, Panzee’s mean correct-trial performance, was calculated for each talker and averaged for the two sessions for each different talker type, ranged from 75.6% (MJB, UCF) to 81.3% (FAM). Word recognition was significantly above chance level for all talker types, $p < 0.001$. 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 performance for all talker-types was similar to annual performance levels (see Figure 12), 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). Results revealed no overall difference in performance among the collapsed categories, $\chi^2(4) = 0.58, ns$.

![Bar chart showing percentage correct for different talker types](chart.png)

**Figure 12.** Experiment 3 chimpanzee word-recognition performance across talkers. Talker-type groups were as follows: MJB is the familiar male researcher, FAM is other familiar adult males,
FAF is familiar adult females, UAM is unfamiliar adult males, UAF is unfamiliar adult female, UCM is unfamiliar boys, and UCF is unfamiliar females. The dashed line signifies chance-rate performance.

5.7 Discussion

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. Infants demonstrate at least some talker normalization ability by the age of ten and a half months (Houston & Jusczyk, 2000), and 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 (Dooling & Brown, 1990; Kuhl & Miller, 1975; Loebach and Wickesberg, 2006; Ohms et al., 2010), talker normalization has not previously been systematically investigated in nonhumans.

In the current experiment, Panzee was tested more deliberately, 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 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
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. Her performance provides evidence in support of the Auditory Hypothesis, rather than the SiS view.

6. GENERAL DISCUSSION

Insights into the generality of the auditory and cognitive processes involved in speech perception are fundamental to resolving the SiS versus Auditory Hypothesis debate. Although arguments by Galantucci et al. (2006) and others almost rule out the possibility of meaningful animal experiments, comparisons between humans and nonhumans are a necessity. Mammalian and close primate relatives are of greatest interest, and the current work demonstrates that comparisons with these animals can be very informative. While many animal studies have investigated perception of rudimentary speech sounds, none have previously presented meaningful words, synthetic versions of those items, or tested a wide range of different talkers. The current work was possible specifically because of Panzee’s language-comprehension abilities (Beran et al., 1998; Brakke & Savage-Rumbaugh, 1995; Heimbauer et al., 2011; Rumbaugh & Savage-Rumbaugh, 1996).
6.1 Current results

Experiment 1 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 also attends more to these particular tones in the synthetic speech, with implications for sensitivity to the corresponding formants in natural speech. Similar results occurred with noise-vocoded speech, showing that both Panzee and humans performed best with stimuli that included four or five noise bands, less well with three, and poorly with two. These outcomes match earlier findings with noise-vocoded speech, showing differences in relative intelligibility of these synthesis forms by humans.

As discussed earlier, it is difficult to specify exactly which acoustic cues are critical in these types of synthesized speech, or what both have in common (Remez et al., 1994). However, 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). Experiment 2 tested temporal cuing differently, using a synthetic speech form Panzee had no experience with. The time-reversed speech used poses a unique perceptual problem in that it disrupts acoustic patterning over time. Intelligibility of time-reversed speech depends on the length of reversed segments relative to typical phoneme length. Differential performance has been attributed to the fact that phonetic content will be relatively undisturbed if the length of the reversal window is within the time frame of typical phoneme length, which is estimated to be
50 to 100 ms (Crystal & House, 1988). Data from both Panzee and the humans reveals a strong relationship between window-length and percentage correct identification, with similar intelligibility thresholds in each case. It was concluded that Panzee was sensitive to phoneme-related cues in time-reversed, as well as in natural, speech.

Finally, Experiment 3 tested Panzee’s ability to solve the lack of invariance problem created in identifying speech across a variety of talkers, including both familiar and unfamiliar individuals, adults and children, and a variety of dialects. Despite the high acoustic variability of human voices (Evans & Iverson, 2004; Hillenbrand et al., 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.

6.2 Implications of experimental results

It is evident from Panzee’s performance in the three experiments that a language-trained chimpanzee can provide a unique animal model for investigating speech perception. Overall, these experimental results contradict the SiS perspective and support the Auditory Hypothesis. In addition, results reveal the likely speech perception capabilities of an ape-human common ancestor.

6.2.1 The SiS view versus the Auditory Hypothesis. The SiS view proposes that humans possess evolutionary specializations for speech perception in the form of a speech module or at least a speech mode of perception (Mann & Liberman, 1983; Trout, 2001). In contrast, the Auditory Hypothesis claims that general auditory capabilities are sufficient to process speech in the absence of uniquely human specializations (Kuhl, 1988)—at least with the necessary
experience with speech input. Many previous animal studies have investigated speech-perception issues based on rudimentary speech elements, and overall results have favored the Auditory Hypothesis. However, the current work and earlier data from experiments testing Panzee (Heimbauer et al., 2011) argue much more strongly in the same direction. In addition to showing evidence of using the available spectro-temporal cues in sine-wave and noise-vocoded speech as humans do, Panzee has also now demonstrated apparent attention to phoneme organization, and a level of talker normalization that arguably solves the lack of invariance problem—at least for individual words.

These outcomes make the SiS view (Liberman & Mattingly, 1989; Mann & Liberman, 1983; Whalen & Liberman, 1987) appear highly unlikely—at least in any strict interpretation. For example, the SiS approach to modularity argues that this specialization is innate, and evolutionarily unique to humans. In this perspective, Panzee cannot have such a module; but she still shows evidence of processing speech as if she does. If a speech module does exist in humans, it is likely to have emerged based on latent speech-perception capabilities that were already present in the common ancestor of humans and chimpanzees, rather than from scratch. The claim of a speech mode of perception is less extreme. A speech mode does not mean that speech-perception capabilities are innate or uniquely human, only that the listener be able to learn important properties of speech from experience. Such experience is routine in human development, and impossible in exactly the same form in any other animal. However, Panzee did receive consistent exposure to human speech almost from birth, and her demonstrated speech-perception capabilities highlight the critical importance of such experience in the context of the SiS versus Auditory Hypothesis debate.
6.2.2 Top-down processing and speech perception experience. Panzee’s abilities also provide evidence pertaining to top-down processing in speech perception. Each of the four tasks presented to her require top-down processing, or at least is considered to in the context of human speech perception. 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. Based on her performance, it is more likely that the 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 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 & Desjardins, 1995).

Perceiving in a speech mode may be more a matter of experience than innateness. Each of the tasks Panzee was able to perform based on top-down processing may also represent a form of speech-mode perception. While some classic demonstrations of the hypothesized speech mode are not applicable to her, Panzee’s ability to perceive fundamentally altered synthetic stimuli as having lexical properties, to compensate or correct for distortions introduced by time-reversal windows, and to map between variable speech acoustics and phonetic features are speech-mode, and top-down, functions.

Panzee’s speech-perception abilities may 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 nonhumans, 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 now 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.

These speech-perception abilities are also a testament to Panzee’s language-rich and enculturated rearing history (Rumbaugh & Savage-Rumbaugh, 1996). A few other apes have been similarly raised, including the now-adult bonobos Kanzi and Panbanisha. These animals have shown strong evidence of understanding spoken English words (Beran et al., 1998; Brakke & Savage-Rumbaugh, 1995; Rumbaugh & Savage-Rumbaugh, 1996), including when ordered syntactically in meaningful sentences (Savage-Rumbaugh et al., 1993). In contrast, apes raised using the same lexigrams and photographs to communicate but without early or extensive functional speech input do not show any notable ability to identify spoken words when tested annually (M. J. Beran, personal communication, January 2010).

6.3 Cognitive processing and language

Although Panzee can recognize familiar words in natural and synthetic versions, she is unlikely to be processing speech exactly as humans do. For instance, she is clearly less efficient than humans with speech in any of the forms tested. Her best performance never approached routine levels of speech recognition in humans, as demonstrated in Experiment 3 when she
never performed above 84% correct with exclusively natural words. There is likely an effect of relatively unimportant factors, such as Panzee may have experienced frustration or boredom by repeatedly being tested with the same limited set of words and with potentially difficult versions presented unpredictably. Although Panzee recognizes the spoken words corresponding to lexigrams and photographs, her inability to produce speech herself sets her apart from all human talkers. In humans, critical aspects of speech knowledge continue to develop throughout childhood, with an individual’s own speech production perceptual knowledge playing a role. Growing awareness of the phonology of language similarly increases both recognition of and manipulation of speech sounds (Goswami, 2006, 2008). While Panzee’s performance demonstrates that speech production is not necessary for speech perception, production may nonetheless be a critical component of human speech processing.

Although Panzee’s performance supports the Auditory Hypothesis and a general auditory model of speech perception, the specific cognitive processes involved are largely unknown. Humans are proposed to have a “phonological loop” in working memory that stores, rehearses, and manipulates auditory input in speech form (Baddeley & Hitch, 1974). No information is available concerning Panzee’s working-memory capabilities, and she may be processing speech in phonological rather than purely auditory form in short-term storage. Baddeley, Gathercole, and Papagno (1998) have suggested that the phonological loop in humans might serve as a language-learning device with an integral role for both spoken and written language acquisition. It is possible that differences between the speech perception abilities of Panzee and those of humans may be due to differences in the development of working memory and possibly even other cognitive processes. Similarly, it is not known how
Panzee maps the words she hears onto corresponding meanings in long-term memory. Differences between chimpanzee and human long-term memory may be responsible for the decrease in her efficiency when responding to word recognition tasks, without reflecting important discrepancies in processing and categorizing the speech sounds themselves.

Panzee may even be at a disadvantage simply due to being unable to take language use to the level of reading. Brain-imaging studies investigating brain development in literate versus illiterate adults provide evidence that reading influences brain structure, specifically being correlated with increased white matter and connectivity in the left hemisphere. In the specific areas involved—the corpus callosum, inferior parietal regions, and parieto-temporal regions—are all involved in reading and verbal working memory. Illiterate adults are often more right lateralized and do not show corresponding white matter and connectivity effects. These results indicate learning to read is linked to brain plasticity and aids in development of left-lateralization (Carreiras et al., 2009; Petersson, Silva, Castro-Caldas, Ingvar, & Reis, 2007). Musical training also leads to stronger left lateralization of the perisylvian brain areas associated with language (i.e., Broca’s and Wernicke’s regions; Limb, Kemeny, Ortigoza, Rouhani, & Braun, 2006), and immature neural responses to rhythmic cues in dyslexic children potentially impede speech development (Goswami, 2006). Increases in overall left lateralization during development may be important for speech-perception efficiency as well.

Other experiments by Petersson and colleagues (2007) demonstrate that literate versus illiterate listeners may engage in different types of cognitive processing when listening to and repeating speech. They proposed that while literate listeners relied solely on language processing, illiterate participants in their study also engaged in visual-spatial processing. Speculatively,
a language-trained ape showing less hemispheric lateralization may also be utilizing more visually based processing and “seeing” lexigrams when hearing English words, instead of focusing on their acoustic properties. Panzee may be processing speech differently because of a lack of the experience that humans obtain by learning to read, which then allows them to take advantage of brain plasticity for enhancement of perception (Carreiras et al., 2009; Petersson et al., 2007). In this case, it may be that development of a specific cognitive ability affects the development of another. All the humans tested for comparison to Panzee may have had an advantage due to a combination of genetic and environmental factors contributing to functional hemispheric specialization (Petersson et al., 2007), which is likely the case for many human language abilities.

6.4 Future directions

The current work has only scratched the surface with respect to potential language-related experimental work with Panzee. For example, additional experiments conducted with both Panzee and other language-trained chimpanzees could provide information about speech-perception and the underlying mechanisms involved, and could contribute to the discussion of the evolution of associated cognitive processes. As a follow-up to Experiment 1, experiments with Panzee and humans could be conducted to investigate mapping of spectro-temporal acoustic cues onto phonetic features and to lexical identity more precisely. Specifically, the effects of small changes at points in the stimuli that are critical for perception by humans could be compared to comparable changes made at non-critical points.

Results of Experiment 2 also provide opportunities for investigating more detailed aspects of Panzee’s speech perception. As demonstrated, time-reversed speech became more
distorted and unintelligible as reversal-window length increased. It appears that Panzee, like humans, has had adequate early developmental experience with language to be able to use phoneme-length information to perceive speech. However, Saberi and Perrott (1999) also reported that perception of time-reversed speech was robust to temporal shifts made to entire segments. This manipulation created apparent reverberation in the speech sounds, but did not importantly disturb phonetic perception. Future experiments could investigate Panzee’s ability to recognize time-reversed speech using this “delay method” to map the robustness of the top-down processing mechanisms she utilizes.

Panzee’s talker-normalization abilities can also be followed up on. Results of this experiment did not shed much light on how normalization occurs, although the topic is still being debated in the speech research community. Here, Panzee may be storing talker-specific, speech-sound variants, as human do (Bradlow, Nygaard, & Pisoni, 1999; Creel & Tumlin, 2009; Sumner, 2011), and then using these representations to generalize to novel, unfamiliar instances of these sounds. Experiments with Panzee could introduce her to new words spoken either by multiple talkers or a single individual. After word learning, her ability to perceive those items from novel talkers, as synthetic replicas, or in altered and reduced versions would reveal more details about how she remembers and represents phonetic information.

Finally, Panzee’s possible use of visual cues to speech has never been investigated. Speech perception is typically a multi-modal event, with cues available both from the acoustic signal and corresponding talker articulation movements. Attending to these visual cues is known to aid human speech perception in noisy environments (Rosenblum, 2005), and facilitates pre-linguistic deaf children in acquiring some aspects of phonological awareness.
through lip reading (Dodd & Hermelin, 1977). Visual cues also help sighted children in learning to distinguish the functional units of spoken language, while the absence of this information is a detriment to language learning in blind children (Mills, 1987).

Patterson and Werker (2003) have argued that the ability of even young infants to integrate auditory and visual phonetic information is evidence of uniquely human speech mechanisms, a proposal that is well suited to testing with Panzee. Izumi and Kojima (2007) have provided suggestive evidence by testing a chimpanzee with auditory and visual information from vocalizing conspecifics. Their study was limited in the number of vocalizations and visual stimuli that could be presented—limitations that are not applicable in Panzee’s case. In fact, she could be tested quite extensively for evidence of the integration of auditory and visual information using a variety of speech sounds, articulatory movements, and talkers.

These are just a few of the questions relating to speech perception capabilities that future research with a language-trained chimpanzee such as Panzee could address. Future studies could investigate a wide range of pertinent issues regarding the details of both perceptual and other cognitive aspects of speech processing, shedding light not only on her particular abilities, but also on those of humans and ancestral apes. The most compelling motivation for pursuing such work, however, may be that Panzee is one of a very small number of animals with these unique speech perception abilities and, therefore, should be involved in such research to the greatest extent possible.

REFERENCES


