Published in final edited form as:
Ear Hear. 2013 ; 34(3): 313–323. doi:10.1097/AUD.0b013e31826fe79e.

Effect of Speaking Rate on Recognition of Synthetic and Natural Speech by Normal-Hearing and Cochlear Implant Listeners

Caili Ji[1,2], **John J. Galvin III**[2], **Anting Xu**[1], and **Qian-Jie Fu**[2]

Caili Ji: caili_ji@yahoo.com; John J. Galvin: jgalvin@hei.org; Anting Xu: xu11110000@yahoo.com.cn; Qian-Jie Fu: qfu@hei.org

[1]Qi Lu Hospital, Shandong University, Jinan, People's Republic of China

[2]House Research Institute, 2100 West Third Street, Los Angeles, California

Abstract

Objective—Most studies have evaluated cochlear implant (CI) performance using "clear" speech materials, which are highly intelligible and well-articulated. CI users may encounter much greater variability in speech patterns in the "real-world," including synthetic speech. In this study, we measured normal-hearing (NH) and CI listeners' sentence recognition with multiple talkers and speaking rates, and with naturally produced and synthetic speech.

Design—NH and CI subjects were asked to recognize naturally produced or synthetic sentences, presented at a slow, normal, or fast speaking rate. Natural speech was produced by one male and one female talker; synthetic speech was generated to simulate a male and female talker. For natural speech, the speaking rate was time-scaled while preserving voice pitch and formant frequency information. For synthetic speech, the speaking rate was adjusted within the speech synthesis engine. NH subjects were tested while listening to unprocessed speech or to an 8-channel acoustic CI simulation. CI subjects were tested while listening with their clinical processors and the recommended microphone sensitivity and volume settings.

Results—The NH group performed significantly better than the CI simulation group, and the CI simulation group performed significantly better than the CI group. For all subject groups, sentence recognition was significantly better with natural than with synthetic speech. The performance deficit with synthetic speech was relatively small for NH subjects listening to unprocessed speech. However, the performance deficit with synthetic speech was much greater for CI subjects and for CI simulation subjects. There was significant effect of talker gender, with slightly better performance with the female talker for CI subjects and slightly better performance with the male talker for the CI simulations. For all subject groups, sentence recognition was significantly poorer only at the fast rate. CI performance was very poor (~10% correct) at the fast rate.

Conclusions—CI listeners are susceptible to variability in speech patterns due to speaking rate and/or production style (natural vs. synthetic). CI performance with clear speech materials may over-estimate performance in real-world listening conditions. The poorer CI performance may be due to other factors besides reduced spectro-temporal resolution, such the quality of electric stimulation, duration of deafness, or cortical processing. Optimizing the input and/or training may improve CI users' tolerance for variability in speech patterns.

Keywords

cochlear implant; speaking rate; synthetic speech

Send Correspondence to: Anting Xu, M.D., Qi Lu Hospital, Shandong University, Jinan, People's Republic of China, xu11110000@yahoo.com.cn.

INTRODUCTION

The cochlear implant (CI) has restored hearing sensation to many individuals with severe to profound sensorineural hearing loss. CI users are capable of good speech understanding under optimal listening conditions such as speech presented in quiet (Spahr & Dorman 2004). However, CI users have difficulty understanding speech in noise (Shannon et al. 2004; Fu & Nogaki 2005), or speech produced by different talkers (Chang & Fu 2006; Liu et al. 2008). CI users encounter great variability in speech patterns in the "real-world," including various speaking rates, emotional speech, accented speech, telephone speech, and synthetic speech. However, CI performance is typically evaluated using clear speech materials, which are highly intelligible and well-articulated. Examples of clear speech materials used in CI studies include CNC words (Peterson & Lehiste 1962), BKB sentences (Bench et al. 1979), HINT sentences (Nilsson et al. 1994), and AZ-Bio sentences (Spahr et al. 2004). Although clear speech may offer good stimulus control, it may greatly overestimate CI performance outside the listening booth.

Different speaking styles, such as clear speech, conversational speech, or accented speech, can significantly affect speech understanding. Clear speech, as opposed to conversational speech, is characterized by slower speaking rates, an expanded vowel space, better temporal modulation and fundamental frequency (F0) variation. Clear speech has been shown to be more intelligible than conversational speech for a variety of listening conditions (e.g., quiet, noise, and reverberation) in normal hearing (NH) listeners and hearing impaired listeners (Picheny et al. 1985; Payton et al. 1994; Bradlow & Kraus 2003). Liu et al. (2004) measured speech reception thresholds (SRTs), defined as the signal-to-noise ratio (SNR) that produces 50% correct sentence recognition, using clear and conversational speech. CI and NH subjects were tested while listening to unprocessed speech and NH subjects were tested while listening to acoustic CI simulations. Acoustic analysis revealed a deeper temporal envelope and slower speaking rate for clear speech than for conversational speech. Results showed that SRTs were 3.1 dB (NH), 3.2 dB (CI simulation), and 4.2 dB (CI) better with clear speech than with conversational speech.

Many listeners regularly encounter synthetic speech in everyday experiences. Text-to-speech (TTS) synthesis converts arbitrary text into audible speech. There are many applications of TTS, including telephone voice service, car navigation information, and audio books. Increasingly, customer service over the telephone involves listening to and responding to TTS prompts. TTS synthesis models generally consist of two parts: (1) sound generation, which handles voiced and unvoiced vowels and consonants, and (2) voice control. The TTS output is aimed at being highly intelligible and natural sounding. Logan et al. (1989) measured NH subjects' speech recognition using natural speech or speech produced by 10 different TTS systems. Results showed that recognition of synthetic speech was consistently poorer than recognition of natural speech, and that there were significant differences in speech intelligibility across the different TTS systems.

Although speech intelligibility with synthetic speech has improved, speech quality generally remains poor, especially in terms of replicating the prosody of natural speech. The poor prosodic quality can negatively affect intelligibility of synthetic speech (Paris et al. 2000). Bunton and Story (2009; 2010) reported that vowel confusions typically occurred across vowel categories when speech was synthesized using a wave-reflection type of vocal tract model coupled to a voice source. Although recognition of synthetic speech has been extensively studied in NH listeners, little is known about CI users' understanding of synthetic speech.

Variability in speaking rate can also affect speech understanding. Many acoustic properties vary with speaking rate, such as vowel or consonant duration, duration of adjacent syllables or phonemes, and transition duration between stops and glides, and F0 range (Lisker, 1957; Miller & Baer 1983; Kohler, 1986; Miller 1987; Crystal & House, 1988; Newman & Sawusch, 1996). Fast-rate speech contains fewer prosodic units than normal or slow rate speech, due to shorter pauses between or within sentences (Lass, 1970). Many previous studies with NH listeners show poorer speech understanding at relatively slow or fast speaking rates (Picheny et al., 1989; Uchanski et al., 1996; Krause and Braida, 2002), In these studies, normal clear speech was uniformly or non-uniformly time-scaled to obtain the target rates, which can produce deleterious artifacts.

Little is known regarding the effect of speaking rate on CI users' speech understanding. Liu and Zeng (2006) found a 2–3 dB advantage in NH listeners' SRTs with clear speech vs. conversational speech, even at fast (time-scaled) speaking rates. Recently, Li et al. (2011) measured the recognition of naturally produced Mandarin sentences with slow (2.5 words per second, or wps), normal (3.7 wps) and fast speaking rates (5.7 wps) in Mandarin-speaking CI users. Results showed that CI performance gradually worsened with increasing speaking rate. In synthetic speech, the speaking rate can be easily modified. At present, little is known about CI users' understanding of synthetic speech at different speaking rates.

In this study, NH and CI listeners' sentence recognition was measured with naturally produced or synthetic speech, presented at slow, normal or fast speaking rates. Natural speech at the normal rate was uniformly time-scaled to obtain the slow and fast rates. The synthetic speech rate was directly modified within the TTS engine to obtain the target rates. To see any talker gender effects, sentence recognition was measured with one male and one female talker at each rate, with both natural and synthetic speech. CI subjects were tested while listening with their clinical processors. NH subjects were tested while listening to unprocessed speech or to an acoustic CI simulation.

MATERIALS AND METHODS

Subjects

Ten CI users (3 male and 7 female) and 14 NH listeners (7 female and 7 male) were recruited for this study. All CI users were post-lingually deafened adults, with an average age of 65.2 years (range: 24–81 years). CI subject demographic information is shown in Table 1. NH subject thresholds were lower than 15 dB HL at octave frequencies from 250Hz to 8000 Hz in both ears. The average age of the NH subjects was 40.1 years (range: 20–57 years). Because of time constraints, two of the NH subjects were able to participate in only one of the processor conditions. As such, there were 13 NH subjects in the unprocessed and CI simulation test conditions. All participants were native speakers of American English. All subjects were paid for their participation in the study. Informed consent was obtained from each subject prior to participation, in accordance with the local Institution Review Board standards.

Materials

Word-in-sentence recognition was measured in quiet using IEEE sentences (Rothauser et al. 1969). IEEE materials consisted of 72 lists of sentences of moderate difficulty. Sentences were naturally produced (1 male and 1 female talker) or synthesized (1 male and 1 female talker) using the Natural Voices TTS engine 1.2.1 (AT&T Lab). The Natural Voices TTS engine uses pre-recorded speech segments for "unit selection synthesis," in which the speech segments are divided into half-phones and then categorized, allowing for robust diphone- and phone-based synthesis, and mixtures thereof. Unit selection synthesis allows for

intelligibility comparable to that of diphone synthesis while increasing the "naturalness" of the synthetic speech (Conkie 1999; Beutnagel et al. 1999).

Speech was presented at three different rates: slow (half-speed), normal (as recorded/synthesized), and fast (double-speed). Table 2 shows the mean F0 and speaking rates (across 720 sentences) for the natural and synthetic talkers. For natural speech, the slow- and fast-rate stimuli were generated using the "time stretch" algorithm in Adobe Audition (Moulines and Laroche 1995; Liu & Zeng 2006). Naturally produced sentences were either time-compressed (for the fast rate) or time-expanded (for the slow rate) while maintaining fundamental frequency (F0) and formant frequency information. In the algorithm, the input speech was first uniformly divided into short time signals. Next, these short-term signals were removed or duplicated, depending on the target rate. The splicing frequency was optimized to preserve the talker F0. Finally, the modified signals were added together in sequence to obtain the time-compressed or – expanded signals. The rate of the synthetic speech was reduced or increased to obtain the targeted fast and slow speaking rates within the adjustment in the speech synthesis engine (AT&T Natural voices™ Text-To-Speech Engines: System Developer's Guide, 2001). The rate adjustment values were -6 (half the default rate), 0 (the default rate), and +6 (twice the default rate).

NH subjects were tested while listening to unprocessed speech (NH group) or to an 8-channel, sinewave-vocoded, acoustic CI simulation (CI simulation group). In the simulation, the input speech signal was first processed through a pre-emphasis filter (+6 dB/octave above 1200 Hz). The input acoustic frequency range (200–7000Hz) was then divided into 8 frequency-analysis bands, distributed according to Greenwood's (1990) formula. The temporal envelope from each frequency analysis band was extracted by half-wave rectification and low-pass envelope filtering (160 Hz). The temporal envelopes were then used to modulate sinusoidal carriers whose frequencies were equivalent to the center frequencies of the carrier bands. The modulated signals from all frequency channels were summed and then normalized to have the same long-term root-mean-square (RMS) level as the input speech signal (65 dBA).

Procedure

For all testing, speech was presented in sound field at 65 dBA from a single loudspeaker (Tannoy Reveal). Subjects were seated directly facing the loudspeaker placed 1 m away. During testing, a list was randomly selected (without replacement) from among the 72 IEEE sentence lists, and a sentence was randomly selected (without replacement) from among the 10 sentences within the list and presented to the subject. The subject responded by repeating the sentence as accurately as possible and the experimenter scored all words correctly identified. Performance was scored in terms of the percent of words correctly identified in sentences. The speech quality (natural or synthetic) and three rate conditions (slow, normal, and fast) were randomized and counterbalanced within and across subjects. No trial by trial feedback was provided. CI subjects were tested while listening with their clinical CI devices and settings. Unilateral CI users were tested with one CI, bilateral CI users were tested with both CIs, and or bimodal CI users (combined use of a CI and hearing aid, or HA) were tested with the CI only (HA turned off).

RESULTS

Figure 1 shows individual and mean CI subject data with the different speaking rates, for natural (top panel) and synthetic speech (bottom panel). A three-way repeated measures (RM ANOVA) was performed on the CI data, with speaking rate, speech quality, and talker gender as factors; the results are shown in Table 3. All pairwise comparisons are reported with Bonferroni adjustment for multiple comparisons. Performance was significantly poorer

with the fast speaking rate than with the slow ($p<0.001$) or normal rates ($p<0.001$); there was no significant difference between the slow and normal rates ($p=0.067$). Performance was significantly better with natural speech than with synthetic speech ($p<0.001$). Performance was significantly better with the female taker than with the male talker ($p=0.021$), though mean performance (across all conditions) was quite similar (female: 54.0% correct; male: 51.3% correct).

Figure 2 shows individual and mean CI simulation data with the different speaking rates, for natural (top panel) and synthetic speech (bottom panel). A three-way RM ANOVA was performed on the CI simulation data, with speaking rate, speech quality, and talker gender as factors; the results are shown in Table 4. All pairwise comparisons are reported with Bonferroni adjustment for multiple comparisons. Performance was significantly poorer with the fast speaking rate than with the slow ($p<0.001$) or normal rates ($p<0.001$); there was no significant difference between the slow and normal rates ($p=0.656$). Performance was significantly better with natural speech than with synthetic speech ($p<0.001$). Performance was significantly better with the male talker than with the female talker ($p=0.039$), though mean performance (across all conditions) was quite similar (female: 71.2% correct; male: 73.4% correct).

Figure 3 shows individual and mean NH data with the different speaking rates, for natural (top panel) and synthetic speech (bottom panel). For the slow and normal rates, performance was near ceiling. A three-way RM ANOVA, was performed on the NH data, with speaking rate, speech quality, and talker gender as factors; the results are shown in Table 5. All pairwise comparisons are reported with Bonferroni adjustment for multiple comparisons. Performance was significantly poorer with the fast speaking rate than with the slow ($p<0.001$) or normal rates ($p<0.001$); there was no significant difference between the slow and normal rates ($p>0.999$). Performance with natural speech was significantly better than with synthetic speech ($p<0.001$). There was no significant difference in performance between the male and female talkers ($p=0.936$).

Figure 4 shows mean performance with natural and synthetic speech for the CI (left panel), CI simulation (middle panel), and NH subject groups (right panel), as a function of speaking rate. A Kruskal-Wallis one-way ANOVA on ranked data showed a significant effect for subject group [$H(2)=142.147$, $p<0.001$]. Pairwise comparison (Dunn's method) showed that NH performance was significantly better than CI simulation ($p<0.05$) or CI ($p<0.05$) performance, and that CI simulation performance was significantly better than CI performance ($p<0.05$).

DISCUSSION

The present results show that CI users are susceptible to variability in speech patterns due to speaking rate or from speech production styles. With naturally produced speech, mean CI performance was slightly poorer than CI simulation performance at the slow and normal rates. At fast rates, CI performance was very poor. With synthetic speech, mean CI performance was nearly 18 percentage points poorer than with the CI simulation at the slow and normal rates, and 32 points poorer at the fast rate. Below we discuss the results in greater detail.

Effects of talker gender

Similar to previous studies (Luo & Fu 2005; Liu et al. 2008), there was a significant effect of talker gender for CI subjects ($p=0.021$) and CI simulation subjects ($p=0.039$); there was no significant effect for NH subjects ($p=0.936$). This suggests that CI signal processing may interact with talker characteristics (e.g., F0, vocal tract length, and oral cavity shape). Note

that while significant, the difference in performance between male and female talkers was small (2.71 points for CI subjects; 2.20 points for CI simulation subjects). The present results may be idiosyncratic to the particular natural and synthetic talkers used. Other talkers might have increased the effect of talker gender.

Effect of speech quality

For all three subject groups, performance was significantly better with natural than with synthetic speech (all $p<0.05$; see Tables 3–5 for exact p-values). However, significant interactions were observed between speaking rate and speech quality for all three subject groups (all $p<0.05$; see Tables 3–5 for exact p-values). Pairwise comparisons (with Bonferroni adjustment) revealed somewhat different patterns across subject groups. CI performance was significantly better with natural speech than with synthetic speech at the slow ($p=0.003$) and normal rates ($p<0.001$), but not at the fast rate ($p>0.999$). Likewise, NH performance was significantly better with natural speech than with synthetic speech at the slow ($p=0.040$) and normal rates ($p=0.007$), but not at the fast rate ($p>0.999$). CI simulation performance was significantly better with natural speech than with synthetic speech at the slow ($p<0.001$) and normal rates ($p<0.001$), but better with synthetic speech than with normal speech at the fast rate ($p<0.001$). At the fast rate, the artifacts associated with temporal distortion to the natural speech may have been more prominent than those associated with speech synthesis, and may have interacted with CI simulation parameters related to temporal envelope extraction.

The performance deficit with synthetic speech was greater with CI signal processing (whether simulated or real), most likely due to reduced spectral resolution. Most previous CI and CI stimulation studies (Fishman et al. 1997; Friesen et al. 2001) have examined the effect of spectral resolution using clear speech materials, such as vowels from Hillenbrand & Gayvert (1993), consonants from Shannon et al. (1999), and HINT sentences (Nilsson et al. 1994). Although CI users may function as if receiving 4–8 spectral channels (Shannon et al. 2004), the functional spectral resolution may be even poorer when the variability in speech patterns is considered.

CI performance with synthetic speech may also have been affected by distortion to temporal cues. Stone et al. (2010) showed that, for noise-band vocoders, temporal envelope cues between 12.5 and 50 Hz contributed most strongly to speech understanding in a competing speech task. Stone et al. (2010) also found temporal periodicity cues between 50 and 200 Hz significantly contributed to performance. With synthetic speech, these temporal envelope cues may have been distorted relative to natural speech.

CI users' difficulty with synthetic speech may be partly explained by the resource-sharing model of spoken language comprehension from Duffy and Pisoni (1992). According to the model, even high quality synthetic speech may require considerable cognitive resources to decode the acoustic-phonetic structure, leaving fewer resources for high-level processing needed to understand the meaning of the word. Synthetic speech lacks the dynamic acoustic structures that give rise to redundant, co-articulated cues found in natural speech. This lack of redundancy can give rise to, for example, longer response times in a competing attention task. Given the limited spectro-temporal resolution associated with CI signal processing, the diminished quality of synthetic speech may have placed an even greater processing load on CI and CI simulation subjects. The present data suggest that CI users are sensitive to distortions to cues that are poorly represented by CI signal processing (e.g., F0, spectral fine structure). This is similar to previous studies that show significant talker variability effects in CI users (Chang & Fu 2006), even though CI users have difficulty identifying or segregating talkers (Stickney et al. 2008; Cullington & Zeng 2011).

Effects of speaking rate

For natural speech, pairwise comparisons (with Bonferroni adjustment) showed no significant difference between the slow and normal speaking rates for CI ($p=0.062$), CI simulation ($p>0.999$), or NH subjects ($p>0.999$). This suggests that any artifacts associated with time expansion of natural speech had only a minor effect on performance. Although not significant, the performance deficit with slow rate was substantial (10.7 points), relative to the normal rate. Performance sharply dropped with the fast rate, relative to the normal rate. Pairwise comparisons (with Bonferroni adjustment) showed significantly poorer performance with the fast rate for CI ($p<0.001$), CI simulation ($p<0.001$), or NH subjects ($p<0.001$). In this study, the effects of speaking rate were evaluated using a time-scaling algorithm. Given the greater variability in speech patterns when naturally produced at different speaking rates, the present data may under-estimate the effects of speaking rate on CI users' speech understanding.

The synthetic speech rate was varied by adjusting the speaking rate of the TTS engine. The unit selection synthesis used by the TTS engine allowed for combinations of phones, di-phones and tri-phones, and mixtures thereof. These units were most likely time-scaled to achieve the target speaking rates, as opposed to the uniform time-scaling used with natural speech. Although overall performance was significantly poorer with synthetic than with natural speech ($p<0.001$), the effects of speaking rate were comparable to those with natural speech, suggesting that differences in the time-scaling algorithms had only a modest effect on performance. Interestingly, CI simulation performance at the fast rate was significantly better with synthetic than with natural speech ($p<0.001$), suggesting that the different time-scaling algorithms may have interacted with CI simulation parameters.

A more moderate range of speaking rates might have been more sensitive to the effects of speech quality and CI signal processing. The floor performance effects with the fast rate could not reveal any differences in speech quality for CI subjects. Likewise, ceiling performance effects could not reveal any differences in speech quality for NH subjects. Given that synthetic speech is often accompanied by some degree of noise (e.g., telephone line noise), additive noise may have helped to differentiate the effects of speech quality at the slow and normal rates. It is likely that a more optimal range of speaking rates and additive noise might show greater effects of speech quality and more interactions with CI signal processing.

It is important to note differences between the experimental fast-rate speech and fast-rate speech in the "real world." As discussed in the Introduction, dynamic acoustic properties such as formant transition, formant transition duration, syllable duration, voice onset time (VOT), prosody, and F0 range may be altered at different speaking rates (Lass 1970; Miller & Liberman 1979; Kohler 1986; Miller 1987). In this study, natural speech was uniformly time-compressed or -expanded. As a result, all consonant and vowel durations, VOTs, silent periods between words were time-scaled by the same amount, which may have produced distorted acoustic and phonetic cues. Krause and Braida (2002) found better speech understanding in NH listeners for sentences naturally produced at fast rates than for normal-speech that was time-scaled to target a fast rate. Liu and Zeng (2006) tested the same time-stretch algorithm used in this study and found audible artifacts when speech was time-compressed then expanded to restore the original sentence duration, but not when the sentence was first time-expanded then compressed.

Effects of CI signal processing

The effect of CI signal processing (whether real or simulated) was more deleterious for synthetic than for natural speech. CI simulation performance with natural speech (black bars

in Fig. 4) was 6.2, 4.1, and 45.1 percentage points poorer at the slow, normal, and fast speaking rates, respectively, relative to NH performance. CI simulation performance with synthetic speech (gray bars in Fig. 4) was 11.1, 12.3, and 37.9 percentage points poorer at the slow, normal, and fast rates, respectively, relative to NH performance. Thus, for the slow and normal rates, CI simulation performance was poorer with synthetic than with natural speech, relative to NH performance.

CI performance was significantly poorer than CI simulation performance ($p<0.05$). CI performance with natural speech (black bars in Fig. 4) was 14.9, 4.1, and 27.7 percentage points poorer at the slow, normal, and fast speaking rates, respectively, relative to CI simulation performance. CI performance with synthetic speech (gray bars in Fig. 4) was 16.0, 20.0, and 32.6 percentage points poorer at the slow, normal, and fast rates, respectively, relative to CI simulation performance. Thus, the CI simulation seems to have overestimated the real CI performance. Factors associated with CI stimulation (e.g., functional spectral resolution, temporal processing limits, current spread, and channel interaction) may have contributed to the deficit in CI performance. CI performance may also have been limited by individual CI subject factors such as duration of deafness, which was significantly correlated with speech performance for some conditions (see below for further discussion). Alternatively, CI simulation performance may have been better than CI performance because NH listeners were listening to degraded speech with a healthy auditory system. As discussed below, CI performance in many conditions was correlated with duration of deafness, which may be related to the health of the impaired auditory system.

The effects of CI signal processing were most severe at the fast speaking rate. As discussed above, the combination of fast speech and CI signal processing may have required additional cognitive processing resources, especially for CI subjects. Increasing the spectral resolution may increase tolerance for fast speaking rates, similar to improvements in speech understanding in noise with increased spectral resolution (Friesen et al. 2001; Shannon et al. 2004). Alternatively, adjusting the speaking rate may offset some of the deficits associated with CI signal processing. Such adjustments are more feasible with synthetic speech, as the speaking rate and voice characteristics can be easily modified. With natural speech, adjustments would be more challenging, especially if performed in real or near-real time. Such signal processing would have to differentiate between speech and non-speech sounds, which might be differently affected by time-scaling.

Subject demographic factors

Linear regression analysis showed no significant correlation between age at testing and any of the speech measures for CI (r^2 range: 0.029–0.320; p-value range: 0.696–0.096), CI simulation (r^2 range: 0.004–0.152; p-value range: 0.825–0.187), or NH subjects (r^2 range: 0.009–0.161; p-value range: 0.677–0.088). These results do not agree with previous studies that showed significant effects of aging on speech performance (Roring et al. 2007; Schvartz et al. 2008; Getzmann & Falkenstein 2011; Hopkins & Moore 2011; Schvartz & Chatterjee 2012). In most of these studies, isolated words or phonemes were used for testing. In the present study, sentences were used for testing. The availability of contextual cues may have diminished aging effects. Also, the age range of the present subjects may not have been sufficient to observe aging effects.

CI subjects' duration of deafness was significantly correlated with natural speech performance at the normal ($r^2=0.46$, $p=0.032$) and slow rates ($r^2=0.61$, $p=0.008$), but not at the fast rate ($r^2=0.06$, $p=0.502$). Duration of deafness was significantly correlated with synthetic speech at the normal ($r^2=0.51$, $p=0.020$), but not at the slow ($r^2=0.37$, $p=0.060$) or fast rates ($r^2=0.12$, $p=0.339$).

Although there are too few subjects to make any strong statements, there was no clear effect of device type on performance. There were 6 users of Cochlear devices and 4 users of Advanced Bionics. A Kruskal-Wallis one-way ANOVA on ranked data, with device type as factor, showed no significant effect of device type [$H(1)=0.166$, $p = 0.684$].

Implications for CI users

Compared to NH or CI simulation subjects, CI subjects were more negatively affected by synthetic speech. Optimizing the acoustic input may improve recognition of synthetic (or even natural) speech. For example, speaker normalization can transform acoustic features (e.g., F0 and/or vocal tract size) from one talker to another. Given that CI users often perform better with one talker than another, speaker normalization may improve understanding with difficult talkers. Liu et al. (2008) reported that speaker normalization improved CI users' recognition of less-intelligible talkers. Similarly, Luo and Fu (2005) reported significantly better Chinese vowel recognition with speaker normalization for NH subjects listening to a 4-channel CI simulation. For synthetic speech, it is unclear which acoustic features may need to be adjusted, especially in the context of electric hearing in which the spectro-temporal resolution is reduced.

CI users were also more susceptible to speaking rate. With TTS readers, it is often possible for listeners to adjust the speaking rate as needed. Similarly, for telephone prompts, which often consist of TTS or a TTS-real speech hybrid, it seems possible to allow for some adjustment of the speaking rate. Though not significant ($p=0.067$), CI performance with synthetic speech was poorer at the slow rate than at the normal rate (see Fig. 4). This suggests that if the synthetic speech rate is too slow, CI performance may worsen. For live, naturally produced speech, it seems easiest to ask a talker to slow their speaking rate. For recorded, naturally produced speech, the signal processing may be quite complex, especially if real-time processing is required or the audio is associated with visual information.

Auditory training may also be a good option to offset these deficits. Auditory training has been shown to significantly improve CI users' speech and music perception, even after years of experience with their device and signal processing (Fu et al. 2005; Galvin et al. 2007; Fu & Galvin 2008; Stacey & Summerfield 2008). Such training has also been shown to improve NH performance while listening to acoustic CI simulations (Rosen et al. 1999; Nogaki et al. 2007; Stacey & Summerfield 2008). These training methods have primarily targeted adaptation to spectral degradation and/or frequency mismatch. Such training may also improve CI users' perception of synthetic speech. Koul and Hester (2006) reported that NH subjects with severe intellectual impairments were able to better recognize synthetic speech after repeated exposure ("passive" learning). "Active" training may accelerate this learning process.

Although training with fast speaking rates has not been explicitly tested in adult CI users, time-scaling techniques have been used to train children with learning disabilities. For example, Fast ForWord™ (Scientific Learning Corporation, Berkeley, CA) slows the speaking rate during training, with mixed success. Strong et al. (2011) found no significant improvement in pediatric NH listeners' language perception after training with Fast ForWord™. However, Schopmeyer et al. (2000) found that Fast ForWord™ significantly improved pediatric CI users' language perception skills; note that there were only 4 subjects in the study. It is unclear whether training with time-scaled speech would benefit adult CI users, especially for understanding of fast, rather than slow or normal speech.

Conclusions

In this study, recognition of naturally produced and synthetic speech presented at slow, normal, and fast speaking rates was measured in NH subjects listening to unprocessed speech or to an acoustic CI simulation and in CI subjects listening with their clinical processors. Results showed:

1. Performance was significantly better with natural speech than with synthetic speech.
2. Performance was significantly poorer only with the fast speaking rate.
3. CI performance was significantly poorer than CI simulation performance, suggesting additional deficits associated with electric hearing and/or deafness

Acknowledgments

The authors thank all subjects who participated in the study. We greatly appreciate their ongoing support for our research. The authors also thank three anonymous reviewers for their helpful comments. This study was supported by NIH grant 5R01DC004993.

ABBREVIATIONS

ACE™	advanced combination encoder
ANOVA	analysis of variance
BKB	Bamford, Kowal, and Bench
CI	cochlear implant
CIS	continuous interleaved sampling
CNC	consonant-nucleus-consonant
HA	hearing aid
HINT	hearing in noise test
IEEE	Institute of Electrical and Electronic Engineers
F	female
F0	fundamental frequency
M	male
NAT	natural
NH	normal hearing
RM	repeated measures
RMS	root-mean-square
SIM	simulation
SNR	signal-to-noise ratio
SPEAK™	spectral peak coding
SRT	speech reception threshold
SYN	synthetic
TTS	text-to-speech

VOT	voice onset time
WPS	words per second

References

AT&T Natural voices™ Text-To-Speech Engines: System Developer's Guide. 2001. Release 1.4. http://www.naturalvoices.att.com

Bench J, Kowal A, Bamford J. The BKB (Bench-Kowal-Bamford) sentence lists for partially-hearing children. Br J Audiol. 1979; 13:108–112. [PubMed: 486816]

Beutnagel, A.; Conkie, J.; Schroeter, Y.; Stylianou, et al. The AT&T next-gen TTS system. Proc. 137th meet. ASA/Forum Acusticum; Berlin. March 1999; 1999. Joint Meeting of ASA, EAA, AND DAGA; March; Berlin, Germany.

Bradlow AR, Kraus NHE. Speaking clearly for children with learning disabilities: sentence perception in noise. J Speech Lang Hear Res. 2003; 46:80–97. [PubMed: 12647890]

Bunton K, Story BH. Identification of synthetic vowels based on selected vocal tract area functions. J Acoust Soc Am. 2009; 125:19–22. [PubMed: 19173389]

Bunton K, Story BH. Identification of synthetic vowels based on a time-varying model of the vocal tract area function. J Acoust Soc Am. 2010; 127(4):146–152.

Chang YP, Fu QJ. Effects of talker variability on vowel recognition in cochlear implants. J Speech Lang Hear Res. 2006; 49:1331–1341. [PubMed: 17197499]

Conkie, A. A robust unit selection system for speech synthesis. Proc. 137th meet. ASA/Forum Acusticum; Berlin. March 1999; 1999.

Crystal TH, House AS. Segmental duration in connected speech signal: Current results. J Acoust Soc Am. 1988; 83:1553–1573.

Cullington HE, Zeng FG. Comparison of bimodal and bilateral cochlear implant users on speech recognition with competing talker, music perception, affective prosody discrimination, and talker identification. Ear Hear. 2011; 32:16–30. [PubMed: 21178567]

Dickstein DL, Kabaso D, Rocher AB, et al. Changes in the structural complexity of the aged brain. Aging cell. 2007; 6:275–284. [PubMed: 17465981]

Duffy SA, Pisoni DB. Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. Lang Speech. 1992; 35:351–389. [PubMed: 1339919]

Fishman KE, Shannon RV, Slattery WH. Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. J Speech Lang Hear Res. 1997; 40(5): 1201–15. [PubMed: 9328890]

Friesen LM, Shannon RV, Baskent D, et al. Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. J Acoust Soc Am. 2001; 110(2):1150–63. [PubMed: 11519582]

Fu QJ, Galvin J, Wang X, et al. Effects of auditory training on adult cochlear implant patients: a preliminary report. Cochlear Implants Int. 2005; 5(Suppl 1):84–90. [PubMed: 18792249]

Fu QJ, Nogaki G. Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. J Assoc Res Otolaryngol. 2005; 6(1):19–27. [PubMed: 15735937]

Fu QJ, Galvin JJ 3rd. Maximizing cochlear implant patients' performance with advanced speech training procedures. Hear Res. 2008; 242(1–2):198–208. [PubMed: 18295992]

Galvin JJ III, Fu QJ, Nogaki G. Melodic contour identification in cochlear implants. Ear Hear. 2007; 28(3):302–319. [PubMed: 17485980]

Getzmann S, Falkenstein M. Understanding of spoken language under challenging listening conditions in younger and older listeners: A combined behavior and electrophysiological study. Brain Res. 2011; 1415:8–22. [PubMed: 21880303]

Greenwood DD. A cochlear frequency-position function for several species – 29 years later. J Acoust Soc Am. 1990; 87:2592–2605. [PubMed: 2373794]

Hillenbrand J, Gayvert RT. Identification of steady-state vowels synthesized from the Peterson and Barney measurements. J Acoust Soc Am. 1993; 94:668–674. [PubMed: 8370872]

Hopkins K, Moore BC. The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise. J Acoust Soc Am. 2011; 130(1): 334–349. [PubMed: 21786903]

Kohler KJ. Parameters of speech rate perception in German words and sentences: Duration, F0 movement and F0 level. Lang Speech. 1986; 29:115–139. [PubMed: 3657349]

Koul R, Hester K. Effects of repeated listening experiences on the recognition of synthetic speech by individuals with severe intellectual disabilities. J Speech Lang Hear Res. 2006; 49:47–57. [PubMed: 16533072]

Krause JC, Braida LD. Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. J Acoust Soc Am. 2002; 112:2165–2173. [PubMed: 12430828]

Lass NJ. The significance of intra- and intersentence pause times in perceptual judgements of oral reading rate. J Speech Lang Hear Res. 1970; 13:777–784.

Li Y, Zhang G, Kang HY, et al. Effects of speaking style on speech intelligibility for Mandarin-speaking cochlear implant users. J Acoust Soc Am. 2011; 129(6):EL242–7. [PubMed: 21682359]

Lisker L. Closure duration and the intervocalic voiced-voiceless distinction in English. Language. 1957; 33:42–49.

Liu S, Del Rio E, Bradlow AR, et al. Clear speech perception in acoustic and electric hearing. J Acoust Soc Am. 2004; 116(4 Pt 1):2374–83. [PubMed: 15532668]

Liu S, Zeng FG. Temporal properties in clear speech perception. J Acoust Soc Am. 2006; 120(1):424–432. [PubMed: 16875238]

Liu C, Galvin JJ III, Fu QJ, et al. Effect of spectral normalization on different talker speech recognition by cochlear implant users. J Acoust Soc Am. 2008; 123:2836–2847. [PubMed: 18529199]

Logan JS, Greene BG, Pisoni DB. Segmental intelligibility of synthetic speech produced by rule. J Acoust Soc Am. 1989; 86(2):566–81. [PubMed: 2527884]

Luo X, Fu QJ. Speaker normalization for Chinese vowel recognition in cochlear implants. IEEE Tran, Biome Eng. 2005; 52(7):1358–1361.

Mahncke HW, Bronstone A, Merzenich MM. Brain plasticity and functional losses in the aged: Scientific bases for a novel intervention. Prog Brain Res. 2006; 157:81–109. [PubMed: 17046669]

Miller JL, Baer T. Some effects of speaking rate on the production of /b/ and /w/. J Acoust Soc Am. 1983; 73:1751–1755. [PubMed: 6863754]

Miller JL, Liberman AM. Some effects of later-occurring information on the perception of stop consonant and semivowel. Percept Psychophys. 1979; 25:457–465. [PubMed: 492910]

Miller, JL. Rate-dependent processing in speech perception. In: Ellis, AW., editor. Progress in the psychology of language. 1987. p. 119-157.

Moulines E, Laroche J. Non-parametric techniques for pitch-scale and time-scale modification of speech. Crystallogr Rep. 1995; 16:175–205.

Nadol JB Jr, Young YS, Glynn RJ. Survival of spiral ganglion cells in profound sensorineural hearing loss: Implications for cochlear implantation. Ann Otol Rhinol Laryngol. 1989; 98:411–416. [PubMed: 2729822]

Newman RS, Sawusch JR. Perceptual normalization for speaking rate: Effects of temporal distance. Percept Psychophys. 1996; 58:540–560. [PubMed: 8934686]

Nilsson M, Soli SD, Sullivan JA. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am. 1994; 95:1085–1099. [PubMed: 8132902]

Nogaki G, Fu QJ, Galvin JJ III. The Effect of Training Rate on Recognition of Spectrally Shifted Speech. Ear Hear. 2007; 28(2):132–140. [PubMed: 17496666]

Paris CR, Thomas MH, Gilson RD, et al. Linguistic cues and memory for synthetic and natural speech. Hum Factors. 2000; 42(3):421–431. [PubMed: 11132803]

Payton KL, Uchanski RM, Braida LD. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. J Acoust Soc Am. 1994; 95:1581–1592. [PubMed: 8176061]

Peterson GE, Lehiste I. Revised CNC lists for auditory tests. J Speech Hear Disord. 1962; 27:62–70. [PubMed: 14485785]

Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing: Intelligibility differences between clear and conversational speech. J Speech Hear Res. 1985; 28:96–103. [PubMed: 3982003]

Picheny MA, Durlach NI, Braida LD. Speaking clearly for the hard of hearing. III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. J Speech Hear Res. 1989; 32(3):600–603. [PubMed: 2779204]

Roring RW, Hines FG, Charness N. Age differences in identifying words in synthetic speech. Hum Factors. 2007; 49(1):25–31. [PubMed: 17315840]

Rosen S, Faulkner A, Wilkinson L. Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. J Acoust Soc Am. 1999; 106(6):3629–36. [PubMed: 10615701]

Rothauser EH, Chapman WD, Guttman N, et al. IEEE recommended practice for speech quality measurements. IEEE Trans Acoust. 1969; 17(3):225–246.

Schopmeyer B, Mellon N, Dobaj H, et al. Use of Fast ForWord to enhance language development in children with cochlear implants. Ann Otol Rhinol Laryngol Suppl. 2000; 185:95–98. [PubMed: 11141025]

Schvartz KC, Chatterjee M, Gordon-Salant S. Recognition of spectrally degraded phonemes by younger, middle-aged, and older normal-hearing listeners. J Acoust Soc Am. 2008; 124:3972–3988. [PubMed: 19206821]

Schvartz KC, Chatterjee M. Gender identification in younger and older adults: Use of spectral and temporal cues in noise-vocoded speech. Ear Hear. 2012; 33(3):411–420. [PubMed: 22237163]

Shannon RV, Jensvold A, Padilla M, et al. Consonant recordings for speech testing. J Acoust Soc Am. 1999; 106(6):71–74.

Shannon RV, Fu QJ, Galvin JJ III. The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. Acta Otolaryngol, Suppl. 2004; 552:50–54. [PubMed: 15219048]

Spahr AJ, Dorman MF. Performance of subjects fit with the Advanced Bionics CII and Nucleus 3G cochlear implant devices. Arch Otolaryngol Head Neck Surg. 2004; 130:624–628. [PubMed: 15148187]

Stacey PC, Summerfield AQ. Comparison of word-, sentence-, and phoneme-based training strategies in improving the perception of spectrally distorted speech. J Speech Lang Hear Res. 2008; 51(2): 526–538. [PubMed: 18367694]

Stickney GS, Zeng FG, Litovsky R, Assmann P. Cochlear implant speech recognition with speech maskers. J Acoust Soc Am. 2004; 116:1081–1091. [PubMed: 15376674]

Stone MA, Füllgrabe C, Moore BC. Relative contribution to speech intelligibility of different envelope modulation rates within the speech dynamic range. J Acoust Soc Am. 2010; 128:2127–2137. [PubMed: 20968383]

Strong GK, Torgerson CJ, Torgerson D, et al. A systematic meta-analytic review of evidence for the effectiveness of the ‘Fast for Word’ language intervention program. J Child Psychol Psychiatry. 2011; 52(3):224–35. [PubMed: 20950285]

Uchanski RM, Choi SS, Braida LD, et al. Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. J Speech Hear Res. 1996; 39(3):494–509. [PubMed: 8783129]

This study investigated the effect of speaking rate on recognition of synthetic and natural speech by normal-hearing (NH) and cochlear implant (CI) subjects. NH and CI subjects were asked to recognize naturally produced or synthetic sentences, presented at a slow, normal, or fast speaking rate. Sentence recognition was significantly better with natural than with synthetic speech and significantly poorer at the fast rate. The results suggest that CI users are more susceptible to variability in speech patterns due to speaking rate or speech production style. Testing with clear speech may over-estimate CI performance in real-world listening conditions.

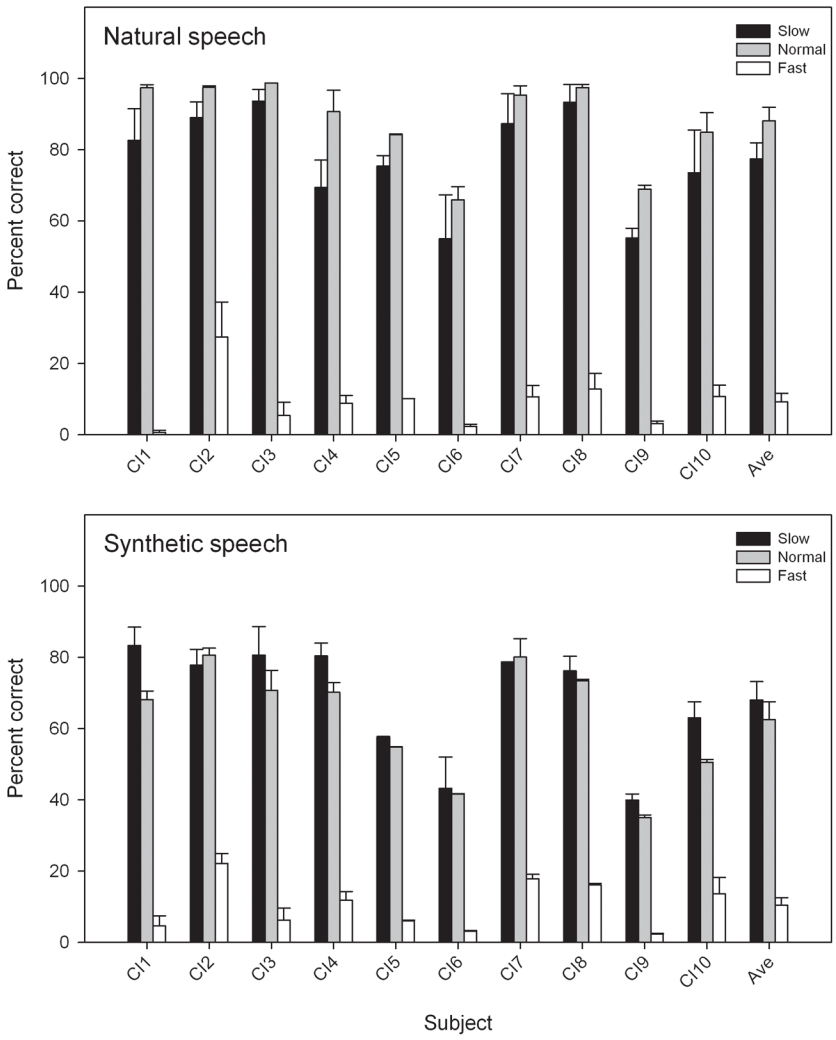

Figure 1.
Individual and mean CI performance with the different speaking rates, for natural (top panel) and synthetic speech (bottom panel). Data are averaged across talker gender. Subjects are ordered according to age, from youngest to oldest. The error bars show the standard error.

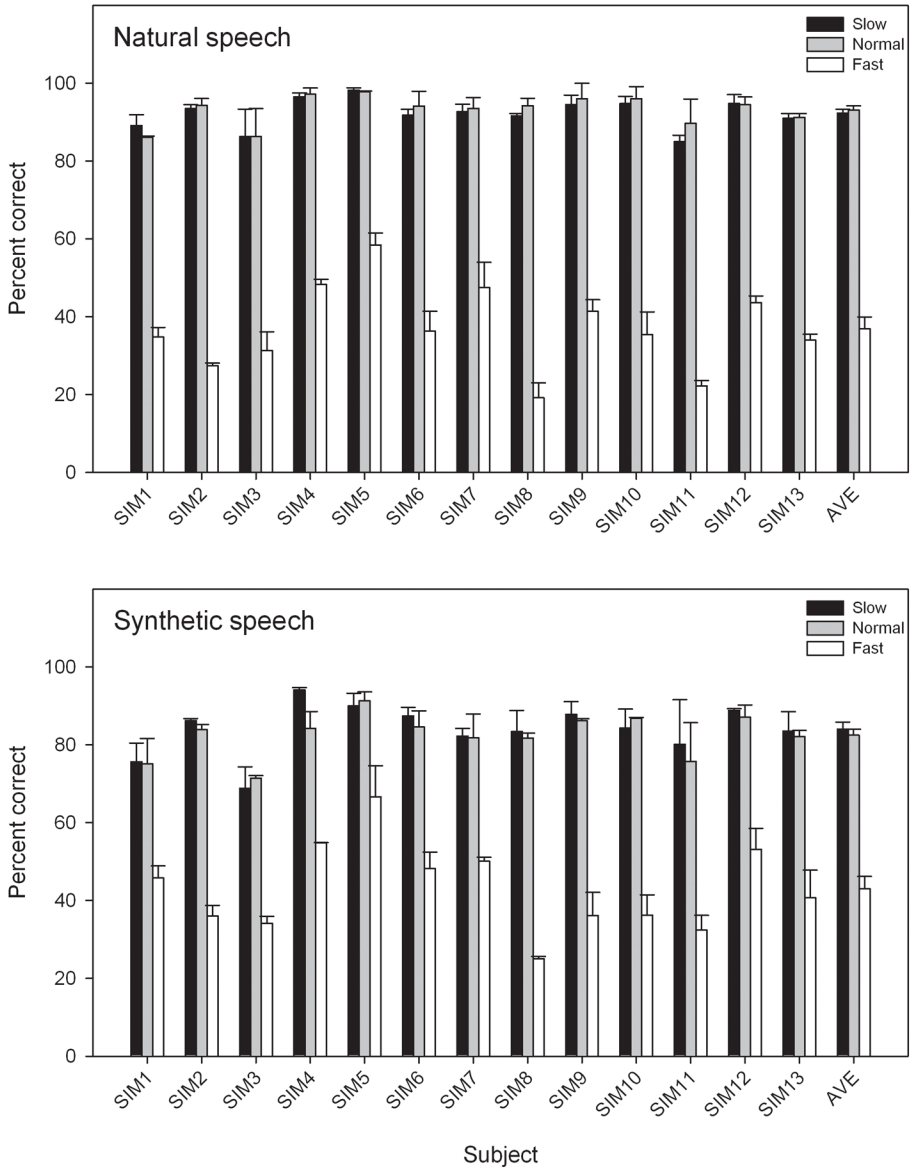

Figure 2.

Individual and mean CI simulation performance with the different speaking rates, for natural (top panel) and synthetic speech (bottom panel). Subjects are ordered according to age, from youngest to oldest. Data are averaged across talker gender. The error bars show the standard error.

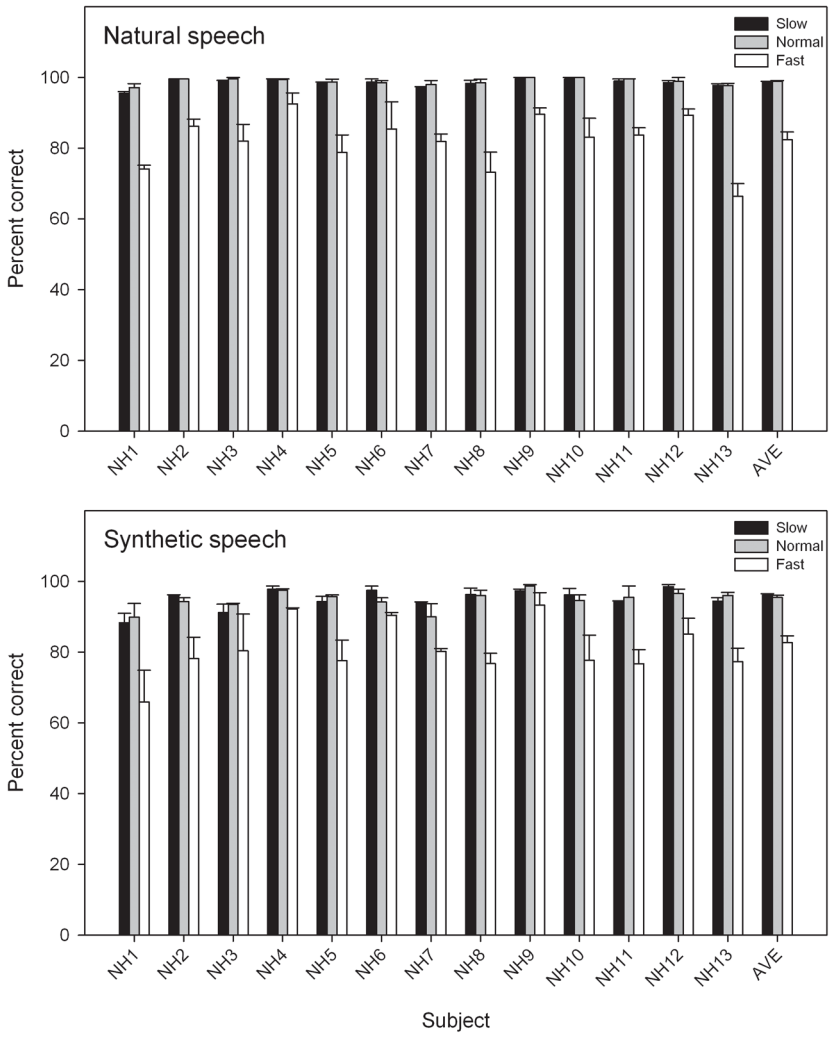

Figure 3.

Individual and mean NH performance with the different speaking rates, for natural (top panel) and synthetic speech (bottom panel). Subjects are ordered according to age, from youngest to oldest. Data are averaged across talker gender. The error bars show the standard error.

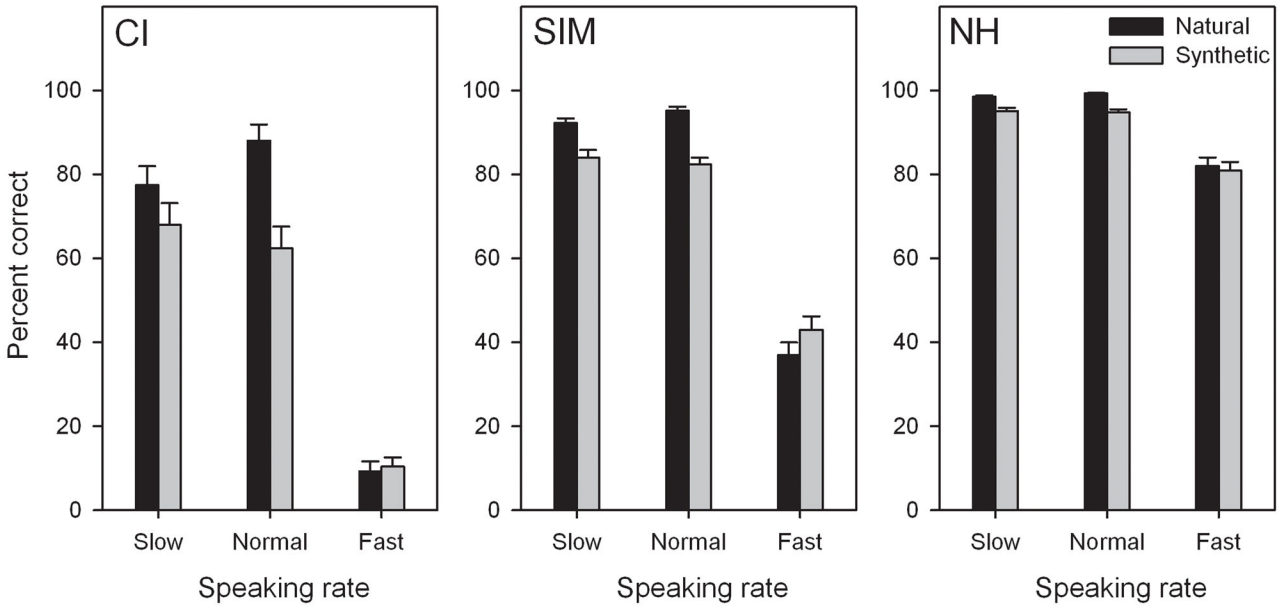

Figure 4.
Mean IEEE sentence recognition with natural (black bars) and synthetic speech (gray bars) for CI subjects (left panel), CI simulation subjects (middle panel), and NH subjects (right panel), as a function of speaking rate. Data are averaged across subjects and talker gender. The error bars show the standard error.

Table 1

Demographic information for CI subjects.

Subject	Gender	Age at testing (yrs)	Duration of deafness (yrs)	CI experience (yrs)	CI device; strategy
CI1	F	24	2	L:5	Freedom®; ACE™
CI2	F	57	3	L:4.5	HiRes® 90K/HiFocus®; Fidelity 120™
				R:1.5	HiRes® 90K/HiFocus®; Fidelity 120™
CI3	M	59	1	L:6	Nucleus® 22; SPEAK™
				R:2	Freedom®; ACE™
CI4	F	62	4	L:4	HiRes® 90K/HiFocus®; Fidelity 120™
				R:16	Clarion® I; CIS
CI5	F	68	20	R:8	Nucleus® 24; ACE™
CI6	F	74	55	L: 10	Clarion® II; Fidelity 120™
				R:2	HiRes® 90K; Fidelity 120™
CI7	M	74	2	L	HA
				R:2	Freedom®; ACE™
CI8	M	75	1	L:12	Clarion® II; Fidelity 120™
				R	HA
CI9	F	78	12	L	HA
				R:11	Nucleus® 24; ACE™
CI10	M	81	1	L:16	Nucleus® 22; SPEAK™

L=left, R=right, ACE™=advanced combination encoder, SPEAK™ =spectral peak coding, CIS=continuous interleaved sampling, HA=hearing aid.

Ear Hear. Author manuscript; available in PMC 2014 May 01.

Table 2

Mean F0 and speaking rates (slow, normal and fast) across 720 sentences for naturally produced and synthetic speech sentences.

		Mean speech rate (words per second)		
	Mean F0	**Slow**	**Normal**	**Fast**
Natural M	111 Hz	1.65	3.18	6.65
Natural F	188 Hz	1.72	3.31	6.93
Synthetic M	114 Hz	1.71	3.45	6.64
Synthetic F	180 Hz	1.64	3.3	6.36

Table 3

Results of 3-way RM ANOVA on CI subject data. The shaded cells indicate significant effects. dF=degrees of freedom; res=residual.

Factor	dF, res	F-ratio	p-value	power
Rate	2, 18	287.057	<0.001	>0.999
Quality	1, 9	56.852	<0.001	>0.999
Gender	1, 9	7.728	0.021	0.697
Rate*Quality	2, 18	52.639	<0.001	>0.999
Rate*Gender	2, 18	0.701	0.509	0.150
Gender*Quality	1, 9	31.011	<0.001	0.998
Rate*Quality*Gender	2, 18	9.938	0.001	0.964

Table 4

Results of 3-way RM ANOVA on CI simulation subject data. The shaded cells indicate significant effects. dF=degrees of freedom; res=residual.

Factor	dF, res	F-ratio	p-value	power
Rate	2, 24	351.553	<0.001	>0.999
Quality	1, 12	43.347	<0.001	>0.999
Gender	1, 12	5.374	0.039	0.568
Rate*Quality	2, 24	88.893	<0.001	>0.999
Rate*Gender	2, 24	0.123	0.885	0.067
Gender*Quality	1, 12	22.084	0.001	0.990
Rate*Quality*Gender	2, 24	3.422	0.049	0.587

Table 5

Results of 3-way RM ANOVA on NH subject data. The shaded cells indicate significant effects. dF=degrees of freedom; res=residual.

Factor	dF, res	F-ratio	p-value	power
Rate	2, 24	87.382	<0.001	>0.999
Quality	1, 12	15.594	0.002	0.591
Gender	1, 12	0.007	0.936	0.051
Rate*Quality	2, 24	4.047	0.031	0.664
Rate*Gender	2, 24	0.033	0.967	0.051
Gender*Quality	1, 12	7.320	0.019	0.700
Rate*Quality*Gender	2, 24	5.417	0.011	0.796