Perception of Synthetic Speech in Adult Users of Cochlear Implants

Kyoko Nagao, Mark Paullin, James B. Polikoff, Jason Lilley, and H. Timothy Bunnell

1Center for Pediatric Auditory and Speech Sciences, Nemours Biomedical Research, Wilmington, Delaware, USA
2Department of Linguistics & Cognitive Science, University of Delaware, Newark, Delaware, USA

{ nagao, paullin, polikoff, lilley, bunnell }@asel.udel.edu

Abstract

Text-to-speech (TTS) synthesis technology has great potential to be implemented in aural rehabilitation because of the flexibility it affords for stimulus generation. However, little is known about how listeners with cochlear implants (CI) perceive and respond to synthetic speech. This study investigated how adult CI users perceive and respond to synthetic speech of varying qualities in order to determine if they respond to synthetic speech in a different manner than individuals with normal hearing (NH). Sentence recognition performance of 16 adult CI users and 30 NH listeners was compared on a semantically unpredictable sentence task using synthetic speech generated at different levels of quality. The results indicated that while the CI group had an overall lower performance than the NH group, their pattern of responding was similar to NH listeners across the speech stimuli. These parallel sentence recognition results of synthetic stimuli in both groups are promising for the application of a concatenative TTS synthesis system for generating speech training material in auditory rehabilitation software for CI users.

Index Terms: speech synthesis, semantically unpredictable sentences, cochlear implants, speech perception, intelligibility, aural rehabilitation

1. Introduction

1.1. Aural rehabilitation in adult CI users

The number of cochlear implant (CI) recipients has continuously risen worldwide for both adult and pediatric populations. In the US, approximately 42,600 adults and 28,400 children have received CIs as of 2010 [1].

After surgery, both adult and child recipients of CIs need to go through a learning/adaptation period to learn how to interpret the sounds generated by the CI system. This learning period varies from a few months to a few years [2]. Generally, speech-language therapists or audiologists provide aural (re)habilitation to help CI users adapt to electric hearing. For children with CIs, there are widely used therapy approaches such as Auditory-Verbal therapy [3]. However, despite the increasing number of adult CI users, there is no established aural rehabilitation (AR) protocol specific to adult CI users. In practice, many adult recipients with CIs do not receive any formal aural rehabilitation training after device activation [4].

1.2. Aural rehabilitation software

The lack of AR sessions with clinicians for adults with cochlear implants is often due to constraints on time, finances, and other resources [5]. To respond to the increasing demand for effective AR programs [4], there is a need for software that patients can use at home which will aid in the adaptation of cochlear implant usage. Recently several computer-based AR software programs have been made available online [6] and the effectiveness of these programs has been documented [7, 8]. However, existing AR software typically uses a limited number of prerecorded speech stimuli. For example, CasperSent, a program designed for sentence-level speech perception training, includes 720 sentences recorded from 2 talkers [9] and the LACE (Listening and Communication Enhancement) program employs about 2000 sentence stimuli. To provide appropriate feedback in an interactive dialogue format, AR software would require a large amount of pre-recorded speech, including story-length voice recordings. In many instances, however, appropriate feedback responses cannot be known in advance. Therefore, having a trained speaker record all of the speech that is necessary for effective AR software would be impractical.

Some of these obstacles in AR software development may be overcome through the use text-to-speech (TTS) synthesis. Modern large database unit selection TTS systems have been shown to produce speech that rivals natural speech in intelligibility. Even relatively small database unit selection systems can produce highly intelligible speech. Moreover, small database systems can be created rapidly to capture individual talker differences, regional dialects, and other natural forms of speech variability that are presumed to be important for AR. However, there is an intelligibility tradeoff between small database TTS systems that are relatively easy and inexpensive to create, and large database TTS systems that are time-consuming and relatively expensive to create. At present, very little is known about the perception of synthetic speech by CI users, and crucially, no study has examined the relationship between synthetic speech quality and its perception by CI listeners. It is possible that as the intelligibility of synthetic speech decreases from that of natural speech, it quickly becomes unintelligible to CI listeners, meaning that only very large database unit selection TTS systems can be used for AR.

1.3. Speech perception in adult CI users and materials used in common tests

To test the speech perception abilities of CI users, a variety of speech materials have been used. Common speech recognition tests use single words (e.g., consonant-nucleus-consonant (CNC) monosyllabic word test [10]) or short sentences (e.g., Hearing in Noise Test (HINT) sentences [11], CUNY sentences [12], Central Institute for the Deaf (CID) sentences, TIMIT sentences [13]) with or without competing noise. Semantically
unpredictable sentences (SUSs) are commonly used for assessing the intelligibility of synthetic speech. As far as we know, however, there are no previously published studies that have used SU sentences to test speech perception in CI users.

Various factors, such as duration of deafness, age at implantation, age at onset of deafness, and duration of implant use, have been found to affect speech perception performance in listeners with CIs [14]. Although CI listeners tend to perform poorly compared to normal hearing (NH) listeners, their results pattern similarly to NH listeners [15].

1.4. Purpose

The current paper reports on the perception of synthetic speech stimuli by adult users of CI devices. This study extends results reported for NH listeners in [16] where the quality of the synthetic speech was varied in a controlled fashion by varying the size of the speech database used to construct unit selection TTS voices. Here we compare the sentence perception results for CI listeners with those previously obtained for NH listeners at five quality levels of synthetic stimuli in addition to natural speech stimuli.

For synthetic speech to be considered useful for AR software, we would expect CI listeners to show a pattern of responding similar to that of NH listeners at different quality levels, but with poorer overall performance.

2. Method

2.1. Participants

Sixteen adults (7 females and 9 males), with at least one year of experience with their CI(s), participated in the study. CI systems varied within the group. Seven participants used Nucleus Freedom, four used Nucleus 5, four were Harmony users, and one was a MED-EL Opus 2 user. Participants received $25 for their participation. Demographic information is summarized in Table 1.

Table 1. Demographic information of the CI listeners

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age</th>
<th>Sex</th>
<th>Implantation Ear(s)</th>
<th>CI Use (years)</th>
<th>Cause of Hearing Loss (HL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01</td>
<td>69</td>
<td>F</td>
<td>Left</td>
<td>11</td>
<td>Genetic</td>
</tr>
<tr>
<td>C02</td>
<td>64</td>
<td>M</td>
<td>Right</td>
<td>6</td>
<td>unknown</td>
</tr>
<tr>
<td>C03</td>
<td>54</td>
<td>F</td>
<td>Both</td>
<td>8</td>
<td>Genetic or rubella</td>
</tr>
<tr>
<td>C04</td>
<td>70</td>
<td>F</td>
<td>Right</td>
<td>2</td>
<td>unknown</td>
</tr>
<tr>
<td>C05</td>
<td>67</td>
<td>M</td>
<td>Both</td>
<td>14</td>
<td>unknown</td>
</tr>
<tr>
<td>C06</td>
<td>70</td>
<td>M</td>
<td>Both</td>
<td>11</td>
<td>unknown</td>
</tr>
<tr>
<td>C07</td>
<td>66</td>
<td>M</td>
<td>Both</td>
<td>5</td>
<td>unknown</td>
</tr>
<tr>
<td>C08</td>
<td>71</td>
<td>M</td>
<td>Right</td>
<td>5</td>
<td>unknown</td>
</tr>
<tr>
<td>C09</td>
<td>72</td>
<td>M</td>
<td>Left</td>
<td>4</td>
<td>unknown</td>
</tr>
<tr>
<td>C10</td>
<td>67</td>
<td>F</td>
<td>Left</td>
<td>1</td>
<td>unknown</td>
</tr>
<tr>
<td>C11</td>
<td>43</td>
<td>F</td>
<td>Right</td>
<td>10</td>
<td>Rubella</td>
</tr>
<tr>
<td>C12</td>
<td>56</td>
<td>F</td>
<td>Left</td>
<td>11</td>
<td>Progressive HL</td>
</tr>
<tr>
<td>C13</td>
<td>55</td>
<td>F</td>
<td>Left</td>
<td>4</td>
<td>unknown</td>
</tr>
<tr>
<td>C14</td>
<td>82</td>
<td>M</td>
<td>Left</td>
<td>2</td>
<td>Occupational HL</td>
</tr>
<tr>
<td>C15</td>
<td>73</td>
<td>M</td>
<td>Right</td>
<td>2</td>
<td>unknown</td>
</tr>
<tr>
<td>C16</td>
<td>85</td>
<td>M</td>
<td>Right</td>
<td>5</td>
<td>unknown</td>
</tr>
</tbody>
</table>

2.2. Stimuli

Stimuli for the current experiment consisted of a subset of the SUSs used in [16]. These stimuli comprised 100 SUSs at six levels of speech quality (5 quality levels of synthetic speech and natural speech). The synthetic stimuli were generated by the ModelTalker unit selection TTS system. Five voices were created based on databases containing 200, 400, 800, 1600, and 3165 sentences recorded by a single male speaker. These corresponded to the first 5 quality levels respectively with quality level 6 corresponding to natural speech [16]. The same male talker also recorded the same 100 SUSs so that the natural speech condition contained stimuli that were precisely equivalent to the synthetic stimuli in speaker-specific voice quality factors.

Five sentence frames were used to create the SUSs. The sentence frames were 6 to 8 words in length and were limited to include no more than 10 syllables. The number of syllables in the words ranged from 1 to 4, but most words were monosyllabic. Examples of each sentence frame can be found in Table 2.

Table 2. Examples of Semantically Unpredictable Sentences

<table>
<thead>
<tr>
<th>Frame</th>
<th>Example Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The bed dwelt from the thin plane.</td>
</tr>
<tr>
<td>2</td>
<td>The camp saved the force that walked.</td>
</tr>
<tr>
<td>3</td>
<td>The good head poured the horse.</td>
</tr>
<tr>
<td>4</td>
<td>Why does the night bless the wrong house?</td>
</tr>
<tr>
<td>5</td>
<td>Suspect the law or the cell.</td>
</tr>
</tbody>
</table>

From these materials, five lists of 120 stimuli were created. Each list contained 20 sentences of each synthetic quality level (4 sentences of each frame type) plus an additional 20 natural sentences. Because there were only 100 different SUSs, the natural speech stimuli were repetitions of 20 synthetic stimuli in each list. The five lists differed in terms of which synthetic stimuli were presented at which quality level. The 20 natural stimuli always repeated the 20 sentences at the lowest quality level. Thus, each list consisted of 80 unique sentences and 20 sentences that were presented twice: once as natural speech and once as synthetic speech. Analyses to examine possible learning effects in these 20 sentences are planned, but not presented here.

2.3. Procedure

The experiment was conducted on a computer in a sound-attenuated chamber. Participants listened through loudspeakers. Prior to beginning the experiment, a most comfortable level was determined for each CI listener using an adaptive procedure.

On each trial, participants heard a single presentation of an SUS and then typed the sentence. They were told to guess at words even if they were uncertain about what they heard, and instructed to use XXX to replace any word that they were unable to understand or remember. Participants had ten practice trials. All practice stimuli were natural speech, and the voice and sentence frames were 6 to 8 words in length and were limited to include no more than 10 syllables. The number of syllables in the words ranged from 1 to 4, but most words were monosyllabic.

Examples of each sentence frame can be found in Table 2.

From these materials, five lists of 120 stimuli were created. Each list contained 20 sentences of each synthetic quality level (4 sentences of each frame type) plus an additional 20 natural sentences. Because there were only 100 different SUSs, the natural speech stimuli were repetitions of 20 synthetic stimuli in each list. The five lists differed in terms of which synthetic stimuli were presented at which quality level. The 20 natural stimuli always repeated the 20 sentences at the lowest quality level. Thus, each list consisted of 80 unique sentences and 20 sentences that were presented twice: once as natural speech and once as synthetic speech. Analyses to examine possible learning effects in these 20 sentences are planned, but not presented here.

2.4. Data analysis

Word-level edit distance was computed for each response. This edit distance tallies the total number of insertions, deletions, and substitutions needed to map a response to the stimulus sentence. The scoring takes into account homophones and unambiguous misspellings (e.g., rows, rose, and roze, or light and lite). A
normalized edit distance was calculated for each response by dividing the raw edit distance by the number of words in the stimulus sentence. These normalized edit distances were used for subsequent analyses.

Analyses presented here combine data from the corresponding conditions of the previously completed study [16] to show how performance of the CI listeners in the present study relates to performance on the same stimuli by 30 NH listeners.

3. Results

A 3-way mixed ANOVA (2 levels of group x 6 levels of stimulus quality x 5 levels of sentence frame) was run on the normalized edit distances. Significant main effects were observed for group, $F(1, 44) = 162.20$, $p < 0.001$, and quality, $F(5, 220) = 120.61$, $p < 0.001$. The main effect of sentence frame was also significant, $F(4, 176) = 25.11$, $p < 0.001$. All 2-way interactions significant, group x quality, $F(5, 220) = 1.39$, $p < 0.001$, group x frame, $F(4, 176) = 20.04$, $p < 0.001$, and frame x quality, $F(20, 880) = 7.03$, $p < 0.001$, as was the 3-way interaction, $F(20, 880) = 1.68$, $p = 0.032$.

Because of the large amount of variability among the CI listeners, as well as the disparity in sample size for the two groups, another 3-way mixed ANOVA was run using the 100 sentences as the sampling units. This analysis also revealed significant main effects for group, $F(1, 95) = 2092.76$, $p < 0.001$, quality, $F(5, 475) = 59.67$, $p < 0.001$, and frame, $F(4, 95) = 8.27$, $p < 0.001$. Significant 2-way interactions were also observed in this analysis, group x quality, $F(5, 475) = 4.73$, $p < 0.001$, group x frame, $F(4, 95) = 9.90$, $p < 0.001$, and frame x quality $F(20, 475) = 1.82$, $p = 0.016$, but the three-way interaction was not significant, $F(20, 475) = 0.75$, $p = 0.771$.

Means and standard errors (M and SE) of normalized edit distances for the six levels of quality by group are shown in Figure 1. The overall mean difference between NH and CI listeners is obvious (overall $M = 0.12$ and 0.52 respectively for HN and CI listeners). For both groups, mean normalized edit distance is highest at quality level 1 (voice based on just 200 recorded sentences) and lowest for quality level 6 (natural speech). At quality 1, the mean normalized edit distance for CI listeners was 0.61 ($SE = 0.07$) and for NH listeners 0.24 ($SE = 0.037$). At the best synthetic quality level (level 5) these means decreased to 0.50 and 0.094 respectively ($SE = 0.07$ and 0.014, respectively). For natural speech, the means were 0.37 and 0.027 for CI and NH listeners, respectively.

Two other noteworthy features shown in Figure 1 are, the trend for standard errors to decrease with improvement in speech quality for NH listeners, but not for CI listeners (whose SE is about 0.07 in all conditions), and the size of the increase in edit distance between natural speech and the best quality synthetic speech. Regarding the former, the relatively constant SE for CI listeners suggest that the listeners themselves are contributing the largest part of the error variance while for NH listeners, the error may be more related to stimulus factors. Regarding the second point, the difference in mean edit distance for natural stimuli and the highest quality synthetic stimuli is larger in absolute terms for CI than for NH listeners, but in relative terms, this difference is substantially larger for NH listeners. For CI listeners, normalized edit distance increased by a factor of 1.35 from natural speech to best quality synthetic speech. For NH listeners, edit distance increased by a factor of 3.3.

To further explore the relationship of synthetic speech perception in CI and NH participants, Pearson product-moment correlation coefficients were calculated using the average sentence scores for each listener group within each quality level (see Figure 2). The correlation was greatest for the lowest synthesis quality and generally decreased as quality increased. Significant correlations were found in all quality levels, $r(98) = 0.212 \sim 0.483$, $p < .05$.

4. Discussion

One of the motivating concerns behind this line of research was the possibility that synthetic speech stimuli — especially at
lower quality levels — might be too difficult for CI users. When compared to young listeners with normal hearing, CI listeners performed more poorly on this task. Crucially, however, CI listener performance mirrored NH listener performance in showing graded differences in intelligibility with graded differences in synthetic speech quality. Consistent with the finding in normal hearing listeners [16], recognition errors decreased as quality of speech stimuli increased. Although the sample size of the current study is small, the result suggests that the intelligibility of TTS synthesis in normal hearing listeners remains predictive of CI listeners’ performance throughout the range of quality levels examined.

This conclusion is further supported by the significant correlations between NH and CI listeners’ responses to individual sentences within quality levels as shown in Figure 2. The correlation is particularly strong at the lowest synthetic quality level where variation in sentence intelligibility is largest. As the quality of the synthetic speech improves, this correlation decreases, but remains significant. One likely reason for the decrease in correlation is that as quality increases, stimulus effects are increasingly overwhelmed by differences among listeners (especially for CI listeners), random effects, and ceiling effects (especially among NH listeners).

The SUS task is difficult even for normal hearing listeners [17]. Performance levels for CI listeners at the lowest quality exceeded 0.6 edits per word, indicating that on average listeners made some sort of perceptual error more often that every other word. Although it seems as though this error rate is too high to support effective communication, our best estimate of what this would correspond to in terms of short meaningful sentences (based on [18]) is an error rate of around 20%. That, in turn, is within the range considered to support communication. Hence, we expect that small to moderate sized unit selection voices would be usable in AR if meaningful sentence materials are used, as is typically the case.

5. Summary

The results of this study indicate a potential usefulness for synthetic speech in the development of AR software. Despite an overall higher error rate, users with CIs show a pattern of responding to SUSs that is similar to that of NH listeners over a range of synthetic speech quality levels. Because even small to moderate-sized unit selection inventories can be used, it should be possible to inexpensively create tailored synthetic voices that would correspond to in terms of short meaningful sentences. Therefore, even small to moderate-sized unit selection voices would be usable in AR if meaningful sentence materials are used, as is typically the case.

6. Acknowledgements

This work is supported by NIDRR grant #H133E080003 to the RERC on Hearing Enhancement. The authors thank Dr. Michael Teixido for his support with data collection and Dr. Jobayer Hossain for his advice on statistical analyses.

7. References