The effects of talker accent on recognition memory for vocoded sentences

Sjors Bech (s2399369)

Bsc Biomedical Engineering

Faculty of Science & Engineering

Rijksuniversiteit Groningen

Supervisor: Dr. Terrin N. Tamati

2nd Supervisor: Prof. Dr. Deniz Başkent

July 22nd, 2018

Contents Pages

Introduction

A cochlear implant (CI) is a sophisticated medical device designed to give a sense of hearing to patients with severe or profound cochlear damage, who do not receive a benefit from hearing aids. Since its initial appearance in the nineteen-eighties, over 219,000 people have undergone cochlear implantation, resulting in an improved quality of life for many deaf-born and post-lingually deafened patients.[1] A CI's internal component consists of a thin tube that spirals into the cochlea via the *scala tympani* and contains an array of electrodes to stimulate the auditory nerve at various intervals of the frequency spectrum, thus generating a sense of hearing. It is connected through the temporal bone to an external microphone equipped with vocoder audio processing.[2] A schematic representation of the orientation and function of a CI's components is displayed in figure 1. The general design hasn't changed drastically in the past decades, though the produced sound quality has improved over the years thanks to novel coding strategies and an increase in the amount of channels due to developments in electronics and signal processing. This trend is expected to continue in future models.

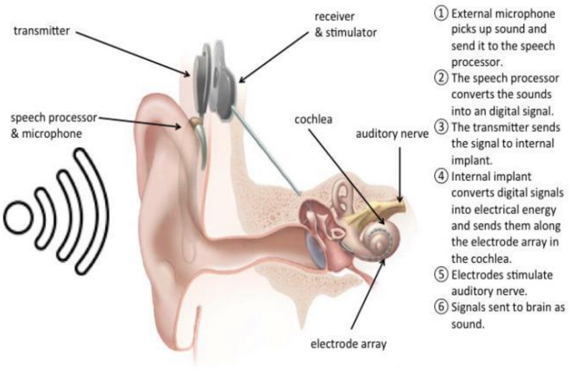
transmitter
receiver
& stimulator
speech processor
& microphone
cochlea
auditory nerve
electrode array
① External microphone picks up sound and send it to the speech processor.
② The speech processor converts the sounds into an digital signal.
③ The transmitter sends the signal to internal implant.
④ Internal implant converts digital signals into electrical energy and sends them along the electrode array in the cochlea.
⑤ Electrodes stimulate auditory nerve.
⑥ Signals sent to brain as sound.

Figure 1: Description of the workings of a CI. Courtesy of Pisoni & Tamati

Presently, CI implantation has become a reliable 'golden standard' treatment of superior quality, providing new hearing for infants as well as adults. Its large-scale application has raised an interest towards the patients' post-implantation perceptual system, especially regarding the effects on the mechanisms of speech perception.

Sound transmitted via the CI device is distorted and differs greatly from natural acoustic sound. The degraded input can impede a listener's ability to comprehend speech, as vocoding of the auditory signal causes loss of the spectral fine structure and frequency resolution.[3] Future developments in the device's design could potentially allow for more acoustic details to remain intact upon signal vocoding. It is essential for patients to learn to adapt to this new signal in order to show benefits in post-implantation speech and language outcomes. However, in spite of its frequent application, little is known about the manner in which the patient's brain adapts to this altered and initially unrecognizable sensory input, particularly when it comes to the learning and memory processes involved in speech perception.[4] Furthermore, post-surgery clinical hearing tests are often insufficiently representative to real life situations. While patients learn to understand speech in a controlled environment, outside of the lab they are confronted with complications such as noise, 'conversational speech', multiple talkers speaking simultaneously, etc. Materials used in clinical testing could be made more representable to realistic conditions by involvement of such factors. Accents and other types of acoustic variability can also pose significant obstacles for CI users to understand speech. Currently, our understanding of various effects on speech recognition and retention from a CI users' perspective is incomplete.

Speech Perception by normal hearing listeners

A fundamental question in the multidisciplinary field of speech perception is how listeners are able to perceive and understand speech in these adverse and variable conditions. A categorical distinction between linguistic and indexical (also described as non-linguistic) aspects of speech emerged initially, which suggested that the perceptional system treats them as separate classes of information. Contrary to this early view, recent research suggests that indexical information is not disregarded and plays an important role in speech perception and spoken word recognition.[5] Some have even argued that detailed episodic memories form the foundation of our entire mental lexicon.[10] While hearing a talker for the first time, a listener perceives and memorizes the talkers' characteristics like gender and age (i.e. indexical properties), affecting long term memory and aiding in adjusting to acoustic patterns that are characteristic to the individual talkers' speech. Linguistic experiments have indicated that this type of learning and memory encoding allow for improved intelligibility by a listener that is exposed to speech repeated by the same or acoustically-similar talker[5] - this phenomenon is generally referred to as the talker repetition effect. The adaptive ability to record and reuse a detailed speaker assessment appears vital for mapping acoustic signals onto lexical representations stored in the mental lexicon - in other words, for recognizing speech.

Since new talkers introduce new indexical details to be retained, one could conjecture that talker variation can have a negative effect on speech intelligibility. In essence, it would mean that listeners would be more accurate and/or faster to recognize speech produced by a single talker compared to speech produced by multiple talkers, as more talkers entail more indexical information. This notion has been confirmed by a study from Mullennix et al.[8], which found that spoken word intelligibility was inversely related to the amount of talkers that produced the stimuli. That is, word recognition decreased with an increasing number of talkers within a block. Improved intelligibility by a listener enables more accurate future recognition via long-term memory, and more accurate explicit judgement about earlier perceived words. This long term benefit has also been demonstrated by experiments on talker variability, speaking rate and perceptual learning.[5, 7, 12]

Talker accent introduces much acoustic variation to speech, and has been shown to affect spoken word recognition. Lexical processing (i.e. recognizing and encoding the lexical aspects of speech), unsurprisingly, is exercised quickest and most accurate for speech presented in the listener's native accent, since the listener has heard speech in this accent most frequently and over an extended period.[9] Hearing a word in a less familiar accent requires the listener to connect an acoustically deviating variant to its known lexical 'definition', which requires effort. Regardless if the acoustic signal is successfully mapped onto the target lexical category, it often results in a less robust updating of the lexical representation of the specific lexical item, meaning its association to the lexical representation is weaker than its native variant. Adapting to an unfamiliar accent takes effort and time, and eventually leads to a more robust lexical representation of the accented words, in a process called normalization. The normalization of non-native accents is a long term process that has been argued to take place at the expense of acute encoding processes that aid in speech recognition.[10]

Talker voice and accent variability are some examples of factors known to influence stimulus intelligibility that can be introduced as parameters in linguistic experiments. Variable characteristics such as speaking rate and speaking style (i.e. how clearly speech is pronounced) have also shown to have substantial effect on recognition, with slower speaking rates and more clearly articulated speech resulting in greater intelligibility.[8, 12] These outcomes strongly portend the contribution of memory in understanding and recognizing speech. Adapting to unintelligible auditory input seems to occupy the memory, partially preventing it to aid in word recognition. Similarly, semantic properties of the stimuli show a strong effect on intelligibility; sentences with high semantic context information are more intelligible than sentences with low semantic context information. Van Engen et al.[12] found that anomalous sentences, syntactically legal but void of context, yielded lower intelligibility scores than semantically meaningful sentences - this effect being even more profound than that of speech clarity. An increase in intelligibility due to these factors also influenced the subject's recollection, shown by improved scores on a recognition memory task. More easily intelligible speech appears to be less cognitively cumbersome to process, saving more capacity for encoding it into episodic memory and making it easier to recognize afterwards, if repeated. A study which looked at adults with NH and with poor hearing, implied that sentence predictability, specifically the predictability of a keyword at the end of the sentence, had significant effect on intelligibility.[13] A distinction was made between sentences in which the last word was a likely follow-up to the prior two words (high predictability), and those in which it didn't (low predictability). Recall scores for both groups of subjects were high for the last word of the former type of sentences. These findings further strengthen the notion regarding the crucial role of memory in word recognition.

Speech Perception by CI Users

Hearing and speaking outcomes can vary greatly after CI placement, and knowledge about the nature of CI users' perceptual system and the contribution of memory is limited and fragmentary.[4] It appears, for NH perceptual systems, that any source of variability or subjective abnormality puts strain on the memory, which is constantly adapting to these variations. This process of normalization leads to better recognition over time, but requires great investments of cognitive resources. One theory that addresses the subject, the 'Effortfulness Hypothesis', speculates that understanding vocoder simulated or otherwise degraded speech requires more processing effort than normal, leaving less cognitive capacity for memory encoding processes to utilize.[8] This idea indicates that CI users' word or sentence recognition would benefit less from the talker repetition effect than NH listeners. The implications that impaired encoding have for CI users are largely debatable but it has been considered as a possible cause for obstructing episodic memory functions and perceptual learning processes.[7]

Considering both the input degradations imposed by the CI and the demanding normalization process, it is fathomable why CI users struggle to cope with talker voice and accent variability. Research has repeatedly demonstrated that CI users are less adept than NH listeners at perceiving important talker voice and gender characteristics.[14] However, findings from Tamati and Pisoni also suggest that, in spite of CI users' weakened ability to take in talker details, they were still somewhat sensitive to foreign-accent variability. Considering that CI users have been thought to rely on different acoustic cues for discerning speech than NH listeners, and as signal vocoding diminishes the notability of some cues, recognizability of some accents could be affected more than others.[15]

Due to the limitations in perceiving indexical details, vocoded speech recognizability could be even more negatively impacted than unprocessed speech when confronted with varying talker effects like speaking rate, style, as well as talker variability. The latter might be most influential since adjusting to new talkers relies heavily on implicit memory and learning. Contrarily, however, one could argue that, if talker details are less evident through vocoder simulated speech, it restricts the perception and encoding of the indexical properties. In theory, this would leave more resources to the lexical assessment of speech. In other words, if the perceptual system commits less resources into encoding, i.e. long term adjustment, it might benefit acute word recognition performance.

A better understanding of CI users' speech perception in real-world listening conditions could help researchers and biomedical engineers to optimize implant customization and auditory training programs. Conforming clinical testing to more realistic standards, combined with more advanced CI models and a more condensed acclimatization period, could lead to enhanced implant success rates, especially outside the lab or clinic. Research discussed in this report aims to investigate the contribution of memory processes in vocoder simulated speech perception through empirical testing, and to determine how this relates to NH patterns. It serves as a pilot study to gain insights for more thorough, focussed, and large scale studies to be conducted in the future. Conclusions may offer insight into the perceptual systems of CI users and could eventually help limit the struggles they currently face in real-world speech perception.

The Current Study

While the role of talker and accent variability in speech recognition and memory has been investigated extensively for NH cases, the implications for CI users are still quite unclear. In the current study, acoustic noise-vocoder simulations were used to represent CI hearing. Presenting CI-simulated speech to NH listeners allowed us to take a first step in understanding the effects of these sources of variability for CI users' speech perception and, more broadly, on their speech perception abilities in real-world situations. Specifically, it could lead to solutions to the current problem that CI users experience in dealing with acoustically varying speech types.

The primary focus of this study was to evaluate the contribution of encoding to speech recognition and retention. Talker accent and sentence predictability were introduced as variables in a recognition memory experiment to illuminate memory processes. In an exposure phase, listeners were first exposed to high and low predictability sentences by hearing and reproducing them verbally in succession. Sentences were produced by native Dutch speakers and two groups of non-native Dutch speakers, who were native speakers of Frisian and German. If encoding was to take place, it would occur during the presentation of these sentences. Subsequently, in the testing phase, their explicit memory for sentences was tested via presentation of both newly introduced and old sentences repeated from the exposure phase, produced by either the same talkers, talkers with the same accent, or talkers with a different accent as during the initial exposure. This variability in talkers and accents from the exposure to the test phase is further referred to as the *repetition condition*. Subjects were asked to make explicit judgements about these sentences; the accuracy and time to make their judgements were recorded and assessed.

Based on findings of research on both NH and CI implanted subjects discussed earlier, the sentences should be highly intelligible in the exposure phase, especially the high predictability types. Relatively uncomplicated stimulus recognition could facilitate encoding of information. In theory, high predictability sentences should be easier to recognize for the listeners than low predictability sentences, possibly leading to increased recognition memory performance.[12]

Because CI users are thought to be sensitive to accent features in speech, recognition accuracies in the exposure phase are expected to vary by accent. This could result in more accurate judgements in the testing phase, depending on the amount of encoding that is allowed by the increased ease of processing. If this effect of accent does not appear, the input was too degraded to perceive accent specific features, or the recognition scores were too high for accent effects to appear.

Sentences produced by a native talker are considered most likely to allow intake of talker details to take place, which could assist in recognition memory tasks[11, 14]. Hence, it is within this category that the highest judgement accuracies were expected during the testing phase. This hypothesis is reinforced by earlier results on the intelligibility of foreign-accented speech among CI users.[5, 9] Outcomes resembling this expectation would signify that CI users' perceptional systems involve at least some encoding of lexical details - and in somewhat similar fashion to their NH counterparts - in spite of being exposed to input that is limited in acoustic-phonetic detail. If so, these details would aid in judgement tasks in the testing phase, when the stimulus is repeated by an identical talker, resulting in enhanced accuracy.

However, since sentence recognition in CI users is facilitated with carefully articulated speech with slow speaking rates, it is also possible that the results would show non-native speech to be more intelligible than native speech. This counter intuitive idea builds on the assumption that native speakers follow more conversational speaking patterns, whereas some non-native accented speakers pronounce certain sounds more slowly and notably. This means that non-native talkers' speech contains more evident acoustic cues for the subjects to notice, resulting in better sentence recognition. Better recognition in the exposure phase could allow for (more) encoding to occur, which would lead to improved performance in the testing phase for these sentences. These findings would be restricted to talker specific effects such as speaking rate, style or accent, and call for future research with a larger amount of subjects. If this effect does not emerge, then increased recognition accuracy for accented speech apparently did not lead to 'easier' processing, therefore encoding was obstructed.

The main question that would follow is whether the apparent absence of encoding benefits would be due to lacking availability of information in the vocoded speech, or because the information was available, but the conditions did not allow for the listener to encode the available indexical information. One way to explore this issue in the current study is to compare repetition effects for repeated talkers and accents separately. If increased judgement scores are found for stimuli that are repeated by talkers with the same accent (which include identical talkers), it is likely that encoding of accent details has taken place. If benefits only show for repetition of identical talkers, it would strongly suggest that the listeners have encoded talker specific (indexical) details. While previous research has shown that the perception of talker information is limited in CI users (e.g., Cleary et al., 2005), CI users are relatively sensitive to regional and foreign accent differences (Tamati, Gilbert, &

Pisoni, 2014; Tamati & Pisoni, 2015). If bottom-up cues are determining the repetition benefit, then limited talker information may result in a limited benefit in recognition memory for repeated talkers but available accent information should result in a robust benefit in recognition memory for sentences repeated with the same accent. If the repetition benefit is primarily influenced by top-down cognitive factors, then effortful processing of accented speech in the exposure phase may lead to a limited repetition benefit for sentences produced with the same accent. Further, if we assume that Frisian-accented speech is relatively easier to process than a non-native German speech, since the speakers learned Dutch at a young age and speak a variety of Frisian-accented Dutch, then we might also expect stronger encoding (and stronger repetition effects) for Frisian-accented sentences than for German-accented sentences. Thus, while limited talker information may restrict talker repetition benefits, effortful processing might restrict accent repetition benefits, especially for German-accented sentences.

If repetition effects are absent or limited, it would suggest a diminished ability of the listeners to encode talker details.[8] In other words, if talker or accent variability does not influence recognition memory accuracy scores, then little or no encoding has taken place. This would suggest that there is a fundamentally different mechanism active in CI users' perceptual systems, since there would be no apparent benefit for recognizing sentences when hearing an acoustically similar talker in comparison to a different one - unlike NH cases.

It would raise the question as to how CI users employ their encoding mechanisms, if at all. It would also mean that the subjects were not just unable to perceive talker details, suggested by absence of repetition benefits, but also failed to distinguish between the various accents, as there would be no apparent effects of stimulus repetition by different accented talker on memory recognition. In this case, the likely cause for this effect would be that the information is available to the subjects, but that they weren't able to process the information due to demanding conditions.

Methods

Listeners

As listeners, a group of 12 students participated in this study. All listeners were native Dutch speakers (mean age = 22.9, range = 21.6 - 24.6), exhibiting normal hearing (NH). Though most resided within the city of Groningen, the Netherlands, their birthplace within the Netherlands varied. People of German or Frisian origin were excluded to take part as listeners, as familiarity with the non-native accents was deemed undesirable. Also, questionnaires were given to all subjects with questions about their linguistic background. It was conformed by these means that none would likely exhibit bias towards any of the accents. All were found to exhibit normal hearing pure-tone thresholds of 20 dB or below at the frequencies 250, 500, 1000, 2000, 4000, 6000, and 8000 Hz. Two participants served as pilot subjects for the recognition memory experiment; their data was not included in the final data analysis.

Stimulus materials

Materials for the current study were selected from a larger corpus of spoken Dutch. A total of 20 talkers was assembled, consisting of native and non-native (accented) Dutch speakers. They filled out the same questionnaire as had been given to the listeners. All talkers were students in Groningen during the recordings, or had been in the past. 16 native talkers (8 male and 8 female), who were of mixed descent within the Netherlands formed the native, or standard group (ST). These native Dutch speakers were judged by the author to speak a standard variety of Dutch. The second group consisted of two females, who speak Frisian (F) as a native language. The remaining two talkers, who were native German (G) speakers, made up the third group. Sentences which were spoken by, for instance, a German talker will be further referred to as G or G-type sentences, while sentences produced by a native Frisian talker will be referred to as F or F-type sentences. The four non-native Dutch participants were born in their respective regions and expressed an audible regional/foreign accent when speaking Dutch. All talkers were of ages 18 to 30 and demonstrated NH thresholds. Financial compensation was offered to all participants at a rate of 8 euros/hour.

Four categories of materials were recorded from each talker in a quiet, noise-isolated booth. The materials consisted of 250 sentences of varied predictability, 300 words, 300 non-words and 150 anomalous sentences. The anomalous sentences were syntactically correct but lacked semantic context. They had been composed by substituting the content words between the semantically correct sentences. Lastly, the 'words' and 'non-words' were all single syllable materials. The non-words were formed by arranging common vowels into phonologically probable non-existing words, resembling examples from Dutch single syllable words.

Only the semantic sentences were used in the current study, taken from two talkers of each language group, and six talkers in total. The remaining materials were collected for other studies. Each semantically correct sentence contained 5-8 words, of which four were single-syllable 'content words', the last of these being a high-frequency keyword. The keyword in each sentence was, regarding the predisposed context, either highly predictable given the preceding context (HP) - e.g., 'De ongelukkige vrouw scheidde van haar *man'* (The unhappy woman divorced her *husband*) - or relatively unpredictable given the preceding context (LP) - e.g., 'De arme man wil graag de *kroon'* (The poor man would like the *crown*). Each keyword was used twice, appearing in both an HP and an LP sentence. For this experiment, a total of 96 sentences (48 HP and 48 LP) were selected. 8 sentences (4 HP/4 LP) per talker appeared in the exposure phase and were repeated in the testing phase ('old' sentences). In addition, 48 sentences (8 sentences (4 HP/4 LP) per talker) that were not presented in the exposure phase ('new' sentences) also appeared in the testing phase.

Vocoding

The original signals were filtered into 8 bands between 150 and 7000 Hz (Greenwood, 1990). Filtering stimuli and white noise was achieved with eighth order, zero-phase, bandpass filters produced analysis and synthesis carrier bands, respectively. The temporal envelope in each analysis band was extracted using half-wave rectification and fourth order, zero-phase, low-pass filter. The synthesis carrier bands were first modulated with envelopes, then summed to produce the vocoded stimulus and finally adjusted to the same level as the unprocessed stimulus.

Procedure

Listeners were seated in an anechoic room in front of a touch-screen computer monitor, approximately 1 meter away from a speaker. The experiment consisted of two phases, starting with the 'exposure phase' followed by the 'testing phase'. A short spoken text (North Wind and the Sun, IPA) with 8-channel vocoder simulation was played to the subjects beforehand to familiarize them with the acoustic-phonetic aspects of 8-channel noise-vocoded speech.

The exposure phase served to familiarize the listeners with a set of 48 sentences. Listeners were presented with a single sentence and were requested to orally repeat what they understood. Responses were recorded, and afterwards, were scored word for word by a native speaker of Dutch (the author) to determine their overall accuracy. The accuracy score was determined by the percentage of words that was repeated correctly out of each sentence.

During the following 'testing phase', 96 stimuli were presented, which had either been played during the exposure phase (old), or had not been presented before (new). The listener was asked to judge whether they had heard the sentence before or not by choosing from options 'old' and 'new' as fast as possible without sacrificing accuracy. These options appeared as boxes on the right and left side of the touch-screen and thus could be selected by being tapped. A correct answer resulted in a score of 1, and an incorrect answer a score of 0.

Within the 'testing phase', repeated stimuli could be of the same (identical) talker (IT), a different talker from the same accent group (SA), or a different talker from a different accent group (DA). Half of the sentences presented in the testing phase was 'new' (48), 25% was of the IT condition (24), and the remaining 25% consisted of 12 DA- and 12 SA-talker sentences in regard to the exposure phase. As said previously, the sentences used in each experiment consisted of equal parts of the LP and HP types. If the talker differed but the lexical content of the sentence did not, it was still to be considered an old sentence.

Results

Exposure phase

The mean accuracy for each accent is represented in figure 2 in proportior correct. The grand mean amounted to 92.8% word accuracy (SD = 3.3). A two-way ANOVA test on the subjects' scores with the exposure accent (G, F, ST) and sentence predictability (HP, LP) as within-subject factors revealed a significant main effect of accent (F(2,45) = 9.685, *p*<.001) but no significant main effect of sentence predictability on accuracy. Sentences produced by Frisian talkers yielded a mean score of 96.5% (SD = 2.7), whereas the German accented stimuli resulted a mean score of 90.2% (SD = 1.3). ST intelligibility ratings averaged 91.7% (SD = .8).

A Tukey mean-difference post-hoc test with 95% family-wise confidence level (α = .05%) resulted in the following comparisons. For exposure accents, F and G mean outcomes showed a significant difference, (*p* = .001), as well as ST compared to F (*p*=.017). In both these cases, F sentences produced higher mear recognition accuracy. ST scores compared to G did not reveal a significant difference.

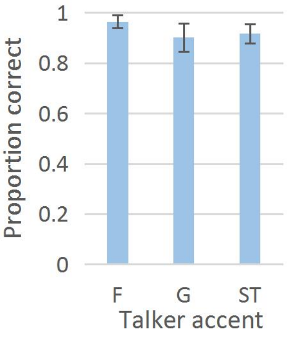

Figure 2 - Mean percentages of words correct by accent, standard deviation included (grand mean = 0.93, SD = .033)

Testing phase

Only 'old' sentences that were introduced in the exposure phase, thus presented for the second time in the testing phase, were studied. Two three-way ANOVAs on the intelligibility ratings and response times were executed to assess the results of the testing phase. Exposure accent (G, F, or ST), sentence predictability (HP or LP) and repetition condition (IT, SA, or DA) were the within-subject factors.

For accuracy scores, a significant main effect was found for sentence predictability (F(1, 153) = 4.697, *p* = .032), The interaction of sentence predictability with exposure accent (F(2, 153) = 5.938, *p* = .003), and the three-way interaction of sentence predictability x exposure accent x repetition condition (F(4, 153) = 2.785, *p* = .029) were also significant. HP and LP sentence scores averaged .739 (SD = .207) and .651 (SD = .201) proportion correct, respectively. Results are shown separately for exposure accent, sentence predictability and repetition condition in Figure 3 below, amounting to 18 (3x3x2) separate means.

Post-hoc analysis of the testing phase data was, like the exposure phase results, achieved using a Tukey mean-difference test. Significant difference in means was found between HP and LP conditions (*p* = .061), a relation that was also indicated by the ANOVA main effect outcomes. Two-way interaction outcomes revealed the influence of sentence predictability to be significant only within ST condition items (ST:HP-ST:LP, *p* < .01), where ST:HP scores were on average 0.28 points higher than ST:LP. Within a set HP condition, ST scores were higher than G by .17 points (ST:HP-G:HP, *p* = .32), though this difference in means was not of statistical significance.

Of the three-way condition comparisons, DA:ST:HP was significantly higher than DA:ST:LP (*p* = .048). DA:ST:HP mean scores were .4 points higher than DA:ST:LP conditions, and an equal difference was found between DA:ST:LP and DA:G:LP conditions. The significance of the latter two comparisons is less statistically compelling, however (*p* < 0.3). To provide an overview of the results of testing phase scores analysis, Table 1 has been added.

New-condition sentences were added to provide variance of stimuli, and the possibility of discerning form the earlier presented sentences, and to check if listeners were not biased in their responses. A two-way ANOVA was applied on the accuracy scores of new sentences to act as control for repetition condition effects. Main effects were found for sentence predictability (F(1,45) = 9.271, *p* = .004) and sentence predictability x exposure accent (F(2, 45) = 4.39, *p* = .0244).

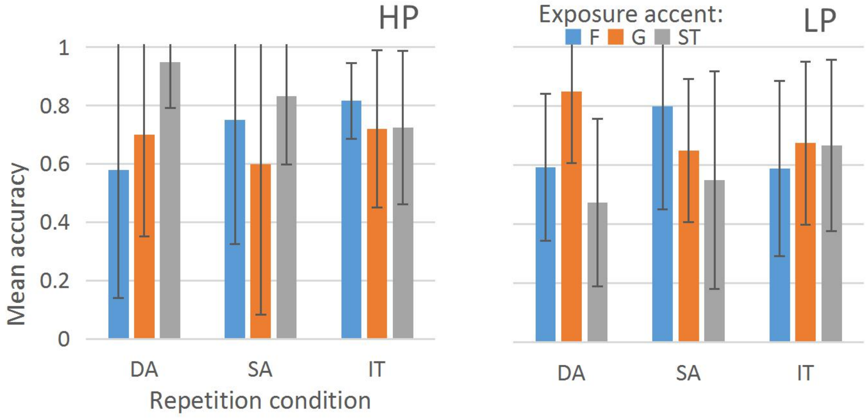

Figure 3 - Mean accuracy for the testing phase task according to exposure accent and repetition condition. Shown separately for HP and LP sentence predictability (grand mean = 0.695, SD = 0.284). Standard deviations are also displayed in the bars.

Conditions	Difference	*p* value
ST:HP - ST:LP	+0.28	**0.009**
ST:HP - G:HP	+0.17	0.324
DA:ST:HP - DA:ST:LP	+0.5	**0.048**
DA:ST:HP - DA:F:HP	+0.4	0.298
DA:ST:LP - DA:G:LP	-0.4	0.298

Table 1: Tukey test analysis of some relevant testing phase accuracy scores. 2- and 3-way condition mean differences are shown. A positive difference between conditions A-B means conditions A had a higher average than conditions B.

For the assessment of the response times, only the results of the correct answers were taken into account, because sources of inaccuracy could cause variability. The analysis of these results revealed significant main effects of exposure accent (F(2, 153) = 7.533, *p* < .001) and of sentence predictability by repetition condition (F(2, 153) = 5.168, *p* = .007). Due to the relatively less apparent effect of sentence predictability on this dataset, no distinction was made according to this condition, and both sentence types were processed into a single chart (figure 4). The means for exposure accents F, G and ST were 3641.9ms (SD = 326.7), 3267.2ms (SD = 320.7) and 3274.2ms (SD = 310.5), respectively. Post-hoc analyses revealed significance between G and F (*p* = .004) and between ST and F (*p* = .003). In both these cases, F items yielded the longer mean response time.

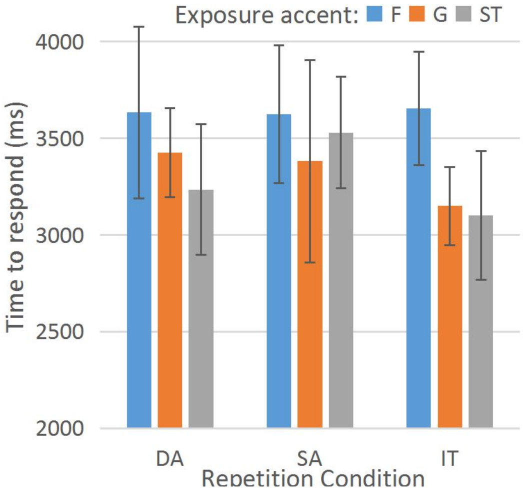

Figure 4 - Mean response times for the testing phase task according to exposure accent and repetition condition. (grand mean = 3395.1ms SD = 355.9)

Discussion

Exposure phase

This study aimed to investigate the contribution of encoding in CI simulated speech recognition of varying accents. A recognition memory task was used to realize this goal. The exposure phase served as a familiarization process during which encoding could take place. Effects of accent variability and sentence predictability were analyzed. Subjects were expected to demonstrate high recognition accuracies. Variation was expected to lie mostly in the effort to recognize the various types of sentences. This variation would express itself in the testing phase outcomes.

Recognition accuracy scores were high, showing that the subjects had indeed no problem with intelligibility under vocoder simulation. This was predicted in the hypothesis. No clear effect of sentence predictability was found on intelligibility, likely because accuracy was near ceiling. However, talker accent influenced accuracy scores in the exposure phase. F type sentences were recognized with the highest overall accuracy. It was hypothesized that this would probably be due to the distinct acoustic cues conveyed by accented speech. These features were apparently sufficiently evident for the listeners to notice, leading to higher sentence recognition accuracy.

Despite this supposed effect of talker accent, the other accented sentences, namely, the G type, yielded no exceptionally high accuracy scores. This opposite outcome could be explained by the notion that German accented Dutch speech contains less noticeable acoustic entities or that the cues are less likely to be conveyed through vocoder simulation. Another likely explanation is that the German speakers are less proficient and fluent in Dutch, which leads to less predictable speaking patterns and lower recognition scores. Since Friesland is a province of the Netherlands, Dutch is spoken alongside Frisian in most areas, so the Frisian talkers started learning Dutch at an earlier age than the German born talkers. Therefore, while speaking with an audible accent, they possessed greater fluency in Dutch than the German talkers.

Furthermore, talker specific variance in speech characteristics such as speaking rate and general clarity were known to vary between the talkers, and could also be a contributing factor for the divergence in accent intelligibility. The relatively small size of the talker groups (2 each) might have amplified these talker specific effects. Perhaps these effects could have been avoided or limited by increasing the size of the talker groups. A way to minimize talker speaking variability in future research could be to instruct the talkers more specifically as to how they are supposed to pronounce the materials, e.g., clearly and slowly, or naturally and conversationally. During the recordings of the current study, talkers were asked to speak both clearly and naturally, possibly leading to varied speaking styles. The found high accuracy scores and the varying accuracies between accents were predicted. Nonetheless, results would have been more accurate and interpretable in regard to accent specific effects if talker specific effects were less prominent. This can be achieved in future studies by compiling more talkers with greater within-accent variability, and provide them with clear instructions.

Testing phase

The testing phase consisted of memory recognition tasks to test for signs of encoding benefits in explicit judgement performance. The variables exposure accent, sentence predictability, and repetition condition were examined. Varying judgement scores between repetition conditions would indicate possible talker repetition effects and thus provide evidence for presence of encoding processes. It was anticipated that an improved judgement accuracy for IT items would suggest encoding benefits. The same applied for ST:HP sentences, which were the group of stimuli predicted to be most easy to process, hence allowing encoding to occur, resulting in higher scores.

While sentence predictability did not influence sentence recognition accuracies in the exposure phase, it appeared to have influenced the accuracy of old/new judgements in the testing phase. Though only in ST conditions, HP type sentences, resulted in significantly higher scores than LP items. HP sentences presented in the exposure phase were recognized most accurately. This possibly led to improved retention of the lexical content, thus explaining the increased scores for these sentences in the testing phase. LP sentences likely required more resources to recognize, at the expense of encoding processes. A similar finding has been described by van Engen et al.[12] This effect of sentence predictability was expected for all sentences, but was found to be significant only in same talker situations. Moreover, an interconditional interaction was revealed by the ANOVA, as mutual influence of exposure accent and sentence predictability on testing phase scores suggested that the effects of accent on explicit recognition memory varied by stimulus predictability.

Contrarily to the exposure phase results, it appears from the testing phase that the type of exposure accent does not influence accuracy scores on its own, rather its effect varies by sentence predictability. HP sentences that were presented by native talkers did not produce the highest recognition accuracy in the exposure phase, but, interestingly, were judged most accurately in the testing phase. Though accent can provide listeners with acoustic cues that aid in direct, short term reproduction (which appears to have happened occasionally during the exposure phase), native spoken sentences seem to allow for better scores in the explicit judgement tasks. This enhanced performance could be explained by the ease of processing, once again presuming that easier to recognize speech allows for more encoding processes to take place. HP, ST sentences are easiest to process, saving the most capacity for encoding to occur. This effect would not have happened for non-native exposure sentences as subjects were occupied with identifying acoustic accent details. High native accent mean judgement scores could however just as well be caused by more robust access of accompanying lexical information, meaning no talker details were stored.

Elaborating on the influence of exposure accents on accuracy, the analysis on ‘new’ item scores in the testing phase yielded no significant main effect of this variability. This rules out presence of a bias towards certain accents, as they all yielded similar accuracies when presented without initial exposure. It was found that exposure accents did have substantial influence on response times. This effect manifested itself most notably for F type sentences, which produced the longest mean response times.

These two mentioned factors, sentence predictability and exposure accent, exhibited a significant three-way main effect in interaction with repetition condition, meaning the effect of repetition condition on accuracy varied by the other two conditions. Though some post-hoc findings were significant and consistent with more earlier observed trends on predictability, many supposed accent effects are statistically not very sound and come across as dispersive. They could be caused by talker specific variations, as mentioned in the discussion of the exposure phase.

No relevant independent influence of repetition condition was observed on accuracy. Although some significant interactions with repetition condition emerged, i.e. the previously mentioned two- or three-way interactions, these were sporadic and a clear picture of the role of repetition condition did not emerge. It was hypothesized that the absence of a concrete main effect of repetition condition would strongly insinuate a deficiency in talker detail perception by the listeners. If it makes no apparent difference for judgement accuracy whether the talker has been heard before or not, then no talker repetition effect has occurred; if it did, it would have resulted in a positive influence on accuracy in IT conditions. Both the hints of accent sensitivity and limitations in in talker perception under vocoder simulations have been found before by Tamati and Pisoni.[14]

On response times, one specific set of conditions did reveal an effect of repetition condition. Within HP:ST conditions, SA presentation of stimuli resulted in longer response times than DA repetition. This trend is in accordance with earlier findings by Clopper et al[9]. However, response times could be considered unsuitable for being subject to any conclusive deductions. Since sentences are thought to be recognized via specific keywords, and these keywords are situated at various locations within the sentences, the moment of recognition can vary between sentences. Therefore, response times could be very stimulus-specific.

Implications for CI users

As mentioned previously, this study utilized 8 channel vocoding to simulate CI processed hearing. Any conclusions derived from vocoder simulated experiments are speculative and not necessarily applicable to CI cases. They could, however, provide insights into more optimal future research design with actual CI users.

If NH subjects cannot access indexical talker details in testing conditions, it is fathomable that CI users struggle with this in real-world situations. This means that CI users are generally unable to distinguish between people relying solely on their hearing, leading to difficult situations. Moreover, speech perception is used for more than just understanding a talker's words, as accompanying emotion and intonation can be very suggestive towards underlying context of speech.

The current view on auditory training programs remains that clinical testing should be focussed on patient habituation towards real life situations. The complexities entailing various accents, speaking styles and other circumstances remain obscure, and should be studied more thoroughly in order to allow CI users to overcome them.

Research in the field of speech perception and adjustments to acclimatization processes are not enough however, as the device in question should keep technologically developing at the steady rate it has been doing for decades. It appears that vocoded auditory signals does not keep the necessary acoustic-phonetic fine structure to fully accommodate the transfer of indexical information without some loss. Future models should allow its users access to more detailed sound for them to be able to detect the additional information carried by speech. Perceptional limits in everyday situations currently pose a communicative obstacle for CI users. Attempting to overcome these is an attempt to grant CI users the same benefits of hearing as NH people do, meaning for them to be fully sensitive to the subtle features and nuances used in speech.

Conclusion

Listeners have demonstrated to be sensitive to accent specific acoustic details of speech, which in some conditions allowed for more accurate sentence recognition. Regional accent was found to provide notable acoustic features that aided in vocoded speech recognition. Despite of increased recognition performance, no evidence was found to reveal encoding of these acoustic accent-specific details by the listeners, since judgement scores for these sentences appeared unaffected. The results of this pilot study suggest that the listeners also did not encode indexical information, possibly because it was not well conveyed in the noise-vocoder simulations of CI hearing. No talker repetition benefits were revealed by the analyses, as repeating a stimulus by an identical talker did not result in higher judgement scores. It was found that future research requires larger talker groups with greater within-accent variability to allow examination of talker repetition effects among various accents.

Encoding of lexical information occurred most evidently if a sentence was easy to process, meaning its linguistic content was highly predictable and presented by a native speaker, leading to increased performance in judgement tasks. A reduced effort necessary to process a sentence did not lead to higher recognition accuracies. The conditions present during exposure determined the extent to which listeners can perform encoding. Recognizing unpredictable or non native speech has been found to demand cognitive processing capacity, obstructing the benefits of encoding. Real life situations entail many similar demanding conditions, which makes encoding of indexical information a very unlikely faculty for CI users. Future research should further investigate memory for spoken sentences under realistic conditions, and within CI users.

References

1. American Speech-Language Hearing Association (ASHA) http://www.asha.org/

2. M. Hainarosie, V. Zainea, and R. Hainarosie. The evolution of cochlear implant technology and its clinical relevance. J Med Life. 2014; 7(Spec Iss 2): 1–4.

3. JL Loebach. Cochlear Implant Simulation Tutorial. Sp. Res. Lab. (2005)

4. Pisoni DB, Kronenberger WG, Chandramouli SH, Conway CM. Learning and Memory Processes Following Cochlear Implantation: The Missing Piece of the Puzzle. Front Psychol. 2016 Apr 8;7:493.

5. Pisoni DB. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. David B. Pisoni. Speech Commun. 1993 Oct; 13(1-2): 109–125.

6. Martin CS, Mullennix JW, Pisoni DB, Summers WV. Effects of talker variability on recall of spoken word lists. J. Exp. Psychol.: Learning, Memory and Cognition. 1989; Vol. 15:676–684.

7. Goldinger SD, Pisoni DB, Logan JS. On the locus of talker variability effects in recall of spoken word lists. J. Exp. Psychol: Learning, Memory, and Cognition. 1991; Vol. 17(No. 1):152–161.

8. Mullennix JW, Pisoni DB, Martin CS. Some effects Of talker variability on spoken word recognition. J. Acoust. Soc Amer. 1989; Vol. 85:365–378.

9. Clopper CG, Tamati TN, Pierrehumbert JB. Variation in the strength of lexical encoding across dialect. J. Phon: v.58, 2016;87(17):0095-4470

10. Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychol Rev. 1998 Apr;105(2):251-79.

11. Ji C, Galvin JJ, Chang Y, Xu A, Fu QJ. 2014. Perception of speech produced by native and non-native talkers by listeners with normal hearing and listeners with cochlear implants. J. Speech, Language, and Hearing Res. 57:532-554.

12. van Engen KJ, Chandrasekaran B, Smiljanic R. Effects of speech clarity on recognition memory for spoken sentences. PLoS One. 2012;7(9):e43753

13. McCoy SL, Tun PA, Cod LC, Colangelo M, Stewart RA, et al. (2005) Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. J. Exp. Psychol: 58A: 22–33.

14. Tamati TN, Pisoni DB. The perception of foreign-accented speech by cochlear implant users. ICPhS (2015).

15. Moberly AC, Lowenstein JH, Tarr E, Caldwell-Tarr A, Welling DB, Shahin AJ, Nittrouer S. Do adults with cochlear implants rely on different acoustic cues for phoneme perception than adults with normal hearing? J Speech Lang Hear Res. 2014 Apr 1;57(2):566-82.