2007

Evaluation of sentence list equivalency for the TIMIT sentences by cochlear implant recipients

Sarah E. King

EVALUATION OF SENTENCE LIST EQUIVALENCY FOR THE TIMIT SENTENCES BY COCHLEAR IMPLANT RECIPIENTS

by

Sarah E. King

A Capstone Project
submitted in partial fulfillment of the
requirements for the degree of:

Doctor of Audiology

Washington University School of Medicine
Program in Audiology and Communication Sciences

May 16, 2008

Approved by:
Jill B. Firszt, Ph.D., Capstone Project Advisor; Ruth M. Reeder, M.A., Second Reader; Laura Holden, M.A.; Margaret Skinner, Ph.D.

Abstract: The equivalency of 34 TIMIT sentence lists was evaluated using adult cochlear implant recipients to determine if they should be recommended for future clinical or research use. Because these sentences incorporate gender, dialect and speaking rate variations, they have the potential to better represent speech recognition abilities in real-world communication situations.

ACKNOWLEDGEMENTS:

I would like to thank the following people for their contributions and hard work in helping me complete this Capstone project.

Jill B. Firszt, Ph.D., Capstone Project Advisor

Ruth M. Reeder, M.A., Second Reader

Laura Holden, M.A.

Margaret Skinner, Ph.D.

This Capstone Research was supported by:

NIH/NIDCD K23DC 05410

Department of Otolaryngology at WUSM

TABLE OF CONTENTS

LISTS OF TABLES AND FIGURES

Introduction

Cochlear implants use electrical current to stimulate the auditory nerve of individuals with severe-to-profound hearing loss. With recent changes to candidacy guidelines, hearing-impaired persons with more residual hearing can take advantage of the benefits from cochlear implantation. Many research and clinical studies have shown that cochlear implant users can obtain high levels of speech recognition in the auditory-only condition (Skinner, Holden, Holden, Demorest, & Fourakis, 1997; Firszt et al., 2004). New processing strategies and advancements in implant technology have further improved performance for recipients. Additionally, bilateral cochlear implantation is becoming more prevalent and will provide distinct benefits compared to those of unilateral implantation. Therefore, speech recognition measures used to evaluate candidacy, technology advancements and bilateral effects must represent how the individual performs in real-world listening situations.

Current Speech Recognition Measures

The Hearing in Noise Test (HINT) is frequently used to assess open-set sentence recognition in both quiet and noise (Nilsson, Soli, & Sullivan, 1994). The HINT sentences were developed in the early 1990's by revising the Bamford-Kowal-Bench (BKB) sentences developed for British children (Bench & Bamford, 1979). They were modified for American English speakers, evaluated for naturalness and recorded by a single male talker. From this analysis, 25 equivalent lists of 10 sentences were developed and normalized using normal hearing listeners. The HINT was created as an adaptive procedure to measure speech recognition thresholds of sentences in quiet and in noise. This procedure avoids floor and ceiling effects by varying the signal-to-noise ratio to determine the point at which the listener can identify the material 50% of the time (Nilsson et al., 1994). Because the sentences were

originally evaluated with young, normal hearing subjects, one cannot assume that the lists will be equivalent in persons with hearing loss. In a study by Hanks and Johnson (1998), the HINT sentence lists were evaluated for equivalency when administered to older listeners with hearing impairment. Participants ranged from 60 to 70 years old and had pure tone averages (500, 1000 and 2000 Hz) of no more than 40 dB HL. Results revealed that individual list means fluctuated within 2 dB of the mean for all three noise conditions (speech in noise at 0 degree azimuth, 90 degree azimuth and 270 degree azimuth). In the original study by Nilsson et al. (1994), list means were within 1 dB of the overall mean; therefore, results in the study for older listeners are comparable to the results for young, normal hearing listeners. This study emphasized the importance of ensuring that speech recognition tests are normalized for the particular population under evaluation.

Other common speech recognition tests for cochlear implant evaluations include the City University of New York (CUNY) Topic Sentences (Boothroyd, Hnath-Chisolm, Hanin, & Kishon-Rabin, 1988) and the Consonant - Vowel Nucleus - Consonant (CNC) monosyllabic word test (Peterson & Lehiste, 1962). In a study by Skinner et al. (1997), the CUNY sentences were used to assess speech recognition in cochlear implant recipients at three stimulus levels. The sentences were recorded by Cochlear Corporation using one male speaker with a Midwestern American English dialect. Results revealed group mean scores at 70, 60 and 50 dB SPL to be 87%, 72% and 29% respectively. The CNC words are more difficult than sentence tests because they lack contextual cues. This monosyllabic word test was originally recorded with one male speaker from a Midwestern American English dialect region, and therefore lacks speaker variation. Because ceiling effects were noted with the original CNC words, Skinner et al. (2006) used new recordings of CNC words from the University of Melbourne to assess

cochlear implant performance. That study revealed that the new lists were more difficult than the original lists and results were not limited by ceiling effects. However, the new lists were still recorded by only one speaker.

To recognize speech from signals that contain multiple variations, normal hearing listeners use the process of perceptual normalization. This process involves extracting the meaning from speech with varying acoustic features related to gender, speaking rate and dialect (Kaiser, Kirk, Lachs, & Pisoni, 2003). Use of a single speaker eliminates the need for the listener to use perceptual normalization and therefore creates an unrealistic listening task that does not represent everyday communication (Loizou, Dorman, & Tu, 1999). Over the past decades, research has demonstrated some of the effects of age, gender, familiarity and dialect variations on speech recognition with normal hearing listeners (Lass, Hughes, Bowyer, Waters, & Borne, 1976; Mullennix, Pisoni, & Martin, 1988; Sommers & Barcroft, 2006). However, further research is needed with other clinical populations such as hearing impaired listeners and cochlear implant recipients.

Speaker Variation Research

In a series of experiments by Mullennix et al. (1988), the use of multiple speakers rather than a single speaker greatly affected the subject's performance on speech recognition tasks. Throughout the four experiments, subjects in mixed talker conditions performed significantly poorer than subjects in single talker conditions. Another finding was that greater processing time is required to recognize speech when the talker is different from trial to trial. Sommers (1997) assessed the effects of speaker variability on speech recognition with three subject groups including normal hearing young listeners, normal hearing elderly listeners and hearing-impaired elderly listeners. Across all subject groups, performance was poorer when multiple talkers were

used. In addition, there was a significant decrease in performance with multiple talkers in the normal hearing elderly group and a further decrement in performance in the hearing-impaired elderly group. When speech recognition tests use only one speaker, increased familiarization with a speaker occurs. The use of multiple, unfamiliar speakers in a speech recognition test better represents every day communication situations in which speaker variations are common.

Age and gender of the speaker affect speech recognition because they influence the fundamental frequency and formant frequency transitions. In a study by Lass et al. (1976), normal hearing listeners were able to identify the gender of the speaker with 96% accuracy for voiced vowel stimuli. Even when the stimuli were low pass filtered or whispered, accuracy was 91% and 75%, respectively. From these results, the authors concluded that for gender identification, the fundamental frequency is a more important part of the acoustic signal than the formants.

In 2004, Spahr and Dorman used test measures that included gender and speaking style variations to determine performance differences in subjects with the Advanced Bionics CII and the Nucleus 3G cochlear implant devices. Stimuli for this study included the AzBio sentences spoken by two male and two female speakers using a conversational speaking style rather than a clear speech style (as in the HINT sentences). The sentence intelligibility was evaluated with normal hearing subjects listening to simulations of five-channel cochlear implant processing. Results revealed that mean scores for the AzBio sentences in quiet and noise were always poorer than mean scores for the HINT and CUNY sentences in similar quiet and noise conditions. In addition, within-sex speaker discrimination was consistently more difficult than between-sex speaker discrimination. From these findings, it can be concluded that gender variation is needed to more completely assess performance in cochlear implant users.

Regional dialect variations can notably decrease speech recognition performance. In a study by Clopper and Pisoni (2004a), the TIMIT speech database was used to provide information about discrimination of regional dialects. The authors found that normal hearing subjects were only able to categorize the dialect region of an unfamiliar speaker with 31% accuracy. In another study by Clopper and Pisoni (2004b), sentences from the TIMIT speech database that represented six dialect regions were used as test stimuli. Results revealed that subjects exposed to a single speaker from one dialect region performed better in training and testing phases than subjects exposed to multiple talkers from the same dialect region; therefore, performance decreases with multiple speakers. However, subjects exposed to multiple speakers were able to generalize the dialect patterns more easily to a group of unfamiliar talkers in later trials. These results indicate that dialect is a critical part of the acoustic speech signal.

Finally, speaking style affects speech recognition. Sommers and Barcroft (2006) found that speech recognition was poorer when speech was presented by multiple speaking styles rather than a single speaking style. The HINT sentences were produced using a 'clear' speech style and the CUNY sentences were recorded using an 'exaggerated clear' speech style (Spahr & Dorman, 2004). For test measures to simulate natural listening conditions, speaking styles should not over exaggerate articulation patterns.

In a typical listening situation, it is likely that multiple speaker variations are present along with background noise. Hearing-impaired listeners, including cochlear implant recipients would have greater difficulty with these variations due to the degraded signal they receive compared to normal hearing listeners. Therefore, it is essential that speech recognition measures incorporate realistic speaker variables (Sommers, Kirk, & Pisoni, 1997). Currently used

measures in the clinic appear to overestimate how cochlear implant recipients feel they perform in everyday situations.

The TIMIT Sentences

In 1986, the TIMIT acoustic-phonetic speech database was developed as a joint effort between researchers at the Massachusetts Institute of Technology (MIT), the Speech Research Institute (SRI) and Texas Instruments (TI) to evaluate factors related to acoustic variability in speech. The database consists of three types of sentences that represent phonetic, contextual and speaker variations that are present in American English. The first type of sentences is the calibration sentences spoken by every talker for a total of 1,280 sentences. These sentences were used to represent phonemes that would be spoken with the greatest amount of dialect variation. The second type of sentences is the phonetically compact sentences spoken by several speakers for a total of 3,150 sentences. These sentences were created to represent the phonetic pairs in the English language. Finally, the third type of sentences is randomly selected sentences to represent alternative occurrences of phonemes. The entire database consists of 2,342 different sentences spoken by 630 talkers (10 sentences per speaker) for a total of 6,320 sentences. These speakers represent eight different American English regional dialects which include New England, New York City, North, North Midland, South Midland, South, West and Army Brat. In addition, gender variations are incorporated whereby 70% of the speakers are male and 30% are female (Lamel, Kassel, & Seneff, 1986). These speaker variations represent the unpredictability of speech in everyday communication situations and therefore have the potential to better assess speech recognition than currently used measures.

To date, the TIMIT sentences have been used in few research studies involving hearing-impaired populations. More commonly, the HINT sentences are used due to their ease of

administration and universal acceptance. However, research has shown a possible ceiling effect with the HINT sentences resulting from a lack of speaker variability. Shannon, Zeng, and Wygonski (1995) and Dorman, Loizou, and Rainey (1997) reported HINT sentence recognition scores in quiet to be 90% for normal hearing individuals listening through four channels of simulated cochlear implant processing. Loizou, Dorman, and Tu (1999) used 135 TIMIT sentences to assess speech recognition performance with four channels of simulation and normal hearing subjects. The sentences chosen were half spoken by males and half spoken by females from the North Midland American English dialect region. Their results revealed 63% recognition which suggests that a greater number of channels are needed to reach higher levels of performance for more difficult speech recognition measures. In addition, it is possible to speculate that performance may have been poorer if the TIMIT sentences used in this study incorporated dialect variations.

These studies show that the variability of the speaker greatly impacts speech recognition in normal hearing individuals listening through simulated cochlear implant processing strategies. These simulations are beneficial for research purposes because they eliminate much of the variability between cochlear implant recipients, such as processing strategy differences, electrode insertion depth, and the history of the hearing loss. However, simulations fail to represent current spread and the interaction of channels represented by electrical stimulation in the impaired cochlea, as well as patient demographic variables (Dorman, Loizou, Fitzke, & Tu, 1998). To best represent performance of cochlear implant recipients, it is essential to evaluate actual recipients.

In a study by Fu, Shannon, and Galvin III (2002), the effects of adaptation following changes in the frequency-to-electrode assignment were analyzed for three cochlear implant

subjects using four different test materials. The HINT sentences were used to represent a low to moderate difficulty measure and the TIMIT sentences were used to represent a moderate to extreme difficulty measure. The sentences used for both these measures were randomly chosen from pseudo-randomly chosen sentence lists. Results revealed that with a shifted frequency assignment, scores for the HINT sentences returned to near normal after adaptation while scores for the TIMIT stayed significantly lower than baseline.

Currently, candidacy guidelines for cochlear implantation require an assessment of open-set sentence recognition in the best aided condition. In 2005, the Center for Medicare and Medicaid Services expanded coverage for cochlear implantation to include hearing-impaired individuals that score 40% or less in their best aided condition on an open-set sentence recognition test (Department of Health & Human Services & Centers for Medicare and Medicaid Services, 2005). In addition, the FDA uses open-set sentence recognition tests for their guidelines of cochlear implant candidacy. For these evaluations, the HINT is the most frequently used open-set speech recognition test to determine candidacy as recommended by the Minimum Speech Test Battery for post-lingually deafened adult cochlear implant patients (Luxford & Ad Hoc Subcommittee, 2001). However, as previously discussed, this speech recognition test may provide an unrealistic measure of the patient's actual performance.

Because of changing candidacy guidelines, increased implementation of bilateral cochlear implants and improvements in speech coding strategies, speech recognition tests must be sensitive to test conditions. In addition, with these advancements, more cochlear implant users are reaching higher levels of speech recognition; therefore, assessment tools must incorporate speaker variability to best represent real world communication. The TIMIT sentences have this potential but list uniformity needs to be determined to guarantee that use of

different lists will produce comparable results. The present study was designed to evaluate the equivalence of intelligibility of 34 TIMIT sentence lists with adult cochlear implant recipients. We hypothesized that the TIMIT sentence lists administered to adult cochlear implant recipients users would be equivalent. Furthermore, we anticipated that the findings would determine whether the TIMIT sentences could be recommended for future clinical and research purposes, and if so, which lists were comparable.

Method

The research protocol and informed consent for this study were reviewed and approved by the Institutional Review Board and the Human Studies Committee at Washington University School of Medicine.

Subjects

22 adult cochlear implant (CI) recipients participated in this research study. The sample size was based on the sample analysis used by Skinner et al. (2006) in a study of CNC word list equivalency. Subjects in the current study were included based on the following criteria: willingness to participate, age greater than 18 years, device use greater than three months, and English as their primary language. In addition, subjects needed to score greater than 30% on their most recent CNC word test when presented at 60 dB SPL in quiet.

All CI subjects were recruited from the Washington University School of Medicine Adult Cochlear Implant Program. A pool of subjects was created by reviewing charts and identifying potential subjects based on the inclusion criteria. Then, potential subjects were sent a letter briefly explaining the study and the requirements of their participation. A copy of the informed consent was included in the mailing. Subjects who responded to these recruitment mailings were

scheduled for one 3-hour test session. At the beginning of the session, the examiner reviewed the informed consent with the subject and provided an opportunity for questions prior to it being signed by the subject and examiner.

The mean age of CI subjects was 58 years (SD = 13 years) with a range of 25 years to 78 years. Length of severe-to-profound hearing loss before implantation ranged from 0.4 years to 33.9 years with a mean of 10 years (SD = 8.3). The mean length of device use for the subjects was 3.9 years (SD = 2.5) with a range of 0.8 years to 10.6 years. In addition, the range of CNC scores for CI subjects was 44% to 92% with a mean of 73% (SD = 13%). These CNC scores represent a group of CI recipients that perform well above average. In a study by Firszt et al. (2004) that included a more representative sample, CI subjects' mean CNC scores when presented at 60 dB SPL were 39% (SD = 21%).

The causes of deafness for the 22 subjects were as follows: Genetics (n=6), Genetics – Autoimmune Inner Ear Disease (n=1), Otosclerosis (n=2), Noise (n=1), Ototoxicity (n=1), Unknown (n=9), Multiple Sclerosis (n=1), and Usher Syndrome II (n=1). Two of the subjects were pre-linguistically deafened (before the age of 2 years old) and the other 20 subjects were post-linguistically deafened (after the age of 4 years old). The mean age of onset of hearing loss was 20 years (SD = 19) with a range of 0 to 60 years. The above information is summarized in Appendix 1.

All CI subjects had previously worn hearing aids and only one subject did not consistently use amplification in the ear that was implanted before implantation. Currently, 11 subjects continue to use a hearing aid in their contralateral ear; however, for this study, the hearing aids were removed during testing. The mean age at implantation was 54 years old (SD = 14) with a range of 23 to 74 years old. One subject was re-implanted due to a device failure. For

this subject, the age of implantation and duration of use was based on the surgery of the first implant because device was not used for only one month. Information about the cochlear implant device used and processing strategies for each subject can be found in Table 1.

Table 1: Cochlear Implant Device Information

Subject	Ear	Internal Device	Processor	Strategy	Rate (pps/ch)
1	L	N24	ESPrit 3G	SPEAK (m=9)	250
2	R	N24	ESPrit 3G	SPEAK (m=10)	250
3	L	N24C	ESPrit 3G	ACE (m=8)	1800
4	L	N24C	ESPrit 3G	ACE (m=8)	900
5	R	ABCII	Auria	HiRes-S	2175
6	R	N24	ESPrit 3G	ACE (m=12)	900
7	L	N24	Sprint	ACE (m=8)	1800
8	R	N22	ESPrit 3G	SPEAK (m=8))	250
9	L	N24C	Sprint	ACE (m=8)	1800
10	L	N24C	ESPrit 3G	ACE (m=8)	1800
11	R	N24CA	Freedom	ACE(RE) (m=10)	1800
12	R	NF	Freedom	ACE (m=10)	1200
13	R	AB90K	Auria	HiRes-S	1406
14	R	N24C	ESPrit 3G	ACE (m=8)	900
15	L	N24C	Freedom	ACE (m=10)	1200
16	L	N24C	ESPrit 3G	ACE (m=8)	900
17	L	N24C	ESPrit 3G	ACE (m=8)	1800
18	L	N24C	Freedom	ACE (m=11)	1200
19	R	NF	Freedom	ACE (m=10)	1200
20	R	AB90K	Auria	HiRes-S	2855
21	L	N24CA	ESPrit 3G	ACE (m=12)	1200
22	R	ABCII	PSP	HiRes-S	1024

Note: N24 = Nucleus 24; N24C = Nucleus 24 Contour; ABCII = Advanced Bionics CII; N22 = Nucleus 22; N24CA = Nucleus 24 Contour Advance; NF = Nucleus Freedom; AB90K = Advanced Bionics HiRes 90K; m = maxima

Eight normal hearing (NH) subjects also participated in this research study. Inclusion criteria consisted of normal hearing, willingness to participate, age greater than 18 years and English as their primary language. The mean age of NH subjects was 23 years (SD = 2) with a range of 18 years to 25 years. Demographic information is summarized in Appendix A.

Procedure:

The testing began with the measurement of detection thresholds in the sound-field at 250 Hz, 500 Hz, 1000 Hz, 2000 Hz, 3000 Hz, 4000 Hz and 6000 Hz using warble tones. All testing for CI and NH subjects was completed using the standard Hughson-Westlake procedure and 2 dB increments in a double-walled soundproof booth. The subjects were seated at one meter from the loudspeaker and at a 0 degree azimuth. The CI subjects used their daily program, volume and sensitivity settings. These detection thresholds were obtained to ensure that the subject's processor settings allowed audibility of the speech frequencies (i.e. thresholds were below 34 dB HL). For NH subjects, sound-field thresholds were obtained bilaterally to ensure normal hearing.

After sound-field testing, the TIMIT sentence lists were administered. The 34 sentence lists used were those created and normalized for equal intelligibility by Dorman, Loizou, Spahr and Dana (2003) using normal hearing subjects listening to simulations of five-channel cochlear implant processing. Each list was presented at 60 dB SPL with the subject seated at one meter and a 0 degree azimuth from the loudspeaker (Firszt et al., 2004). The sentence lists were randomly presented to each subject except for List 1. List 1 was administered as the first list for practice to minimize learning effects between lists. In addition, List 1 was presented as the final list for each subject. For all data analysis, an average of scores from List 1 as practice and as the final list was used (See Appendix B for a sample of the TIMIT sentences).

A total of 700 TIMIT sentences were presented during the testing session with a mean length of time to administer one list of 4 minutes. The subjects were asked to repeat the sentence and were encouraged to guess if they were unsure. Frequent breaks were given to alleviate

fatigue and boredom. The lists were scored by total number of words correct for each list. The mean number of words per list was 128 words (SD = 6 words) with a range of 113 to 142 words.

Results

Sound-field Testing

Group mean warble tone detection thresholds for CI subjects and NH subjects are shown in Figure 1. The group mean threshold across all frequencies and across CI subjects was 23 dB HL (SD = 0.61) with a range of 21 to 25 dB HL. These thresholds indicate that all subjects were appropriately mapped to ensure that speech frequencies were audible. Warble tone thresholds for NH subjects ranged from 2 to 7 dB HL with a mean of 5 dB HL (SD = 0.41) indicating normal hearing sensitivity. Mean thresholds represent similar audibility between all subjects.

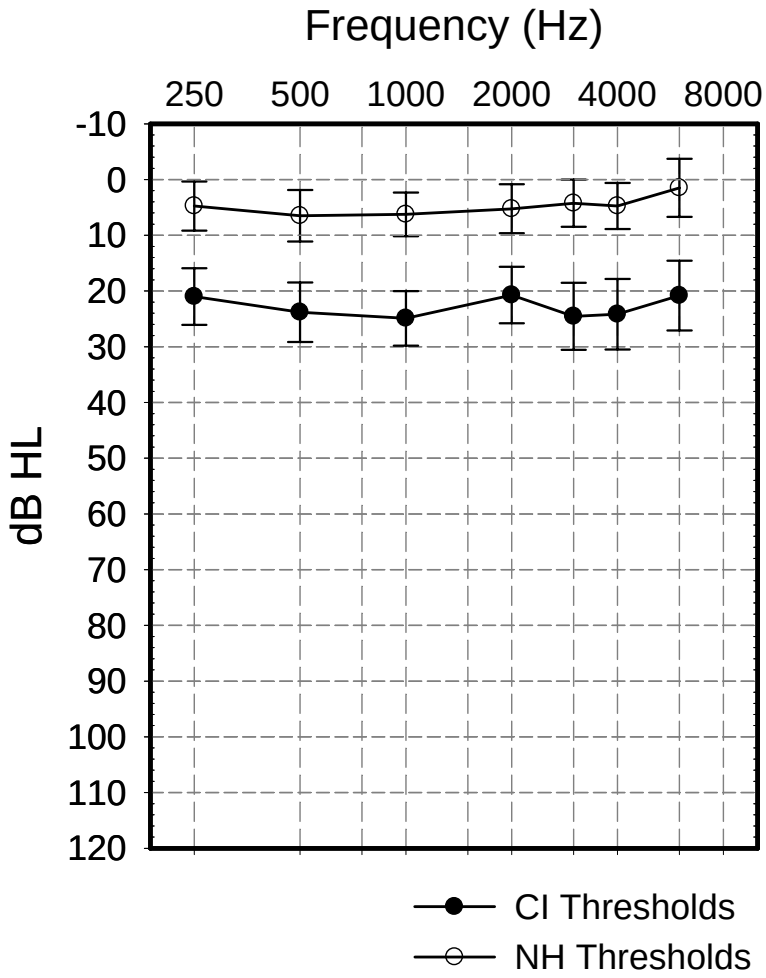

Figure 1: Group mean sound-field thresholds (dB HL) from 250Hz to 6000 Hz. The filled circles represent thresholds for the CI subjects. The open circles represent thresholds for the NH subjects. Error bars are +/- 1 standard deviation of the mean.

Mean TIMIT Scores Across Subjects By Sentence List

Group mean scores for the 34 TIMIT lists by rank and by list number order across CI and NH subjects are presented in Figures 2 and 3, respectively. For NH subjects, the mean score across all lists and subjects is 98% (SD = 0.01) with a range of scores from 96% to 100%. These results indicate excellent speech recognition abilities for the NH subjects. The mean score across lists for CI subjects is 73% (SD = 0.04) with a range of scores from 66% to 81%. Upon visual inspection, these results reveal that single TIMIT lists were not equivalent with each other due to

variable mean list scores. To verify that the lists were not equivalent, a Friedman nonparametric two-way ANOVA by ranks was performed. This analysis method was chosen because it is used to compare the distributions of two or more variables when the data is non-normally distributed. When performed, pairwise Friedman's tests ($p < 0.0001$) rejected the null hypotheses and indicated that the group mean scores for the 34 lists were not equivalent.

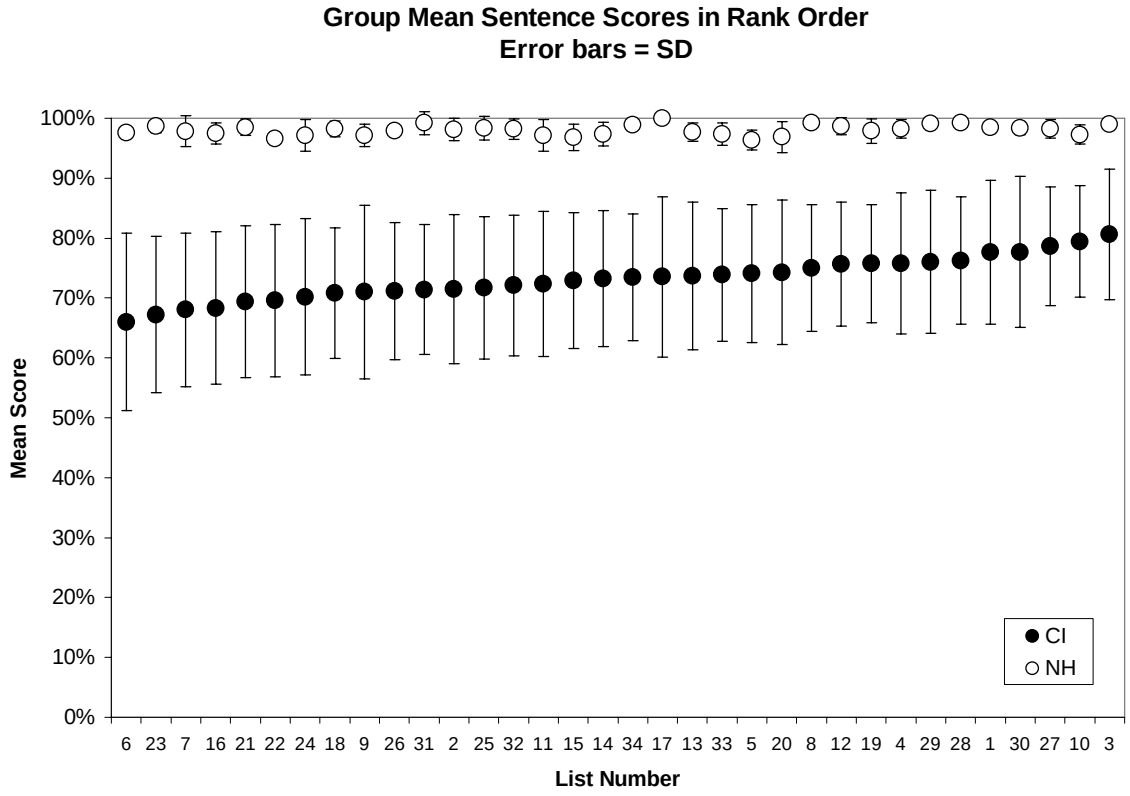

Figure 2: Group mean sentence scores across subjects in percent correct for each of the 34 TIMIT lists in rank order. The 22 CI subjects are shown with filled circles and the 8 NH subjects are shown with open circles. Error bars are +/- 1 standard deviation of the mean.

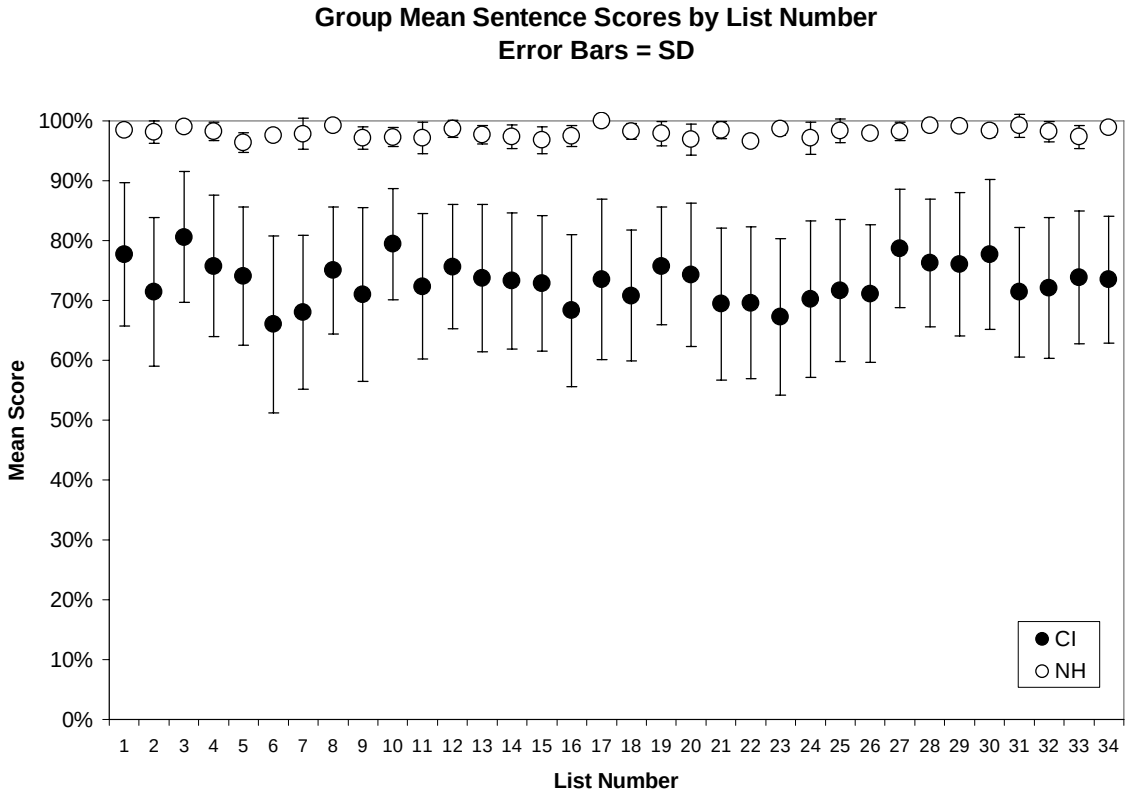

Figure 3: Group mean sentence scores across subjects in percent correct for each of the 34 TIMIT lists in list number order. The 22 CI subjects are shown with filled circles and the 8 NH subjects are shown with open circles. Error bars are +/- 1 standard deviation of the mean.

Individual Subject's Mean Scores

Mean scores across all 34 TIMIT lists for each subject are shown in Figure 4. Mean CI subjects' scores range from 54% to 89% with an overall mean of 73% (SD = 0.11). Mean NH subjects' scores range from 96% to 99% with an overall mean of 98% (SD = 0.01). To examine a correlation between mean score and range of scores, a scatter plot comparing the mean scores across all 34 lists compared to the subjects' range of scores was created (Figure 5). As this plot shows, the subjects with the highest mean scores (NH subjects) had the smallest range of scores. In addition, the plot shows that the largest range of scores is associated with the lowest mean average. A Pearson's correlation coefficient revealed a large negative correlation between the

range of scores and the mean score for all lists (Pearson's = -0.94). The range of scores for each subject was also compared with subject variables and demographic information but no significant correlations were noted.

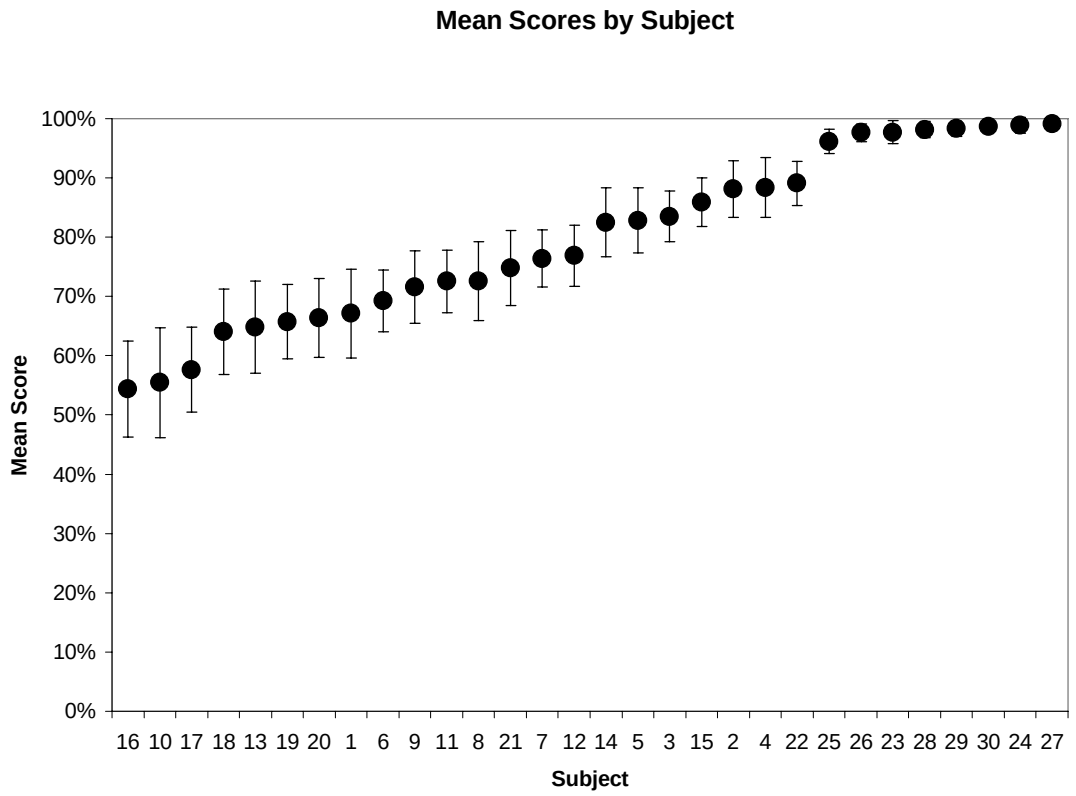

Figure 4: Individual subjects' mean scores across all 34 TIMIT sentence lists for CI subjects (filled circles) and NH subjects (open circles) in rank order. Error bars are +/- 1 standard deviation of the mean.

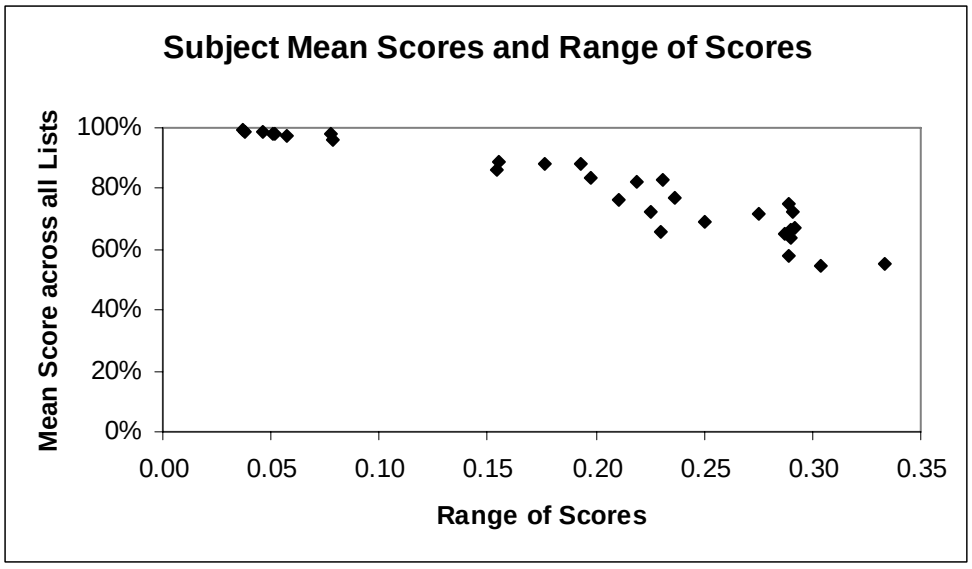

Figure 5: Scatter plot of CI and NH subjects' ranges of scores across lists compared to their mean score across all 34 TIMIT sentence lists.

Mean Scores of Paired Lists

Because the mean list scores are not equivalent between individual lists, it would be difficult to use single lists to assess and compare a CI recipient's speech recognition abilities over time or with other recipients. Paired lists of the TIMIT sentences were created by pairing the lists with the highest and lowest mean scores, the second highest and second lowest mean scores and so forth. With these list pairs, the overall mean is 73% (SD = 0.002) with scores ranging from 72.9% to 73.5% (Figure 6). These results reveal minimal variability between mean scores of list pairs. Furthermore, a Friedman nonparametric two-way ANOVA by ranks was performed and supported the null hypotheses. These results indicate that the group mean scores for the 17 paired lists were equivalent ($p = 0.99$).

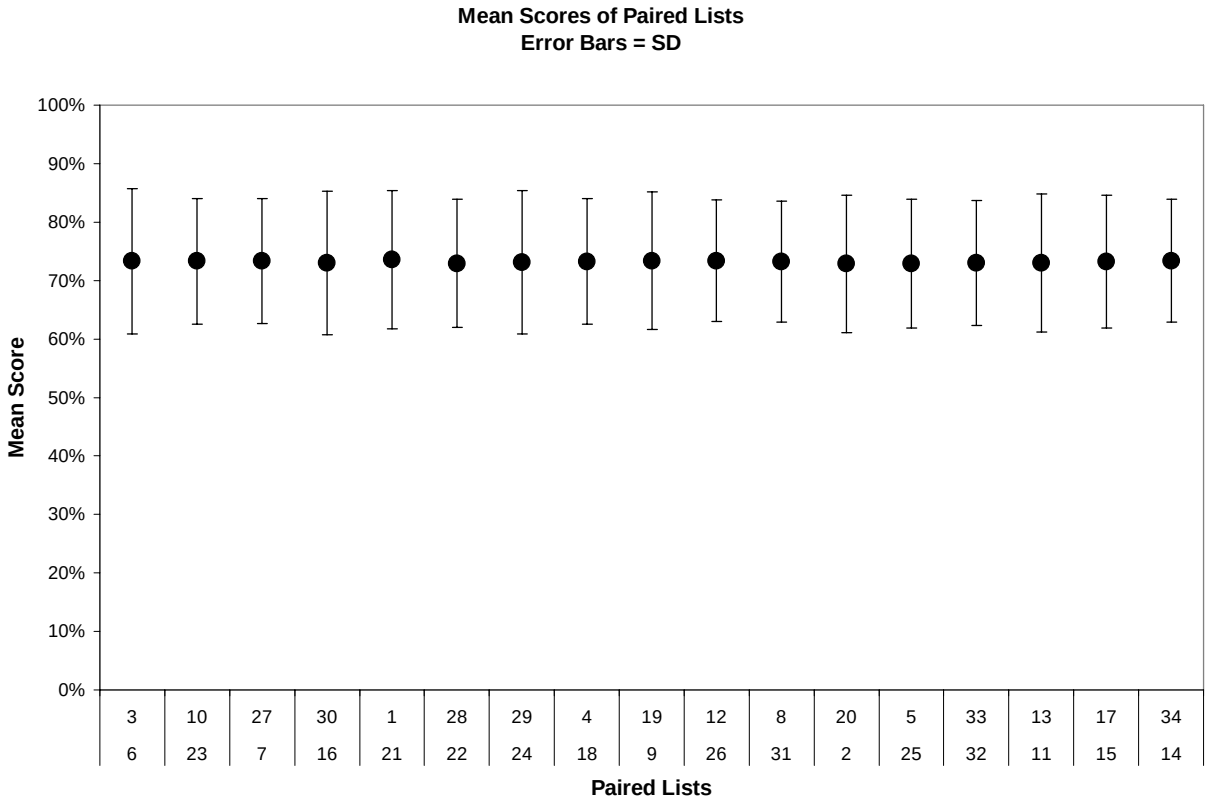

Figure 6: Group mean scores for CI subjects for paired TIMIT sentence lists. Error bars are +/- 1 standard deviation of the mean.

Discussion

Because the TIMIT sentences incorporate gender, dialect and speaker rate variations, they have the potential to represent speech recognition abilities in everyday communication situations. To date, research with the TIMIT sentences has focused on normal hearing listeners and normal hearing subjects listening through simulated cochlear implant processing. This study was carried out to determine if the 34 TIMIT sentence lists were equivalent in order to recommend them for clinical and research purposes. When the lists were evaluated with CI recipients, the results revealed that the 34 single TIMIT sentence lists were not equivalent with mean list scores across subjects ranging from 66% to 81%.

In an attempt to decrease the variability, the lists were paired. These 17 new list pairs had equivalent mean scores near 73%. The benefit of creating list pairs with similar means is that test-score variability and learning effects can be minimized while increasing test reliability. It is necessary to have uniformity of lists within the test measure to accurately assess performance within subjects and between subjects. The TIMIT list pairings (Appendix C) found to be equivalent in this study are suggested as a measure to evaluate speech recognition performance with more representative speaker variability. They are recommended for use in the clinic and for research to evaluate cochlear implant recipients.

The mean score across all lists and all CI subjects for this study was 73%. These results may appear to be high but they represent a group of subjects that perform above average. The aim of the study was not to assess the range of scores on the TIMIT sentence lists but rather to evaluate the equivalency between lists. When these lists were created by Dorman et al. (2003), the goal was to achieve a uniform mean score of around 70% to 75% for each list. Using normal hearing subjects listening through five channel simulated cochlear implant processing, mean sentence lists scores for individual lists were near 70%. The similarity between our results from the paired list means and those of Dorman et al. (2003) confirm the selection of sentences from the original database.

Data analysis indicated that mean scores across all 34 TIMIT lists for CI subjects were correlated to their most recent CNC score (Pearson's correlation coefficient = 0.70) indicating a strong relationship between performance on the two tests. Demographic factors such as length of use and onset of hearing loss did not correlate to TIMIT scores. In addition, the CI subjects' mean scores did not correlate with the NH subjects' mean scores (Pearson's correlation coefficient = 0.24).

Many cochlear implant recipients criticize currently used speech recognition tests in the clinic because they do not represent how they actually perform in the real world. The subjects in this study were enthusiastic about having a sentence test that better represents the difficulty they encounter outside the clinic. In addition, subjects reported that these sentence lists were more difficult than other test sentences. Some frequent comments from the subjects were that the speakers spoke fast, the dialect variations were difficult, and the unpredictability of the speaker between sentences made the task more challenging, which was more like their own everyday listening experience.

Conclusion

In conclusion, because the TIMIT sentences incorporate speaker variations, they better represent real world performance of cochlear implant recipients. It is recommended that the sentence lists be presented in list pairs based on this equivalency study. Currently, these TIMIT sentences pairings as described are being implemented in clinical research studies at Washington University School of Medicine. Because of changing candidacy guidelines and new technology, these sentences will be beneficial in evaluating performance changes. In addition, with the implementation of bilateral implantation, these sentences will be valuable in evaluating performance of subjects that encounter ceiling effects on currently used speech tests.

References

Bench, J., & Bamford, J. (1979). *Speech-hearing tests and the spoken language of hearing-impaired children*. New York: Academic Press.

Boothroyd, A., Hnath-Chisolm, T., Hanin, L., & Kishon-Rabin, L. (1988). Voice fundamental frequency as an auditory supplement to the speech-reading of sentences. *Ear and Hearing*, 9(6), 306-312.

Clopper, C.G., & Pisoni, D.B. (2004a). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, 32, 110-140.

Clopper, C.G., & Pisoni, D.B. (2004b). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, 47(3), 207-239.

Department of Health and Human Services (DHHS) & Centers for Medicare and Medicaid Services (CMS). CMS Manual System; Pub. 100-04 Medicare Claims Processing. *Cochlear Implantation*. No. 3796. July 1, 2005.

Dorman, M.F., Loizou, P.C., & Rainey D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America*, 102(4), 2403-2411.

Dorman, M.F., Loizou, P.C., Fitzke, J., & Tu, Z. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels. *Journal of the Acoustical Society of America*, 104(6), 3583-3585.

Dorman M.F., Loizou P.C., Spahr A., & Dana C.J. (2003). Simulations of combined acoustic/electric hearing. Engineering in Medicine and Biology, 25 th Annual International Conference of the IEEE, Cancun, Mexico September 17-21, 2003, pp. 1999-2001. Stoughton, Wisconsin: IEEE.

Firszt, J.B., Holden, L.K., Skinner, M.W., Tobey, E.A., Peterson, A., Gaggl, W., Runge-Samuelson, C.L., & Wackym, P.A. (2004). Recognition of speech presented at soft to loud levels by adult cochlear implant recipients of three cochlear implant systems. *Ear and Hearing*, 25(4), 375-387.

Fu, Q.J., Shannon, R.V., & Galvin III, J.J. (2002). Perceptual learning following changes in the frequency – to – electrode assignment with the Nucleus-22 cochlear implant. *Journal of the Acoustical Society of America*, 112(4), 1664-1674.

Hanks, W.D., & Johnson, G.D. (1998). HINT list equivalency using older listeners. *Journal of Speech, Language, and Hearing Research*, 41, 1335-1340.

Kaiser, A.R., Kirk, K.I., Lachs, L., & Pisoni, D.B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, 46, 390-404.

Lamel, L., Kassel, R., & Seneff, S. (1986). *Speech database development: design and analysis of the acoustic-phonetic corpus.* Proceedings of DARPA Speech Recognition Workshop, 100-109.

Lass, N.J., Hughes, K.R., Bowyer, M.D., Waters, L.T., & Bourne, V.T. (1976). Speaker sex identification from voiced, whispered and filtered isolated vowels. *Journal of the Acoustical Society of America*, 59(3), 675-678.

Loizou, P.C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America*, 106(4), 2097-2103.

Luxford, W.M., & the Ad Hoc Subcommittee of the Committee on Hearing and Equilibrium of the American Academy of Otolaryngology-Head and Neck Surgery. (2001). Minimum speech test battery for postlingually deafened adult cochlear implant patients. *Otolaryngology – Head and Neck Surgery*, 124, 125-126.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, 85(1), 365-378.

Nilsson, M., Soli, S.D., & Sullivan, J.A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, 95(2), 1085-1099.

Peterson, G.E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders*, 27(1), 62-70.

Shannon, R.V., Zeng, F., & Wygonski, J. (1995). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America*, 104(4), 2467-2476.

Skinner, M.W., Holden, L.K., Holden, T.A., Demorest, M.E., & Fourakis, M.S. (1997). Speech recognition at simulated soft, conversational, and raised-to-loud vocal efforts by adults with cochlear implants. *Journal of the Acoustical Society of America*, 101, 3766-3782.

Skinner, M.W., Holden, L.K., Fourakis, M.S., Hawks, J.W., Holden, T., Arcaroli, J., & Hyde, M. (2006). Evaluation of equivalency in two recordings of monosyllabic words. *Journal of the American Academy of Audiology*, 17, 350-366.

Sommers, M.S. (1997). Stimulus variability and spoken word recognition. II. The effects of age and hearing impairment. *Journal of the Acoustical Society of America*, 101(4), 2278-2288.

Sommers, M.S., Kirk, K.I., & Pisoni, D.B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I. The effects of response format. *Ear and Hearing*, 18, 89-99.

Sommers, M.S., & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *Journal of the Acoustical Society of America*, 119(4), 2406-2416.

Spahr, A.J., & Dorman, M.F. (2004). Performance of subjects fit with the Advanced Bionics CII and Nucleus 3G cochlear implant devices. *Arch Otolaryngol Head Neck Surg*, 130, 624-628.

Appendices

Appendix A: Subject Demographic Information

Demographic Information - CI subjects

Subject	Gender	Etiology	AAT	LOU (yrs)	LOD (yrs)	CNC
1	M	Usher II	51	5	34	44%
2	F	unknown	46	7	13	66%
3	F	Otosclerosis	58	4	10	78%
4	F	Unknown	46	3	13	86%
5	F	Genetic	52	4	5	90%
6	F	Genetic	57	6	13	71%
7	M	Unknown	72	8	4	78%
8	M	Genetic	55	11	6	84%
9	M	Otosclerosis	74	5	14	82%
10	F	Genetic (AIED)	48	4	6	48%
11	F	unknown	38	1	3	74%
12	F	Genetic	70	1	10	83%
13	F	unknown	72	2	15	81%
14	M	MS	49	3	4	86%
15	M	Genetic	71	2	4	92%
16	F	unknown	54	2	29	59%
17	M	Noise	65	3	4	58%
18	F	unknown	78	4	2	62%
19	F	unknown	75	3	0	58%
20	F	unknown	58	1	7	68%
21	F	Ototoxicity	25	2	12	79%
22	F	genetic	57	4	11	80%

Demographic Information - NH Subjects

Subject	Gender	AAT
23	M	18
24	F	24
25	F	23
26	F	25
27	F	24
28	F	25
29	M	20
30	F	24

Abbreviations used in Appendix 1: AAT: Age at test; LOD: Length of auditory deprivation; LOU: length of use; CNC: most recent score on Consonant Vowel-Nucleus Consonant test; AIED: Autoimmune Inner Ear Disease; MS: Multiple Sclerosis

Appendix B: Sample of TIMIT Sentences – List 1

1. Just long enough to make you feel important.
2. Your leg muscles and back muscles feel weary.
3. Kinda like a zombie?
4. You always come up with pathological examples.
5. The hallway opens into a huge chamber.
6. His voice seemed thick and purposeless.
7. Make it come off all right.
8. I know I didn't meet her early enough.
9. Cut off every building at the seventh floor.
10. But she suffered in her off-duty hours.
11. Destroy every file related to my audits.
12. Challenge each general's intelligence.
13. Toothpaste tube should be squeezed from the bottom.
14. But such cases were, in the past, unusual.
15. He murmured to himself, with firmness: no surrender.
16. What it does: aids in preventing foamy bloat.
17. Each is still glorified as a national hero.
18. Suppose he ran up the white flag altogether?
19. Bake slowly at least one-half hour longer.
20. These men were without capital or experience.

Appendix C: Recommended TIMIT List Pairs

Recommended List Pairings					
List Number	Mean Score	SD	Paired Lists	Mean Score	SD
6	0.66	0.12	6 & 3	0.73	0.12
23	0.67	0.12	23 & 10	0.73	0.11
7	0.68	0.11	7 & 27	0.73	0.11
16	0.68	0.12	16 & 30	0.73	0.12
21	0.69	0.12	21 & 1	0.74	0.12
22	0.70	0.15	22 & 28	0.73	0.11
24	0.70	0.13	24 & 29	0.73	0.12
18	0.71	0.11	18 & 4	0.73	0.11
9	0.71	0.15	9 & 19	0.73	0.12
26	0.71	0.09	26 & 12	0.73	0.10
31	0.71	0.12	31 & 8	0.73	0.10
2	0.71	0.10	2 & 20	0.73	0.12
25	0.72	0.12	25 & 5	0.73	0.11
32	0.72	0.11	32 & 33	0.73	0.11
11	0.72	0.11	11 & 13	0.73	0.12
15	0.73	0.13	15 & 17	0.73	0.11
14	0.73	0.13	14 & 34	0.73	0.11
34	0.73	0.11			
17	0.74	0.10	*Mean*	0.73	
13	0.74	0.12	*SD*	0.00	
33	0.74	0.13			
5	0.74	0.13			
20	0.74	0.13			
8	0.75	0.13			
12	0.76	0.12			
19	0.76	0.11			
4	0.76	0.10			
29	0.76	0.11			
28	0.76	0.12			
1	0.78	0.13			
30	0.78	0.11			
27	0.79	0.12			
10	0.79	0.11			
3	0.81	0.11			