# LEVERAGING AUTOMATIC SPEECH RECOGNITION IN COCHLEAR IMPLANTS FOR IMPROVED SPEECH INTELLIGIBILITY UNDER REVERBERATION

*Oldooz Hazrati*, Shabnam Ghaffarzadegan, John H.L. Hansen*

Center for Robust Speech Systems (CRSS),
The University of Texas at Dallas, Richardson, TX 75080-3021, USA
{hazrati, shabnam.ghaffarzadegan, john.hansen}@utdallas.edu

## ABSTRACT

Despite recent advancements in digital signal processing technology for cochlear implant (CI) devices, there still remains a significant gap between speech identification performance of CI users in reverberation compared to that in anechoic quiet conditions. Alternatively, automatic speech recognition (ASR) systems have seen significant improvements in recent years resulting in robust speech recognition in a variety of adverse environments, including reverberation. In this study, we exploit advancements seen in ASR technology for alternative formulated solutions to benefit CI users. Specifically, an ASR system is developed using multi-condition training on speech data with different reverberation characteristics (e.g., $T_{60}$ values), resulting in low word error rates (WER) in reverberant conditions. A speech synthesizer is then utilized to generate speech waveforms from the output of the ASR system, from which the synthesized speech is presented to CI listeners. The effectiveness of this hybrid recognition-synthesis CI strategy is evaluated under moderate to highly reverberant conditions (i.e., $T_{60}$ = 0.3, 0.6, 0.8, and 1.0s) using speech material extracted from the TIMIT corpus. Experimental results confirm the effectiveness of multi-condition training on performance of the ASR system in reverberation, which consequently results in substantial speech intelligibility gains for CI users in reverberant environments.

***Index Terms***— Automatic speech recognition, cochlear implants, multi-condition training, reverberation

## 1. INTRODUCTION

Although hearing is restored in profoundly deafened individuals by the aid of the cochlear implants (CI), they still encounter speech perception difficulties in challenging listening environments where a background masker is present (e.g. noise and/or reverberation) [1, 2]. Advancements in digital signal processing have resulted in improvements in speech understanding for CI users in the presence of noise and/or reverberation [1, 3–12], in the forms of modified speech coding strategies or front-end signal enhancement [5,10]. Despite the effectiveness of these techniques for improving the quality and/or intelligibility of speech in the presence of noise and/or reverberation, there still exists a large gap between performance of CI recipients in an anechoic quiet environment and in noisy and/or reverberant environments.

On the other hand, the performance of automatic speech recognition (ASR) systems drop substantially in mismatched train and test conditions where the system is only trained with the anechoic neutrally spoken clean speech, but tested on speech masked with noise and/or reverberation or spoken in a different mode [13–15]. Several techniques have been proposed to reduce word error rate (WER) of ASR systems in mismatched conditions where reverberation or noise exists [13, 14].

Taking into account both WER improvements in ASR systems under adverse environments, and also the ineffectiveness of the few proposed signal processing strategies in highly reverberant environments, here advancements in ASR systems are exploited in favor of CI listeners in moderate to highly reverberant conditions.

In this study, ASR systems are trained with anechoic clean speech, as well as different amounts of speech in the presence of reverberation. The output of the ASR engine tested in various reverberant conditions is then submitted to a text-to-speech (TTS) system, and the synthesized speech is then presented to the CI listeners. In order to evaluate the effectiveness of this recognition-synthesis strategy on speech understanding of CI users, four moderate to highly reverberant conditions ($T_{60}$ = 0.3, 0.6, 0.8, and 1.0s) are considered using the TIMIT speech corpus.

## 2. SYSTEM DESCRIPTION

The block diagram of the proposed hybrid recognition-synthesis strategy is shown in Fig.1. In the training stage, the speech recognizer is trained using anechoic, as well as a subset of the reverberant training speech. In the test stage, the ASR system output text transcription is submitted through a TTS synthesizer. Finally, the synthesized waveform is presented to the CI listener to evaluate speech intelligibility.

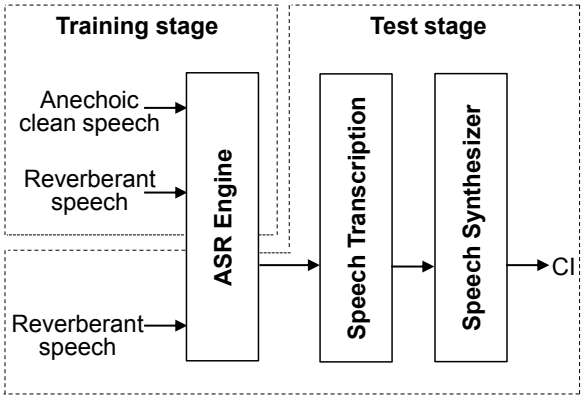


**Fig. 1**. Block diagram of the proposed hybrid strategy.

The state-of-the-art speech recognition toolkit, Kaldi [16], was used to train acoustic models and to decode test data. The trained model for each triphone is a three-state left-to-right HMM. A gender independent speech recognizer was trained on 3.5 hours of audio from the anechoic clean TIMIT database, and a portion of the reverberant version of TIMIT depending on the experiment, using Kaldi s5 recipe [16]. The forced alignments were generated using Linear Discriminate Analysis (LDA) [17] with Maximum Likelihood Linear Transform (MMLT) [18] to reduce the dimensionality. Finally, a global Feature space Maximum Likelihood Linear Regression (fMLLR) [19] was applied to normalize inter-speaker variability. 39-dimensional Mel Frequency Cepstral Coefficients (MFCC) [20] feature vectors were used, consisting of 13 statistic features along with their delta and acceleration coefficients. These features were extracted using 25ms speech frames with a 10ms frame shift between successive frames for 16kHz speech signals, using Kaldi toolkit. In all experiments, Cepstral Mean Normalization (CMN) was applied in an effort to minimize channel distortion. A trigram language model was used in the decoder. All experiments were carried out on an open speaker test set. An off-the-shelf TTS system was used to synthesize speech waveforms.

## 3. EXPERIMENTS

Performance of the proposed hybrid speech enhancement system is evaluated in the context of speech intelligibility for CI users, as well as WER of the ASR engine. For both ASR system assessment and CI intelligibility tests, speech data from TIMIT [21] was used. Training and test sets of the TIMIT database include 4158 train sentences (326 male and 136 female speakers) and 1512 test sentences (112 male and 56 female speakers), from different speakers with no overlap between training and test sentences.

For the speech synthesizer phase, a text-to-speech system was used to generate audio speech files at 16kHz sampling rate with two default speakers from a Windows-8 machine, "Microsoft David" as male speaker, and "Microsoft Zira" as female speaker (English-United States).

Four room impulse responses (RIR) with reverberation times equal to 0.3, 0.6, 0.8, and 1.0s were convolved with the anechoic clean train and test sentences of the TIMIT corpus in order to generate the reverberant data. The RIRs were recorded in a 10.06m $\times$ 6.65m $\times$ 3.4m (length $\times$ width $\times$ height) room [22] where reverberation time of the room was gradually varied from 1.0s to 0.8, 0.6, and 0.3s by floor carpeting and adding absorptive wall panels. The direct-to-reverberant ratios (DRR) of the RIRs are 1.5, -1.8, -3.0, and -0.5 corresponding to $T_{60}$ = 0.3, 0.6, 0.8, and 1.0s, respectively. The distance between the single-source signal and the microphone is 5.5m, which is beyond the critical distance.

The ASR system is trained using the training set from TIMIT in anechoic quiet, as well as different portions of the reverberant TIMIT train corpus. All TIMIT test material was used in the evaluation of the ASR system performance in anechoic quiet and four reverberant ($T_{60}$ = 0.3, 0.6, 0.8, 1.0s) conditions. For intelligibility listening tests, two types of scenarios were considered: (a) naturally spoken, and (b) synthesized speech.

Four adult post-lingually deafened native speakers of American English CI listeners were tested in anechoic quiet and the four above mentioned reverberant conditions with both naturally spoken and synthesized sentences. For the set of naturally spoken sentences, in each condition twenty sentences (10 spoken by a male and 10 by a female speaker) were presented to the CI users. For the synthesized speech set of tests, based on the WER of the ASR system at that specific condition, twenty synthesized sentences (again 10 spoken by a male and 10 by a female speaker) with the same WER were selected and presented to the CI listeners. All listening tests were conducted in a double-wall anechoic sound attenuating booth through a loud-speaker located in front of the CI listener.

During the speech intelligibility tests, sentences were each presented to the CI listener twice and the listener was asked to repeated the words he/she could hear. The number of correct words identified by the listener in each condition was then divided by the total number of test words used in that condition to compute the speech intelligibility score. The order of the conditions and sentences presented to the CI users was randomized across subjects.

## 4. RESULTS

The ASR results obtained in different multi-condition training scenarios are shown as WERs in Table 1. In order to evaluate the effect of various reverberant conditions used in multi-condition training, in addition to the anechoic quiet train data, different portions of reverberant training data (16%, 33%, and 100%) were also used in the speech recognizer training stage. In the ASR performance evaluation phase, the WER for ane-

**Table 1**. Effect of multi-condition training on the WER of the ASR system. $R_{0.3}$, $R_{0.6}$, $R_{0.8}$, and $R_{1.0}$ stand for reverberation time of $T_{60}$ = 0.3, 0.6, 0.8, 1.0s used in training stage (in addition to the anechoic quiet data) and test stage. 16%, 33%, 100% represent the portion of the reverberant training data used in each condition.

| | **Word Error Rate (WER) in %** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Train** / **Test** | **Clean** | **clean+(x%)$\mathbf{R_{0.3}}$** | | | **clean+(x%)$\mathbf{R_{0.6}}$** | | | **clean+(x%)$\mathbf{R_{0.8}}$** | | | **clean+(x%)$\mathbf{R_{1.0}}$** | | |
| | | **16%** | **33%** | **100%** | **16%** | **33%** | **100%** | **16%** | **33%** | **100%** | **16%** | **33%** | **100%** |
| **Clean** | 0.9 | 0.8 | 0.7 | 0.8 | 0.9 | 0.7 | 0.7 | 0.6 | 0.8 | 0.8 | 0.6 | 0.7 | 0.9 |
| $\mathbf{R_{0.3}}$ | 4.2 | 0.8 | 0.6 | 0.6 | 0.8 | 0.6 | 0.6 | 1.1 | 0.9 | 0.9 | 1.2 | 0.8 | 0.9 |
| $\mathbf{R_{0.6}}$ | 32.5 | 3.8 | 2.2 | 1.2 | 2.6 | 1.6 | 0.9 | 2.7 | 2 | 0.8 | 1.1 | 2.4 | 0.8 |
| $\mathbf{R_{0.8}}$ | 55.2 | 12.0 | 6.7 | 3.2 | 6.4 | 3.8 | 1.5 | 6.2 | 3.2 | 1.6 | 9.8 | 3.9 | 1.9 |
| $\mathbf{R_{1.0}}$ | 64.8 | 23.1 | 15.9 | 9.6 | 16.8 | 7.3 | 3.9 | 14.2 | 6.6 | 3.4 | 10.0 | 4.7 | 3.0 |

choic, as well as all four reverberant conditions ($T_{60}$ = 0.3, 0.6, 0.8, and 1.0s) were computed. The second column (entitled "clean") in Table 1 shows ASR system performance when no reverberation data has been used in the training phase (i.e., WERs range from 0.9% to 64.8%).

As seen from the table, adding more reverberant data in the training phase, reduces WER in all reverberant conditions, especially when reverberation increases (larger $T_{60}$ values). For example, if we compare "clean" train WERs vs. "clean+$R_{0.3}$ (100%)" train WERs, we see that increasing the amount of reverberant training data even in the lowest reverberation time tested here ($T_{60}$ = 0.3s) decreases WERs from 4.2% to 0.6%, 32.5% to 1.2%, 55.2% to 3.2%, and 64.8% to 9.6% in reverberant conditions with $T_{60}$ = 0.3, 0.6, 0.8, and 1.0s, respectively. Two clear trends are observed from the data in Table 1: (1) as the amount of reverberant data increases in the training phase, the WER reduces in all conditions tested, and (2) adding the same amount of reverberant data with larger $T_{60}$ in the training stage results in greater WER reduction in all conditions (e.g., WERs in $T_{60}$ = 1.0s reduced from 23.1% to 16.8%, 14.2% and 10% when only 16% of the reverberant data in $T_{60}$ = 0.3, 0.6, 0.8, and 1.0s was added in the training phase, respectively).

In order to evaluate the effect of the proposed hybrid recognition-synthesis strategy on the intelligibility of the speech, a model trained on $R_{0.3}$ with 16% reverberant data is tested against clean and four reverberant conditions, and the synthesized speech output is presented to the CI listeners. Subjective intelligibility scores from four CI listeners are presented in Fig.2. The CI users were also tested with naturally spoken sentences in all 5 conditions (clean and reverberant with $T_{60}$ = 0.3, 0.6, 0.8, and 1.0s) for comparative purposes. These results are presented in Fig. 2(b).

Evident from the results shown in Fig. 2(a), is that CI users speech intelligibility scores (tested with the output of the speech recognizer) were very close to the baseline ASR performance in that condition. It is worth mentioning that WERs in each condition were converted to accuracy scores by excluding the effects of substitutions and insertions, and only counting the number of words correctly recognized by the ASR system. This is the reason why ASR accuracy scores reported in Fig. 2(a) do not exactly match the WERs in the third column of Table 1.

As seen in Fig. 2(b), speech intelligibility for CI listeners drop significantly as reverberation time ($T_{60}$) increases. The average subjective speech intelligibility scores dropped from an average of 62.04% to 23.63%, 14.37%, 11.54% and 4.38% in anechoic clean, and $T_{60}$ = 0.3, 0.6, 0.8, and 1.0s conditions, respectively. These results although in line, are approximately 15% less than the results of a previous study with CI users [1]. This is due to the perplexity of the TIMIT database used here compare to the IEEE [23] database used in [1], and also the regional dialect of the speakers of TIMIT corpus (the speaker of IEEE sentences has no specific dialects).

The average subjective intelligibility score in anechoic quiet condition obtained from synthesized speech, Fig. 2(a), is approximately 37% higher than that achieved from naturally spoken sentences in the same anechoic clean condition. The following factors contributed to such differences in scores at the same condition: 1) The synthesized speech does not carry any specific regional dialect or accent. In contrast, the naturally spoken sentences of the TIMIT database are collected from speakers with various regional dialects which affect speech perception of CI listeners; 2) A low speaking rate was used in the TTS in order to generate clear speech. However, most speakers in the TIMIT corpus have a faster speaking rate. Even taking these reasons into consideration, the synthesized sentences from the proposed hybrid recognition-synthesis systems are still significantly more intelligible than the naturally spoken sentences in all conditions. Two valuable outcomes can be extracted from the results in Fig. 2. First, using only the multi-condition training, the performance of ASR systems can increase substantially under reverberation. The reverberant data used in the training phase of the speech recognizer acts as a model adaptation tool which considers the effects of reverberation on phonemes in the model training and therefore, results in lower WERs even under higher reverberation. Second, removing the speaker

related characteristics such as accent and dialect from the spoken speech results in significantly greater intelligible speech for hearing impaired listeners with CI devices.

Moreover, feedback from all four CI listeners tested with synthesized speech indicated that none of them were able to determine whether the speech was naturally produced or artificially synthesized. This is promising in the sense that large intelligibility benefits can be obtained with the aid of this hybrid recognition-synthesis strategy in challenging listening conditions, where the context of spoken sentences is of a much greater interest compared to the specific speech characteristics of the speaker.

In both ASR and subjective experiments, using only a small portion of the moderate reverberant data ($T_{60}$ = 0.3s) leads to substantial improvements in ASR accuracy and speech intelligibility, even in highly reverberant environments. This can be easily used in favor of CI listeners as many daily environments (e.g., office, classrooms, living room) have a reverberation time of 0.2 to 0.3s. Therefore, each CI user can easily access such reverberant data in order to train the ASR system linked to her/his own device. Moreover, today, due to large advancements in digital signal processing, PDA based cochlear implant interfaces [24], as well as the extensive use of smart phones, one can easily use a speech recognizer in the implant or connect her/his implant to the smart phone to leverage computing resources for speech recognition.

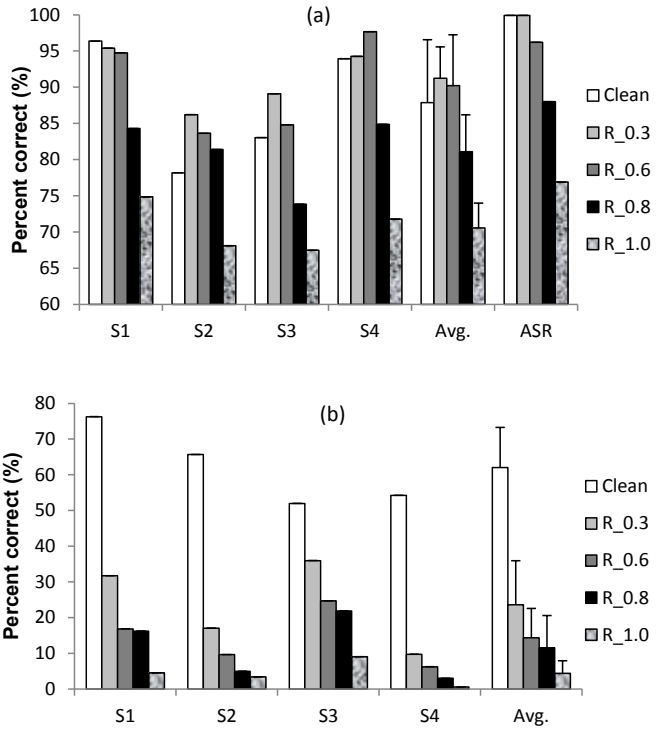


**Fig. 2**. Mean speech intelligibility scores of four CI users in anechoic clean and reverberant ($T_{60}$ = 0.3, 0.6, 0.8, and 1.0s) conditions. Panel (a) represents the intelligibility scores of CI listeners tested with synthesized speech from the recognition-synthesis strategy (the WER results were also presented in terms of ASR accuracy for comparative purposes). Panel (b) demonstrates the intelligibility scores of CI listeners tested with naturally spoken sentences in all conditions. Error bars indicate standard deviations.

## 5. CONCLUSION

This study has proposed a hybrid recognition-synthesis cochlear implant (CI) strategy for reverberant speech intelligibility enhancement. Multi-condition training has been shown to reduce WER of the speech recognizer under moderate to highly reverberant ($T_{60}$ = 0.3, 0.6, 0.8, and 1.0s) conditions. Synthesized speech is then generated from the output of the speech recognizer engine using a text-to-speech (TTS) system. The effect of multi-condition training on speech intelligibility for hearing impaired listeners is evaluated by presenting the output of the recognition-synthesis system to CI listeners using data from TIMIT. The proposed hybrid strategy resulted in significant speech intelligibility gains under mismatched reverberant conditions for CI users. Using only clean data with a small subset of moderate reverberant data (16% of the training corpus) for ASR training, CI users were able to identify re-synthesized speech with over 70% accuracy in most reverberant conditions tested ($T_{60}$ = 1.0s). This high speech intelligibility result was due to the low WER of the recognizer trained with reverberation characteristics, as well as excluding accent/dialect of the spoken speech by the use of speech synthesizer. Speech identification under reverberation is a very challenging task for CI users. Therefore, due to low speech identification performance of CI listeners under reverberation, development of signal processing strategies that improve intelligibility of reverberant speech is of great importance. Incorporating a speech recognizer trained with only about an hour of moderate reverberant ($T_{60}$ = 0.2-0.3s) data, which can be easily accessed in regular rooms, and a simple speech synthesizer will provide a substantial speech intelligibility gain for CI users in moderate to highly reverberant conditions.

## 6. ACKNOWLEDGMENT

This work was supported by a contract from Cochlear Limited to the University of Texas at Dallas.

## 7. REFERENCES


[1] K. Kokkinakis, O. Hazrati, and P.C. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 129, pp. 3221–3232, May 2011.

[2] O. Hazrati and P.C. Loizou, "The combined effects of reverberation and noise on speech intelligibility by cochlear implant listeners," *Int. J. Audiol.*, vol. 51, pp. 437–443, June 2012.

[3] Y. Hu, P.C. Loizou, N. Li, and K. Kasturi, "Use of a sigmoidal-shaped function for noise attenuation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 122, pp. EL128–EL134, September 2007.

[4] Y. Hu and P.C. Loizou, "A new sound coding strategy for suppressing noise in cochlear implants," *J. Acoust. Soc. Am.*, vol. 124, pp. 498–509, July 2008.

[5] O. Hazrati and P.C. Loizou, "Reverberation suppression in cochlear implants using a blind channel-selection strategy," *J. Acoust. Soc. Am.*, vol. 133, pp. 4188–4196, June 2013.

[6] O. Hazrati, S.O. Sadjadi, P.C. Loizou, and J.H.L. Hansen, "Simultaneous suppression of noise and reverberation in cochlear implants using a ratio masking strategy," *J. Acoust. Soc. Am.*, vol. 134, November 2013.

[7] O. Hazrati, J. Lee, and P.C. Loizou, "Blind binary masking for reverberation suppression in cochlear implants," *J. Acoust. Soc. Am.*, vol. 133, pp. 1607–1614, March 2013.

[8] P.S. Jafari, H.Y. Kang, X. Wang, Q.J. Fu, and H. Jiang, "Phase-sensitive speech enhancement for cochlear implant processing," in *Proc. IEEE ICASSP*. May, 2011, pp. 5104–5107.

[9] P.W. Dawson, S.J. Mauger, and A.A. Hersbach, "Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus cochlear implant recipients," *Ear & Hear.*, vol. 32, pp. 382–390, May/June 2011.

[10] A.A. Hersbach, S. Mauger, D. Grayden, J. Fallon, and H. McDermott, "Algorithms to improve listening in noise for cochlear implant users," in *Proc. IEEE ICASSP*. May, 2013, pp. 428–432.

[11] K. Nie, G. Stickney, and F-G. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," *IEEE Trans. Biomed. Eng.*, vol. 52, pp. 64–73, January 2005.

[12] A. Bhattacharya, A. Vandali, and F-G. Zeng, "Combined spectral and temporal enhancement to improve cochlear-implant speech perception," *J. Acoust. Soc. Am.*, vol. 130, pp. 2951–2960, November 2011.

[13] S.O. Sadjadi, H. Boril, and J.H.L. Hansen, "A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort," in *Proc. IEEE ICASSP*. March, 2012, pp. 4701 – 4704.

[14] P.J. Moreno, Ed., *Speech Recognition in Noisy Environments; Ph.D. thesis*, ECE. CMU, PA, USA, 1992.

[15] S. Ghaffarzadegan, H. Boril, and J.H.L. Hansen, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," in *Proc. IEEE ICASSP*. May, 2014, pp. 2544 – 2548.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*. December, 2011.

[17] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. IEEE ICASSP*. March, 1992, pp. 13–16.

[18] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, pp. 75–98, 1998.

[19] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. IEEE ICASSP*. May, 1998, pp. 661–664.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, 1980.

[21] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Comm.*, vol. 9, pp. 351–356, September 1990.

[22] A.C. Neuman, M. Wroblewski, J. Hajicek, and A. Rubinstein, "Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults," *Ear Hear.*, vol. 31, pp. 336–344, June 2010.

[23] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 225–246, 1969.

[24] H. Ali, A. Lobo, and J.H.L. Hansen, "Design and evaluation of a PDA-based research platform for cochlear implants," *IEEE Trans. Biomed. Eng.*, vol. 60, pp. 3060–3073, 2013.