ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747404

INTERNET STREAMING AUDIO BASED SPEECH RECEPTION THRESHOLD MEASUREMENT IN COCHLEAR IMPLANT USERS

Xi Chen[1,2], *Yefei Mo*[3], *Kang Ouyang*[1], *Mingyue Shi*[1], *Huali Zhou*[1],
Yupeng Shi[2], *Wei Xiao*[2], *Shidong Shang*[2], *Qinglin Meng*[3*], *Nengheng Zheng*[1†]

[1]Guangdong Key Lab. of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China
[2]Tencent Ethereal Audio Lab, Shenzhen, China
[3]South China University of Technology, Guangzhou, China

ABSTRACT

Traditional face-to-face subjective listening test has become a challenge due to the COVID-19 pandemic. We developed a remote assessment system with Tencent Meeting, a video conferencing application, to address this issue. This paper presents our work on evaluating the reliability of the remote assessment system. Two speech reception threshold (SRT) experiments were conducted to study the effects of noise suppression and maxima selection number on cochlear implant (CI) hearing. Both experiments were conducted locally and remotely, the correlations between the respective results were analyzed. Results showed that remote tests replicated the differences among testing conditions observed in local tests, but the absolute SRT values for individual conditions varied significantly between the two modes. The variations could be attributed to multiple reasons, such as online data transmission issues, audio playback devices, environmental conditions, and the training of participants. In conclusion, the relative variation of SRTs for CIs can be measured reliably, but the absolute SRT values should be carefully compared and explained according to objective and subjective experimental conditions.

Index Terms— Remote subjective assessment, cochlear implant, speech reception threshold

1. INTRODUCTION

When the COVID-19 pandemic curtailed in-person subjective listening assessments, researchers sought for video conferencing applications (apps) like Zoom to conduct subjective tests [1]. This has been facilitated and accelerated by the rapid development of internet technology. For example, the capacity and stability of internet have been increased significantly: the average transmission rate has increased from 32.2 to 59.1 Mbps since 2016, and by 2020, 89.9 percent of broadband subscribers accessed web service at speed higher than 100 Mbps [2]. Real-time communication apps like Teams, Tencent Meeting, and Zoom boost the services of remote communications. However, the quality of experiences (QoE) is confronting several challenges, most notably packet loss due to transmission errors and compression effects by speech codecs [3], which could degrade the intelligibility of online speech for hearing-impaired listeners. Therefore, it is worthwhile to investigate and improve the accuracy of the remote assessment methodology to mimic the local experiment.

Various web-based platforms are available for state-of-the-art speech intelligibility tests [4–8], in particular, [8] summarizes such platforms for hearing impaired participants. Recent studies have proposed frameworks for remote assessment of the speech intelligibility in normal-hearing (NH) listeners [9] and cochlear implant (CI) users [10], and demonstrated that remote test is feasible to a certain extent. Method in [9] requires subjects installing MATLAB standalone installer and uploading the test results to the cloud, which was time-consuming. In [10], test sounds were served as direct audio inputs (DAI) to the recipients, such bypassed the microphone on participants' devices. However, DAI is an exclusive feature on specific hearing devices and subjects cannot communicate with experimenters in DAI mode, which restricts its practical applications. Therefore, it is necessary to explore and evaluate a convenient and effective remote assessment method.

This paper presents our recent work on evaluating the reliability of remote speech reception threshold (SRT, an important index for speech intelligibility) assessments with CI users. The evaluation was conducted via Tencent Meeting, one of the most widely used video conference platforms in China. Two subjective experiments were conducted, focusing on analyzing the correlations between the local and remote assessments to evaluate the reliability of remote assessments. In Experiment I, SRTs were measured with seven CI users under two noise-masking conditions (with and without noise reduction, NR). In Experiment II, ten NH listeners participated in a CI-simulated listening test. SRTs in babble noise with dif-

*Corresponding email: mengqinglin@scut.edu.cn.
†Corresponding email: nhzheng@szu.edu.cn.

ferent channel-stimulating parameters (i.e., maxima selection numbers) were measured. All participants went through both remote and local assessments.

2. EXPERIMENT I: NOISE MASKING

2.1. Participants

Seven unilaterally implanted CI users (aged 22 to 47) were recruited (see Table 1) as participants (i.e., experiment subjects). The experimenters in Figure 1 are graduate students in audiology who conducted the test. All were native Mandarin speakers. The study protocol was fully approved by the local institutional review board. Written informed consent was obtained before testing and participants were financially compensated for their participation.

Table 1. Clinical data of cochlear implant (CI) users.

Subject	Age	Gender	CI experience (years)	Processor Type	Etiology
C1	23	F	18	Nucleus 6	Drug
C2	23	F	3	Nucleus 6	Congenital
C3	47	F	13	OPUS 2	Drug
C4	22	F	3	Nucleus 6	Congenital
C5	28	M	17	Kanso 1	Drug
C6	24	M	21	Nucleus 7	Drug
C7	24	F	19	Freedom	Drug

2.2. Materials

SRTs under two noise-masking conditions were measured, one without NR (denoted by 'Noisy') and the other with a deep neural network (DNN)-based NR (denoted by 'DNN'). SRTs in noise are the signal-to-noise ratio (SNR) at which the listeners could recognize 50% words in a sentence. The SRT was measured using an adaptive staircase psychophysical procedure with SNR as an adaptive feature. In this study, the NR method in [11] was implemented. That is, an ideal ratio masking (IRM) gain was estimated by a DNN and used for noise suppression. The DNN was trained using the THCHS-30 database [12]. Two types of masking noise were used, i.e., a babble noise (Babble) from NOISEX-92 [13] and a speech-shaped noise (SSN). The speech-shaped noise was computed by filtering white noise through an FIR filter with a frequency response matching the long-term spectrum of the sentences in the HINT database [14]. Sentences from closed version of the Mandarin Chinese matrix (CMNmatrix) corpus [15], which includes 40 lists with 20 sentences in each were used as the target speech.

2.3. Procedures

The assessments were conducted in three scenes, a face-to-face assessment (*Local*) and two remote assessments (*Remote* 1 and *Remote* 2). Figure 1 shows the schematic diagram for

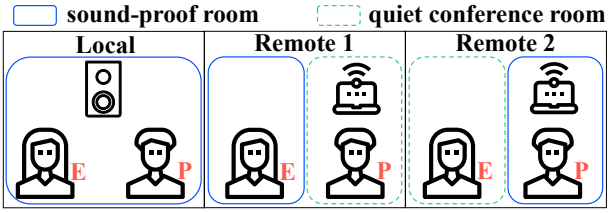

Fig. 1. Scenes schematic set for local and remote scenes (E = experimenter; P = participant).

the three scenes. In Scene *Local*, experimenters and participants both stayed in a sound-proof chamber (noise floor $\leq$ 30 dBA) and communicated face-to-face during the experiments. In Scene *Remote* 1, experimenters were in the sound-proof room and participants were in a quiet conference room (noise floor $\leq$ 40 dBA). The experimenters shared the screen and audio with participants via Tencent Meeting. Scene *Remote* 2 was the same as *Remote* 1 except that experimenters and participants switched their working rooms. *Remote* 1 represented a practical scenario of remote assessments. *Remote* 2 was adopted to evaluate the performance degraded by introducing network transmission error only, while other factors remained unchanged from *Local*.

Stimuli were delivered with a high-quality monitor speaker (Genelec 8030A) connected to a personal computer via a sound card (YULONG Aquila II) for the sound-proof room (*Local* and *Remote* 1), and with the built-in laptop (ThinkPad X1 Carbon Gen 4) loudspeakers via Tencent Meeting for the quiet conference room (*Remote* 2). Sound volume was adjusted to a comfortable level by participants. The internal noise reduction of Tencent Meeting was switched on during the test to remove ambient noise.

Each participant went through three scenes which repeated twice, i.e., six blocks. Test order of the six blocks was randomized across participants. In each block, the order of the four conditions, i.e., two noisy types (SSN vs. Babble) by two noise-masking conditions ('Noisy' vs. 'DNN') were also randomized. Each condition was tested using different CMNmatrix lists and results with each condition were averaged over the two repeated blocks as the final SRT.

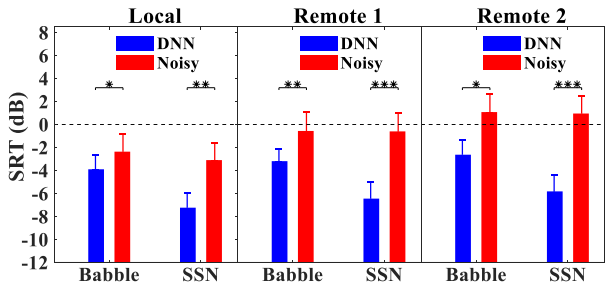

Fig. 2. Mean SRTs for different test conditions. Error bars indicate the standard deviations. Asterisks above indicate the NR effect statistical significance (*p <0.05, **p <0.01, ***p <0.001)

Table 2. ANOVA results of Experiment I.

Source	F value	p value
S	$F(2,12) = 19.849$	$< 0.001^{***}$
NT	$F(1,6) = 32.558$	0.001^{**}
NR	$F(1,6) = 48.669$	$< 0.001^{***}$
S × NT	$F(2,12) = 0.767$	0.486
S × NR	$F(2,12) = 9.258$	0.004^{**}
NT × NR	$F(1,6) = 19.493$	0.004^{**}
S × NT × NR	$F(2,12) = 0.483$	0.628

S represents scene; NT represents noise type; NR represents noise reduction.

2.4. Results and discussions

Figure 2 shows the SRTs for different noise types (Babble vs. SSN) with two NR condition ('Noisy' vs. 'DNN') in three scenes. The asterisks above indicate the significance of NR effect between 'Noisy' and 'DNN' in each condition. As shown, applying NR effectively improves the speech intelligibility (lower SRTs) for both noise types in local and remote assessments. SRTs obtained in *Local* are consistently lower than those obtained remotely. A three-way repeated measures analysis of variance (ANOVA) was conducted to analyze the main effects of the scene, noise type and NR, and their interaction effects on SRTs. (Table 2) Results are followed:

1) Scene effects were statistically significant on SRTs. Note that *Remote* 1 gives lower SRTs than *Remote* 2 under all conditions, even though the latter seems to be closer to *Local*. *Remote* 2 and *Local* had the participant seated in the same sound-proof room, the only difference came from the speech quality degradation caused by network transmission; in contrast, *Remote* 1 had extra environmental mismatches from *Local* since participants were seated in a regular conference room. A possible explanation could be that the conference room provided a more comfortable environment for participants than the sound-booth and there was no substantial noise difference (40 dBA vs. 30 dBA).

2) Noise type effects on the SRTs. One can see from Fig. 2 that participants performed better in SSN than in Babble, which was consistent with previous findings [16].

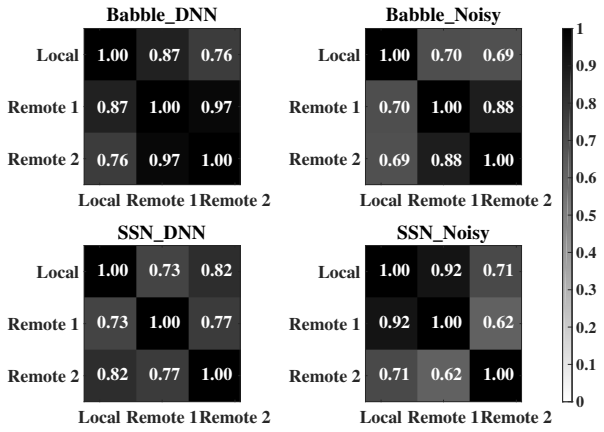

Fig. 3. Correlation matrices of the mean SRTs among different assessment scenes.

3) NR effects on SRTs were statistically significant. As shown in Fig. 2, DNN-based NR significantly improved the intelligibility of denoised speech under all circumstances. These results show that the effect of NR algorithms on CI could be revealed via remote assessments.

4) Interaction between scene and noise type was not significant; besides, interactions among all three factors were not significant, indicating no combined effects for these factors. Two significant interaction effects were observed (scene × NRs, and noise types × NRs), which tells that NR effects differed across scenes and noise types.

Figure 3 demonstrates pair-wise correlations of mean SRTs among the three scenes in four noise-related conditions. As shown, both remote assessments had strong correlations with local assessments regardless of the noise-related conditions. These results show that even remote experiments failed to match the absolute scores obtained in local assessments, remote experiments may replicate critical relative outcomes. For example, both remote and local tests revealed the effect of NR for CI users. The SRTs of remote participants fell short of their laboratory performance in terms of absolute levels, but the results were highly correlated among the three modes.

3. EXPERIMENT II: MAXIMA SELECTION NUMBER

3.1. Rationale

The objective of this experiment was to determine how well the remote subjective assessments match the pattern of results observed in their laboratory counterparts when the parameters of the signal processing strategy change.

In most multichannel CIs, temporal envelopes (TEs) are extracted from each channel of the audio input. Peak-picking strategies picks a certain number of channels with peak TEs (Maxima selection) to generate the modulated electrical pulses at corresponding electrodes along the cochlea [17, 18]. Generally, the more channels picked, the more detailed spectro-temporal information could be presented to the auditory system. This experiment assessed the effect of the maxima selection number on SRTs in babble noise. The assessments were conducted locally and remotely, and the reliability of remote assessments was evaluated.

3.2. Methods

Ten NH native Mandarin speakers were recruited as participants. All had thresholds $\leq$ 25 dB Hearing Level between 0.25 and 8 kHz.

Vocoded speech, a synthesized version of the signal, was used as test stimuli in this experiment. Vocoder simulations are commonly used to simulate the effects of signal processing provided by a CI [19], and to help researchers to evaluate how the auditory system processes degraded auditory signals with controlled conditions.

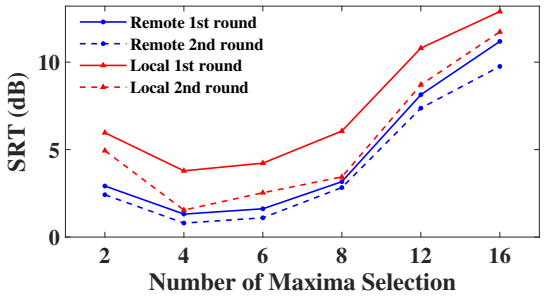

Fig. 4. Mean SRTs for different test conditions.

3.3. Materials and Procedures

The target speech were the same as in Experiment I. A 4-talker babble noise was used. The vocoded speech were generated according to [19]. SRTs were measured to evaluate speech intelligibility processed by a simulated advanced combination encoder (ACE) strategy with the number of maxima selection set to 2, 4, 6, 8, 12 and 16.

The assessments were conducted face-to-face in laboratory (*Local*) and remotely via Tencent Meeting (*Remote*). In *Local*, both the experimenter and participants were seated in a sound booth. Stimuli were presented to the participants via a pair of headphones (Sennheiser HD650) connected to a laptop (HUAWEI MateBook 14) with an external sound card (Focusrite Scarlett 2i4). In *Remote*, the experimenter used the same laptop as in *Local* to run the experiment and share audios with participants via Tencent Meeting. Participants were asked to have the tests in a quiet room (noise floor $\leq$ 40 dBA) at their places, and audios were presented via their own computers and earphones (wired or wireless).

This experiment followed a procedure similar to Experiment I, except that *Local* was always tested first, followed by *Remote* was conducted no less than 24 hours after.

3.4. Results and discussions

Results are shown in Figure 4. SRTs changed as a function of the number of maxima selection, which is consistent with findings in [20–22].

It is clear that the remote and local assessments gave similar trends of SRTs. The overall performance achieved in *Local* were worse than in *Remote*, while the results of the 2nd round in *Local* are similar to those of *Remote*. Previous studies have shown the importance of training in learning to use the information available in degraded signals [23, 24]. Therefore, insufficient training of vocoded speech for NH listeners might explain the higher SRTs of the 1st round than those of the 2nd round in *Local* ($p < 0.05$ except at maxima = 2, paired-sample t-test). Besides, no significant difference was found between SRTs for the 2nd round of *Local* and the 1st round of *Remote* ($p > 0.05$ for all comparisons), suggesting that it is training rather than remote signal transmission issues and background noise that affected the results of cochlear parameter (maxima selection) changes. The Pearson correlation between *Local* and *Remote* indicates a significant correlation ($r = 0.995, p < 0.001$), indicating good reliability of the remote test.

4. CONCLUSIONS

In this paper, the feasibility and reliability of remote subjective speech intelligibility assessments in CI users were evaluated. Even the absolute SRTs attained remotely mismatched with those measured in local tests due to various subjective and objective conditions, the effects of different factors (e.g., noise type, NR, and number of maxima selection) in remote scenes were consistent with those in local. We can conclude that remote subjective assessments could be a reliable alternative to the face-to-face assessments for CI research in the pandemic. The relative variation of specific performance can be measured reliably, but the absolute values should be carefully compared and explained according to experimental conditions.

5. ACKNOWLEDGE

The authors are grateful to all the participants for their kindly cooperation. This work is jointly supported by National Natural Science Foundation of China (61771320), Guangdong Key Area R&D Project (No. 2018B030338001), and Tencent Ethereal Audio Lab. This work was done when the first author worked as intern at Tencent Ethereal Audio Lab.

6. REFERENCES

[1] Deborah Lupton, "Doing fieldwork in a pandemic," *Crowdsourced document.*, https://docs.google.com/document/d/1clGjGABB2h2qbduTgfqribHmog9B6P0NvMgVuiHZCl8/edit, 2020.

[2] China Internet Network Information Center, "The 47th China statistical report on internet development," Report, CNNIC. http://www.cac.gov.cn/2021-02/03/c1613923423079314.htm, 2021.

[3] Pedro Mayorga, Laurent Besacier, Richard Lamy, and J-F Serignat, "Audio packet loss over IP and speech recognition," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2003, pp. 607–612.

[4] P Burgos, EP Sanders, C Cucchiarini, RWNM van Hout, and H Strik, "Auris populi: crowdsourced native transcriptions of dutch vowels spoken by adult spanish learners," *Proceedings of Interspeech 2015*, pp. 2819–2823, 2015.

[5] Charlotte R Vaughn, "Expectations about the source of a speaker's accent affect accent adaptation," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 3218–3232, 2019.

[6] Yevgeniy Vasilyevich Melguy and Keith Johnson, "General adaptation to accented english: Speech intelligibility unaffected by perceived source of non-native accent," *The Jour-*

nal of the Acoustical Society of America, vol. 149, no. 4, pp. 2602–2614, 2021.

[7] Sarah E Yoho and Stephanie A Borrie, "Combining degradations: The effect of background noise on intelligibility of disordered speech," *The Journal of the Acoustical Society of America*, vol. 143, no. 1, pp. 281–286, 2018.

[8] De Wet Swanepoel and James W Hall, "Making audiology work during covid-19 and beyond," *The Hearing Journal*, vol. 73, no. 6, pp. 20–22, 2020.

[9] Kevin M Chu, Leslie M Collins, and Boyla O Mainsah, "Assessing the intelligibility of vocoded speech using a remote testing framework," *arXiv preprint arXiv:2105.14120*, 2021.

[10] Joshua D Sevier, Sangsook Choi, and Michelle L Hughes, "Use of direct-connect for remote speech-perception testing in cochlear implants," *Ear and Hearing*, vol. 40, no. 5, pp. 1162–1173, 2019.

[11] Yuyong Kang, Yupeng Shi, Fushi Xie, and Nengheng Zheng, "A deep learning approach for single-channel speech dereverberation," *National Conference on Man-Machine Speech Communication*, 2019.

[12] Dong Wang and Xuewei Zhang, "THCHS-30: A free Chinese speech corpus," *arXiv e-prints*, pp. arXiv–1512, 2015.

[13] Andrew Varga and Herman JM Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[14] Michael Nilsson, Sigfrid D Soli, and Jean A Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[15] Hongmei Hu, Xin Xi, Lena LN Wong, Sabine Hochmuth, Anna Warzybok, and Birger Kollmeier, "Construction and evaluation of the Mandarin Chinese matrix (CMNmatrix) sentence test for the assessment of speech recognition in noise," *International Journal of Audiology*, vol. 57, no. 11, pp. 838–850, 2018.

[16] Cuncun Ren, Jing Yang, Dingjun Zha, Ying Lin, Haihong Liu, Ying Kong, Sha Liu, and Li Xu, "Spoken word recognition in noise in mandarin-speaking pediatric cochlear implant users," *International Journal of Pediatric Otorhinolaryngology*, vol. 113, pp. 124–130, 2018.

[17] Qian-Jie Fu and John J Galvin III, "The effects of short-term training for spectrally mismatched noise-band speech," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1065–1072, 2003.

[18] Philipos C Loizou, "Signal-processing techniques for cochlear implants," *IEEE Engineering in Medicine and Biology Magazine*, vol. 18, no. 3, pp. 34–46, 1999.

[19] Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[20] Lendra M Friesen, Robert V Shannon, Deniz Baskent, and Xiaosong Wang, "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," *The Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1150–1163, 2001.

[21] Michael Dorman, Philipos C Loizou, Anthony J Spahr, and Erin Maloff, "A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants," *Journal of Speech, Language, and Hearing Research*, vol. 45, no. 4, pp. 783–788, 2002.

[22] Yefei Mo, Huali Zhou, Fanhui Kong, Peina Wu, and Qingling Meng, "Effects of cochlear simulation with different dynamic range and electrode maxima selection on speech understanding in noise," *The 9th Conference on Sound and Music Technology*, 2019.

[23] Matthew H Davis, Ingrid S Johnsrude, Alexis Hervais-Adelman, Karen Taylor, and Carolyn McGettigan, "Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences.," *Journal of Experimental Psychology: General*, vol. 134, no. 2, pp. 222, 2005.

[24] Julia Jones Huyck, Rachel H Smith, Sarah Hawkins, and Ingrid S Johnsrude, "Generalization of perceptual learning of degraded speech across talkers," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 11, pp. 3334–3341, 2017.