Speech Communication 89 (2017) 47–57

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Predicting speech reception thresholds of cochlear implant users using a modified envelope based measure

Vahid Montazeri[a,*], Shaikat Hossain[b], Peter F. Assmann[a]

[a] School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, TX 75080, United States
[b] Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, United States

ARTICLE INFO

Article history:
Received 23 June 2016
Revised 7 January 2017
Accepted 23 February 2017
Available online 28 February 2017

ABSTRACT

An envelope-correlation based measure (ECM), proposed in Yousefian and Loizou (2012), was applied to predict the speech reception thresholds (SRTs) of nine pre- and post-lingual cochlear implant (CI) users in the presence of speech-shaped masker and a female competing talker. A principal component analysis showed that CI users' aided pure-tone thresholds were almost orthogonal to the predictions made by ECM, indicating that the ECM does not take aided hearing loss into account. A modified ECM is proposed which includes pure-tone thresholds into the predictions. The proposed model adjusts the ECM predictions based on the pure tone threshold values, and is able to predict CI users' SRTs with higher correlation ($r = 0.92$) compared to the original ECM ($r = 0.59$). The results reported in this study indicate that information available on MAPs (e.g. T-levels, number of active electrodes) may not be sufficient to predict CI users' SRTs in noise. In additions, results show that lack of audibility (i.e. elevated pure-tone thresholds) can dramatically limit speech intelligibility in noise in (even experienced) CI users.

1. Introduction

Cochlear implantation has been shown to significantly improve different aspects of quality of life in severe-to-profound hearing-impaired individuals. In particular, cochlear implants (CIs) have been clinically demonstrated to improve speech intelligibility (e.g. Faulkner and Pisoni, 2013). Although CI listeners are relatively successful in understanding speech in quiet, often times they experience great difficulties in the presence of background noise (Gifford and Revit, 2010; Caldwell and Nittrouer, 2013). The front-end speech enhancement algorithms implemented in cochlear implants (e.g. Mirzahasanloo et al., 2013) do not provide adequate noise attenuation. This limitation can be attributed in part to the absence of a model that accounts for different aspects of cochlear implant speech perception. Development of such a model would provide engineers with a benchmark to further optimize the existing speech processing strategies for cochlear implants. It would also help audiologists choose the optimum parameters for CI users (i.e. their MAP[1]s) to maximize their speech intelligibility in the presence of background noise.

Several speech perception models have been developed for normal-hearing and hearing-impaired individuals, such as the traditional speech transmission index (STI) (Houtgast and Steeneken, 1985) and speech intelligibility index (SII) (ANSI, 1997). These models, however, do not provide accurate predictions for CI users as the assumptions they are based upon are not valid for CI users. The SII, for instance, assumes that intelligibility varies linearly within a range of 30 dB (from −15 dB to 15 dB). As will be seen later in this paper, this assumption is invalid for CI users as the starting point of the linear segment of CI listeners' psychometric function is highly correlated with their aided hearing loss,[2] and starts from SNR values higher than −15 dB (Friesen et al., 2001).

Given their relative success in predicting speech intelligibility for hearing-impaired population (Payton et al., 1994), modified STI-based measures were expected to be able to predict CI speech intelligibility. This expectation was further supported by the similarities of STI calculation and the signal processing procedures in CI devices. Motivated by this, Goldsworthy and Greenberg (2004) investigated the reasons that the STI-based intelligibility models (and even their modified versions) often fail to predict speech intelligibility when the speech is passed through non-linear operations (such as those used in CI processors). Based on their analyses, they proposed four modifications that could possibly account for the effects of linear and non-linear operations as

* Corresponding author.
E-mail addresses: vahid.montazeri@utdallas.edu (V. Montazeri), shaikath@usc.edu (S. Hossain), assmann@utdallas.edu (P.F. Assmann).

[1] MAPs include all the signal processing parameters programmed in a CI speech processor. These parameters are adjusted by audiologists to maximize CI users' speech recognition scores.

[2] Pure-tone threshold when the CI device is turned on.

http://dx.doi.org/10.1016/j.specom.2017.02.007
0167-6393/

well as additive noise and reverberation on speech intelligibility. Although all the four modifications failed to achieve such a goal, they produced qualitatively reasonable results to predict the effects of non-linear processing on speech intelligibility. The authors, thus, suggested that these modifications could have the potential to predict the intelligibility of CI-processed speech in noise.

The similarity of speech processing in cochlear implants and voice encoders (vocoders) (Shannon et al., 1995) has led researchers to investigate CI listeners' speech intelligibility using vocoder simulations (see Dorman et al. (1998) as an example). This approach has the benefit of limiting the effects of patient-specific confounds. Chen and Loizou (2010) examined the performance of perceptual evaluation of speech quality (PESQ) metric in predicting the intelligibility scores obtained from normal-hearing individuals listening to vocoded speech in noise (tone-excited vocoder). PESQ was originally proposed to predict speech quality over telecommunication systems (ITU-T, 2000). However, studies such as Ma et al. (2009) found that PESQ can also predict speech intelligibility in noisy conditions. Chen and Loizou (2010) reported a minimum correlation of $r = 0.92$ across different vocoder and noise conditions.

Chen and Loizou (2011) evaluated the correlation of several measures with speech intelligibility scores collected from normal-hearing listeners presented with vocoder simulations, where the speech stimuli were degraded with speech-shaped noise or a competing talker. In their study, a Normalized Covariance Metric (NCM) was reported to yield a high correlation of 0.92 with listeners' intelligibility scores.

In a follow-up study by Yousefian and Loizou (2012), an envelope-correlation based measure (ECM) was proposed which is an adaptation of the NCM for CI recipients. Their proposed model computes the covariance between the envelopes of the target and degraded signals (additive noise added to the target) after they are processed based on the CI user's MAP. Using a model training procedure, the ECM then maps the calculated covariance value to speech intelligibility scores ($0 - 100\%$). For example, in their study, an overall covariance of 0.65 corresponded to 50% intelligibility in noise. In other words, the Signal to Noise Ratio (SNR) at which the overall covariance was 0.65 corresponded to the speech reception threshold (SRT), where SRT is defined as the SNR at which a CI user scores 50% intelligibility. The model performance was evaluated by recruiting nine post-lingual CI users in a variety of conditions (i.e. six masker types and three speech enhancement methods). Good agreement between the observed and predicted SRTs was reported ($r = 0.96$). Other studies such as Santos et al. (2013) have also shown high correlation between CI users' actual and predicted speech intelligibility scores, using envelope coherence-based models.

Although the above envelope coherence-based models of speech perception have shown the potential to predict CI users' speech intelligibility in the presence of background noise, it is not clear to what extent these models can capture individual differences between CI users. CI users' intelligibility scores vary depending on a number of factors, including etiology and onset of hearing loss, number and pattern of surviving ganglions cells, electrode array insertion depth, distance of the electrodes from the neurons, etc. The ECM proposed by Yousefian and Loizou was developed based on the assumption that nearly all the individual differences in CI users can be captured from their MAPs. For instance, the distance of the electrodes from the excitable neural elements affects the electrodes' threshold levels (T-levels[3]), available in CI users' MAPs. Thus, incorporating CI users' MAPs into the ECM predictions was able to capture the contribution of individual differences toward speech intelligibility scores. In Yousefian and Loizou, the correlation between the CI users' actual and ECM-predicted SRT scores was found to be $r = 0.96$ across 18 different conditions.

It is, however, possible that the participants in Yousefian and Loizou (2012) were a set of homogenous CI users with a restricted range of aided hearing loss. This is supported by the fact that all the CI users tested by Yousefian and Loizou scored above 50% in quiet. Thus, the model training, performed on a subset of participants, was able to capture all the individual differences among the CI users and account for the ECM prediction errors. On the contrary, if the set of recruited CI users was heterogeneous, in the sense that their intelligibility scores spanned a wide range of scores below and above 50%, the subset chosen for model training would not be fully representative of the overall set, leading to higher prediction errors. If this interpretation is valid, then the accuracy of the ECM predictions would be more likely attributable to the similarity of the individual factors among the CI users in Yousefian and Loizou and not the wealth of neuropathological information encoded in the MAPs.

Plomp (1986) proposed an SRT model which is developed based on two sources of deficits, namely attenuation and distortion. According to Plomp's study, the attenuation factor is concerned with the lack of audibility of the incoming sounds or, in other words, threshold elevations. On the other extreme, the distortion factor is concerned with the linear or nonlinear distortions introduced to the incoming sound signal (suprathreshold). Other studies such as Preece and Fowler (1992) have also reported linear relationships between the listeners' SRT in noise and pure-tone thresholds. Thus, it seems reasonable to hypothesize that information regarding the listeners' pure-tone thresholds may contribute to CI users' SRT scores.

In this paper, we ask whether the CI users' pure-tone thresholds are correlated with their actual SRT scores and whether the SRT scores predicted by ECM account for these thresholds. We demonstrate that the CI users' SRT scores are correlated with their pure-tone thresholds. We also show that variations in the ECM values and pure-tone thresholds are weakly correlated with each other, meaning that the information encoded in CI users' MAPs, shown in the form of ECM values is not adequate to predict SRT scores. As a result, knowledge of coherence (or covariance) between the envelopes of the degraded and target signals only partially explains the variability in the speech intelligibility scores. Following the interpretation of the results, we propose a modified ECM that incorporates CI users' audiometric thresholds with the original ECM.

2. Materials and methods

2.1. Participants

Nine adult CI users were recruited to participate in the experiments. The participants were all native speakers of American English. The average age of the participants was 58 years ($\pm$ 19.46). Three of the nine participants were pre-lingually deafened CI users and six were post-lingual. Participants were paid for their participation. All participants had a minimum of one year experience using their device routinely. Two of the nine participants used 'Freedom' devices and seven had 'Nucleus' devices; all devices are developed by Cochlear Ltd. Biographical data for all participants are presented in Table 1.

2.2. Speech material

The speech material (target sentences) was taken from the IEEE corpus (IEEE, 1969). Each IEEE sentence contains 7 to 12 words, produced by a male speaker in anechoic conditions (mean $f_0 =$ 130.2 Hz). The corpus is organized in 72 lists of 10 sentences

[3] T-level is defined as the softest electrical input level detectable by user for each electrode. Similarly, C-level is defined as the electrical input level that is perceived as loud, but comfortable (Wolfe and Schafer, 2010).