The Effect of Cochlear Implant Processing on Speaker Intelligibility: A Perceptual Study and Computer Model

Lin Lin[1], *Jon Barker*[2] *and Guy J. Brown*[2]

[1]College of Communication Engineering, Jilin University, Changchun 130025, P. R. China
[2]Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

lin_lin@jlu.edu.cn, j.p.barker@sheffield.ac.uk, g.j.brown@sheffield.ac.uk

10.21437/Interspeech.2015-364

Abstract

Cochlear implant (CI) users have great difficulty understanding speech in noise. However, some speakers are found to be more intelligible than others. This paper tests whether a glimpse-based model, that was previously used to successfully explain speaker intelligibility in normal-hearing listeners, can also predict speaker intelligibility for CI users. The model employs a front-end that mimics the effects of energetic masking. This is coupled with a back-end that employs statistical models of CI-processed speech to recognise the unmasked glimpses of the target signal. Listening tests were conducted using signals that simulate the effect of hearing speech mixed with speech-shaped noise through a CI, at signal-to-noise ratios ranging from -4 to 6 dB. The intelligibility of 34 different talkers was measured at each noise level. The model is able to explain the variation in the speaker intelligibilities: the correlation between listener and model intelligibilities varies between 0.87 and 0.91 depending on noise level. This is higher than correlations previously found for normal hearing listeners. Our results have the potential to inform future CI signal processing strategies.

Index Terms: speaker intelligibility, cochlear implant, automatic speech recognition, glimpsing

1. Introduction

There is a great variability across the average speech intelligibility of different talkers. This variability is related to a large number of properties of the signal including average speech rate, average fundamental frequency, spectral balance, etc. These factors interact in complex ways and there are many conflicting findings in the literature (see [1] for a recent review). However, for normal-hearing listeners, noise-free intelligibility variability is of little relevance as listening errors are rare and have little impact on communication.

Unlike normal hearing listeners, cochlear implant (CI) users can find conversational speech hard to understand even in noise-free environments. For example, [2] reports that CI users find talkers using a conversational speech style to be less intelligible than those using a clear speech style. Li et al. [3] found that CI listening performance worsened with increasing speech rate. Green et al. [4] found that judgements of talker intelligibility made by CI-users correlated with those of normal hearing users, suggesting that the primary determinants of intelligibility remained the same. This is a useful finding that may inform CI signal processing, or the design of synthesised voices for CI-users. Unfortunately, however, speech is rarely heard in noise-free environments, and CI-users have to contend with both an impoverished speech signal and the effects of background noise, plus interactions between these two adversities.

The intelligibility of a speaker in noise is governed by different factors, and is not necessarily well predicted by the speaker's noise-free intelligibility. In stationary noise, intelligibility is reduced due to energetic masking causing information to be lost. The degree of information loss can vary between speakers. For example, speakers whose formants have a smaller bandwidth and higher peak energies will be more resistant to masking. In general, intelligibility prediction requires a model that can measure both the amount of signal unmasked, and the amount of information encoded in the unmasked region. Barker and Cooke [5] presented a 'glimpsing' model of speaker intelligibility that combined these factors. Their model employed two stages: a simulation of the auditory periphery that modelled the effects of energetic masking, and a statistical model of each speaker that used missing data techniques to recognise utterances based on the unmasked glimpses. It made successful predictions in high noise levels where only sparse signal regions survived energetic masking; it was less successful in low noise levels.

The current study measures and models CI-processed talker intelligibility in noise. This extends the work of Green et al. [4] in two respects. First, listening experiments are conducted with speech *in noise* rather than with noise-free speech (Sec. 2). Second, we present a complete model, based on that of [5], that is able to successfully predict talker intelligibility (Sec. 3).

2. Listening test

2.1. Cochlear implant simulation

Stimuli were processed by a CI simulation that modelled an n-of-m advanced combination encoder (ACE) strategy [6, 7]. The input signal was pre-emphasised by a high-pass filter with a cut-off of 1200 Hz, and then passed through a 22-channel filterbank implemented by applying the fast Fourier transform (FFT) to 8 ms Hann-windowed frames. The envelope was computed at the output of each filterbank channel and a compression function was applied. Subsequently, the largest 8 spectral peaks were extracted for each time frame. Where a spectral peak was selected, a current pulse was generated in the corresponding frequency channel. The resulting train of interleaved pulses constituted the output from the CI simulator. The parameters of the simulator were set according to those given by [7].

To generate a CI-processed signal that could be presented to listeners, a further resynthesis stage was used. A low-pass filter (cutoff 400 Hz) was applied to the pulse train in each channel in order to estimate the envelope. White noise was then passed through the same filterbank that was used in the analysis phase, and this was modulated in amplitude by the channel envelope. Time-domain functions for the output of each channel were then

Copyright © 2015 ISCA

reconstructed by overlap-add, and summed across all channels to give the final CI-processed resynthesised signal.

2.2. Speech and noise materials

Speech material was taken from the Grid corpus [8, 5]. This consists of a total of 34,000 utterances, comprising 1000 utterances from each of 34 speakers (18 male, 16 female). Each sentence in the Grid corpus is a six-word utterance such as "put red at G 9 now", in which the letter and the digit are keywords. More specifically, the keywords are 10 digits ('0' to '9') and 25 letters ('W' was excluded due to its multisyllabicity). All signals were sampled at a rate of 16 kHz.

Speech-shaped noise was added to each utterance at 6 signal-to-noise ratios (SNRs): -4, -2, 0, 2, 4 and 6 dB. The spectrum of the noise matched the long-term spectrum averaged over all utterances in the Grid corpus. The initial and trailing silence was removed from each utterance prior to the addition of the noise. Subsequently, each noisy utterance was passed through the simulation of CI processing described above.

A cohort of 30 participants was split into 5 blocks of 6 listeners. Within each block, the 6 listeners heard the same CI-processed utterances but with a different permutation of the SNRs. In order to balance the utterances across SNRs, sub-lists were formed consisting of 3 utterances spoken by 34 speakers drawn at random, without replacement, from the Grid corpus. These lists were then rotated around the 6 listeners in each block. Within each block, listeners therefore heard $34 \times 3 \times 6 = 612$ utterances, and a total of $612 \times 5 = 3060$ Grid utterances were used in the study. The utterances within each listener list were presented in random order; however, to limit the number of abrupt changes in SNR each consecutive pair of utterances had the same SNR.

2.3. Procedure

Listeners were seated in an IAC single-walled sound-proof booth, and heard the stimuli over headphones (Sennheiser HD480). Stimuli were presented diotically at a sound level of 70 dB SPL. The experiment was controlled by a computer program that delivered the stimuli to listeners and recorded their responses. For each CI-processed mixture of speech and noise, listeners were asked to identify the letter and digit and type their responses on a conventional computer keyboard.

Prior to the main listening test, each subject completed a pre-test which familiarised them with the stimuli and the computer interface. Listeners heard 20 pairs of utterances, consisting of an unprocessed Grid utterance, followed by the same utterance processed through the CI simulation. No noise was added to the utterances in the pre-test. Each listener completed the pre-test (40 utterances) and main listening test (6 lists of 102 utterances) in a total of approximately 40 minutes.

2.4. Results

Figure 1 shows how the keyword identification rate varies with SNR. In the least noisy condition, 6 dB, 85 % of the digits and 65 % of letters are correctly identified. Across all SNRs, performance for letter recognition is approximately 20 % lower than digit recognition. Both scores fall by about 45% points as the SNR is reduced to its lowest level, -4 dB, but remain substantially above chance. The variation of intelligibility across speakers – the focus of this study – will be presented in Section 4 alongside the results of the model.

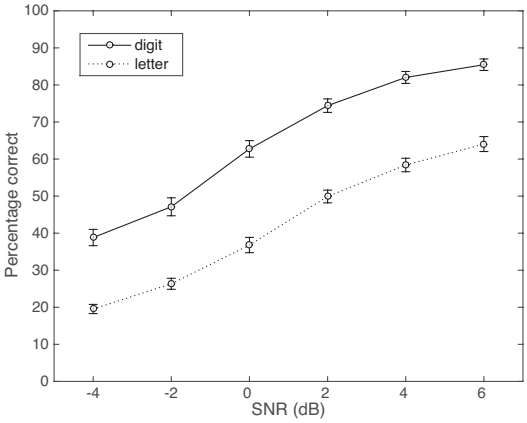

Figure 1: *Percentage of keywords correctly identified by listeners. Error bars denote* ± 1 *standard error.*

3. Modelling

3.1. Glimpse analysis

It has been argued that the robustness of speech perception in noise is due, at least in part, to the ability of listeners to exploit spectro-temporal 'glimpses' of speech that lie above the noise floor. For example, Cooke [8] shows that the proportion of time-frequency units that are glimpsed is a good predictor of the intelligibility of speech in noise. Similarly, a glimpsing model was shown to provide a close fit to the intelligibility of different talkers, particularly in high noise conditions [5]. Here, we use a computer model to investigate whether glimpsing can explain the intelligibility of different speakers in noise, when they are heard through a CI. The computer model combines a glimpsing analysis with automatic speech recognition (ASR), and was presented with the same signals that listeners heard during the perceptual experiment described in Sect. 2.

Speech glimpses in the CI-processed signals were identified as follows. First, a representation of firing rate in the auditory nerve was computed by passing the acoustic signal through a bank of 32 gammatone filters [9]. The centre frequencies of the filters were spaced linearly between 50 Hz and 8 kHz on an ERB-rate scale. Within each frequency channel, the Hilbert envelope was computed and smoothed by a leaky integrator with a 8 ms time constant. The smoothed envelopes were then sampled at a rate of 100 Hz and compressed by a cube root function to give a spectro-temporal excitation pattern (STEP). Separate STEPs were computed for clean CI-resynthesised Grid utterances and utterances to which noise was added before CI processing, as described in Sect. 2.2 above. The clean and noisy STEPs were then compared, and a time-frequency unit in the noisy STEP was marked as 'reliable' if the local SNR exceeded a threshold T dB. Adjacent reliable units were aggregated to form larger regions, which varied in size; if a reliable region contained greater than N time-frequency units, then it was defined to be a glimpse.

3.2. ASR-based glimpsing model

Following [8, 5] a missing-data ASR system was used to recognise the target words in each test utterance, given the STEP features for the utterance and information about the reliable glimpses that it contained. Speaker-dependent statistical mod-

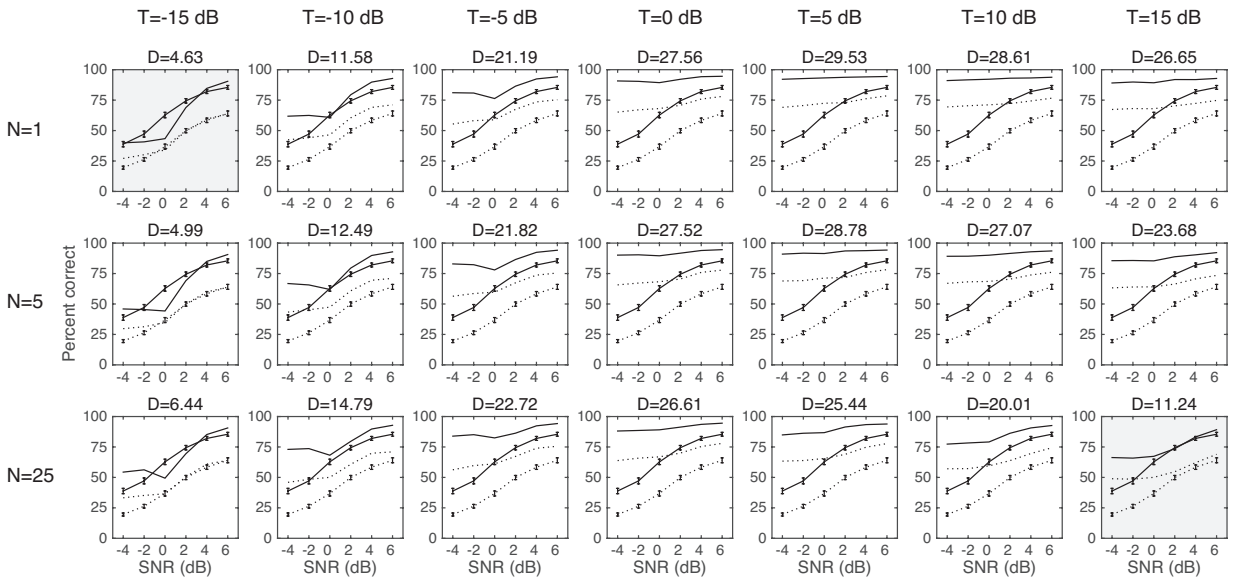

Figure 2: *Listener and model performance for recognition of digits (solid lines) and letters (dotted lines) at a range of SNRs, and for different values of the model parameters T and N. The distance D between listener and model performance is shown for each subplot. Listener results are shown with error bars of $\pm$ 1 standard error.*

els were trained for each of the 34 speakers in the Grid corpus, using 500 utterances drawn randomly from the corpus (but excluding the utterances used in the test set). Statistical models were trained using 64-dimensional spectral vectors. These consisted of 32 STEP features extracted from the CI-processed training speech, together with their temporal derivatives (computed using linear regression over a two-frame window).

HTK [10] was used to train a set of word-level hidden Markov models (HMMs) for each speaker. Word models had 2 states per phoneme, and each state was modelled by a Gaussian mixture model (GMM) with 7 components and diagonal covariance. Initially, the HMMs were trained on unprocessed Grid utterances. Subsequently, the models were adapted to the CI-processed speech by single-pass retraining [11] on speech that had been processed through the CI simulation and resynthesised, as described in Sect. 2.1.

During testing, the glimpse analysis of each noisy utterance was encoded by a binary spectro-temporal mask in which 'reliable' (glimpsed) regions were set to unity and 'unreliable' (masked) regions were set to zero. Each noisy test utterance was recognised from its STEP features and binary mask using the missing data technique of [12]. This used a bounded marginalisation approach to decode the noisy speech features, using the speaker-dependent statistical models of CI-processed clean speech. The output of the ASR system was scored in the same way as listener responses.

3.3. Parameter tuning

There were two free parameters in the glimpse analysis; the local SNR threshold T and the glimpse size N. The values of these parameters were tuned empirically, by minimising the average distance between the digit and letter scores for listeners and the ASR system across all speakers in the experiment. As shown in Fig. 2, model performance was substantially better than listener performance for nearly all values of T and N. The figure also shows that two regions of the parameter space give an acceptable match between listener and model performance. In the top left corner of the plot, a good match is obtained when the local SNR threshold T is low and the glimpse size N is small. This region of the parameter space corresponds to a glimpsing strategy that is not sufficiently conservative (i.e., listeners incorrectly interpret some of the noise-dominated background as being part of the target signal). Conversely, an acceptable match between listener and model performance is also obtained at the bottom-right corner of the figure, particularly for SNRs above 0 dB. Here, the local SNR threshold and minimum glimpse size are large, suggesting a glimpsing strategy that is too conservative (i.e., some target-dominated time-frequency regions are not accepted as belonging to the target).

In the following, two parameter sets are therefore considered which provide an acceptable match between listener and model performance, but represent different glimpsing strategies; one with a high acceptance threshold that is overly conservative (T=15, N=25) and one with a low threshold that is not conservative enough (T=-15, N=1).

4. Results

Listener results for the CI-processed speech (Fig. 1) broadly resemble those obtained by [5] for unprocessed speech. Recognition performance for digits and letters increases monotonically with increasing SNR, and performance was higher for digits than letters at all SNRs. However, our performance curves are shifted substantially to the right compared to those of [5], indicating that the task was more challenging when CI processing was applied to the test signals. The two shaded regions of Fig. 2 show that the computer model gives a reasonable match to the average listener data, particularly at higher SNRs.

Individual speaker intelligibilities were compared following the approach described in [5]. For both the listener and computer model, recognition results were partinioned into three sets according to SNR: a low noise conidition (4 and 6 dB SNR), a medium noise condition (0 and 2 dB SNR) and a high noise condition (-4 and -2 dB SNR). For each noise condition, recog-

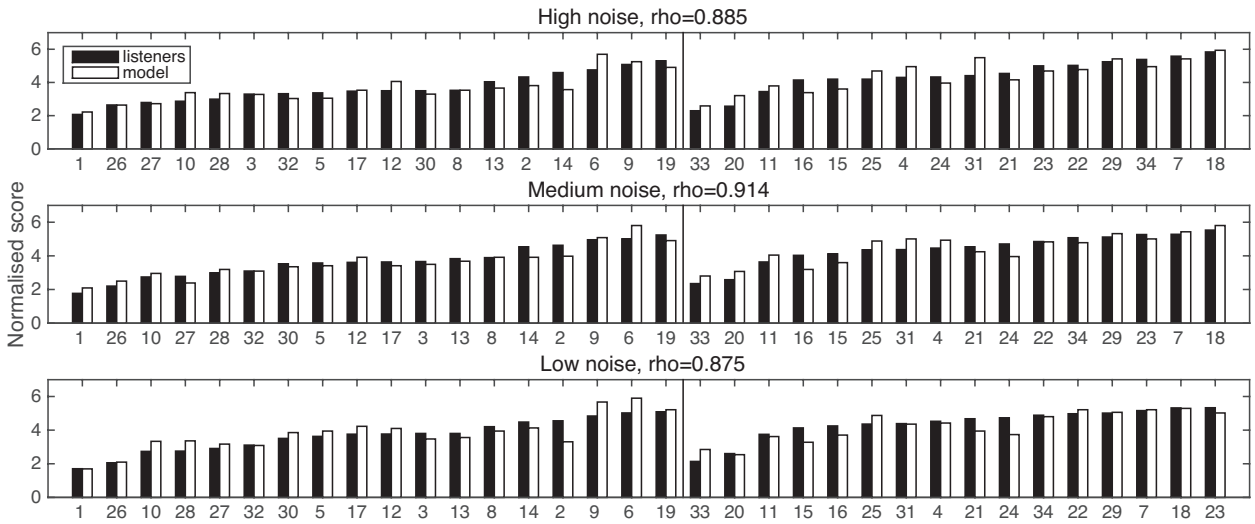

Figure 3: *A comparison of the model (white bars) and listening test scores (black bars) across all speakers in the Grid corpus for the high, medium and low noise level conditions. Male speakers are shown on the left of the vertical line, and female speakers on the right.*

nition scores for each talker were averaged and then arcsine-transformed. Correlation between the listener scores and the model predictions were then computed.

As also found in [5], the average intelligibility of the female speakers is slightly higher than that of the male speakers, but there is a large spread of intelligibilities among the speakers of each gender. The relative intelligibility of speakers is largely independent of the noise level. For the low-threshold tuning (T=-15, N=1) the model-listener correlations were measured to be 0.611, 0.704 and 0.761 for the high, medium and low noise levels respectively. For the high-threshold 'conservative' tuning (T=15, N=25) the correlations rose to 0.875, 0.914 and 0.885 for the high, medium and low noise levels.

Fig. 3 illustrates the correlation for the high-threshold tuning. In the figure, male speakers are shown on the left and female speaker on the right. Within each subplot, speakers are ordered according to increasing intelligibility in the listening test. To make the correlation between the listener and model results more readily apparent, listener and model scores were normalised to have equal mean and variance.

5. Discussion and Conclusions

The correlations between predicted intelligibilities and listening test results are in the range 0.88 to 0.91 across all noise levels (-4 to 6 dB). This contrasts with the natural speech study [5] in which the model only had a good fit (0.93) for the lowest SNR range (-14 to -8 dB). However, in terms of overall keyword intelligibility, it is this -14 to -8 dB SNR range that most closely matches the range of adversity experienced by listeners in the current study. In the previous study it was conjectured that the model works best in situations in which the available spectra-temporal information is sparse. This appears consistent with results presented here.

Of the two alternative tunings that were suggested by matching the average intelligibilities, only the conservative tuning actually provides a good fit to the speaker intelligibility data. This too is consistent with the previous study, where it was found to be necessary to raise the glimpsing threshold above the optimum 0 dB and set a minimum glimpse size in order to fit the listener data. However, for CI-processed speech it was necessary to raise the threshold and minimum glimpse size far higher. This appears to suggest that our listeners were found it harder to extract the CI-processed speech from the noise background, i.e. they could only recover spectra-temporal regions with very high local SNRs. This could be due to impoverished source separation cues, or to weaker internal models of CI-processed speech. If the latter, then it might be expected that experienced CI-users would report high intelligibilities and be better modelled with a lower local-SNR threshold.

The model-listener comparisons shown in Fig. 2 indicate that the model over-estimates listener performance at the lowest SNRs. The glimpse model performance degrades very slowly below 0 dB. This is partly because the extra missing information is compensated by a negative-evidence term in the model that rejects hypotheses that would predict observed speech energy in masked regions. This is a strongly model-driven effect, and it is possible that listeners are not able to benefit from this effect when listening to unfamiliar speech. Again, this could be a limitation of using normal-hearing subjects listening to unfamiliar simulated CI speech, as opposed to real CI users.

In conclusion, the current study shows that a computer model based on glimpsing provides a close fit to the intelligibility of different talkers speaking in a noisy background, and heard through a simulated CI processor. The model allows predictions to be made about the effect of glimpse threshold and size on speaker intelligibility, and might therefore be used to inform signal-processing strategies in CI processors. In ongoing work, experiments are planned with CI-implanted listeners to further validate the model.

6. Acknowledgements

Lin was supported by the China Scholarship Council. Brown was supported by the EU project TWO!EARS under grant agreement 618075. Barker was supported by the EU project INSPIRE under grant agreement FP7-PEOPLE-2011-290000. Many thanks to Tim Jürgens for making available the implementation of the cochlear implant simulator.

7. References

[1] A. Amano-Kusumoto and J.-P. Hosom, "A review of research on speech intelligibility and correlations with acoustic features," Department of Biomedical Engineering Oregon Health and Science University (OHSU), Beaverton, OR, Tech. Rep., 2011.

[2] S. Liu, E. D. Rio, A. R. Bradlow, and F. Zeng, "Clear speech perception in acoustic and electrical hearing," *J. Acoust. Soc. Am.*, vol. 116, no. 4, pp. 2374–83, 2004.

[3] Y. Li, G. Zhang, H. Kang, S. Liu, and D. Han, "Effects of speaking style on speech intelligibilty for mandarin-speaking cochlear implant users," *J. Acoust. Soc. Am.*, vol. 129, no. 6, pp. EL242–EL247, 2011.

[4] T. Green, S. Katiri, A. Faulkner, and S. Rosen, "Talker intelligibility differences in cochlear implant listeners." *J. Acoust. Soc. Am.*, vol. 121, no. 6, pp. EL223–EL229, 2007.

[5] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Communication.*, vol. 49, pp. 402–417, 2007.

[6] A. E. Vandali, L. A. Whitford, K. L. Plant, and G. M. Clark, "Speech perception as a function of electrical stimulation rate: using the nucleus 24 cochlear implant system." *Ear Hear*, vol. 21, no. 6, pp. 608–624, Dec 2000.

[7] S. Fredelake and V. Hohmann, "Factors affecting predicted speech intelligibility with cochlear implants in an auditory model for electrical stimulation," *Hearing Research.*, vol. 287, pp. 76–90, 2012.

[8] M. Cooke, "A glimpsing model of speech perception in noise." *J Acoust Soc Am*, vol. 119, no. 3, pp. 1562–1573, Mar 2006.

[9] G. Brown and M. Cooke, "Computational auditory scene analysis." *Computer Speech and Language.*, vol. 8, pp. 297–336, 1994.

[10] S. Young, G. Evermann, M. Gales, T. Hain, and et al., "The HTK book," http://htk.eng.cam.ac.uk/.

[11] P. Woodland, M. Gales, and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," in *Proc. ICASSP*, 1996, pp. 65–68.

[12] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "A glimpsing model of speech perception in noise." *Speech Communication.*, vol. 34, pp. 267–285, 2001.