# Intelligibility prediction with a pretrained noise-robust automatic speech recognition model

*Zehai Tu, Ning Ma, Jon Barker*

University of Sheffield, Department of Computer Science, Sheffield, UK

{ztu3, n.ma, j.p.barker}@sheffield.ac.uk

## Abstract

This paper describes two intelligibility prediction systems derived from a pretrained noise-robust automatic speech recognition (ASR) model for the second Clarity Prediction Challenge (CPC2). One system is intrusive and leverages the hidden representations of the ASR model. The other system is non-intrusive and makes predictions with derived ASR uncertainty. The ASR model is only pretrained with a simulated noisy speech corpus and does not take advantage of the CPC2 data. For that reason, the intelligibility prediction systems are robust to unseen scenarios given the accurate prediction performance on the CPC2 evaluation.

## 1. Introduction

An accurate intelligibility predictor can be crucial to the development of hearing aid speech enhancement algorithms. CPC2 aims to make comparisons among speech intelligibility prediction approaches for hearing impaired listeners.

The CPC2 database provides a large number of speech and listener recognition performance pairs. The speech signals are simulated in domestic environments and interfered by noises, music, or additional speeches. Various speech enhancement systems are used to process these speech signals and generate binaural outputs to maximise the intelligibility of the target speech for hearing impaired listeners. These processed binaural speech signals and the intelligibility of the corresponding listeners are provided in the database. The CPC2 database is divided into three partitions, each of which consists of a training set and an evaluation set. The predictions on the evaluation sets are used to measure the performance of intelligibility predictors.

The approaches in this paper follow the previous works of taking advantage of state-of-the-art ASR models for intelligibility predictions [1, 2]. In [1], an intrusive approach based on the hidden representations of an ASR model was proposed. Meanwhile, a non-intruisve approach based on ASR recognition uncertainty was proposed in [2]. Unlike the approaches in the previous works, a pretrained noise-robust ASR model is used in this work, thus the training data provided in the CPC2 database is not seen by the ASR model. As a result, the proposed intelligibility predictors show a high potential to generalise to a wide range of scenarios.

## 2. Method

### 2.1. Pretrained noise-robust ASR

For the purpose of building generalised ASR-based intelligibility predictors, the ASR model used in this work was trained with a simulated noisy speech corpus. The Speechbrain [3] transformer ASR recipe for the LibriSpeech [4] was used for training the model, and its released parameters are used to initialise the model[1] before training with the simulated noisy speech corpus.

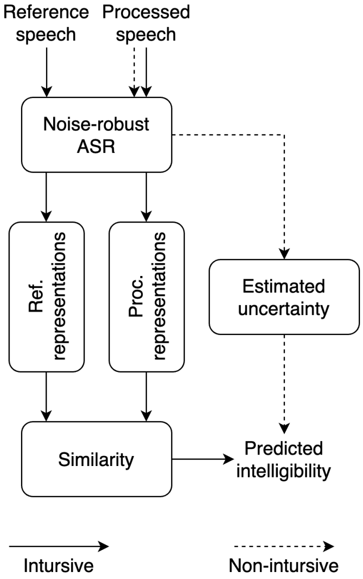

Figure 1: *The work flow of the intrusive and non-intrusive intelligibility prediction methods with a noise-robust ASR model.*

The speech material in the simulated noisy speech corpus is LibriSpeech, which includes 960-hour read English utterances. Each utterance is convolved with a room impulse response (RIR) randomly sampled from the RIR corpus introduced in [5]. In addition, the speech signals are interfered with by the environmental sounds in the ESC database [6]. The speech-weighted signal-to-noise ratio (SNR) of each simulated noisy speech signal is randomly sampled from the range from -6 to 6 dB. The ASR model was trained for 50 epochs and then used for intelligibility prediction.

### 2.2. Intelligibility prediction

An overall work flow of both the intrusive and non-intrusive intelligibility prediction methods are shown in Figure 1.

Given a pair of the processed speech signal and the corresponding reference signal, the intrusive method extracts the representations of the ASR hidden layer. Then the similarity between the reference and processed speech is measured and used to correlate the speech intelligibility. A detailed description of the method can be found in [1]. The decoder representation is used in this work, because it was shown that the language model information can help to achieve more accurate intelligi-

[1] https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/ASR/transformer

Table 1: *Correlation between the intelligibility by listeners and the predicted intelligibility measures on each CPC2 evaluation subset.*

| Subset | RMSE ↓ | NCC ↑ | KT ↑ |
|---|---|---|---|
| Intrusive | | | |
| *1* | 0.250 | 0.790 | 0.579 |
| *2* | 0.277 | 0.713 | 0.543 |
| *3* | 0.242 | 0.801 | 0.632 |
| Non-intrusive | | | |
| *1* | 0.303 | 0.660 | 0.500 |
| *2* | 0.274 | 0.715 | 0.531 |
| *3* | 0.256 | 0.773 | 0.607 |

bility prediction.

When the reference speech signal is not provided, the utterance-level recognition uncertainty can be estimated and then correlated to intelligibility. The detailed derivation is illustrated in [2]. The negative entropy is used in this work as it takes the recognition probabilities of the tokens within the decoding beam into consideration and could produce a more accurate estimation.

The intelligibility predictors in this work do not really take hearing losses into consideration. In the first Clarity Prediction Challenge [7], the evaluation results in [1, 2] showed that the benefit by using the MSBG hearing loss simulation [8, 9, 10, 11] is not significant. Also, for the purpose of building a more general intelligibility predictor, the training of the ASR model does not involve the hearing loss simulation.

Following the convention of evaluating intelligibility prediction, a logistic mapping function $f(x) = 1/[1+\exp(ax+b)]$ is used to re-scale the predicted intelligibility by the pre-trained ASR model. The two parameters of the mapping function were optimised with the training set.

## 3. Results

The evaluation results of the intrusive and non-intrusive methods on the three evaluation subsets are shown in Table 1. The root mean square error (RMSE), normalised cross-correlation (NCC), and the Kendall's Tau (KT) coefficient are reported.

## 4. References

[1] Z. Tu, N. Ma, and J. Barker, "Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners," in *Proc. Interspeech 2022*, 2022, pp. 3488–3492.

[2] ——, "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction," in *Proc. Interspeech 2022*, 2022, pp. 3493–3497.

[3] M. Ravanelli *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[5] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[6] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," 2015. [Online]. Available: https://doi.org/10.7910/DVN/YDEPUT

[7] J. Barker, M. Akeroyd, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter, and R. V. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. Interspeech 2022*, 2022, pp. 3508–3512.

[8] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *JASA*, vol. 94, no. 3, pp. 1229–1241, 1993.

[9] ——, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *JASA*, vol. 95, no. 4, pp. 2277–2280, 1994.

[10] B. C. J. Moore and B. R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *JASA*, vol. 94, no. 4, pp. 2050–2062, 1993.

[11] M. A. Stone and B. C. Moore, "Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear and Hearing*, vol. 20, no. 3, pp. 182–192, 1999.