Fast Speech Intelligibility Estimation using a Neural Network Trained Via Distillation

Trevor J. Cox[1], Yan Tang[1,2] and Will Bailey[1]
[1]*University of Salford, UK.* [2]*University of Illinois*
t.j.cox@salford.ac.uk twitter: @trevor_cox

Aim

Objective measures of speech intelligibility have many uses such as evaluating the degradation of speech during transmission. One intrusive approach is based on the audibility of speech glimpses. The binaural version of the double-ended glimpse method (*BiDWGP*) can provide more robust performance compared to other binaural state-of-the-art metrics [1,2]. But the glimpse method is too slow to allow real-time applications. Knowledge distillation through machine learning will allow BiDWGP to be estimated faster, but what accuracy is achieved?

Knowledge distillation

The slow glimpse method is used to derive a simpler machine-learnt model capable of real-time operation (Figure 1).

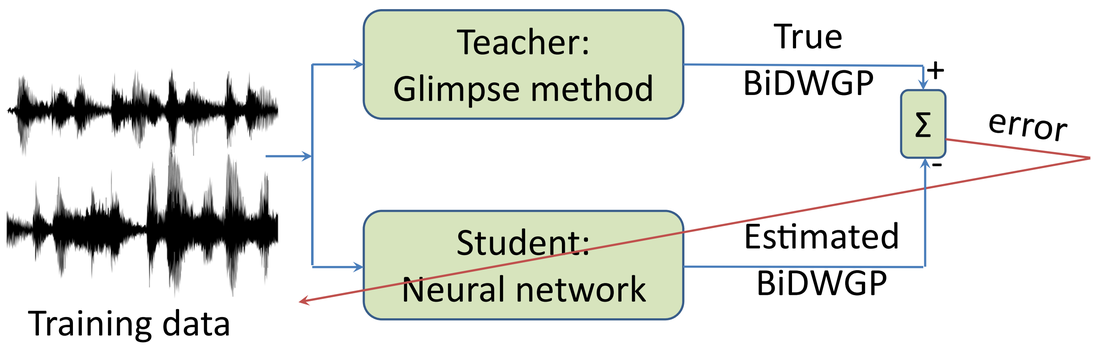

Figure 1. Training the student neural network via knowledge distillation

The *student model* is trained on data generated from the slow *teacher model*, thereby distilling knowledge from teacher to student [3,4]. In this case the teacher is the glimpse-based model (BiDWGP), and the student an artificial neural network. During training, the error between the true and estimated BiDWGP is used to update the weights within the student neural network. Once this network is trained, the student can rapidly estimate the glimpse-based speech intelligibility metric. It is fast enough to allow real-time operation as an intelligibility meter in a Digital Audio Workstation (DAW).

Student neural network

The inputs to the student neural network are cross-correlations between Mel-Frequency Cepstral Coefficients (MFCCs) for the clean and noisy speech. First, MFCCs are calculated in 20 ms windows with 50% overlap [5]. 34 filters from 100 to 7,500 Hz are used. The first eight MFCCs and the energy in the window give the 9 features for each window. The cross-correlation is then evaluated over short sentences to give a reasonable time for intelligibility estimation. Only the largest value of the cross-correlation for the left and right ear signals are used as inputs to simulate better-ear binaural listening (Figure 2).

A feedforward neural network was trained using scaled conjugate training. It had two hidden layers with 16 and 5 neurons respectively.

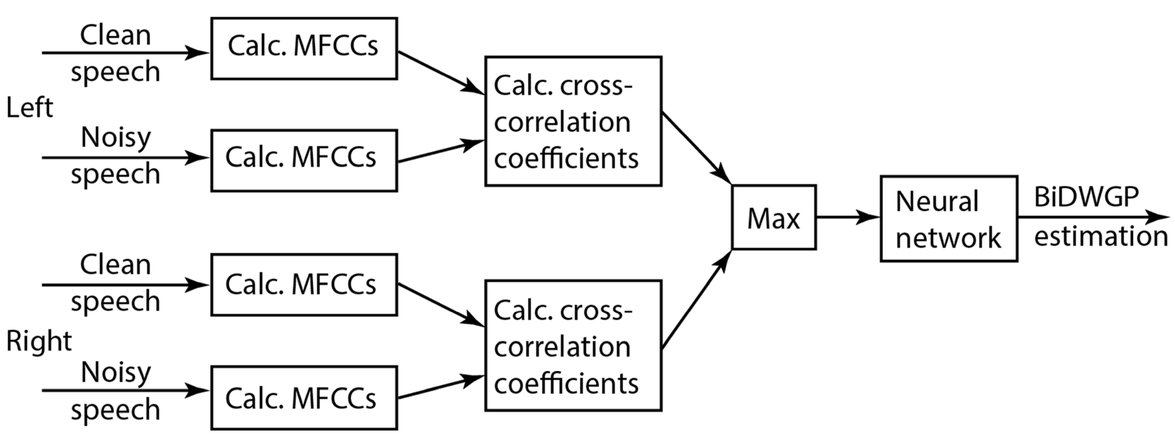

Figure 2. Schematic for BiDWGP estimation via student neural network.

Data

Even for this lightweight artificial neural network, a large amount of training data is necessary to make the distillation robust. 1,200 hours of audio samples were used containing speech from a wide range of sources (SALUC, SCRIBE and r-spin speech corpora and librivox audiobooks). Maskers included speech-shaped noise, competing speech, amplitude-modulated noise, music and sound effects. The signal-to-noise ratio ranged between -35 and +20 dB.

Performance

Performance is evaluated using 736 samples of test data not used in training. A comparison between the estimated speech intelligibility to the full glimpse-based model gives an r^2 of 0.94 (Figure 3). 84% of estimations are within ±0.1 of the correct value. The way the data was generated created more BiDWGP values at the extremes. A sub-set of this test data with a more even distribution of the ground truth data gives r^2 =0.9 and 79% within ±0.1 of the correct value. The next step is to retrain the network with a dataset with a more uniform distribution of BiDWGP.

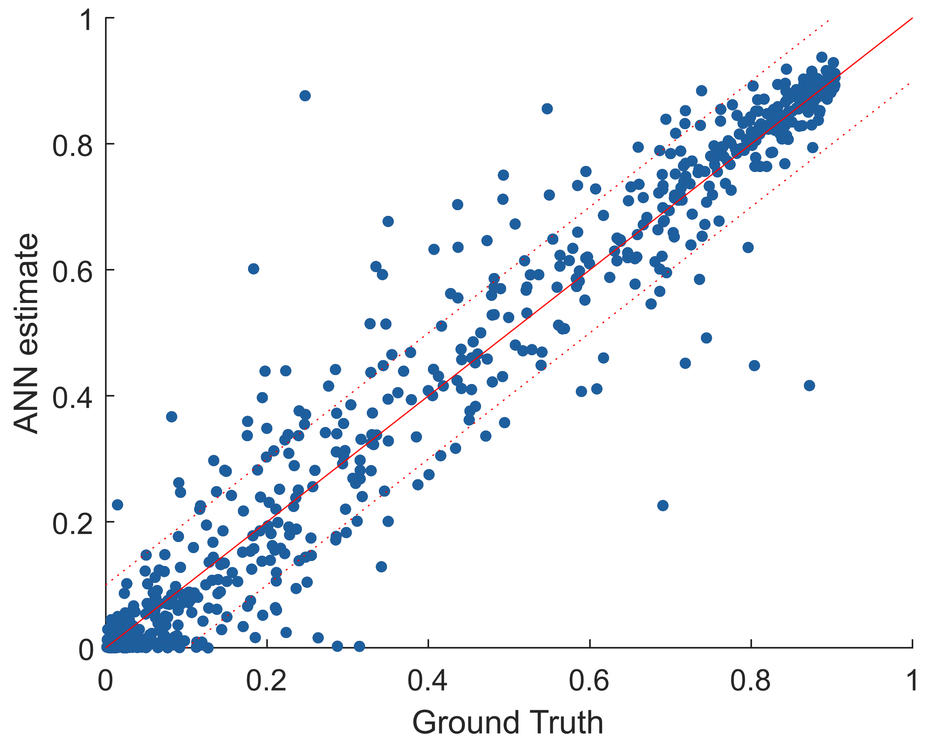

Figure 3. The estimated BiDWGP vs the true value for the test data set. The dashed red lines indicate ±0.1 of the correct value

Open source code

https://github.com/bbc/speech-intelligibility-meter/

Acknowledgements

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

References

[1] Tang, Y., Cooke, M., Fazenda, B. M., and Cox, T. J. (2016). A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers. Journal of the Acoustical Society of America, 140:1858–70.
[2] Tang, Y., Hughes, R.J., Fazenda, B.M. and Cox, T.J., 2016. Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms. Speech Communication, 82, pp.26-37.
[3] Buciluǎ, C., Caruana, R. and Niculescu-Mizil, A., 2006, August. Model compression. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 535-41). ACM.
[4] Hinton, G., Vinyals, O. and Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
[5] http://htk.eng.cam.ac.uk/