Exploiting deep learning to inform spectral contrast enhancement for hearing-impaired listeners

Ning Ma[1], *Guy J. Brown*[1], *Jon Barker*[1], *Michael Stone*[2]

[1]Department of Computer Science, University of Sheffield, UK
[2]School of Psychological Sciences, University of Manchester, UK
{n.ma, g.j.brown, j.p.baker}@sheffield.ac.uk, michael.stone@manchester.ac.uk

Abstract

Hearing-impaired (HI) listeners have great difficulty in understanding speech when there is noise, music, or people talking in the background. Despite of some advances that have been made in the field, current algorithms adopted in modern hearing aids have had limited success in improving speech intelligibility, as background noise is amplified along with people's voice of interest. Previously many studies [1, 2, 3, 4, 5] have focused on enhancing spectral contrast of sound in order to increase speech intelligibility in noise. These techniques typically employ a digital processing method for altering spectral contrast – the difference in amplitude between spectral peaks and valleys. However, the effects of such processing are usually small (for example, [4] report a relative improvement of 8% in intelligibility for one set of parameters only at a signal-to-masker ratio of -6 dB, but not higher).

One of the reasons that such methods are not very effective is that spectral contrast is enhanced without distinguishing target speech from background noise. Such blind processing does not necessarily improve auditory grouping, a likely process listeners use to organise sound mixtures into auditory streams belonging to individual sources. In some cases grouping cues may be degraded by such processing [4].

Recently, machine learning methods that utilise a deep neural network (DNN) have shown breakthrough performance in speech and language processing. DNNs can learn to identify the spectro-temporal regions in which speech dominates the noise (referred to as a "mask")[6, 7, 8]. The speech-dominating parts are amplified whereas those in which background noise dominates are discarded. Such noise-filtering methods have been shown improved speech intelligibility for hearing-impaired listeners [9, 10].

There are two reasons, however, to believe that filtering out all background noise may not be the optimal solution for improving speech intelligibility for hearing-aid users. First, improved signal-to-noise ratios do not necessarily translate into improved speech intelligibility. Identification of reliable speech regions is challenging in daily listening environments where a large and variable number of sound sources are present, and mislabelling of speech-dominating parts can often degrade speech intelligibility. Second, even if we could perfectly filter out the background noise, some of them are often desired to which listeners might want to switch attention. For example, when talking to someone at a train station, the listener might also want to pay attention to the announcement in the background.

In this study we look beyond traditional spectral contrast enhancement and propose an approach in which deep learning is used to inform spectral contrast enhancement. At first, a DNN based method is adopted to identify the spectro-temporal mask dominated by the target speech. Previous studies have used relatively simple DNN architectures and input features [9, 10]. The recent research conducted at Sheffield has demonstrated that incorporation of pitch-related features in a long short term memory (LSTM) network is capable of learning long-term dependencies which is particular effective when the background noise is not stationary as in daily listening environments [8]. In the subsequent spectral contrast enhancement, the time-frequency components belonging to a same source are processed coherently. This could include not just enhancing the spectral contrast for the target speech, but also reducing the spectral contrast. Such a method is analogous to reducing the depth of field of a lens in photography, thus emphasising the target subject while de-emphasising the background.

The aim of this study is to determine the benefit of the proposed deep learning methods on speech intelligibility for hearing-impaired listeners. The objectives include:

- To measure the benefit of improved deep learning methods for speech masks estimation on speech intelligibility for hearing-impaired listeners;
- To determine the effect of mask-informed spectral contrast enhancement.

We plan to conduct two sets of listening experiments. In the first set of experiments, the estimated probabilistic mask will be used to directly resynthesise enhanced signals, by weighting time-frequency (T-F) bins accordingly before overlap-adding signals over all frequencies. Three different DNNs will be used including the state-of-the-art baseline system proposed in [10] and two proposed systems [8]. In the second set of experiments, the best performing mask estimation method from the first experiment will be selected and used to inform spectral contrast enhancement. In this case spectral contrast enhancement is applied only to the T-F bins that are more likely to be dominated by speech. The other T-F bins are left intact so that the background sound are not completely filtered out. The results will be compared with those by direct resynthesis from the first set of experiments.

We will measure the effect of such processing on speech intelligibility by measuring the percent correct identification of keywords in sentences presented in both speech-shaped noise and a second talker. Two groups of listeners will be invited to take part in this experiment: normal-hearing listeners and hearing-impaired listeners. Two signal-to-masker ratios (SMRs) will be used for each set of tests.

Index Terms: spectral contrast enhancement, deep learning, hearing impairment, speech understanding in noise

1. References

[1] M. A. Stone and B. C. J. Moore, "Spectral feature enhancement for people with sensorineural hearing impairment: Effects on speech intelligi- bility and quality," *J. Rehab. Res. Devel.*, vol. 29, pp. 39–56, 1992.

[2] T. Baer, B. C. J. Moore, and S. Gatehouse, "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality and response times," *J. Rehab. Res. Devel.*, vol. 30, pp. 49–72, 1993.

[3] J. M. Alexander, R. L. Jenison, and K. R. Kluender, "Real-time contrast enhancement to improve speech recognition," *PLoS ONE*, vol. 6, no. 9, 2011.

[4] J. Chen, T. Baer, and B. C. J. Moore, "Effect of enhancement of spectral changes on speech intelligibility and clarity preferences for the hearing impaired," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2987–98, 2012.

[5] W. Nogueira, T. Rode, and A. Buchner, "Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants," *J. Acoust. Soc. Am.*, vol. 139, no. 2, pp. 728–739, 2016.

[6] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[7] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. ICASSP*, 2014, pp. 3709–3713.

[8] N. Ma, R. Marxer, J. Barker, and G. Brown, "Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 490–495.

[9] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029–3038, 2013.

[10] J. Chen, Y. Wang, S. Y. S.E., D. Wang, and E. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Am.*, vol. 139, pp. 2604–12, 2016.