# Speaker-informed speech enhancement and separation

*Bence Mark Halpern*[123], *Finnian Kelly*[4], *and Anil Alexander*[4]
[1]*University of Amsterdam, Amsterdam, The Netherlands*
[2]*Netherlands Cancer Institute, Amsterdam, The Netherlands*
[3]*Delft University of Technology, Delft, The Netherlands*
b.m.halpern@uva.nl
[4]*Oxford Wave Research, Oxford, UK*
{finnian|anil}@oxfordwaveresearch.com

In law-enforcement related audio recordings, particularly those made in crowded multi-speaker environments, it is often difficult to understand the speech of one specific speaker amongst all the others. Traditional speech enhancement approaches attempt to extract all speech signals from other non-speech background noise, and do not differentiate between the speech of different speakers. We propose a two-pronged approach using deep neural networks (DNN) to enhance speech intelligibility: one (speech enhancement) that generally attempts to improve the perceptual quality and intelligibility of speech in the presence of noise, and the other (speaker separation) that preferentially tries to improve the intelligibility of the speech from a specific speaker using a model of their voice.

The development of effective tools for speech enhancement and separation can greatly benefit investigations involving audio recordings, especially in the case of covertly-recorded speech. In this paper, we consider a joint approach to speech enhancement and separation that leverages information about the speaker of interest via a speaker embedding (a model of their voice). We use a DNN-based approach (specifically, a Convolutional Neural Network - Long Short-Term Memory framework (Wang et al. 2018)) for enhancement and separation, and explore the use of both x-vector (Snyder et al. 2018, Kelly et al. 2019) and d-vector (Wang et al. 2018) speaker embeddings for speaker separation.

For training and testing the proposed approach, a noisy signal is generated by mixing a clean and an interfering audio signal in an equal energy ratio (see Figure 1). The speech signals are taken from the LibriSpeech (Panayotov et al. 2015) corpus, and the noise signals from MUSAN (Snyder et al 2015.) and WHAM! (Wichern et al. 2019) (for training and testing, respectively). The enhancement and the separation performance are evaluated with two metrics, (1) the signal-to-distortion ratio (SDR), and (2) the word error rate (WER). These metrics are sensitive to different aspects of the performance: the SDR is suitable for the evaluation of perceptual quality, while the WER is suitable for assessing the intelligibility of the enhanced utterances. To calculate the WER, we use an end-to-end automatic speech recogniser (Watanabe et al 2018.).

Our initial results indicate that both types of speaker embeddings bring improvement in SDR and WER for both the enhancement and the separation tasks. For speech enhancement, the d-vector embedding achieves best performance, with an SDR improvement, and a reduction of WER from 56.8% for the noisy signal to 39% for the enhanced signal. For the speech separation task, the x-vector-LDA architecture performs best in terms of SDR (x-vector LDA: 4.58 dB vs d-vector: 4.24 dB). Both the x-vector and the d-vector provide significant improvement in the WER for the speaker separation task (measured with respect to the speaker of interest). The d-vector provides a reduction in WER from 118% for the mixed signal to 73.3% for the separated signal, and the x-vector a similar reduction in WER from 118% for the mixed signal to 82% for the separated signal (note that a WER of greater than 100% is possible due to the way transcription errors are counted).

These initial results are very encouraging and offer the possibility of leveraging automatic speaker recognition algorithms within a DNN-based speech enhancement context to obtain intelligible speech either in general or specifically for a given speaker of interest.

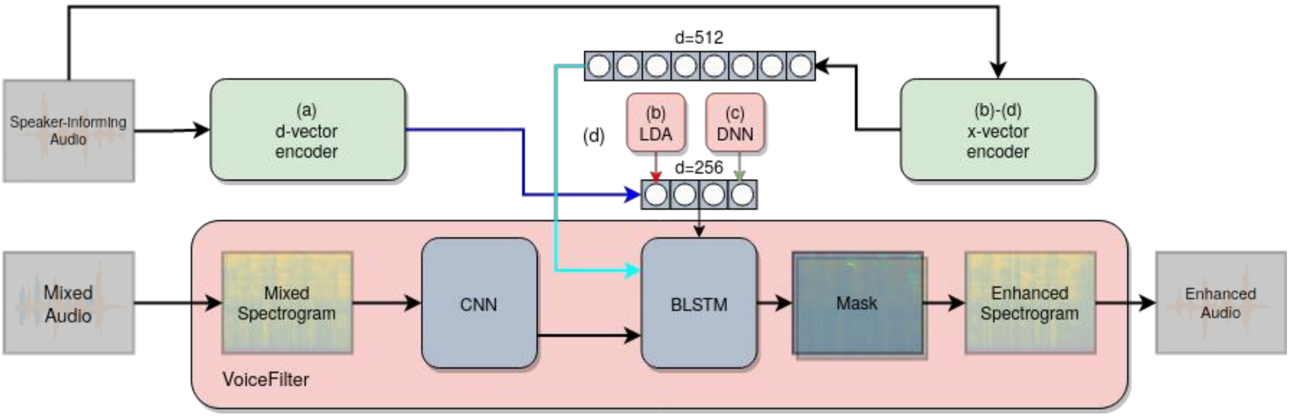

**Figure 1** Experimental setup: the red boxes show elements which are directly trained for the speech enhancement and speaker separation task. The green boxes represent the different encoding flavours for the speaker-informed enhancement considered in our experiments.

## References

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. In Audio Engineering Society Conference: 2019 AES International Conference on Audio Forensics. Audio Engineering Society.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015).Librispeech: an ASR corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5206-5210). IEEE.

Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5329-5333). IEEE.

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4879-4883). IEEE.

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J., Saurous, R.A., Weiss, R.J., Jia, Y., and Moreno, I.L. (2019). VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. Proc. Interspeech 2019, 2728-2732.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N.E.Y., Heymann, J., Wiesner, M., Chen, N., et al. (2018). ESPnet: End-to-End Speech Processing Toolkit. Proc. Interspeech 2018, 2207-2211.

Wichern, G., Antognini, J., Flynn, M., Zhu, L. R., McQuinn, E. Crow, D., Manilow, E., Le Roux, J. (2019). WHAM!: Extending Speech Separation to Noisy Environments. Proc. Interspeech 2019, 1368-1372.