Speaker-informed speech separation and enhancement

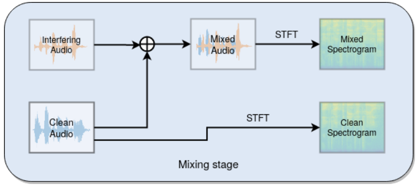
Bence Mark Halpern[1], Finnian Kelly[2] and Anil Alexander[2]
[1]University of Amsterdam, The Netherlands, [2]Oxford Wave Research Ltd., Oxford, U.K.

1. Motivation

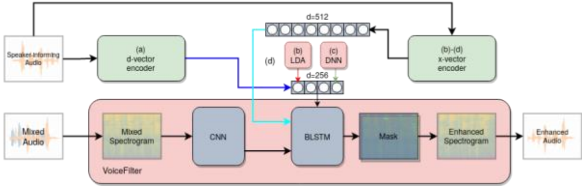
- Law-enforcement-related audio recordings are often made in noisy, multi-speaker environments: understanding the speech of a speaker of interest can therefore be difficult.
- Speech enhancement attempts to improve the intelligibility and perceptual quality of speech; traditional approaches attempt to extract speech signals from background noise, and do not give preference to the speech of one speaker over another.
- We propose a method to **(1)** preferentially extract the speech of a speaker of interest, and **(2)** perform speech enhancement in a single architecture.

2. Data and mixing

- **Speech data – LibriSpeech:** American-English read speech, clean conditions, 460 hours (train-clean-100 and train-clean-360 partitions).
- **Noise data - MUSAN:** Instantaneous noise dataset, 6 hours (noise partition), and **WHAM!**: Environmental noise dataset, 80 hours.
- **Speaker separation data:** mix two Librispeech samples from different speakers with an equal energy ratio, resulting in a total of 100,000 utterances (3–10 s duration).
- **Speech enhancement data:** mix one Librispeech sample and one noise sample with an equal energy ratio, using MUSAN for train and WHAM! for test, resulting in a total of 100,000 utterances (3–10 s duration).



3. Methods



- **Separation and enhancement approach** (based on VoiceFilter, Wang et al. 2018):
  - Feed a DNN with a **speaker-informing audio** sample alongside the **noisy (mixed) audio** sample
  - The DNN predicts a **mask** to reduce non-speaker-relevant spectral interference from the **mixed spectrogram**, resulting in an **enhanced spectrogram**
  - Inverse Fourier Transform the enhanced spectrogram to obtain the **enhanced audio**
- To represent the speaker-informing audio, we use **x-vector** embeddings (Snyder et al. 2018, Kelly et al. 2019) with three different architectural modifications **(LDA (b), DNN (c), LSTM (d))**, and compare these with the use of **d-vector** embeddings **(a)** (Wang et al. 2018).

4. Evaluation metrics

- The enhancement and the separation performance is evaluated with (1) the **signal-to-distortion ratio (SDR),** and (2) the **word error rate (WER)**.
- The SDR measures the perceptual quality of the enhanced speech, while the WER is used as proxy for the intelligibility of the enhanced speech.
- Calculation of the WER requires a speech transcription, for which we use an end-to-end automatic speech recognition system (Watanabe et al 2018.)
- After alignment of the decoded and ground truth sentences, the WER is calculated as below. Note that because there can be more insertion errors than words, the WER can more than 100%.

$$\mathrm{WER} = \frac{\text{insertion} + \text{substitution} + \text{deletion}}{\text{number of words}}$$

5. Results

- Our initial results indicate that both x-vectors and d-vectors improve SDR and WER for the enhancement and the separation tasks.
- In the speaker separation task, the use of x-vector LDA results in the best SDR
- The word transcripts are converted to 'phoneme' sequences using the mapping in CMUdict. The d-vector provides better phoneme recognition than the x-vector, with the exception of /ch/ and /oy/

	speaker separation		speech enhancement	
	SDR (dB)	ΔWER (%)	SDR (dB)	ΔWER (%)
d-vector	4.24	**-44.7**	**10.91**	**-17.8**
BLSTM	3.26	-30.2	10.09	-13.8
Linear	1.20	-21.4	10.73	-10.6
LDA	**4.58**	-36.0	10.09	-14.6
Clean	N/A	-91.6 (26.4)	N/A	-32.1 (24.7)

Initial results: A higher SDR and a lower ΔWER is better. The best results in each condition are in bold. For the clean results,the absolute WER is provided in parentheses.

Speaker separation

- Mixed:
- Speakers A and B:

Speech enhancement

- Mixed:
- Speaker A:

6. Conclusion

- The DNN VoiceFilter approach was successfully adapted to a speech enhancement task, demonstrating that enhancement and separation can be performed within a single model.
- The speech enhancement DNN generalises very well: in the training set, mostly instantaneous noises are present, however, the ambient noises present in the test set are readily suppressed.
- The use of x-vectors improves SDR performance on the speaker separation task compared to the original VoiceFilter d-vector representation.
- Certain 'phonemes' such as /ch/ and /oy/ are enhanced better with the use of x-vectors.
- Future work will aim to further improve the intelligibility of the speech by experimenting with alternative DNN loss functions (SDR, STOI, multi-scale FFT) and a 256-dimensional bottleneck x-vector, expand the phonetic analysis, and evaluate performance on real noisy data.

References

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.

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)

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

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.