DTU Compute
Department of Applied Mathematics and Computer Science

DTU

Speech separation in the waveform domain

Matěj Hoffmann (s221913), Chuansheng Liu (s212661), Sturla Njarðarson (s222676), Dominik Stiller (s221811)

Technical University of Denmark, Kgs. Lyngby, Denmark

1 Introduction

Separating mixed signals into their respective sources is a highly underdetermined problem in machine learning. In this project, we will adapt the Demucs model [1], initially created for stereo music, to speech separation. This corresponds to a mapping F from the true speaker utterances y_1, y_2 and ambient noise n to the estimated speaker utterances $\hat{y}_1, \hat{y}_2$:

$$F\{x\} = F\{y_1 + y_2 + n\} = \begin{cases} \hat{y}_1 \\ \hat{y}_2 \end{cases} .$$

From this perspective, our model has two goals: separating the two overlapping speech signals, and de-noising them.

2 Dataset

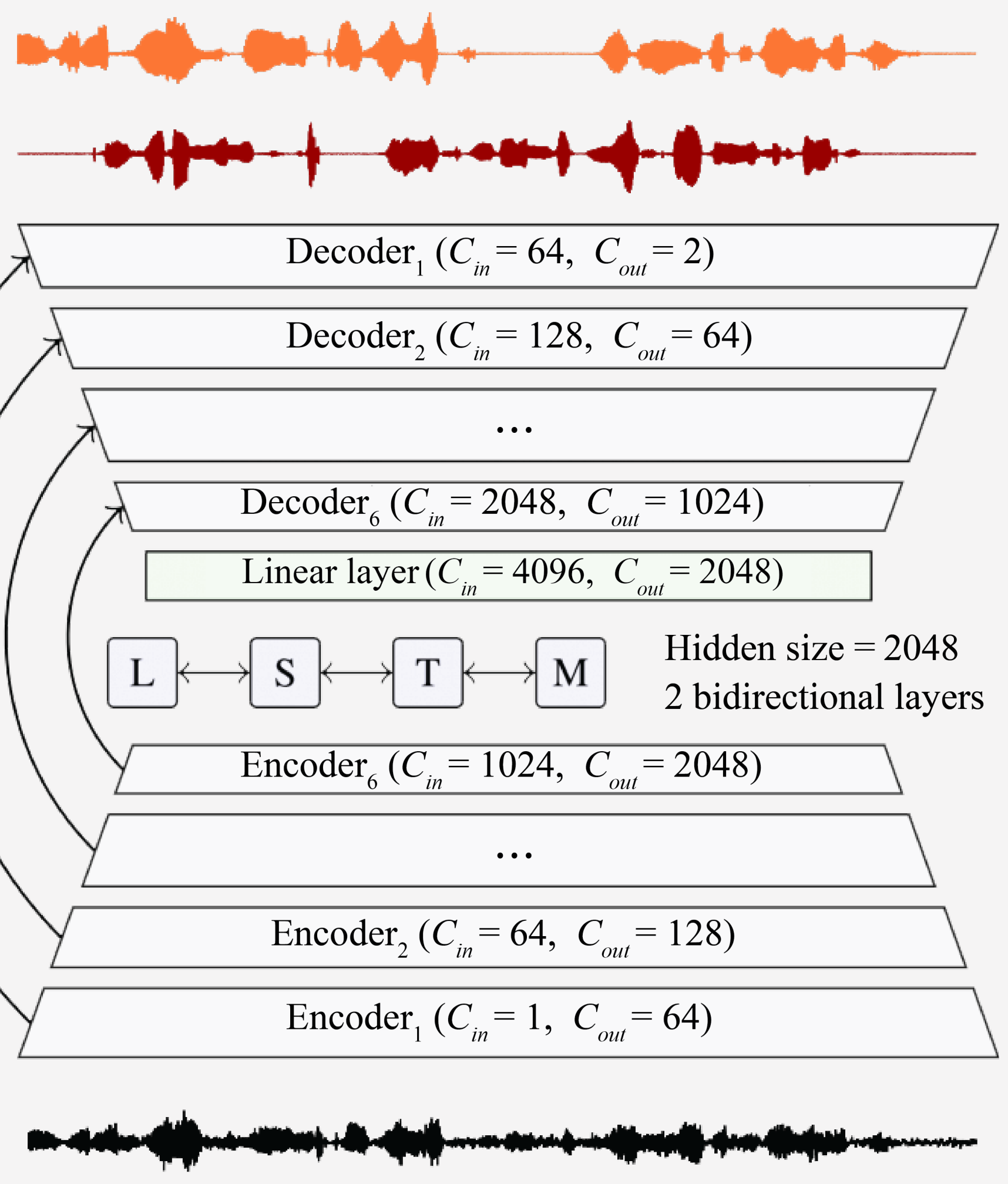
We use the LibriMix dataset [2], a speaker mixture corpus derived from audiobooks which provide clean ground truth speaker utterances y_1, y_2 overlayed with noise n. Statistics of test and validation splits are listed in Tab. 1.

Table 1: Training and validation dataset statistics. Examples from LibriMix dataset were filtered for length > 4 s and sampled at 8 kHz.

Set	Examples	Hours	Split
Train	61619	68.5	96.7%
Val	2096	2.33	3.29%

3 Model architecture

The model (Fig. 1) has a U-net structure with an LSTM at the bottleneck. Each of the 6 encoder blocks consists of a 1D convolution (size 8, stride 4, ReLU), followed by channel-doubling 1D convolution reduced to the original number of channels by GLU activation to learn attention-like masking on the input. The LSTM is bilinear with a hidden size of 2048 and followed by a linear layer to combine the forward and reverse activations. Each of the 6 decoder blocks consists of a 1D convolution (stride 1) with GLU masking and a variable size, which provides context about adjacent timesteps if larger than one, followed by a transpose 1D convolution (size 8, stride 4, ReLU) to successively construct the separated signals. The skip connections provide phase information. The input needs to be padded to a valid length for the convolutions. We train for 360 epochs and validate after every epoch, which takes 25.4 h on 2 Tesla A100 PCIE 40 GB.



Decoder$_1$ (C_{in} = 64, C_{out} = 2)
Decoder$_2$ (C_{in} = 128, C_{out} = 64)
...
Decoder$_6$ (C_{in} = 2048, C_{out} = 1024)
Linear layer (C_{in} = 4096, C_{out} = 2048)
L S T M
Hidden size = 2048
2 bidirectional layers
Encoder$_6$ (C_{in} = 1024, C_{out} = 2048)
...
Encoder$_2$ (C_{in} = 64, C_{out} = 128)
Encoder$_1$ (C_{in} = 1, C_{out} = 64)

Figure 1: Architecture of our model. We take mixed two-speaker utterances with noise and output two clean speech signals. Figure adapted from [1].

4 Results

We consider two metrics on our validation data to evaluate our results in addition to L1 loss minimization: scale-invariant signal-to-distortion ratio (SI-SDR) [3] and short-term objective intelligibility (STOI) [4]. The SI-SDR is a logarithmic quantity, and STOI takes values in [0,1]. For both metrics, higher values are better. The results are shown in Tab. 2.

Table 2: Effect of different batch sizes B and dropout probabilities p on model performance on the training and validation set, collected at end of training. L1 loss is given in $\times 10^{-3}$.

Param.		Train.	Val.		
B	p	L1	L1	SI-SDR	STOI
8	0	5.36	19.9	-7.11	0.488
64	0	4.91	18.2	-5.14	0.492
64	0.2	10.0	19.3	-6.44	0.463

A smaller batch size led to more stochastic behaviour of loss curves but did not change the final result. Dropout increased training loss as expected, but validation results did not improve.

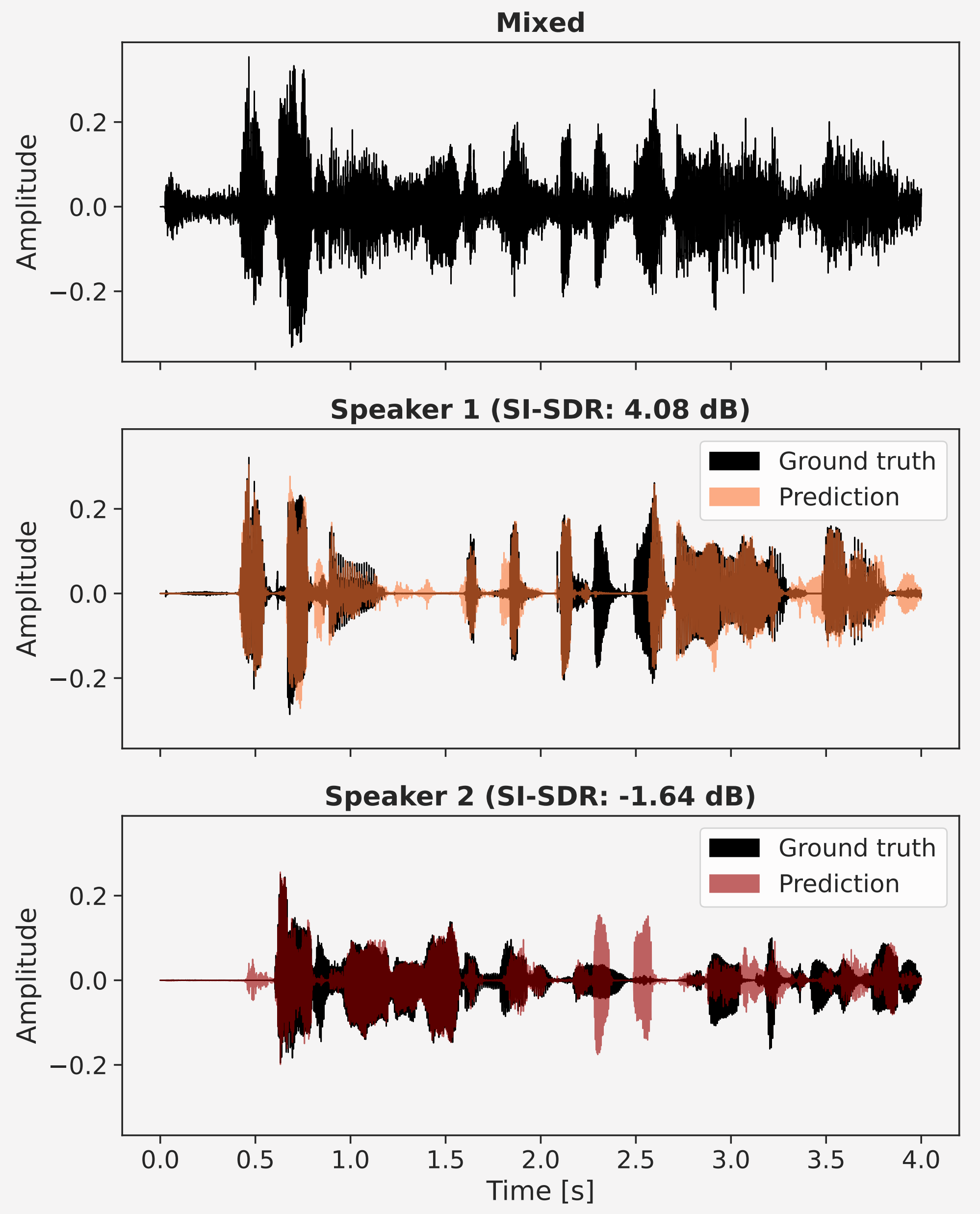


Figure 2: Prediction of a validation example. The top signal shows the input, followed by a comparison between ground truth and predicted signals. Confusion is present around 2.5 s.

5 Discussion

Overall, we had some success with our objective of separating two speech signals. The performance is best when both speakers have similar volume levels. Our model showed skill on the training data but did not generalize well, which manifested in confusion between speakers in validation predictions (Fig. 2). We could not resolve the confusion with wider contexts and more LSTM layers. Regularization with dropout, weight decay and smaller batch sizes did not improve validation performance. In the future, we want to explore more approaches to improving generalization and use three-speaker mixtures.

Code: https://github.com/DominikStiller/dtu-speechsep

References

[1] Alexandre Défossez, Nicolas Usunier, Lèon Bottou, and Francis Bach, "Music source separation in the waveform domain," 2019.

[2] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Librimix: An open-source dataset for generalizable speech separation," 2020.

[3] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR - half-baked or well done?," Nov. 2018, arXiv:1811.02508 [cs, eess].

[4] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, sep 2011.