Target Speaker ASR

Desh Raj

October 2, 2020

Motivation

- “Speech recognition was solved 15 years ago.”
- 1st place WER on CHiME-6: 36%
- Where is this dichotomy coming from?

Motivation

- Where is this dichotomy coming from?
- A lot of ASR research does hill-climbing on well-curated benchmarks.
- Example: 2% WER on Librispeech [1]
- Librispeech is clean single-speaker read speech

[1] Lüscher, Christoph et al. "RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation." *INTERSPEECH* (2019).

Motivation

- Most real world applications of ASR do not involve clean, single-speaker read speech
- Real data:
  - Noisy, reverberant
  - Conversational artifacts
  - Overlapping speakers

Motivation

- Most real world applications of ASR do not involve clean, single-speaker read speech
- Real data:
  - Noisy, reverberant
  - Conversational artifacts
  - **Overlapping speakers** We will look at some methods that seek to solve this problem

The Problem

- How to train a model which can recognize the outputs of 2 speakers speaking at the same time?
- Naive solution: Train a model to produce 2 outputs
- But how do we know which output corresponds to which speaker?

Frame-level permutation problem

Solution 1: PIT

- Permutation-invariant training [2]

- Compute the average loss for all input-output permutations and pick the one with the minimum.

- But how to have consistent output permutation across different utterances, i.e., Speaker A1 and B1 in utt1 vs Speaker A2 and B2 in utt2?

Utterance-level permutation problem

[2] Yu, Dong et al. “Recognizing Multi-talker Speech with Permutation Invariant Training.” *ArXiv* abs/1704.01985 (2017): n. pag.

Target-speaker ASR

- Recognize multi-speaker input one at a time.
- Network takes 2 inputs:
  - Multi-speaker audio
  - Target speaker information (i-vector etc.)
- Produces output corresponding to target speaker

Solves both frame-level and utterance-level permutation problems

Speaker Beam

Delcroix, M. et al. "Single Channel Target Speaker Extraction and Recognition with Speaker Beam." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018): 5554-5558.

Speaker Beam

- One implementation of Target Speaker ASR [3]
- Uses context adaptive DNN (CA-DNN)

2 ways of using adaptation

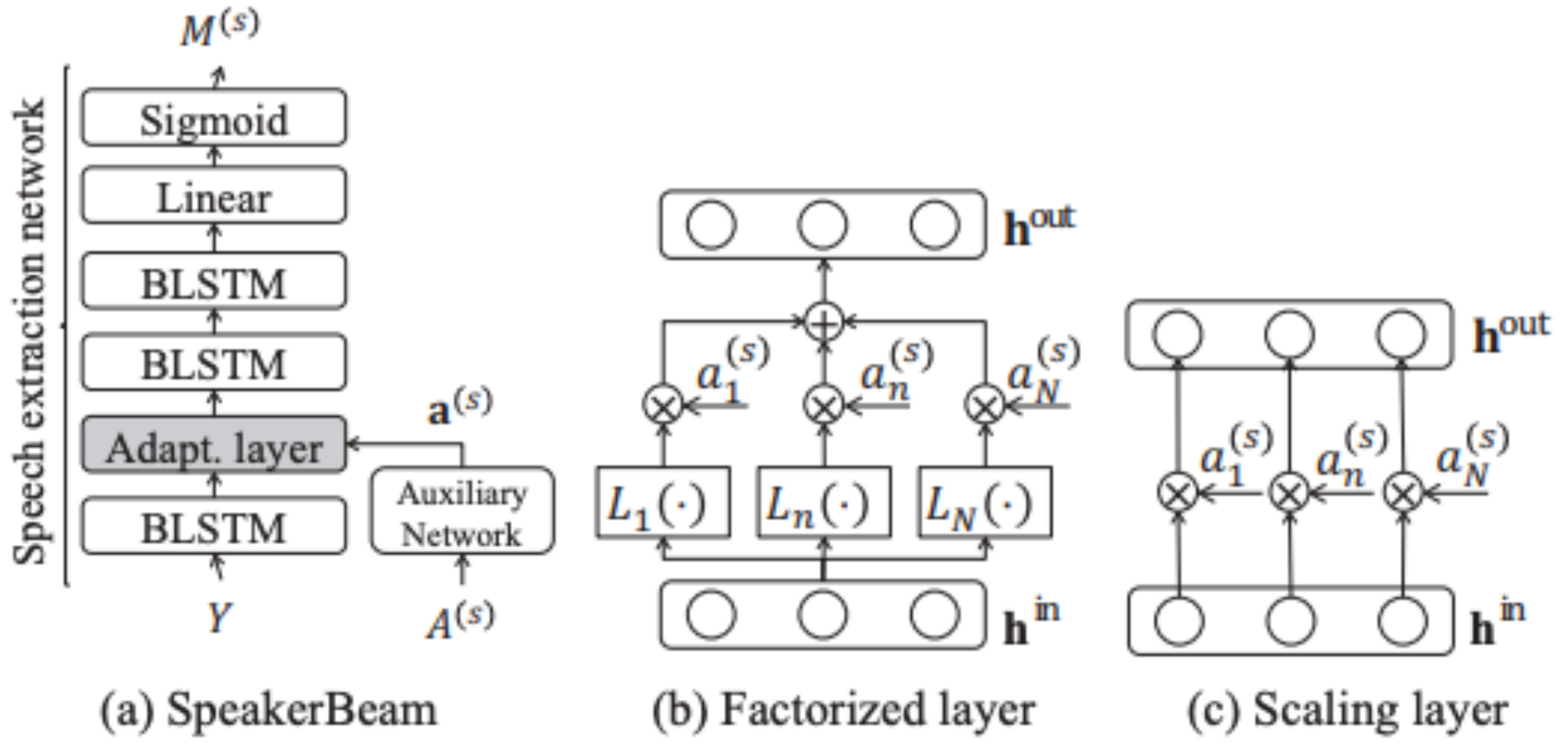


Fig. 1. Network architecture of SpeakerBeam. $A^{(s)} = \{\mathbf{a}_{t'}^{(s)}; t' = 1, \ldots, T'\}$ is the set of amplitude spectrum features of the adaptation utterance

Sequence summary network

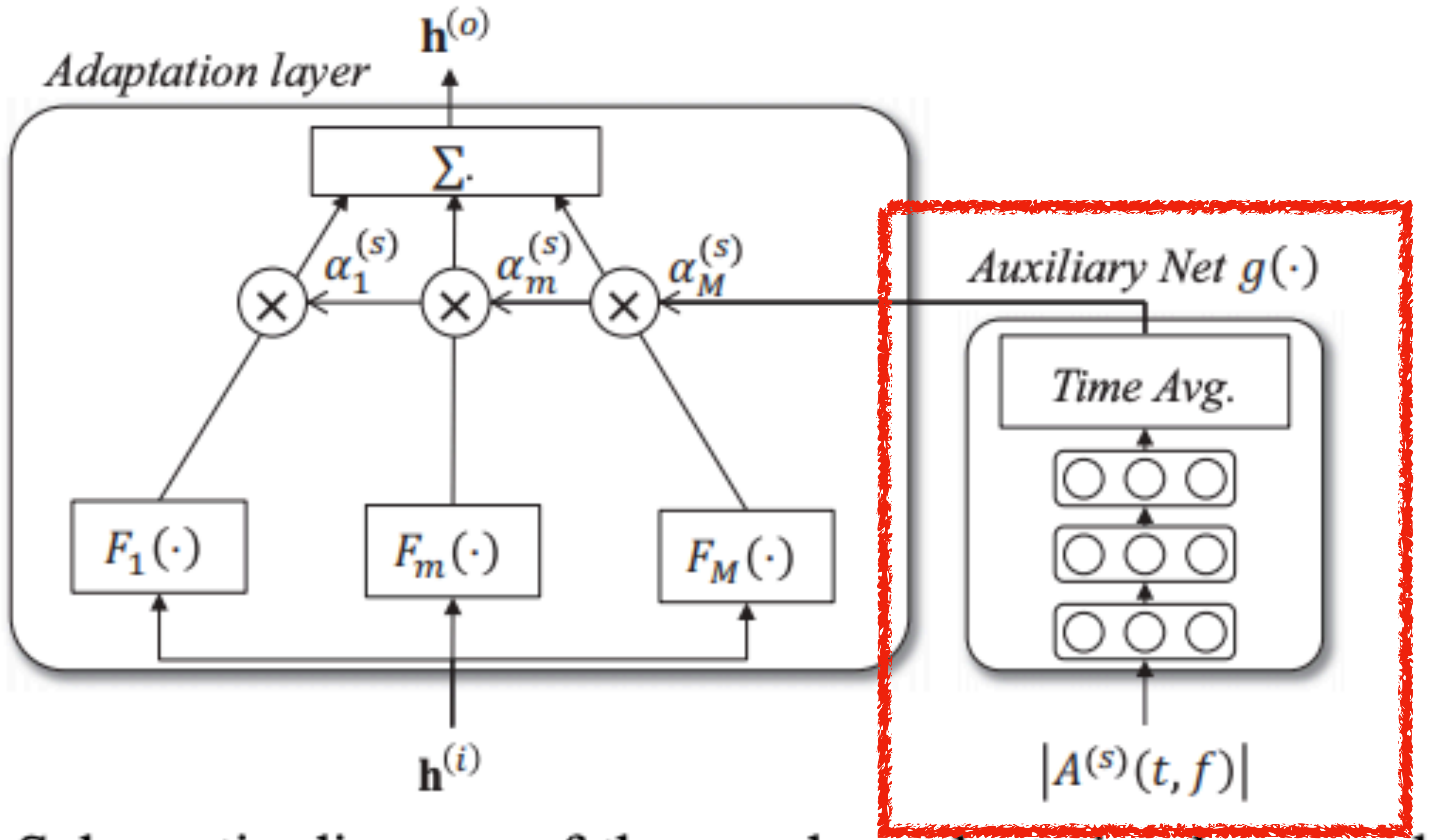
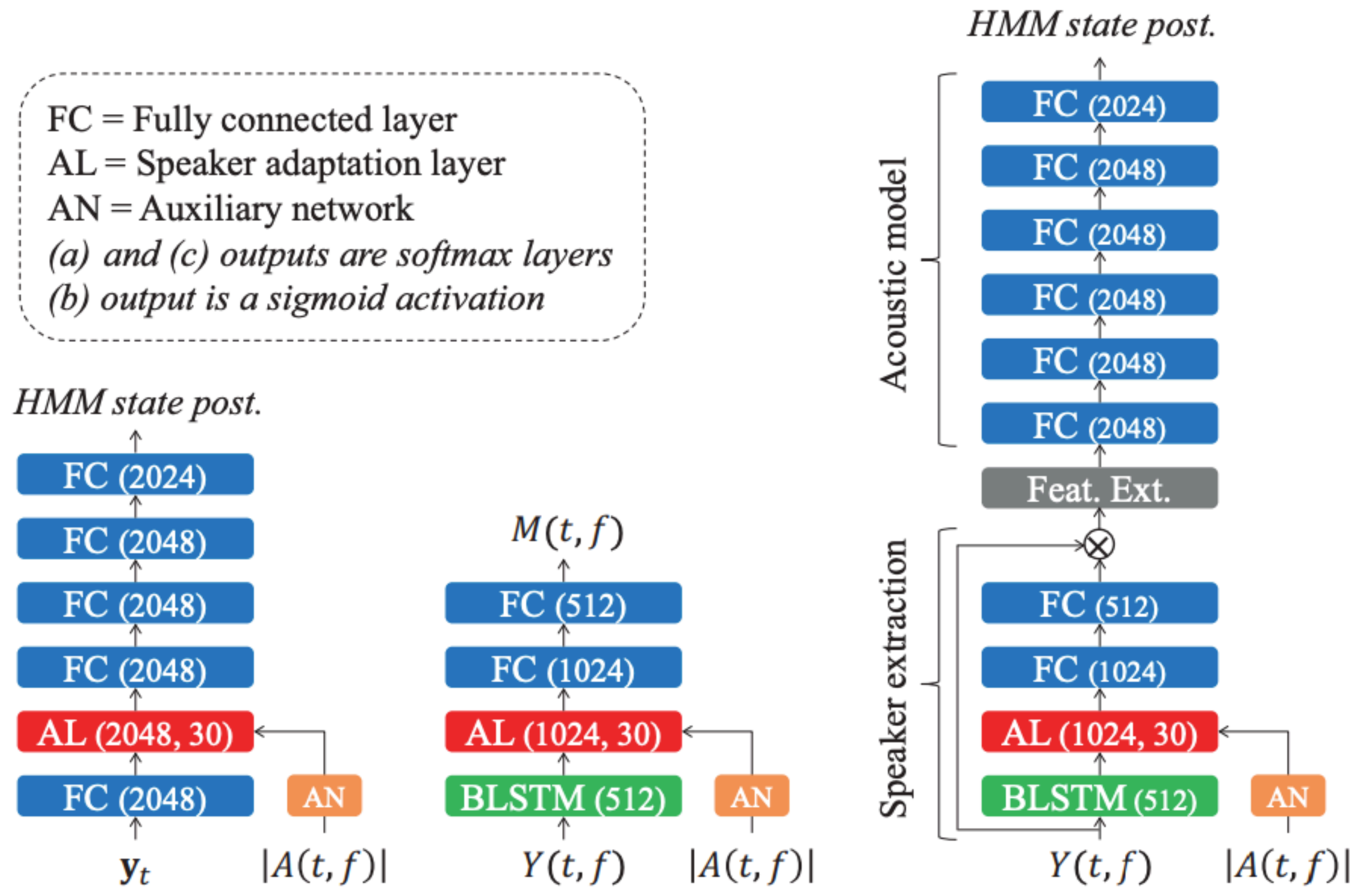


Fig. 1. Schematic diagram of the speaker adaptation layer and the sequence summary auxiliary network.

Sequence summary network

- Speaker adaptation layer
- S.S.N trained jointly with main network

Different training strategies



(a) SpeakerBeam-AM (b) SpeakerBeam-FE (c) SpeakerBeam-JT

Results on WSJ mixed

Table 1. WER as a function of the input SIRs for the eval set. WER a single speaker recognized with the baseline AM was 4.1 %.

	0dB	5dB	10dB	15dB	20dB
Mixture w/ baseline AM	95.7	70.4	40.3	14.0	5.9
Auxiliary input AM	85.2	72.6	66.5	70.5	76.8
SpeakerBeam-AM	45.8	28.3	20.3	18.1	17.3
SpeakerBeam-FE	54.5	39.7	32.8	30.0	29.2
SpeakerBeam-JT	34.0	17.5	9.8	7.5	6.5

Speaker Beam papers

- Delcroix, M. et al. "Improving Speaker Discrimination of Target Speech Extraction With **Time-Domain Speakerbeam**." ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020): 691-695.

- Ochiai, Tsubasa et al. "**Multimodal SpeakerBeam**: Single Channel Target Speech Extraction with Audio-Visual Speaker Clues." INTERSPEECH (2019).

- Delcroix, M. et al. "**End-to-End SpeakerBeam** for Single Channel Target Speech Recognition." INTERSPEECH (2019).

- Delcroix, M. et al. "**Compact Network** for Speakerbeam Target Speaker Extraction." ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019): 6965-6969.

Voice Filter

Wang, Q., Hannah Muckenhirn, K. Wilson, P. Sridhar, Zelin Wu, J. Hershey, R. A. Saurous, Ron J. Weiss, Ye Jia and I. Lopez-Moreno. “VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking.” *INTERSPEECH* (2019).

Wang, Q., I. Lopez-Moreno, M. Saglam, K. Wilson, Alan Chiao, Renjie Liu, Y. He, Wei Li, J. Pelecanos, Marily Nika and A. Gruenstein. “VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition.” *ArXiv* abs/2009.04323 (2020): n. pag.

VoiceFilter pipeline

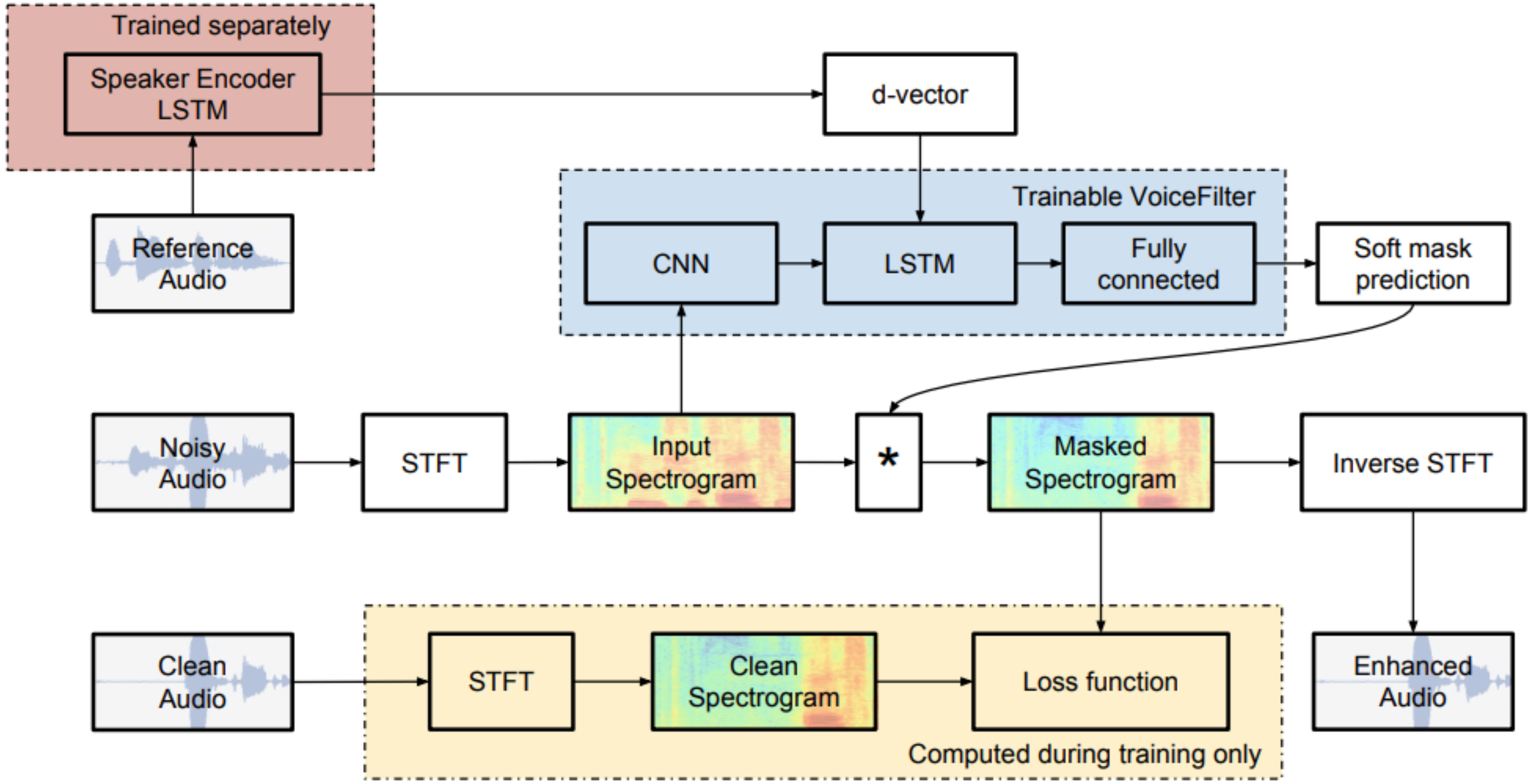


Figure 1: *System architecture.*

WER Results on Librispeech

Table 2: *Speech recognition WER on LibriSpeech. VoiceFilter is trained on LibriSpeech.*

VoiceFilter Model	**Clean WER (%)**	**Noisy WER (%)**
No VoiceFilter	10.9	55.9
VoiceFilter: no LSTM	12.2	35.3
VoiceFilter: LSTM	12.2	28.2
VoiceFilter: bi-LSTM	**11.1**	**23.4**

VoiceFilter-Lite pipeline

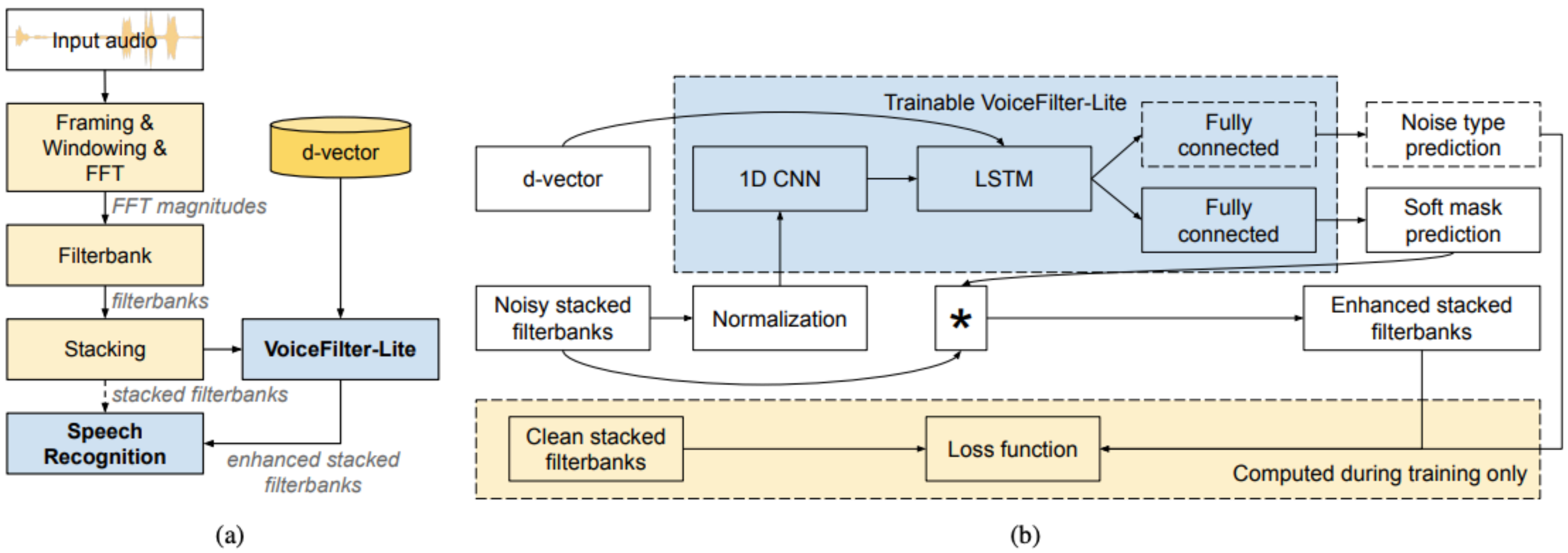


Figure 1: *VoiceFilter-Lite architecture, assuming using stacked filterbank energies as inputs and outputs. (a) Integration with ASR. The dashed arrow indicates the original connection without VoiceFilter-Lite. (b) Neural network topology of the VoiceFilter-Lite model.*

How to prevent degradation on clean speech

- Motivation:
  - modern ASR systems are already-noise robust
  - Don’t want to make performance on clean data worse
  - **Over-suppression problem**

Asymmetric loss

$$L = \sum_t \sum_f \Big(S_{\mathrm{cln}}(t, f) - S_{\mathrm{enh}}(t, f) \Big)^2.$$

+

$$g_{\mathrm{asym}}(x, \alpha) = \begin{cases} x & \text{if } x \leqslant 0; \\ \alpha \cdot x & \text{if } x > 0. \end{cases}$$

=

$$L_{\mathrm{asym}} = \sum_t \sum_f \Big(g_{\mathrm{asym}}\big(S_{\mathrm{cln}}(t, f) - S_{\mathrm{enh}}(t, f), \alpha\big) \Big)^2.$$

Adaptive suppression strength

$$S_{\text{out}}^{(t)} = w \cdot S_{\text{enh}}^{(t)} + (1 - w) \cdot S_{\text{in}}^{(t)}.$$

Obtained from the Noise type prediction branch

Results on Librispeech

Table 1: *WER (%) for VoiceFilter-Lite models. ASR is trained and evaluated on LibriSpeech.*

Feature	Loss	Suppression strength	Clean	Non-speech noise		Speech noise		Size
				Additive	Reverb	Additive	Reverb	
No voice filtering			8.6	35.7	58.5	77.9	79.3	N/A
FFT magnitude	L2	$w = 1.0$	9.1	21.5	48.3	25.5	54.2	6.8 MB
	asym L2, $\alpha = 10$	$w = 1.0$	8.8	24.1	50.8	35.5	60.6	
Filterbank	L2	$w = 1.0$	9.3	23.4	48.9	25.4	55.6	5.8 MB
	asym L2, $\alpha = 10$	$w = 1.0$	8.6	24.8	49.8	30.6	58.4	
Stacked filterbank	L2	$w = 1.0$	8.9	22.2	48.2	23.5	53.7	6.8 MB
	asym L2, $\alpha = 10$	$w = 1.0$	8.8	23.9	49.7	30.6	57.8	
		$w = 0.6$	8.6	24.4	50.7	42.0	60.2	

Key takeaways

- In presence of interference (in the form of noise or other speakers), target speaker information can be used to guide the neural network
- Different implementations:
  - Speaker Beam
  - Voice Filter

Thank You