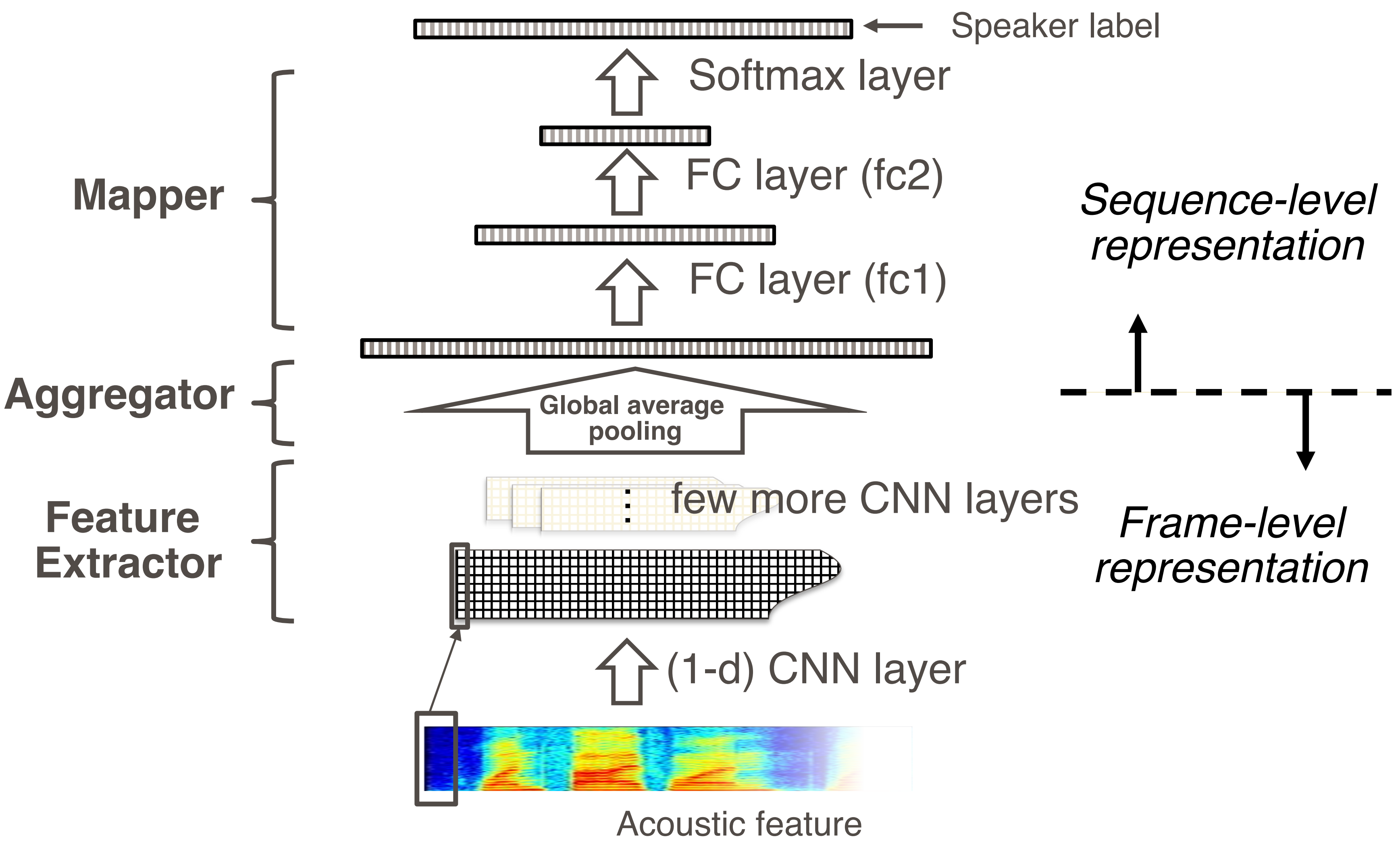
VoiceID Loss :
Speech Enhancement for Speaker Verification

Suwon Shon, Hao Tang, James Glass

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA, USA

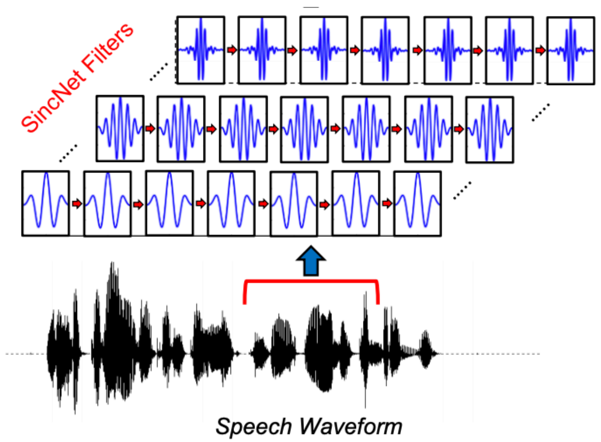
General model based on CNN



Advances in speaker recognition

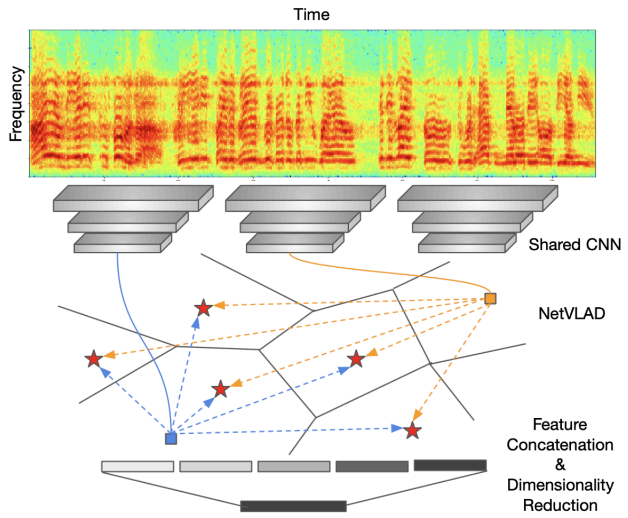
- **Recent studies**

Feature Extractor	Aggregator	Mapper
SLT2018	**ICASSP 2019**	**SP Letter 2018**



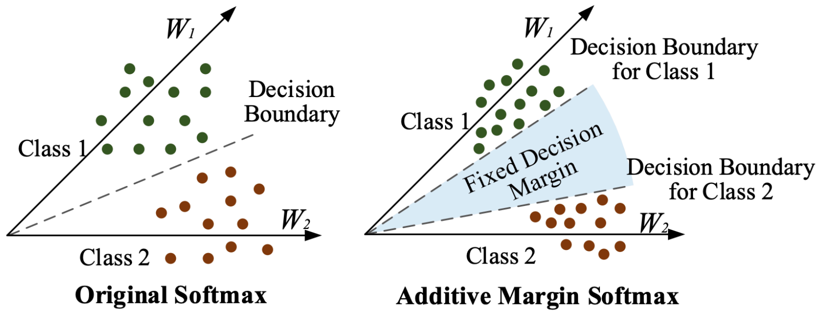
Speaker Recognition from Raw Waveform with SincNet

Ravanelli and Bengio



Utterance-level aggregation for speaker recognition in the wild

Xie, Nagrani, Chung and Zisserman



Additive Margin Softmax for Face Verification

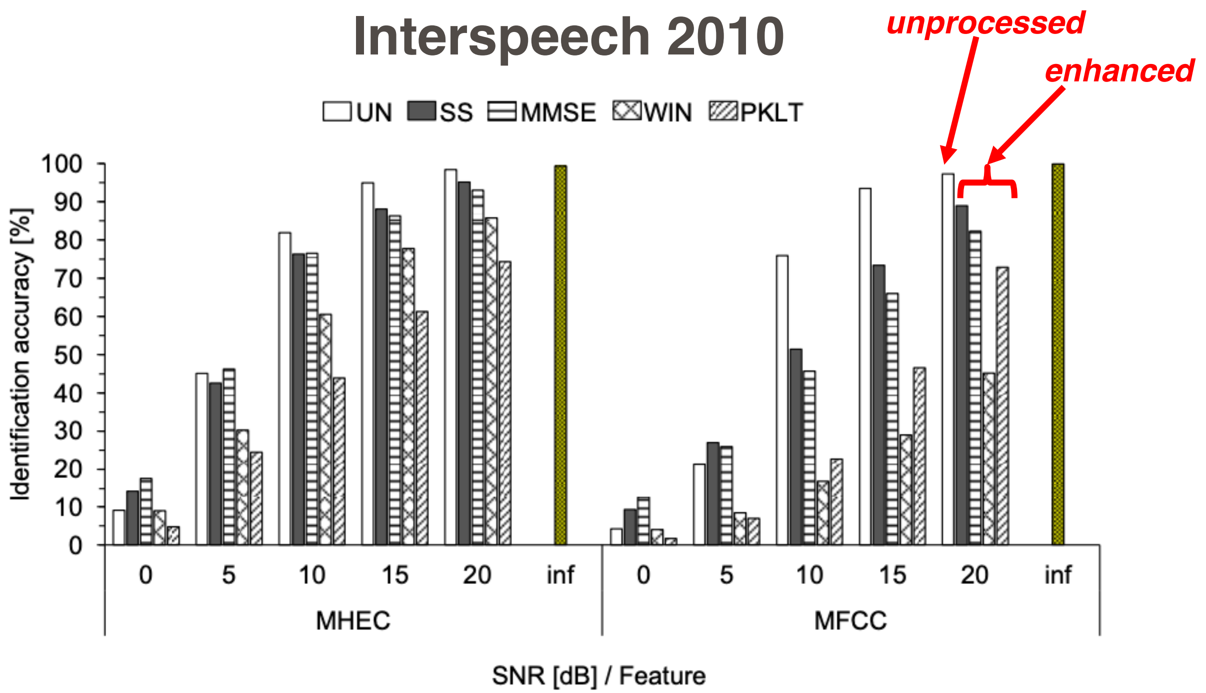
Wang, Cheng, Liu and Liu

Lack of study under noisy condition

- **Most of studies tested on clean or mild noise condition**
- **However, still vulnerable on distant, noise and reverberation**
- **Very few studies of speech enhancement on speaker recognition**
  - Sadjadi and Hansen, *Interspeech* 2010
  - Plchot et al, *ICASSP* 2016
- **Why so few?**
  - Artifacts and distortion make speaker recognition worse

Lack of study under noisy condition

Interspeech 2010



Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions

Sadjadi and Hansen

ICASSP 2016

	PLDA trained on **clean** data					
	Original data			Enhanced data		
Condition	DCF^{min}_{new}	DCF^{min}_{old}	EER	DCF^{min}_{new}	DCF^{min}_{old}	EER
tel-tel	0.372	0.108	2.07	0.370	0.109	2.18
prism,noi	0.415	0.126	2.94	0.364	0.099	2.28
prism,rev	0.408	0.108	2.07	0.224	0.059	1.37
int-int	0.310	0.077	1.74	0.251	0.064	1.68
int-mic	0.244	0.053	1.09	0.216	0.046	1.04
prism,chn	0.307	0.048	0.79	0.178	0.021	0.47

Audio enhancing with DNN autoencoder for speaker recognition

Plchot, Burget, Aronowitz and Matejka

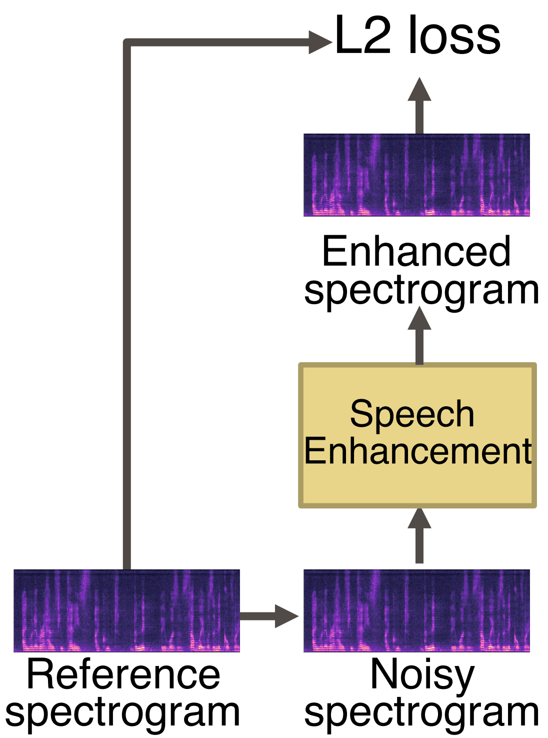
Lack of study under noisy condition

- **Most of studies tested on clean or mild noise condition**
- **However, still vulnerable on distant, noise and reverberation**
- **Very few studies of speech enhancement on speaker recognition**
  - Sadjadi and Hansen, *Interspeech* 2010
  - Plchot et al, *ICASSP* 2016
- **Why so few?**
  - Artifacts and distortion make speaker recognition worse

Let’s expose the downstream task on speech enhancement!

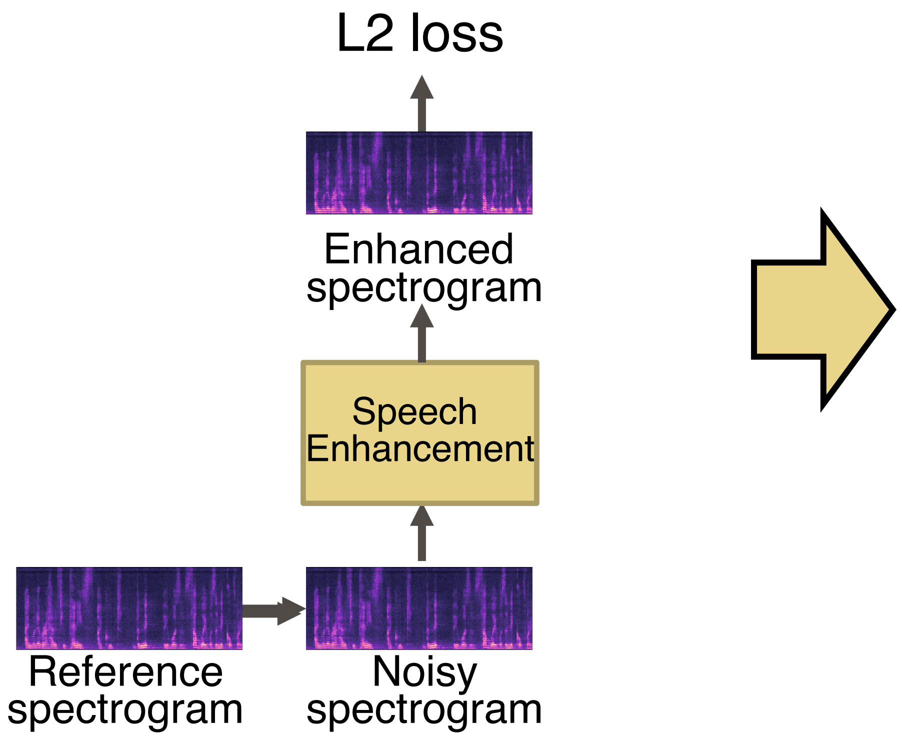
Speech enhancement

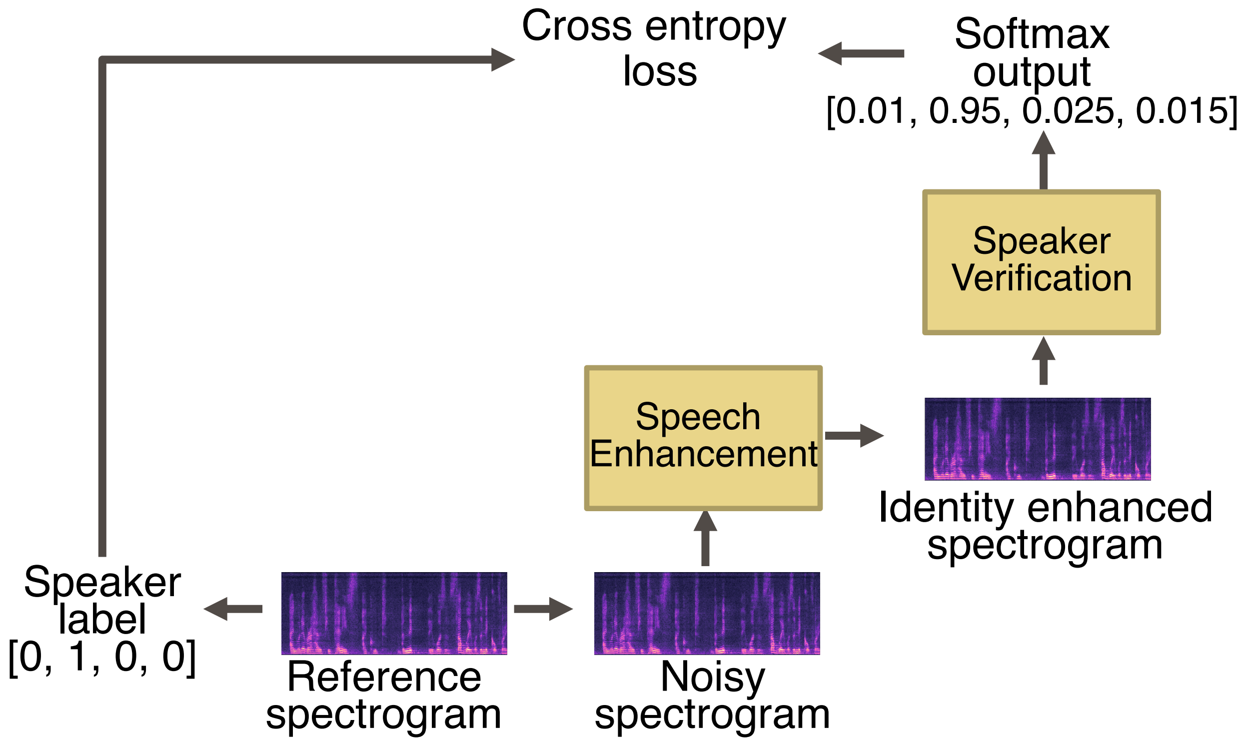
- **Objective : reconstructing original signal from noisy input**
- **Denoising Autoencoder (DAE) structure with L2 loss**



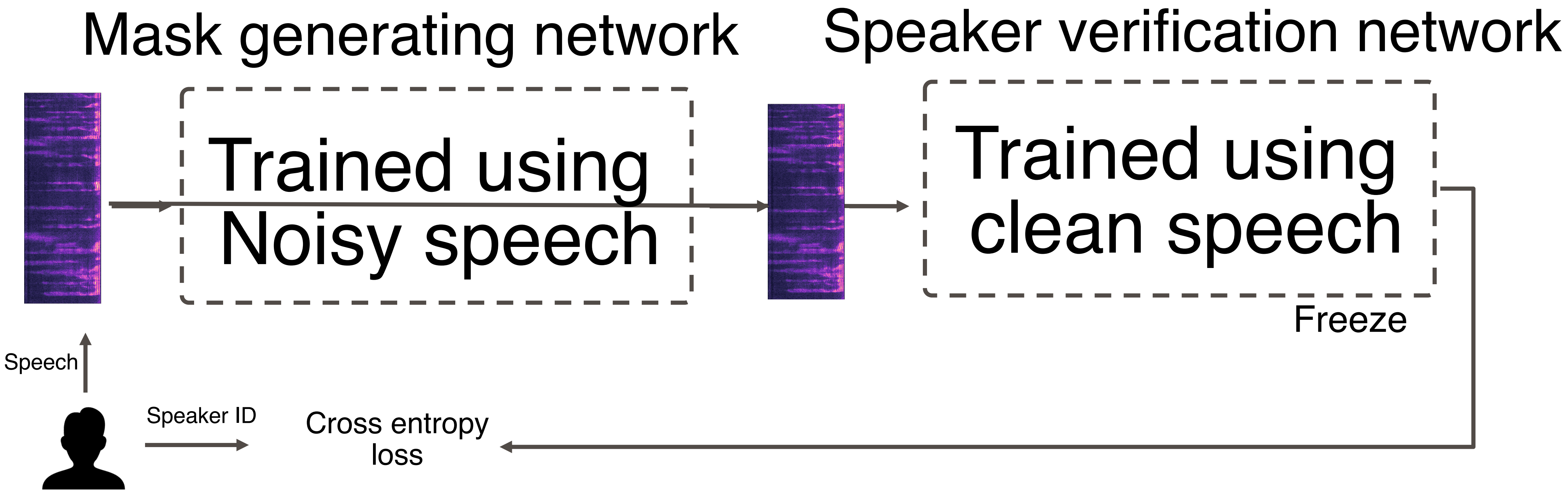
Speech enhancement ~~on~~ *for* speaker recognition

- Objective : ~~Reconstructing original signal from noisy input~~
  Improving verification performance

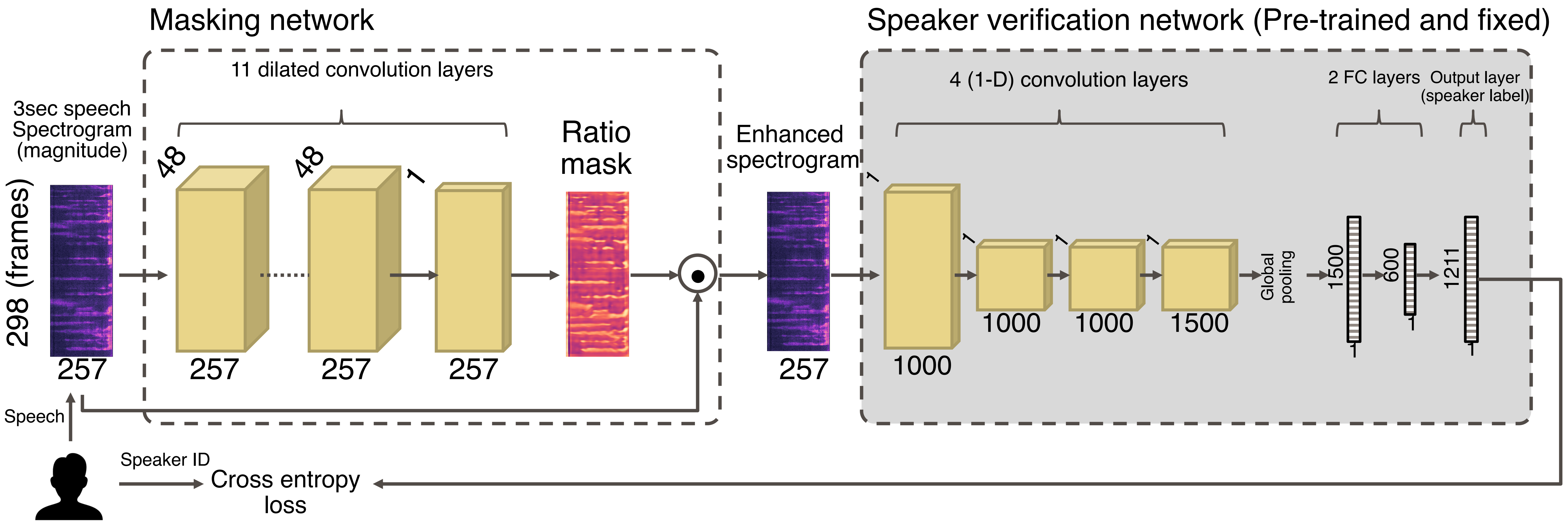




Proposed structure



Proposed structure (detail)



Experiments

- **Implementation**
  - Voxceleb1 dataset : Dev set for training, Test set for evaluation
  - MUSAN for noise augmentation and noisy test set
  - Conducted on two speaker verification model
    * Voxceleb1 dev
    * Voxceleb1 dev + noise augmented Voxceleb1 dev
  - Masking network was trained using noise augmented Voxceleb1
  - Test set was augmented with noise (SNR 0~20)
  - Use magnitude of spectrogram as input, linear scale, power-law compressed with 0.3
  - Using 3sec input for training (298frames)
  - (noisy phase was used only to reconstruct waveform for demo)
  - *DAE used for comparison (8-layer TDNN,1000 hidden units per layer)*

Tested on Voxceleb1-test

(a) Original test set

	EER (%)
SV	7.73
Proposed → Mask+SV	**6.99**
DAE*+SV	7.73

(b) Music (SNR=0dB)

	EER (%)
SV	29.03
Mask+SV	**18.89**
DAE+SV	20.41

(c) Babble (SNR=0dB)

	EER (%)
SV	44.64
Mask+SV	**42.20**
DAE+SV	43.30

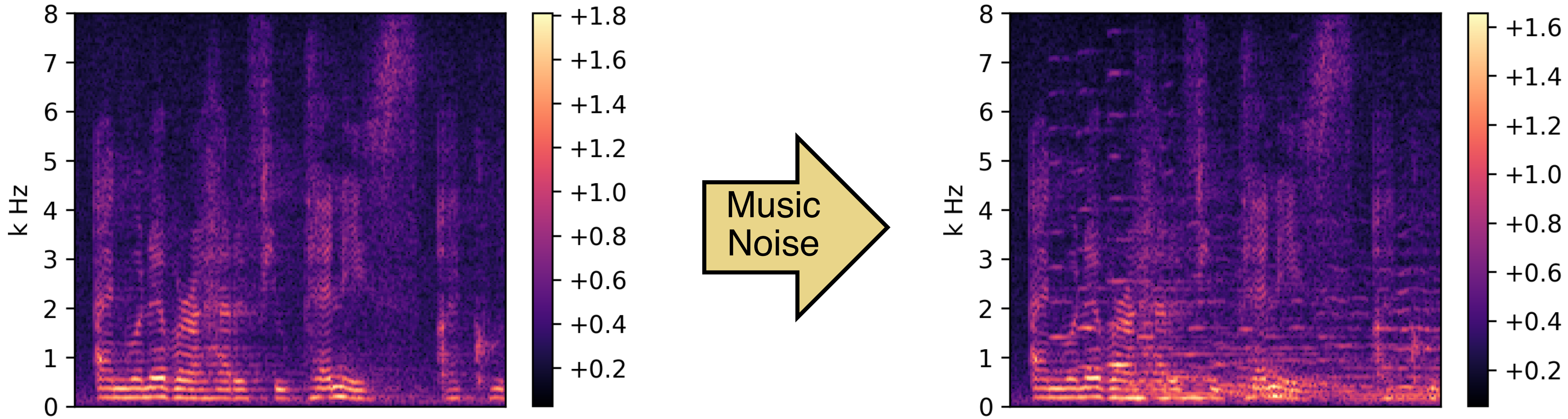
(d) Reverb (small room)

	EER (%)
SV	13.81
Mask+SV	**10.02**
DAE+SV	13.54

*8-layer time-delay neural network (TDNN)
1000 hidden units per layer
Context size is 25 frames

Experiments

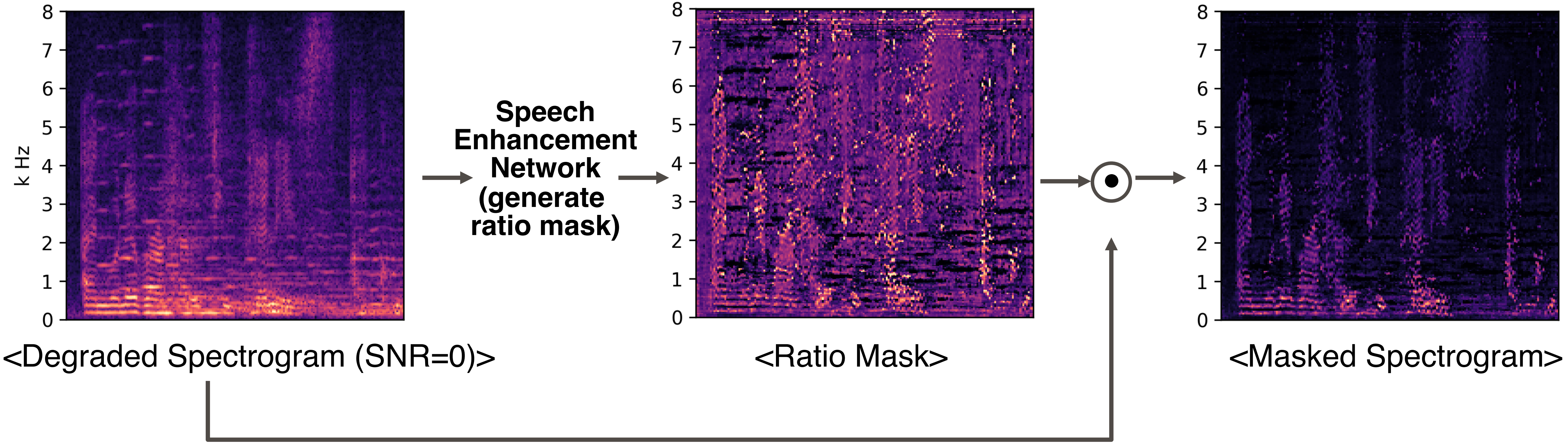
- **Spectrogram samples (degrading)**



<Original Spectrogram>   <Degraded Spectrogram (SNR=0)>

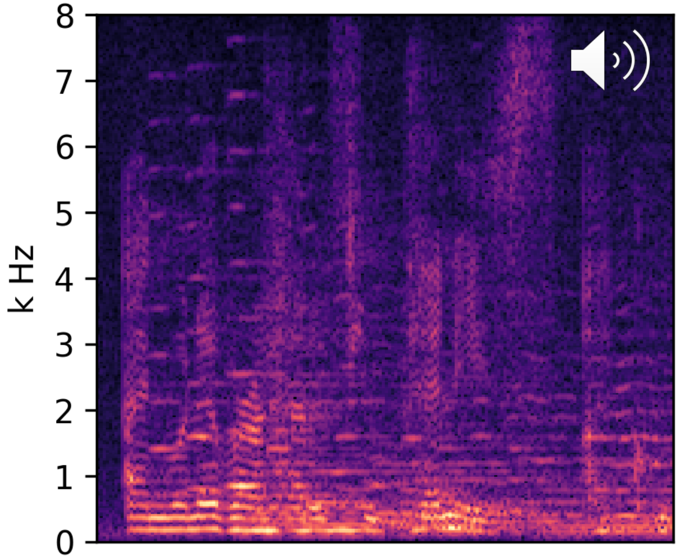
Experiments

- **Spectrogram samples (enhancement)**

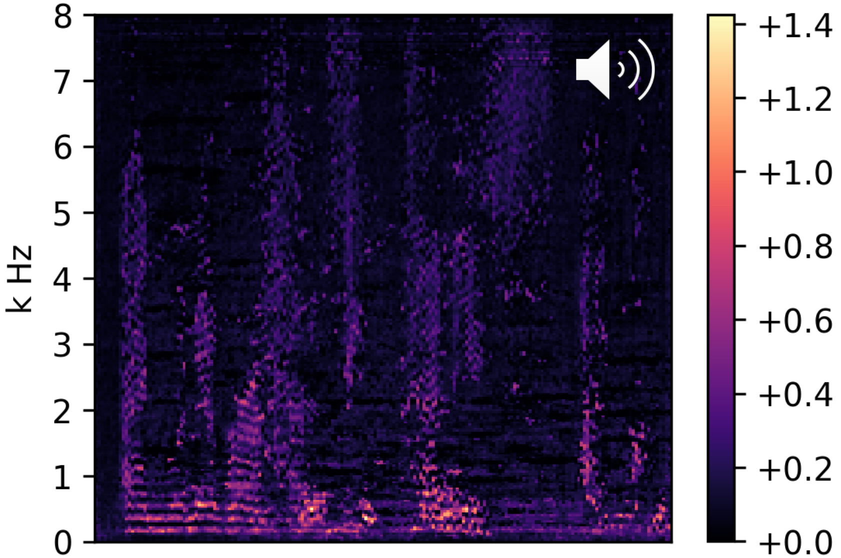


Experiments

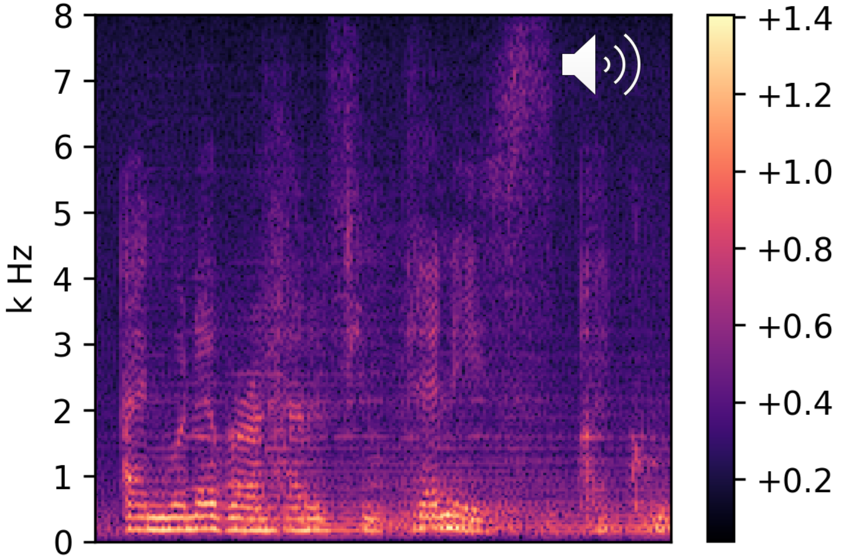
- **Wav samples**



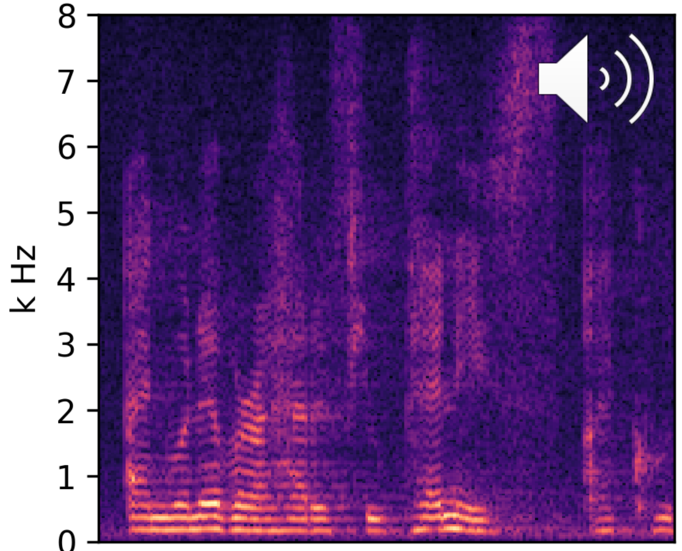
<Degraded Spectrogram (SNR=0)>



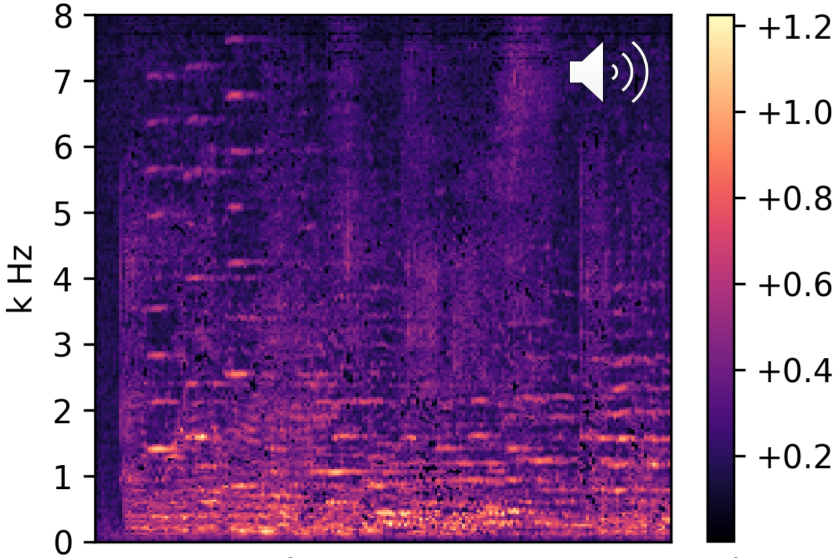
<Proposed (Masked)>



<DAE result>



<Original>



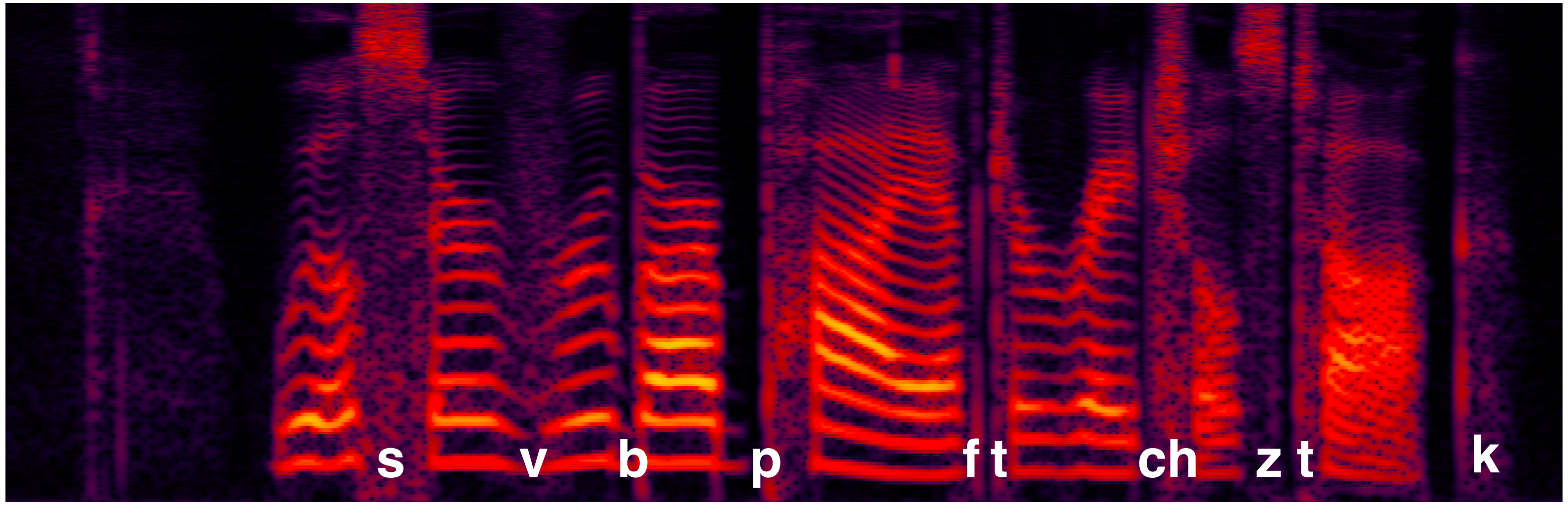
<Residue (Degraded-masked)>

Experiments

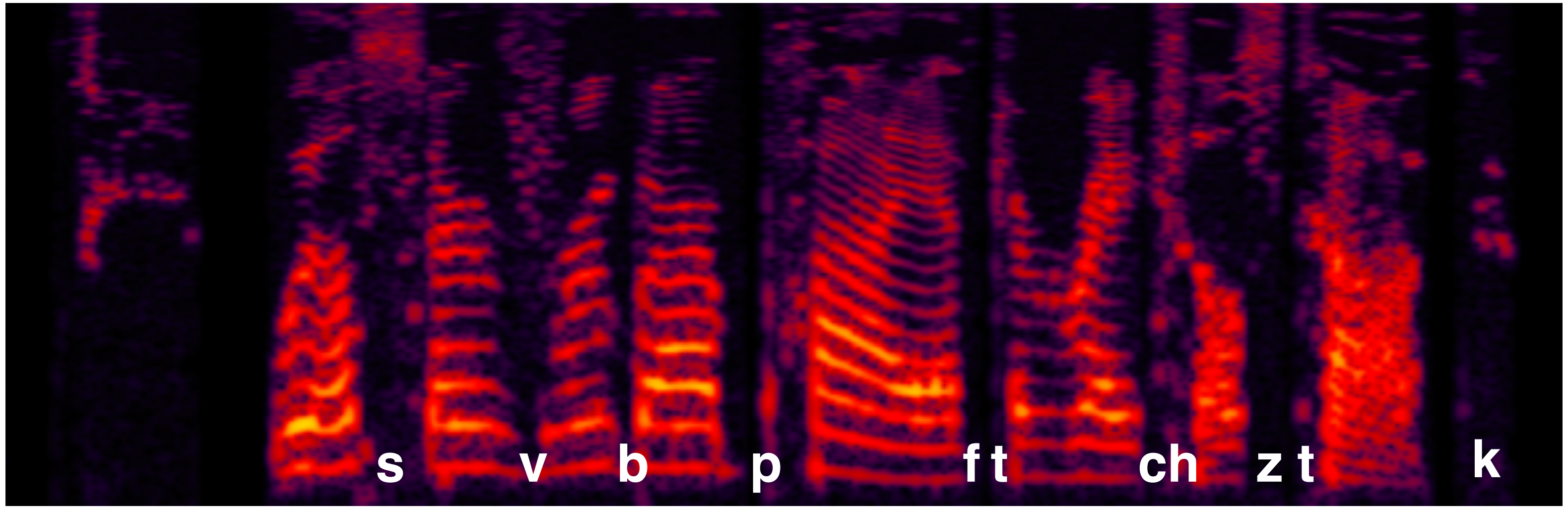
- **Spectrogram samples from TIMIT**

We'll serve rhubarb pie after rachel's talk

Original

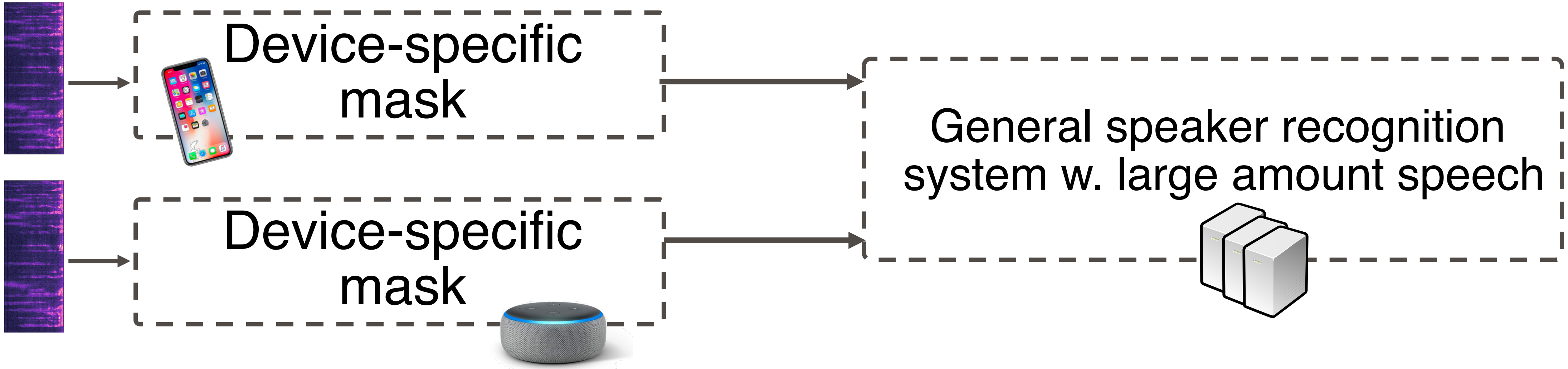


Enhanced
(masked)



Conclusion

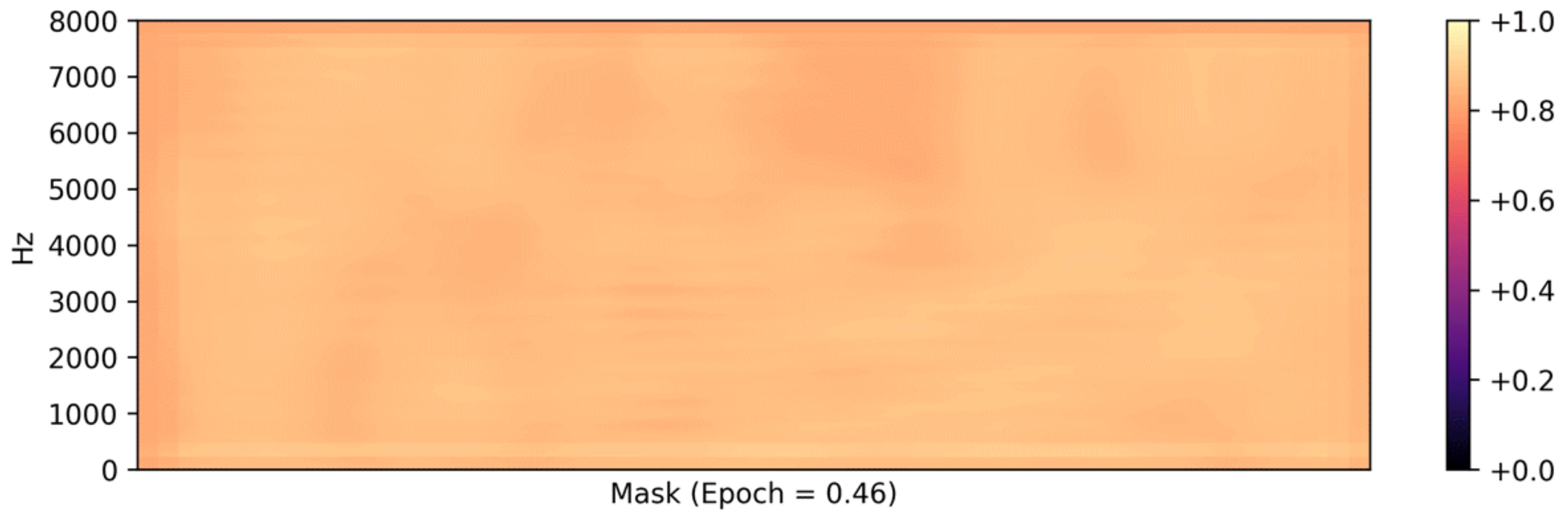
- First speech enhancement attempt only for text-independent speaker verification
- Only use speaker label for speech enhancement
- Speech enhancement for multi-condition scenario



Thank you

Check more samples ->

people.csail.mit.edu/swshon/supplement/voiceid-loss



Appendix

Experiments

Verification model trained with only clean set

Verification model trained with clean and noisy set

Verification network		Using original set $\mathcal{D}$						Using original and augmented set $\mathcal{D}$ +$\mathcal{D}^{\mathcal{N}}$					
Enhancement		-		Proposed		DAE				Proposed		DAE	
Enhancement network		-		Using $\mathcal{D}^{\mathcal{N}}$		Using $\mathcal{D}$+$\mathcal{D}^{\mathcal{N}}$				Using $\mathcal{D}$+$\mathcal{D}^{\mathcal{N}}$		Using $\mathcal{D}$+$\mathcal{D}^{\mathcal{N}}$	
Type	SNR	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Original test set $\mathcal{T}$		7.73	0.608	**6.99**	**0.590**	7.73	0.608	7.01	0.592	**6.79**	**0.574**	6.93	0.589
Noise	20	10.34	0.761	**8.10**	**0.675**	10.02	0.738	8.08	0.659	**7.83**	**0.639**	8.28	0.671
	15	13.05	0.909	**9.32**	**0.699**	11.45	0.833	8.99	0.720	**8.69**	**0.686**	8.96	0.761
	10	17.71	0.987	**11.24**	**0.770**	14.00	0.943	10.36	0.770	**9.86**	**0.747**	10.73	0.869
	5	24.34	0.999	**14.78**	**0.885**	18.01	0.988	12.90	0.851	**12.26**	**0.830**	13.51	0.958
	0	31.76	1.000	**20.82**	**0.983**	23.87	0.998	17.68	0.945	**16.56**	**0.938**	18.32	0.994
Music	20	8.97	0.710	**7.54**	**0.666**	9.32	0.714	7.73	0.670	**7.48**	**0.635**	7.82	0.651
	15	10.60	0.764	**8.23**	**0.715**	10.27	0.743	8.43	0.695	**8.10**	**0.677**	8.42	0.692
	10	14.10	0.883	**9.72**	**0.760**	11.75	0.808	9.73	0.760	**9.13**	0.733	9.54	**0.728**
	5	20.37	0.992	**13.00**	**0.819**	15.15	0.941	12.28	0.833	**11.44**	**0.818**	11.76	0.846
	0	29.03	1.000	**18.89**	**0.937**	20.41	0.993	17.45	0.935	16.24	**0.913**	**15.96**	0.961
Babble	20	12.87	0.837	**10.16**	**0.781**	11.34	0.778	9.17	0.725	**8.99**	**0.705**	9.55	0.723
	15	18.83	0.931	**13.50**	**0.864**	14.45	0.881	11.68	**0.793**	**11.25**	0.807	12.10	0.801
	10	28.78	0.991	**21.18**	**0.944**	21.37	0.969	17.38	**0.922**	**16.66**	0.926	17.41	0.941
	5	38.74	1.000	**33.39**	**0.996**	33.14	0.997	28.21	**0.992**	**27.12**	0.996	29.19	**0.992**
	0	44.64	1.000	**42.20**	1.000	43.30	**0.999**	38.72	1.000	**37.96**	1.000	41.11	**0.999**
Reverb	Small room	13.81	0.835	**10.02**	**0.744**	13.54	0.831	10.52	0.725	**9.94**	**0.708**	11.52	0.814
	Large room	13.74	0.825	**10.11**	**0.756**	14.09	0.999	10.64	0.724	**10.17**	**0.691**	11.47	0.792