The important ”feature” for speaker recognition

Pengqi Li

2022/05/27

Introduction

- Not all position in speech features(MFCC, Fbank, Spectrogram) contribute equally to the speaker recognition system performance.
- Analysis from CAMs.
- Can this "important" information be known in advance? The most intuitive downstream task is to improve the noise robustness of the model

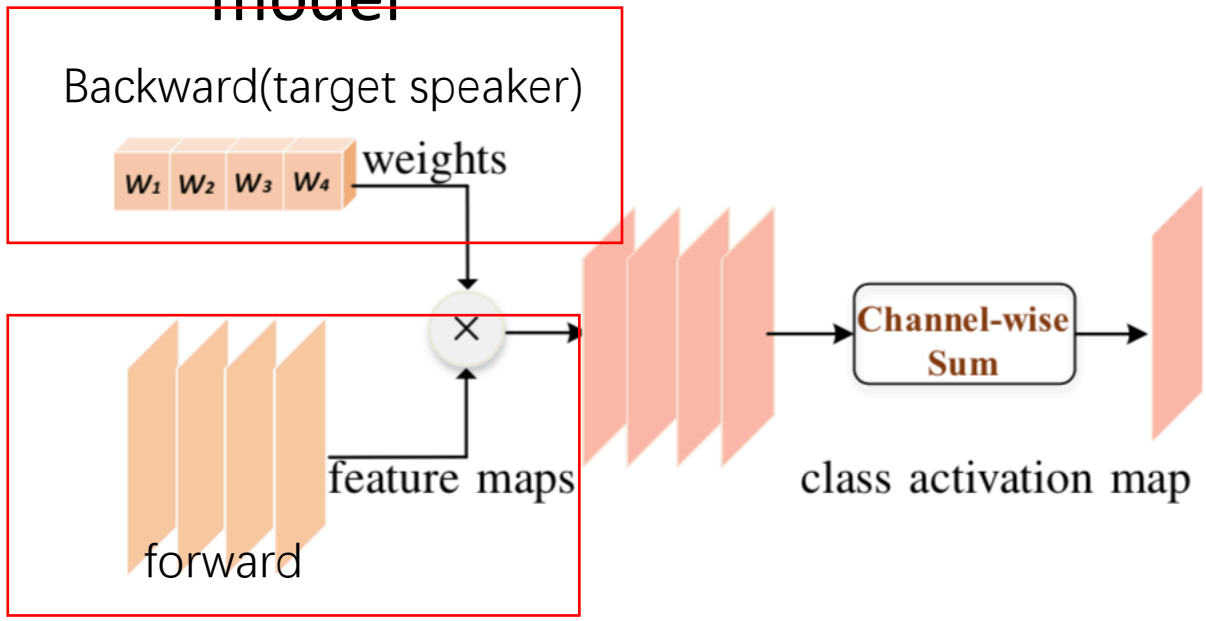


Fig. 3. The process of class activation mapping methods [1], [2], [3].

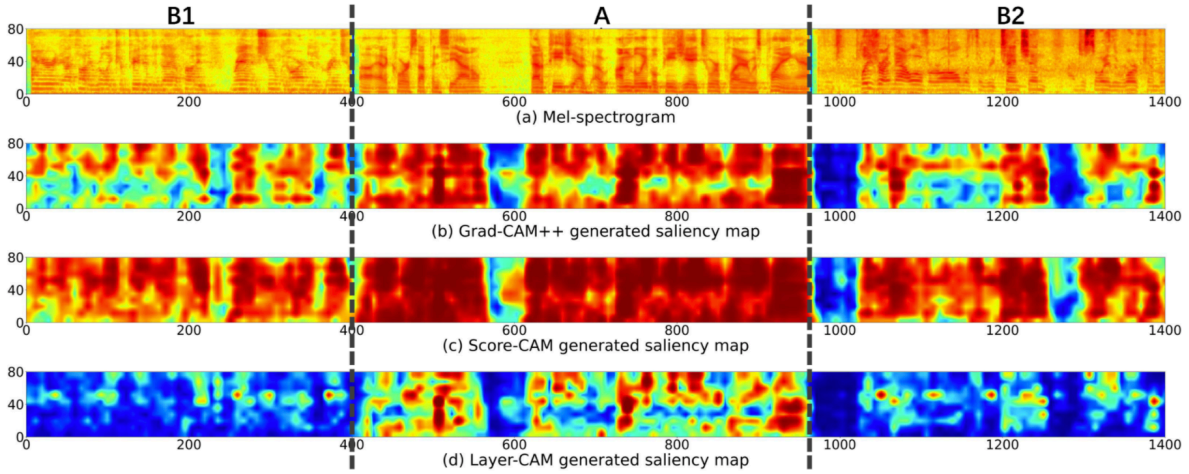


Figure 3: *Saliency maps on a 'B-A-B' test example.*

Robust Speaker Recognition Using Speech Enhancement And Attention Model

*Yanpei Shi**, *Qiang Huang**, *Thomas Hain*

Speech and Hearing Research Group
Department of Computer Science, University of Sheffield
{YShi30, qiang.huang, t.hain}@sheffield.ac.uk

Yanpei Shi, Qiang Huang, and Thomas Hain, "Robust speaker recognition using speech enhancement and attention model," arXiv preprint arXiv:2001.05031, 2020.

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Motivation
  - To increase the robustness against noise.
  - **Highlight** the speaker related features
- Methods
  - speech enhancement and speaker recognition are integrated into one framework by a joint optimisation using deep neural networks.
  - multi-stage attention mechanism(MS)

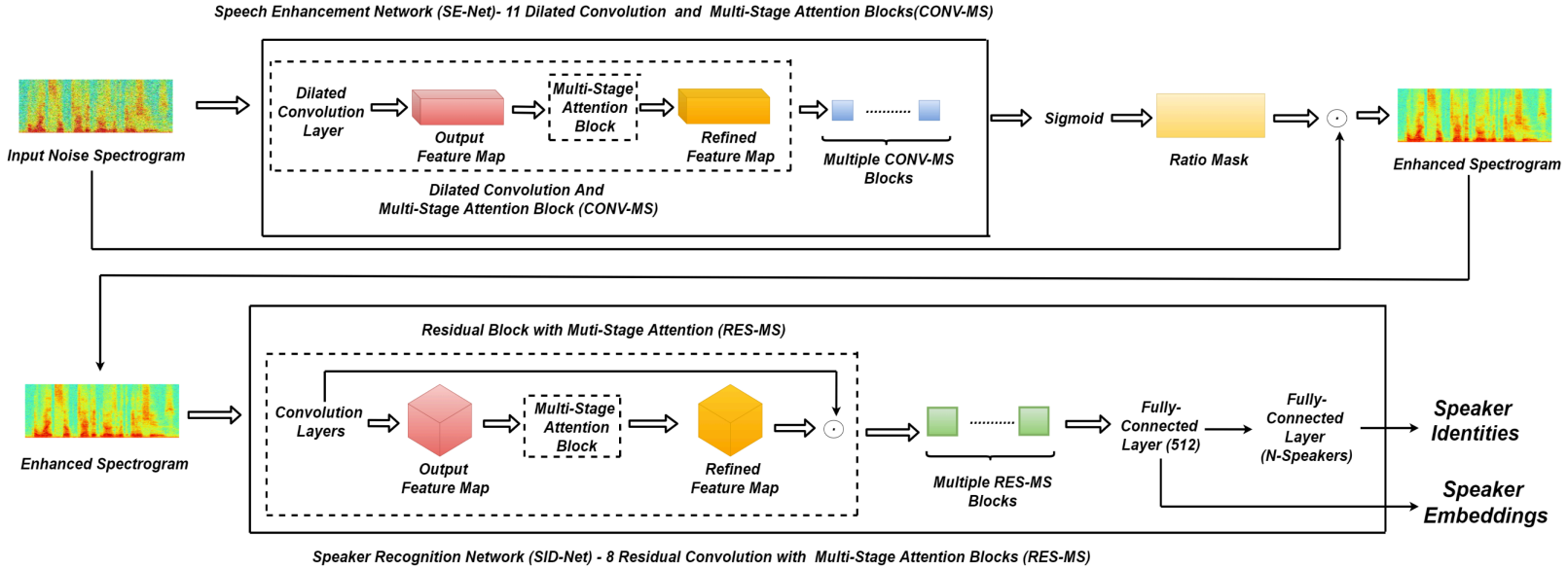


Figure 1: Architecture of proposed approach by cascading speech enhancement and speaker recognition. SE-Net denotes the speech enhancement network with taking noise spectrogram as input and consisting of 11 dilated convolution and multi-stage attention (CONV-MS) blocks. SID-Net denotes the speaker recognition network, with taking the enhanced spectrogram as input and consisting of 8

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Methods
  - multi-stage attention mechanism(MS)

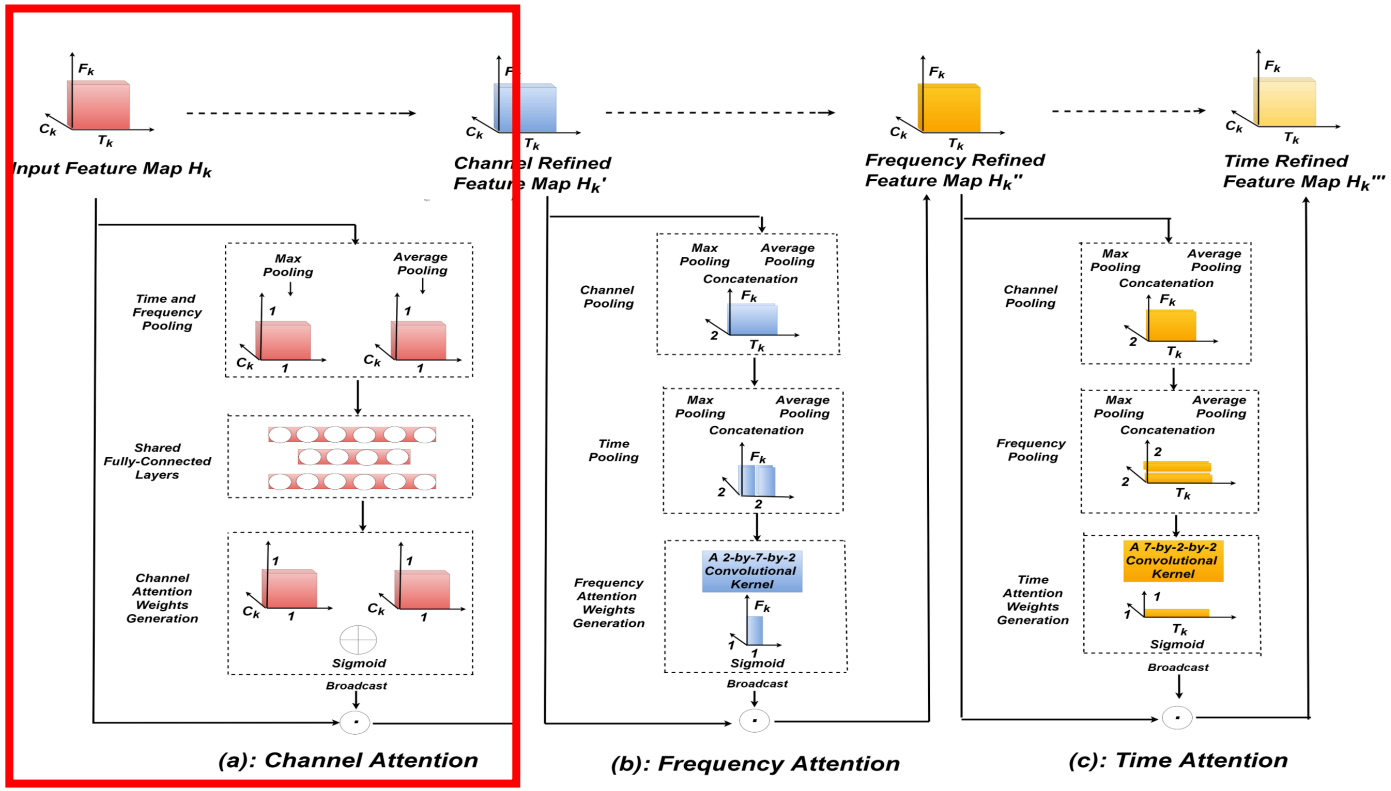


Figure 2: The multi-stage (MS) attention consists of three blocks attention block (a): Channel Attention; (b): Frequency Attention; (c): Time Attention, which are run in a cascading order.

$$H_k^{'} = \alpha_{C,k} \odot H_k$$

$$H_k^{''} = \alpha_{F,k} \odot H_k^{'}$$

$$H_k^{'''} = \alpha_{T,k} \odot H_k^{''}$$

$$H_{k,max}^{C} = max^{T_k \times F_k \times 1}(H_k)$$

$$H_{k,avg}^{C} = avg^{T_k \times F_k \times 1}(H_k)$$

$$S_{max} = \mathrm{Relu}((H_{k,max}^{C})W_0 + b_0)W_1$$

$$S_{avg} = \mathrm{Relu}((H_{k,avg}^{C})W_0 + b_0)W_1$$

$$\alpha^{C,k} = \mathrm{Sigmoid}(S_{avg} + S_{max})$$

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Methods
  - multi-stage attention mechanism(MS)

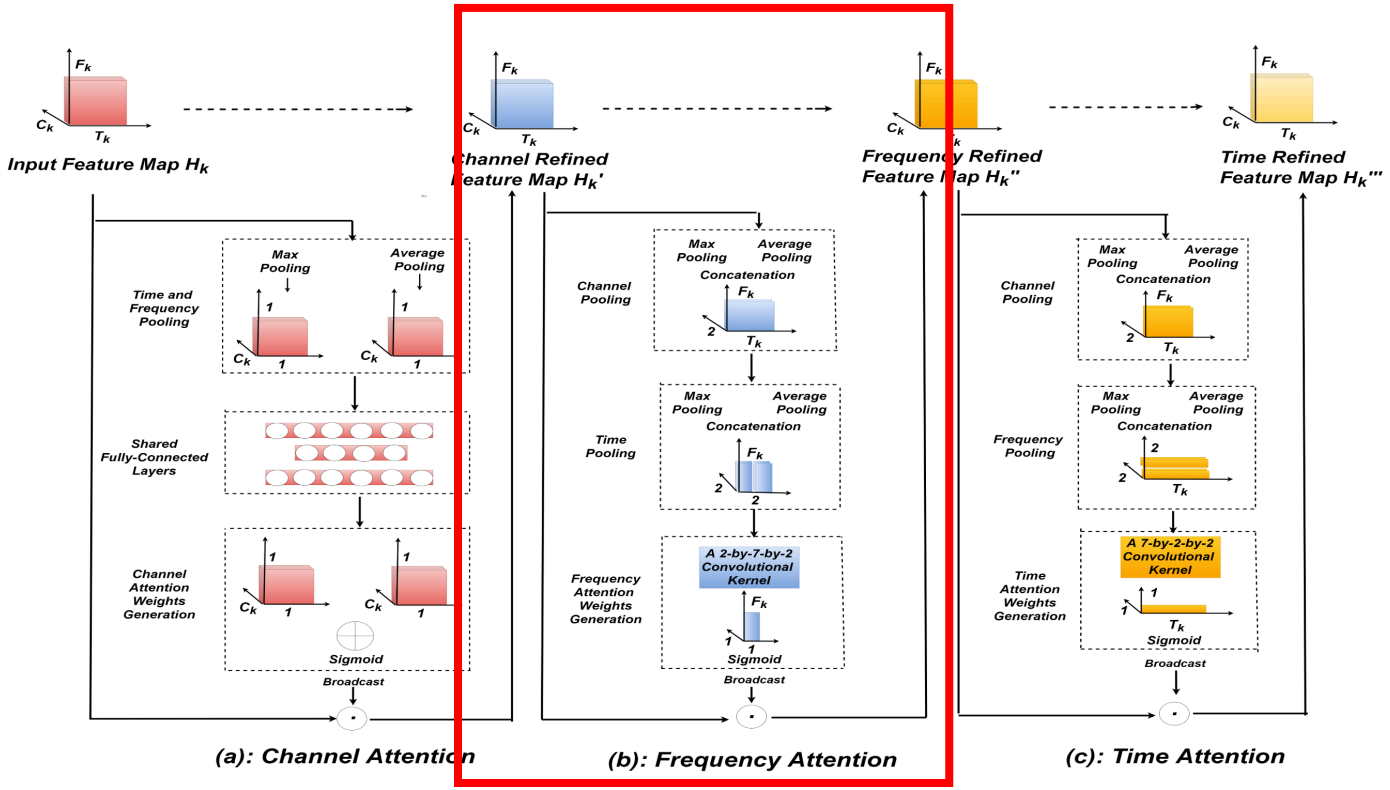


Figure 2: The multi-stage (MS) attention consists of three blocks attention block (a): Channel Attention; (b): Frequency Attention; (c): Time Attention, which are run in a cascading order.

$$H_k' = \alpha_{C,k} \odot H_k$$

$$H_k'' = \alpha_{F,k} \odot H_k'$$

$$H_k''' = \alpha_{T,k} \odot H_k''$$

$$H_{k,max}^{C'} = max^{1\times1\times C_k}(H_k')$$

$$H_{k,avg}^{C'} = avg^{1\times1\times C_k}(H_k')$$

$$H_{k,pool}^{C'} = \text{Concat}[H_{k,avg}^{C'}; H_{k,max}^{C'}]$$

$$H_{k,max}^{T'} = max^{T_k\times1\times1}(H_{k,pool}^{C'})$$

$$H_{k,avg}^{T'} = avg^{T_k\times1\times1}(H_{k,pool}^{C'})$$

$$H_{k,pool}' = \text{Concat}[H_{k,avg}^{T'}; H_{k,max}^{T'}]$$

$$\alpha_k^F = \text{Sigmoid}(f^{2\times7}(H_{k,pool}'))$$

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Methods
  - multi-stage attention mechanism(MS)

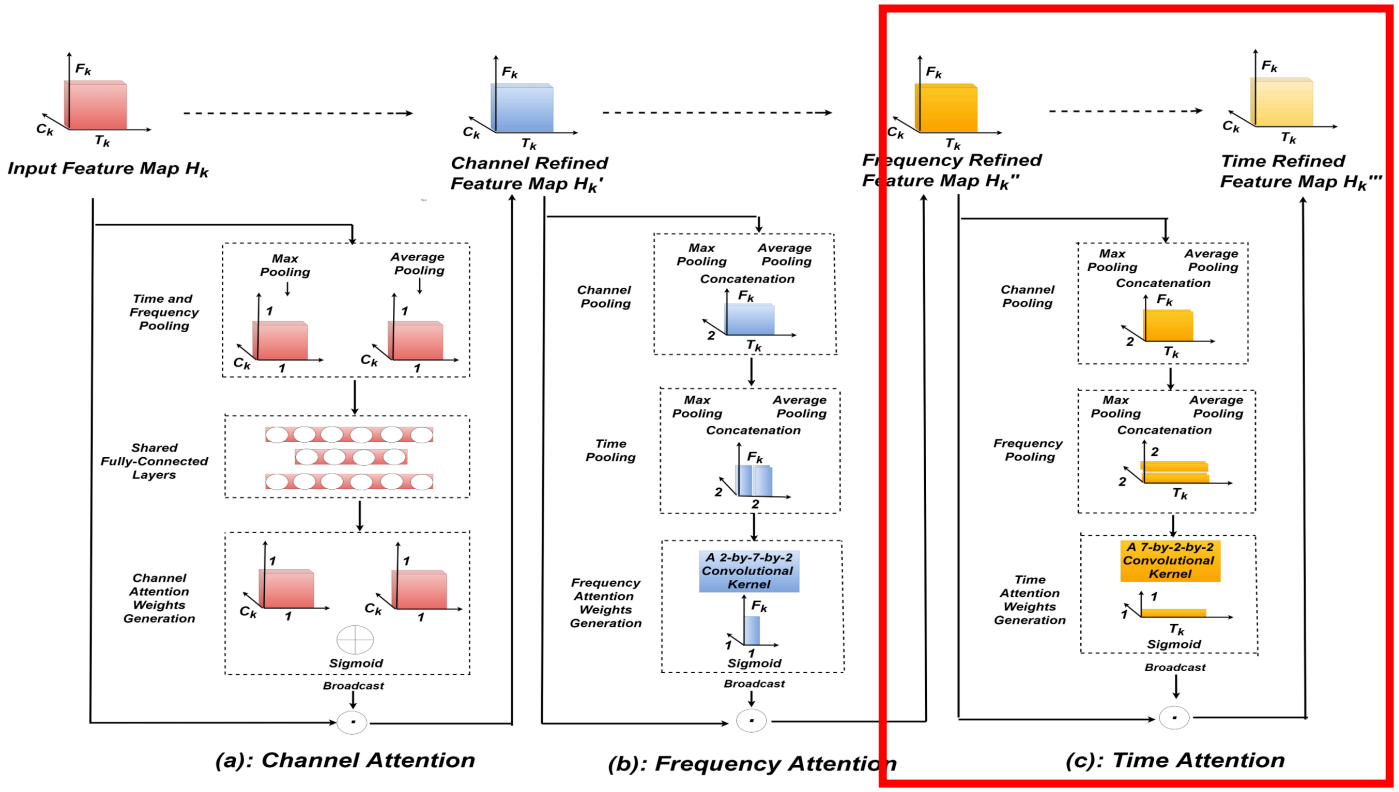


Figure 2: The multi-stage (MS) attention consists of three blocks attention block (a): Channel Attention; (b): Frequency Attention; (c): Time Attention, which are run in a cascading order.

$$H_k' = \alpha_{C,k} \odot H_k$$

$$H_k'' = \alpha_{F,k} \odot H_k'$$

$$H_k''' = \alpha_{T,k} \odot H_k''$$

$$H_{k,max}^{C''} = max^{1\times1\times C_k}(H_k'')$$

$$H_{k,avg}^{C''} = avg^{1\times1\times C_k}(H_k'')$$

$$H_{k,pool}^{C''} = \text{Concat}[H_{k,avg}^{C''}; H_{k,max}^{C''}]$$

$$H_{k,max}^{F''} = max^{1\times F_k\times1}(H_{k,pool}^{C''})$$

$$H_{k,avg}^{F''} = avg^{1\times F_k\times1}(H_{k,pool}^{C''})$$

$$H_{k,pool}'' = \text{Concat}[H_{k,avg}^{F''}; H_{k,max}^{F''}]$$

$$\alpha_k^T = \text{Sigmoid}(f^{7\times2}(H_{k,pool}''))$$

Similar with Squeeze-and-Excitation Module !

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Methods
  - Architecture

Layer Name	Structure	Dilation
CONV-MS Block1	7x1x48 MS	1x1
CONV-MS Block2	1x7x48 MS	1x1
CONV-MS Block3	5x5x48 MS	1x1
CONV-MS Block4	5x5x48 MS	1x2
CONV-MS Block5	5x5x48 MS	1x4
CONV-MS Block6	5x5x48 MS	1x8
CONV-MS Block7	5x5x48 MS	1x1
CONV-MS Block8	5x5x48 MS	2x2
CONV-MS Block9	5x5x48 MS	4x4
CONV-MS Block10	5x5x48 MS	8x8
CONV-MS Block11	1x1x1 MS	1x1

Table 1: Architecture of the speech enhancement network (SE-Net) consists of 11 blocks. In each block, a dilated convolutional layer is followed by a multi-stage attention (MS) layer.

Block Name	Structure	Output
RES-MS Block1	3x3x64 3x3x64 3x3x64 MS-ATT	150x129
RES-MS Block2	3x3x128 3x3x128 3x3x128 MS-ATT	75x65
RES-MS Block3	3x3x128 3x3x128 MS-ATT	75x65
RES-MS Block4	3x3x256 3x3x256 3x3x128 MS-ATT	38x33
RES-MS Block5	3x3x256 3x3x256 MS-ATT	38x33
RES-MS Block6	3x3x256 3x3x256 MS-ATT	38x33
RES-MS Block7	3x3x256 3x3x256 MS-ATT	38x33
RES-MS Block8	3x3x512 3x3x512 3x3x128 MS-ATT	19x17
Pool	19x1	1x17x512
FC	512	

Table 2: Architecture of SID-Net consists of 8 blocks. Within each block, the multiple convolutional layers are followed by a multi-stage attention (MS) layer before a residual connection.

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Experiments
  - Datasets : VoxCeleb1
  - Results

Noise Type	SNR	SID		SE+SID		SE-MS +SID		SE+SID-MS	
		Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)
Noise	0	74.1	86.9	76.3	88.9	**78.5**	**90.0**	77.7	89.2
	5	79.2	90.0	81.1	91.8	**83.4**	**92.1**	81.9	91.8
	10	83.2	93.2	86.0	94.7	**87.3**	**95.6**	86.7	95.1
	15	84.9	94.6	87.3	95.8	**89.5**	**96.7**	88.8	96.0
	20	87.9	95.4	89.1	96.6	**90.9**	**97.5**	90.2	97.0
Music	0	65.8	82.0	67.7	83.7	**70.3**	**84.1**	69.5	83.5
	5	76.9	89.1	80.0	91.0	**81.6**	**91.5**	80.6	90.8
	10	83.8	93.5	85.2	94.7	**86.3**	**95.3**	85.8	94.7
	15	86.1	93.9	88.4	95.6	**89.1**	**96.7**	88.2	95.4
	20	87.4	94.7	89.1	96.0	**90.2**	**97.1**	89.5	96.6
Babble	0	62.4	80.2	65.7	81.5	**67.5**	**83.0**	66.6	81.9
	5	76.2	87.3	78.6	88.9	**80.6**	**89.9**	79.3	89.6
	10	81.4	92.2	84.6	93.6	**86.6**	**94.5**	85.3	83.2
	15	84.0	92.6	86.8	93.9	**88.3**	**94.7**	87.6	94.0
	20	85.8	92.9	87.1	94.6	**89.0**	**95.5**	88.8	95.2
Original		88.5	95.9	89.8	96.5	**91.9**	**97.6**	90.8	97.3

Table 4: Speaker Identification Results on the Voxcebe1 test data when being corrupted by three types of noise (Noise, Music and Babble) at different SNR (0-20 dB) levels. Four different scenarios are tested: SID-Net (SID), the use of both SE-Net and SID-Net without employing a multi-stage attention (SE+SID), a joint system combing SE-Net with SID-Net, but a multi-stage attention is used only in SE-Net(SE-MS+SID); The SE-Net and SID-Net denotes a joint system, with a multi-stage attention layer being used only in SID-Net(SE+SID-MS).

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Experiments
  - Datasets : VoxCeleb1
  - Results

Noise Type	SNR	SID		VoiceID Loss [10]		SE+SID		SE-MS+SID		SE+SID-MS	
		EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF	EER (%)	DCF
Noise	0	16.94	0.933	16.56	0.938	16.20	0.912	**15.95**	**0.901**	16.13	0.908
	5	12.48	0.855	12.26	0.830	11.99	0.819	**11.76**	**0.805**	11.78	0.812
	10	10.03	0.760	9.86	0.747	9.54	0.732	**9.17**	**0.717**	9.29	0.727
	15	8.84	0.648	8.69	0.686	8.48	0.665	**8.08**	**0.639**	8.10	0.641
	20	7.96	0.594	7.83	0.639	7.52	0.629	**7.07**	**0.615**	7.09	0.623
Music	0	17.04	0.940	16.24	0.913	15.96	0.901	**15.58**	**0.899**	15.89	0.904
	5	11.54	0.828	11.44	0.818	11.15	0.805	**10.93**	**0.791**	11.04	0.801
	10	9.69	0.749	9.13	0.733	9.12	0.731	**8.87**	**0.714**	8.97	0.725
	15	8.40	0.689	8.10	0.677	8.08	0.643	**7.62**	**0.621**	7.77	0.629
	20	7.70	0.665	7.48	0.635	7.39	0.619	**7.13**	**0.607**	7.26	0.614
Babble	0	38.90	1.000	37.96	1.000	37.53	0.999	37.55	0.999	**37.46**	**0.998**
	5	28.04	0.998	27.12	0.996	26.97	0.979	26.42	0.981	**26.35**	**0.977**
	10	17.34	0.917	16.66	0.926	16.44	0.911	**16.30**	**0.907**	16.36	0.911
	15	11.31	0.795	11.25	0.807	11.24	0.801	**10.89**	**0.795**	10.94	0.801
	20	9.12	0.720	8.99	0.705	8.77	0.695	**8.39**	**0.677**	8.51	0.688
Original		6.92	0.565	6.79	0.574	6.41	0.541	**6.18**	**0.528**	6.26	0.535

Table 5: Speaker Verification Results on Voxceleb1 test data when it being corrupted by different types of noise (Noise, Music and Babble) at different SNR (0-20 dB). Four different scenarios are tested: only use SID-Net (SID); A joint system combining the SE-Net with the SID-Net without a multi-stage attention (SE+SID); A joint system using both SE-Net and SID-Net, but without being used in multi-stage attention (SE-MS+SID); A joint system consisting of SE-Net and SID-Net, with a multi-stage attention being used in SID-Net (SE+SID-MS). The results of VoiceID Loss [10] is listed and works as a baseline.

Robust Speaker Recognition Using Speech Enhancement And Attention Model

- Conclusion
  - Multi-stage attention model to highlight speaker relevant information.
  - A joint optimisation by cascading the speech enhancement network and speaker recognition network.
  - **Enhance spectral features(**Fine-grained**).**

Knowing What to Listen to: Early Attention for Deep Speech Representation Learning

Amirhossein Hajavi*, **Ali Etemad**
Department of ECE and Ingenuity Labs
Queen's University, Kingston, Canada
{a.hajavi, ali.etemad}@queensu.ca

Amirhossein Hajavi and Ali Etemad, "Knowing what to listen to: Early attention for deep speech representation learning," arXiv preprint arXiv:2009.01822, 2020.

Knowing what to listen to: Early attention for deep speech representation learning

- Motivation
  - Attention models play an important role in improving deep learning models.
  - However current attention mechanisms are unable to attend to fine-grained information items.
- Methods
  - Fine-grained Early Frequency Attention (FEFA) for speech signals.

$$p_i = \frac{exp(index(Spec(x(t), \omega_i), F)) \times W)}{\sum_{j=1}^{|M|} exp(index(Spec(x(t), \omega_j), F)) \times W)}$$

$$AttentionMap = \sum_{i=1}^{|M|} p_i \times Spec(x(t), \omega_i)$$

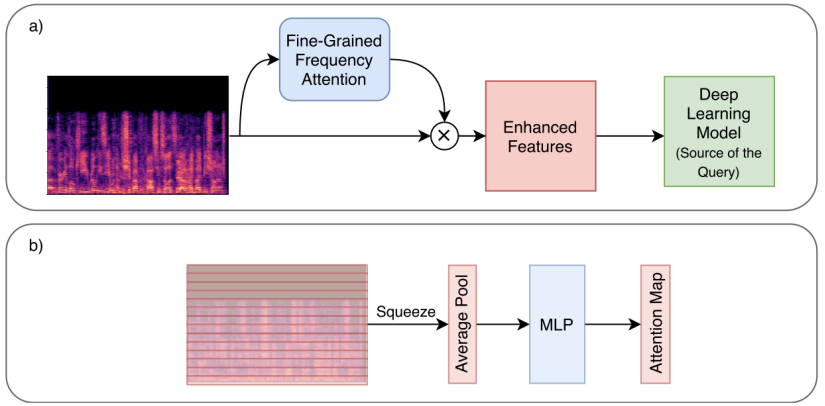


Figure 1: a) The overview of the FEFA model. The model uses the spectrogram representation of the utterance as the memory set and the feature set associated with the early layers of the DNN model as the modality to extract query. b) The modules inside the FEFA model consist of a squeeze function and an MLP module.

Knowing what to listen to: Early attention for deep speech representation learning

- Experiments
  - Datasets : VoxCeleb2(train), VoxCeleb1(test); IEMOCAP
  - 257-Spectrogram; several SOTA backbone
  - Results

Table 1: Speaker Recognition results. (* The result of identification accuracy was not published and is replicated using the trained models provided by the authors.)

Model	FEFA Layers	EER (%)	ΔEER (%)	Acc. (%)	ΔAcc. (%)
ResNet + Self-Attention (Bian, Chen, and Xu 2019)	None	5.4	N/A	N/A	N/A
CNN (unspecified) + Soft-Attention (Okabe, Koshinaka, and Shinoda 2018)	None	3.8	N/A	N/A	N/A
VGG (Nagrani, Chung, and Zisserman 2017)	None	7.8	N/A	80.5	N/A
VGG + FEFA	Single-layer	7.4	+5.1	84.7	+5.2
VGG + FEFA	Multi-layer	7.6	+2.5	82.4	+2.3
ResNet34 (Chung, Nagrani, and Zisserman 2018)	None	4.83	N/A	N/A	N/A
ResNet50 (Chung, Nagrani, and Zisserman 2018)	None	3.95	N/A	N/A	N/A
Thin-ResNet + Ghostvlad (Xie et al. 2019a)	None	3.22	N/A	86.5*	N/A
Thin-ResNet + FEFA	Single-layer	3.12	+3.1	93.6	+8.2
Thin-ResNet + FEFA	Multi-layer	3.18	+1.2	91.7	+6.0
SE-ResNet	None	4.81	N/A	90.5	N/A
SE-ResNet + FEFA	Single-layer	3.68	+19.0	93.8	+3.6
SE-ResNet + FEFA	Multi-layer	4.58	+4.7	91.5	+1.1

Table 2: Speech Emotion Recognition results.

Model	FEFA Layers	Acc. (%)	ΔAcc. (%)
Thin-ResNet	None	59.72	N/A
Thin-ResNet+FEFA	Single-layer	62.32	+4.35
Thin-ResNet+FEFA	Multi-layer	61.57	+3.09
VGG	None	52.48	N/A
VGG + FEFA	Single-layer	56.70	+8.21
VGG + FEFA	Multi-layer	55.36	+5.48
SE-ResNet	None	59.82	N/A
SE-ResNet + FEFA	Single-layer	62.28	+4.11
SE-ResNet + FEFA	Multi-layer	61.63	+3.02

Knowing what to listen to: Early attention for deep speech representation learning

- Experiments(Noise)

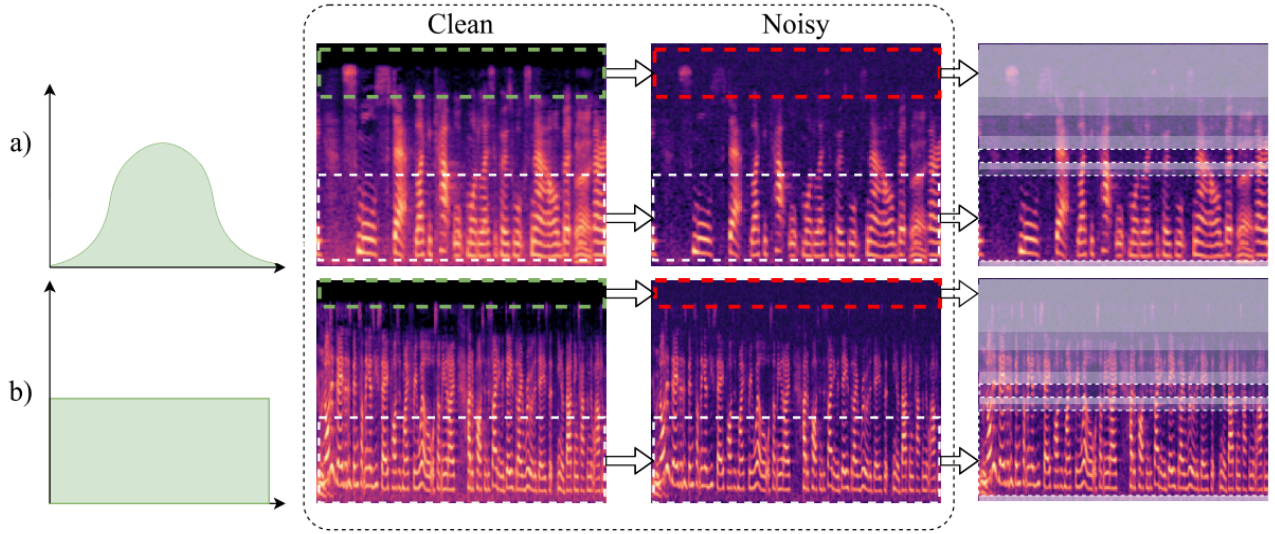


Figure 2: Robustness test by adding synthetic noise to the utterances. (a) the noise selected from a Gaussian distribution. (b) the noise selected from a uniform distribution

Table 3: Robustness test results for SR task. The comparison is performed with the state-of-the-art model GhostVlad (Xie et al. 2019a) with and without FEFA.

Noise	Model	SNR	EER (%)	ΔEER (%)
Normal	w/o FEFA	20db	3.40	-5.5
	w/o FEFA	50db	3.85	-19.5
	w/o FEFA	100db	4.82	-49.6
	+ FEFA	20db	3.12	0
	+ FEFA	50db	3.15	-0.9
	+ FEFA	100db	3.44	-10.2
Uniform	w/o FEFA	20db	3.32	-3.1
	w/o FEFA	50db	3.48	-8.0
	w/o FEFA	100db	3.96	-22.9
	+ FEFA	20db	3.12	0
	+ FEFA	50db	3.14	-0.6
	+ FEFA	100db	3.41	-9.4

Knowing what to listen to: Early attention for deep speech representation learning

- Conclusion
  - The **FEFA** provides a better representation of the spectrogram **by attending to each frequency bin individually**.
  - **Enhance spectral features.**

ON THE IMPORTANCE OF DIFFERENT FREQUENCY BINS FOR SPEAKER VERIFICATION

Aiwen Deng[1], *Shuai Wang*[2*], *Wenxiong Kang*[1], *Feiqi Deng*[1]

[1]South China University of Technology, Guangzhou, China
[2] Shanghai Jiao Tong University, Shanghai, China

A. Deng, S. Wang, W. Kang and F. Deng, "On the Importance of Different Frequency Bins for Speaker Verification," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7537-7541, doi: 10.1109/ICASSP43922.2022.9746084.

ON THE IMPORTANCE OF DIFFERENT FREQUENCY BINS FOR SPEAKER VERIFICATION

- Motivation
  - Naturally, there would be a question of whether all different frequency bins contribute equally to the speaker verification system performance?
- Methods
  - Frequency Reweighting Layer (FRL)

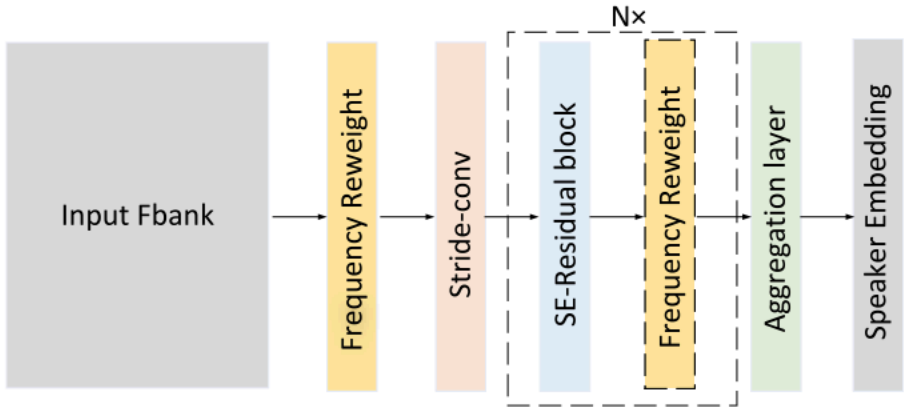


Fig. 2. Overall system structure, the newly proposed frequency reweight layer can be freely inserted into original architecture

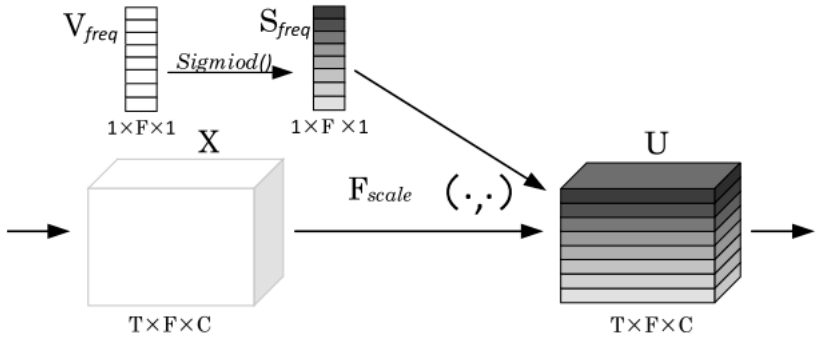


Fig. 1. The frequency reweighting layer architecture

ON THE IMPORTANCE OF DIFFERENT FREQUENCY BINS FOR SPEAKER VERIFICATION

- Experiments
  - Datasets : VoxCeleb1
  - SE-Fast-ResNet34, 80-dimensional Fbank
  - Results

System	EER	minDCF	New Params
Baseline	3.35	0.383	0
+ inp	3.23	**0.367**	80
+ lay1	3.24	0.382	40
+ lay2	3.27	0.370	20
+ inp + lay1 + lay2	**2.99**	0.372	140

Table 1. EER and MinDCF performance of the systems on the standard VoxCeleb1-test. The inp, lay1, lay2 denote the FRL insertion after the input Fbank features, the first SE-Residual block, and the second layer SE-Residual block, respectively.

Max mask length	0	5	10	15	20
Baseline	3.35	5.03	8.25	12.40	16.62
Baseline+inp	3.23	4.73	7.00	10.45	15.06
ΔEER	0.12	0.3	1.25	1.95	1.56

Table 2. Frequency mask robustness test results. The values in the table indicate EER. The inp denote the FRL insertion after the input Fbank features.

ON THE IMPORTANCE OF DIFFERENT FREQUENCY BINS FOR SPEAKER VERIFICATION

- Experiments(Analysis)

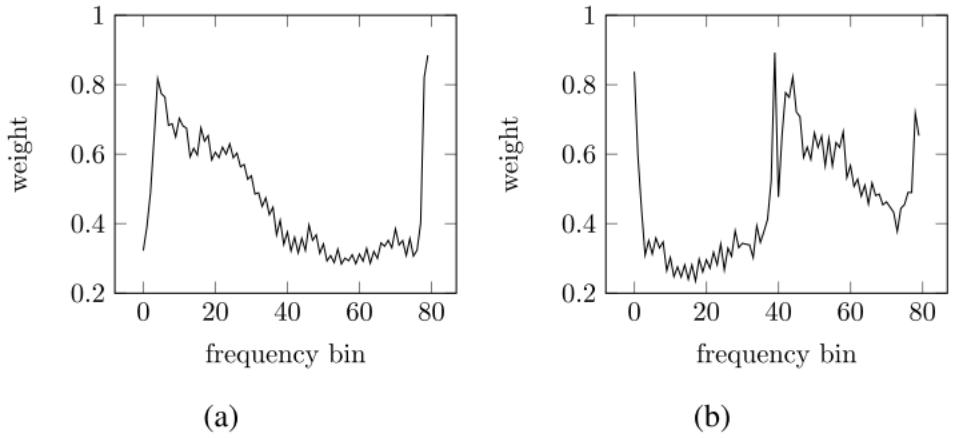


Fig. 3. Visualization of the learned weight. (a) the FRL weight distribution for normal input Fbank features; (b) the FRL weight distribution of the input spectrum by swapping the first 40 dimensions of the input Fbank features with the last 40 dimensions.

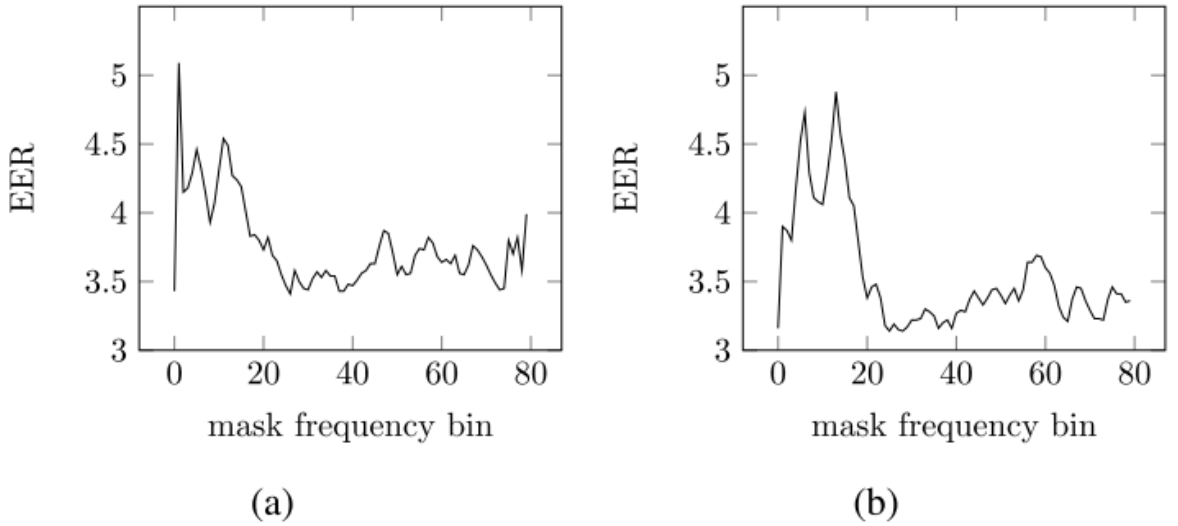


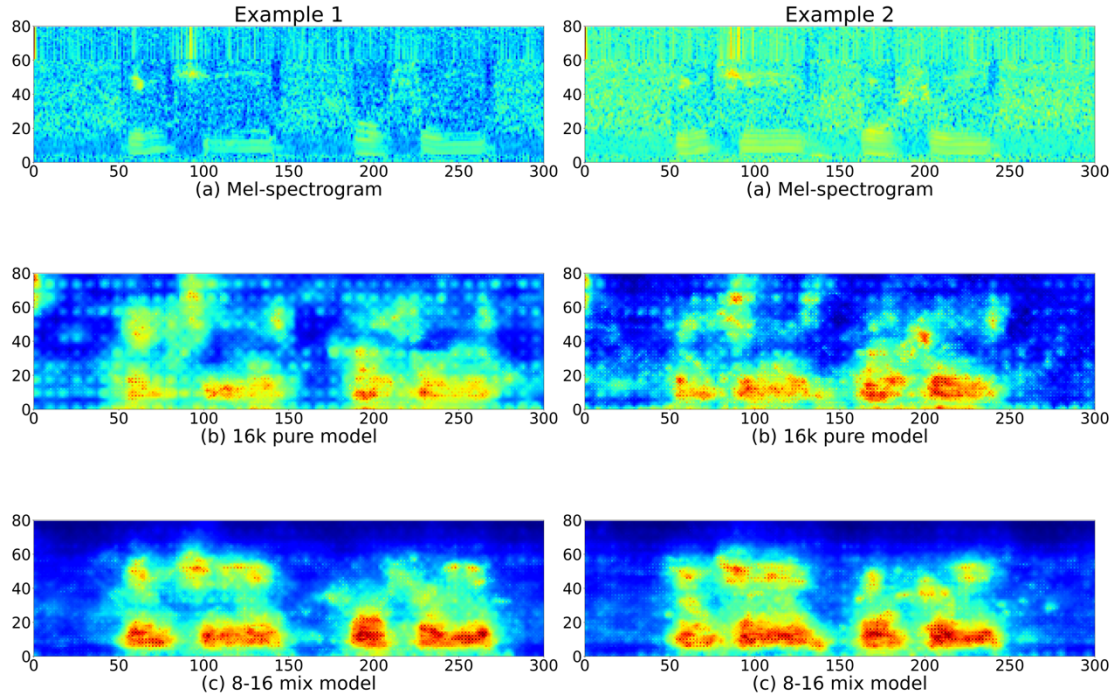
Fig. 4. (a)the baseline model frequency bin masking results; (b)the FRN model frequency bin masking results.

ON THE IMPORTANCE OF DIFFERENT FREQUENCY BINS FOR SPEAKER VERIFICATION

- Conclusion
  - Frequency Reweighting Layer to automatically learn the importance of different frequency dimensions.
  - Show that the system performance attributes more to the lower frequencies.

Conclusion

- Some fine-grained enhancements are made on the spectrum through methods such as the attention mechanism.
- Thereby finding task-relevant features and improving the noise robustness of the model.
- However poor interpretability.



Reference

Method	Year	Paper Title	PDF
MS-attention	2020	Robust Speaker Recognition Using Speech Enhancement And Attention Model	https://arxiv.org/abs/2001.05031
FEFA	2020	Knowing What to Listen to: Early Attention for Deep Speech Representation Learning	https://arxiv.org/abs/2009.01822
FRL	2022	ON THE IMPORTANCE OF DIFFERENT FREQUENCY BINS FOR SPEAKER VERIFICATION	https://ieeexplore.ieee.org/document/9746084