BDEE 2023

2023 THE 3RD INTERNATIONAL CONFERENCE ON
BIG DATA ENGINEERING AND EDUCATION

第三届大数据工程和教育国际会议

August 21-23, 2023 | Chengdu, China 中国成都 | 2023年8月21-23日

Co-Hosted by
联 合 主 办

Enhancing Synthesized Speech Detection with Dual Attention Using Features Fusion

Bo Wang[1], Yanyan Ma[2,*], Yeling Tang[3], Rui Wang[4], MaoZhen Zhang[5]

School of Information and Communication Engineering, Dalian University of Technology, China

大连理工大学
Dalian University Of Technology

DeepFake = Deep Learning + Fake

DeepFake

image deepfake

video deepfake

voice deepfake

- Text-to-speech (TTS)
- Voice conversion (VC)
- Impersonation
- Replay
- Other adversarial attacks

■ Related works

Moustafa Alzantot
Deep Residual Neural Networks
(2019)

Xu Li
SE-Res2Net50
(2021)

Ruchira Ray
Logmel+ResSENet
(2021)

Youxuan Ma
RW-Resnet
(2021)

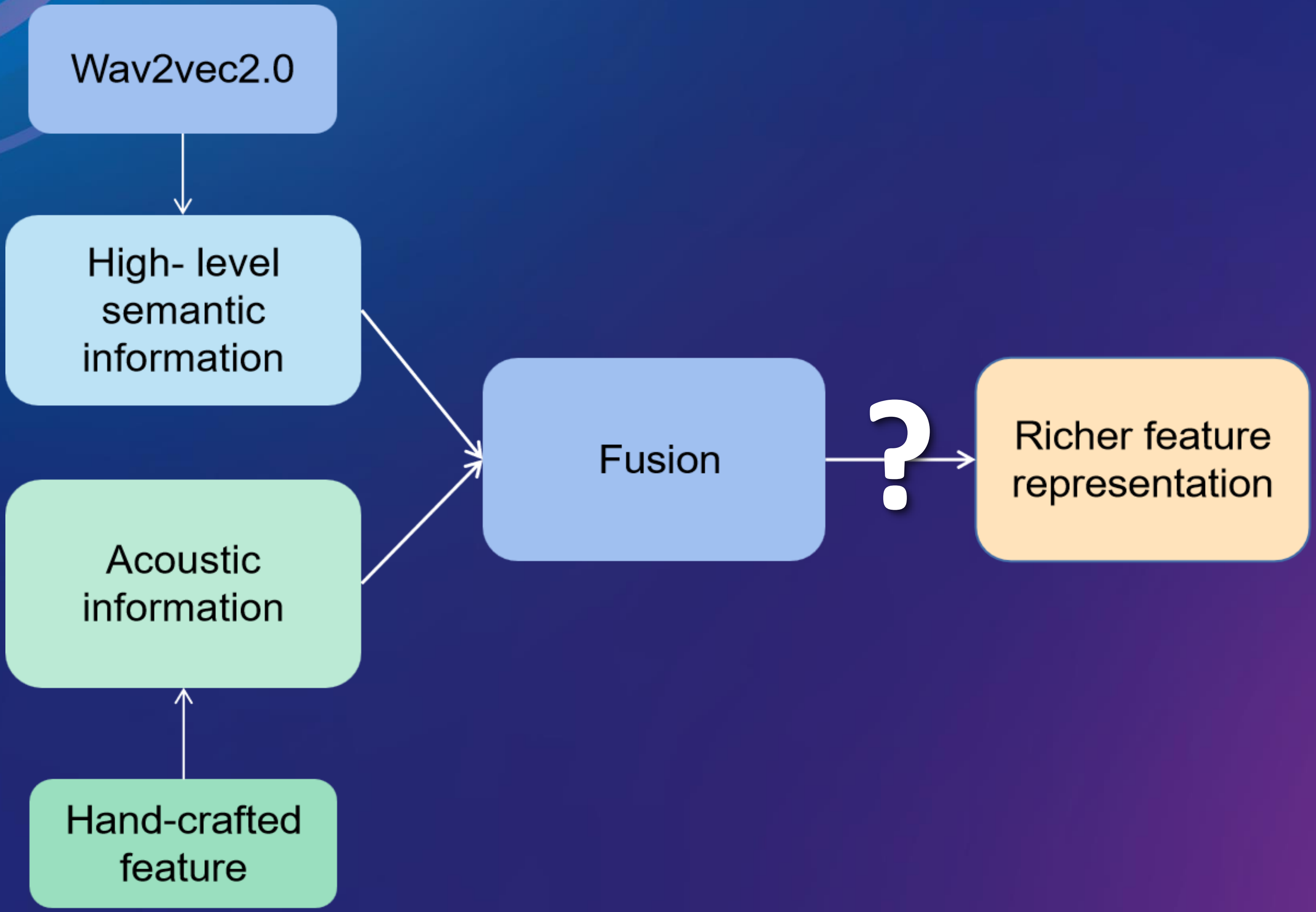
Tak H
End-to-end Rawnet2
(2021)

Hemlata Tak
Graph attention networks
(2021)

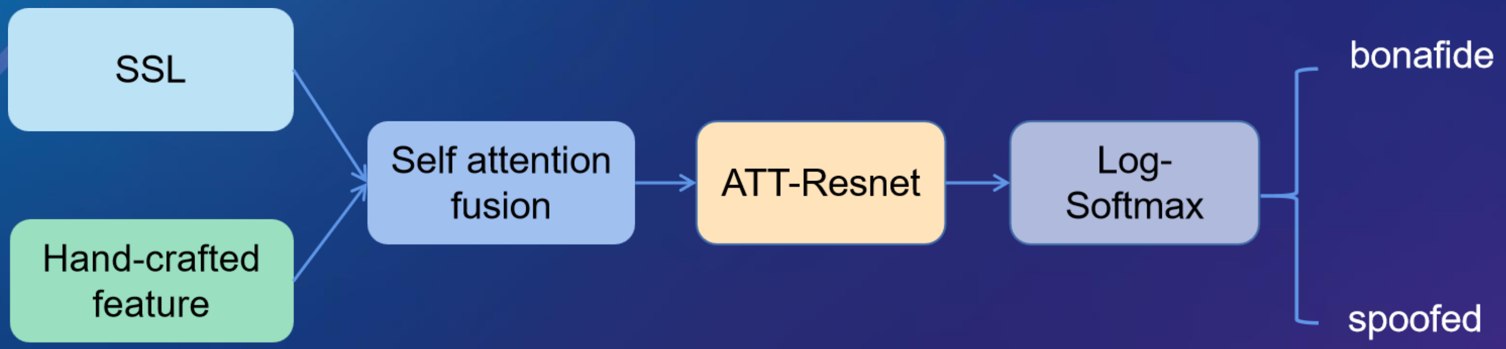
Jee-weon Jung
Integrated spectro-temporal graph attention network
(2022)

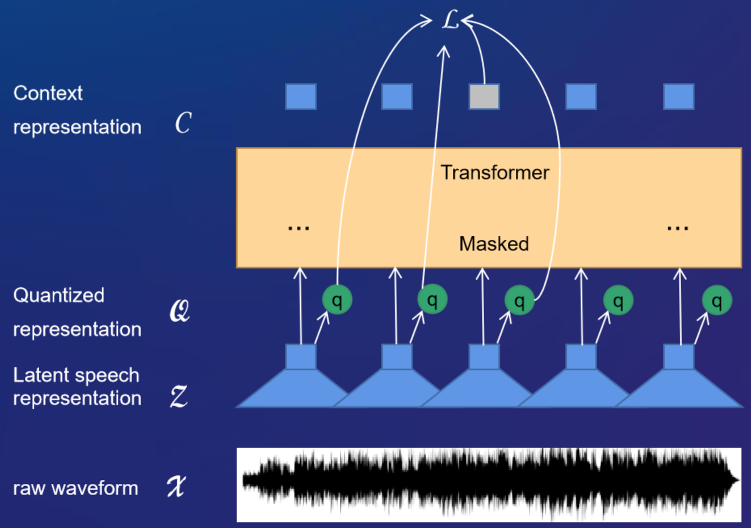
Hemlata Tak
Wav2vec2.0 front-end.
(2023)

Motivation

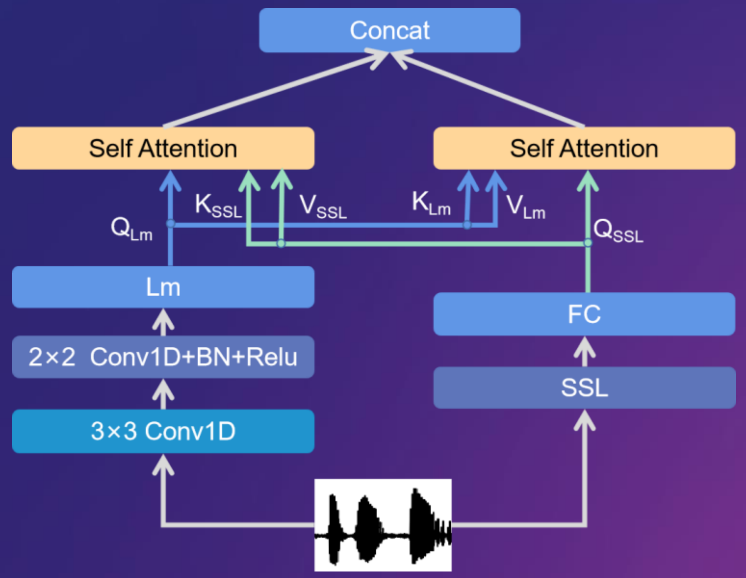

Wav2vec2.0
High- level semantic information
Acoustic information
Hand-crafted feature
Fusion
?
Richer feature representation

■ Architecture overview





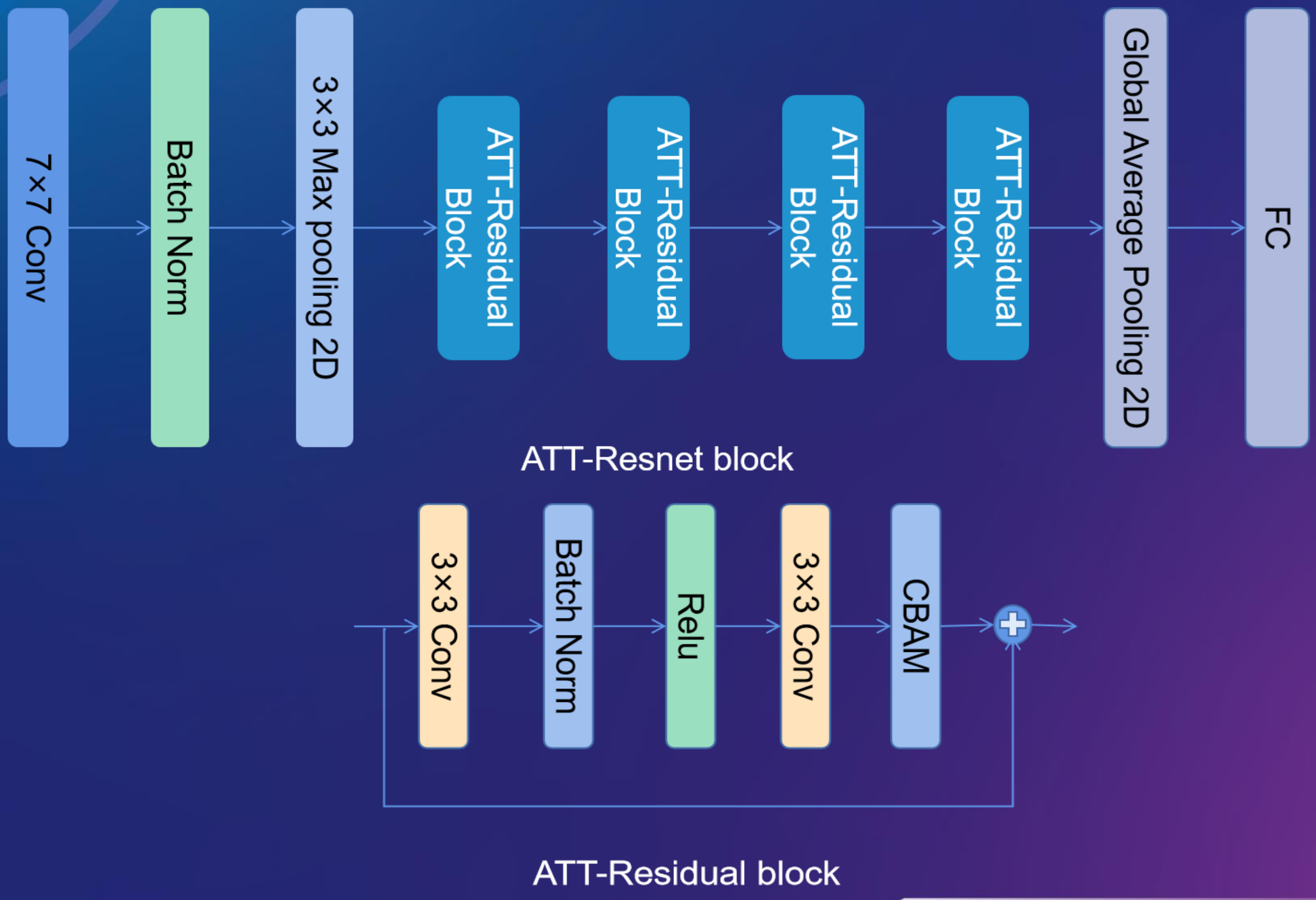
Wav2vec2.0 overall framework



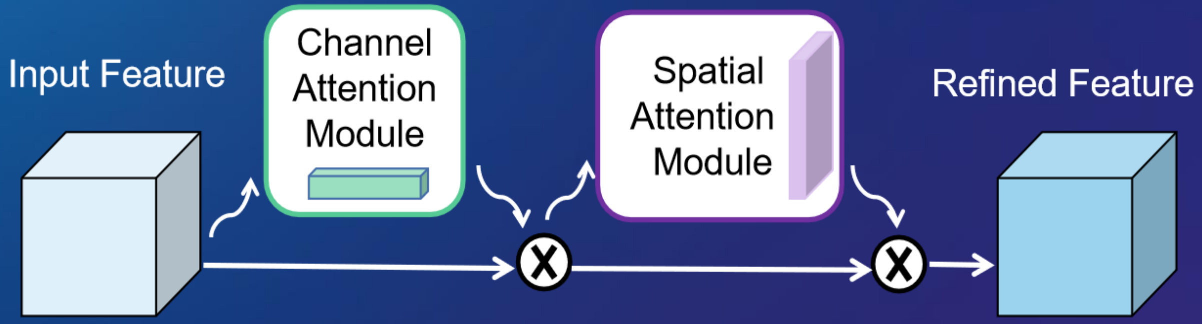
$$C_{Logmel} = Soft\max(\frac{Q_{Logmel} * K_{SSL}}{\sqrt{D}})V_{SSL}$$

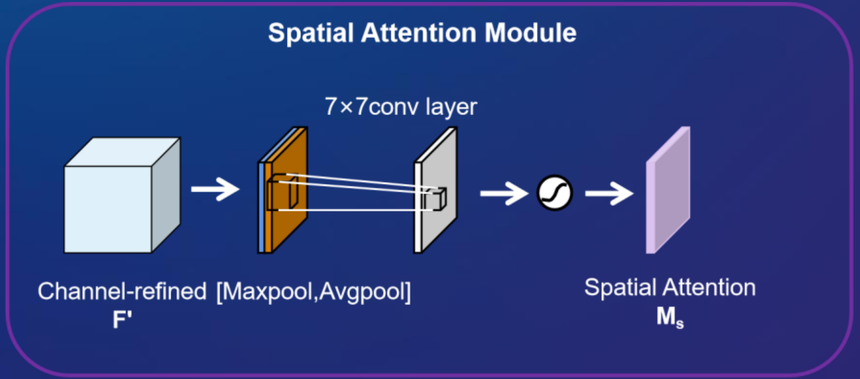
$$C_{SSL} = Soft\max(\frac{Q_{SSL} * K_{Logmel}}{\sqrt{D}})V_{Logmel}$$

■ Architecture overview



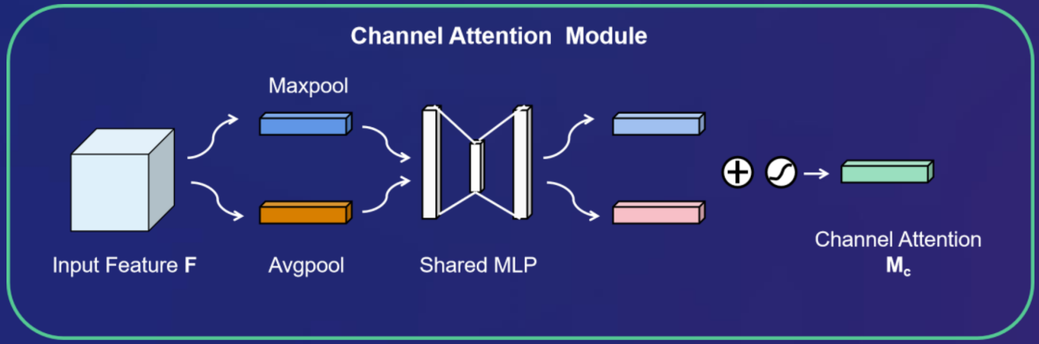
■ Architecture overview





Dynamically adjust the weights of each spatial position feature. The local features of the image can be better extracted by weighting the spatial position.

$$M_S(F) = \sigma(f^{7\times 7}([Avgpool(F); Maxpool(F)]))$$



Dynamically adjust the feature weights of different channels to highlight the importance of each channel in the input feature mapping.

$$M_c(F) = \sigma(MLP(Avgpool(F)) + MLP(Maxpool(F)))$$

Experiments

Experimental settings

Data sets: the ASVspoof2019 data set is trained and tested on the ASVspoof2021 Logical Access (LA) and DeepFake (DF) data sets.

Preprocessing: the input speech is randomly cut into a 4 seconds segment

Traditional feature: 128-dimensional log mel-filterbank

Optimizer: 1e-4 Adam optimizer

loss function: cross entropy loss

learning rate: 1e-6

Experiments

Experimental result

Table I Performance Comparison with Other Single Systems on the Evaluation Set of the Asvspoof 2021 LA

System	EER(%)	t-DCF
CQCC-GMM	15.62	0.4974
LFCC-GMM [31]	19.3	0.5758
LFCC-LCNN [32]	9.26	0.3445
RawNet2[8]	9.5	0.4257
LFCC-ECAPA-TDNN [33]	5.46	0.3094
SSL-AASIST [24]	4.48	0.3094
Our	**4.12**	**0.3008**

Table II Results of Different Modules in Asvspoof 2021 LA Scenarios

System	EER(%)	t-DCF
SSL-ATT-Resnet18	4.52	0.3105
Self attention fusion- Resnet18	4.85	0.3134
Self attention fusion-ATT-Resnet18	**4.12**	**0.3008**

■ Experiments

Experimental result

Table III Comparison of the Results of Each System on the Asvspoof 2021 DF dataset

System	EER(%)
CQCC-GMM	25.56
LFCC-GMM	25.25
LFCC-LCNN	23.48
RawNet2	22.38
LFCC-ECAPA-TDNN	20.33
SSL-AASIST	4.57
Self attention fusion - Resnet18	**3.90**
Selfattention fusion - ATT-Resnet18	5.34

■ Conclusion

- We have implemented a self-attention-based combination of self-supervised features and Logmel features to better capture complex patterns and contextual information in audio signals.
- With the help of CBAM, the performance of self-attentional combination features on ASVspoof 2021 LA is further improved.
- The experimental results show that we achieved a certain improvement in performance on the ASVspoof 2021 LA and DF datasets.

Thank you for listening!

大连理工大学
Dalian University Of Technology