Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement

报告人：陈航
指导老师：杜俊副教授

Outline

- Introduction
- Previous methods
- Place Based Visual Embedding
- Multimodal Embedding Aware Speech Enhancement
- Experiment and Result Analyses

Introduction

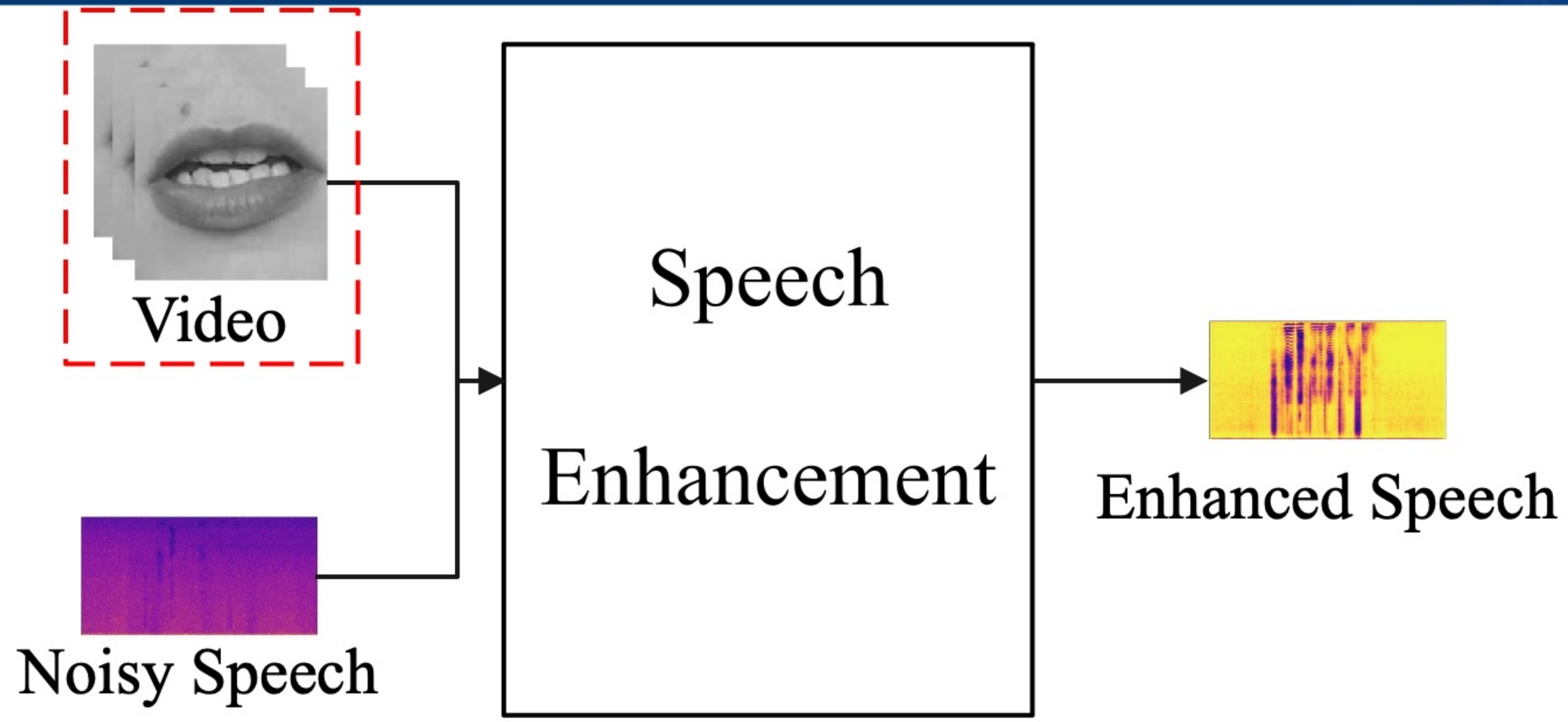


Fig.1. Illustration of Audio-Visual Speech Enhancement (AVSE)

- Speech Enhancement
  - The method generating the enhanced speech with better speech quality and clarity by suppressing background noise components in noisy speech.
- Audio-Visual Speech Enhancement (AVSE)
  - The speech enhancement method utilizing both audio and visual signals.

Previous methods

- **In the term of the visual input:**

Landmarks — Image sequences — Face embedding — Lip embedding …… Time

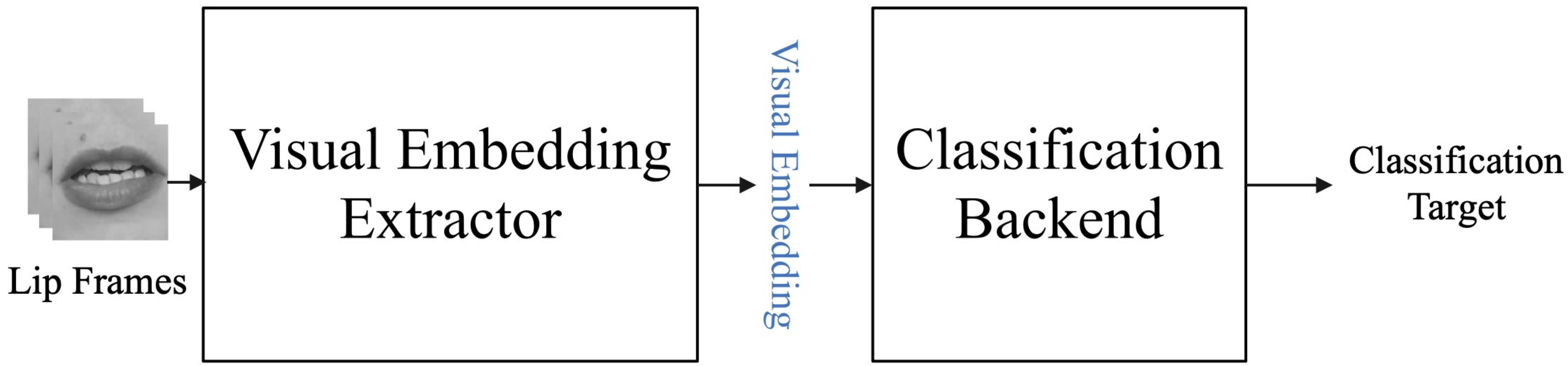


Fig.2. Illustration of A Classification-based Embedding Extracting Framework

- **In the term of the audio-visual fusion:**
  - Channel-wise concatenation at the middle layer of the enhancement network
  - Result fusion

Articulation Place Based Visual Embedding Extraction

Articulation place classes	**CI-phones**
Coronal	d, l, n, s, t, z
High	ch, ih, iy, jh, sh, uh, uw, y
Dental	dh, th
Glottal	hh
Labial	b, f, m, p, v, w
Low	aa, ae, aw, ay, oy
Mid	ah, eh, ey, ow
Retroflex	er, r
Velar	g, k, ng

Tab.1. The Mapping Between Articulation Place Classes and CI-phones

Dental Velar Glottal Coronal Retroflex Low High Mid Labial

Fig.3. 9 lip shapes corresponding to utterance segments representing 9 articulation positions

We propose that the articulation place have a high correlation with the visual acoustic information and can provide a more useful supervisory signal in the training of visual embedding extractor.

Visual Embedding Aware Speech Enhancement

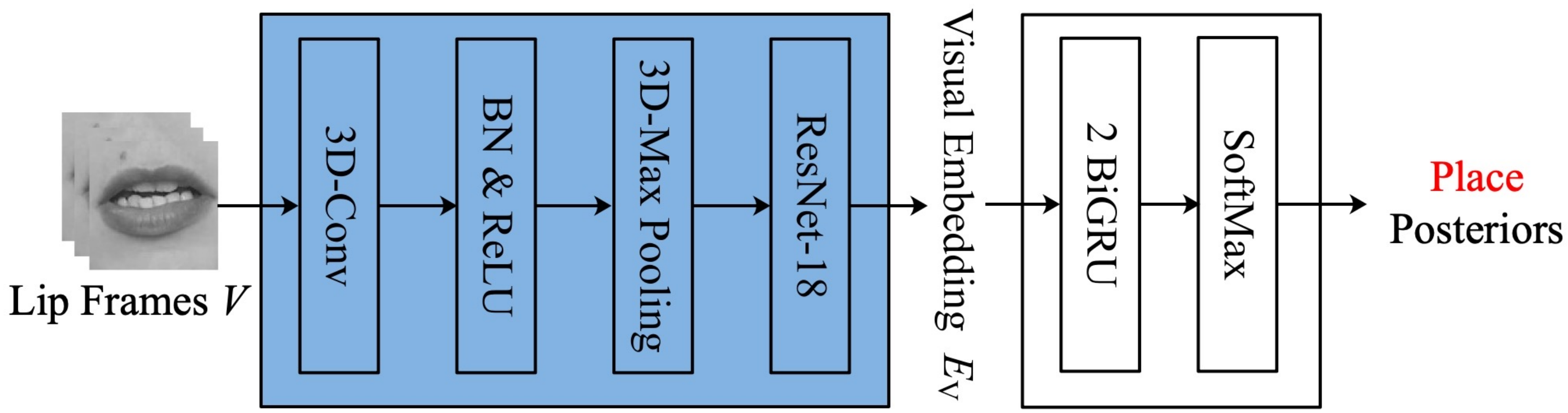


Fig.4. Illustration of a visual embedding extractor

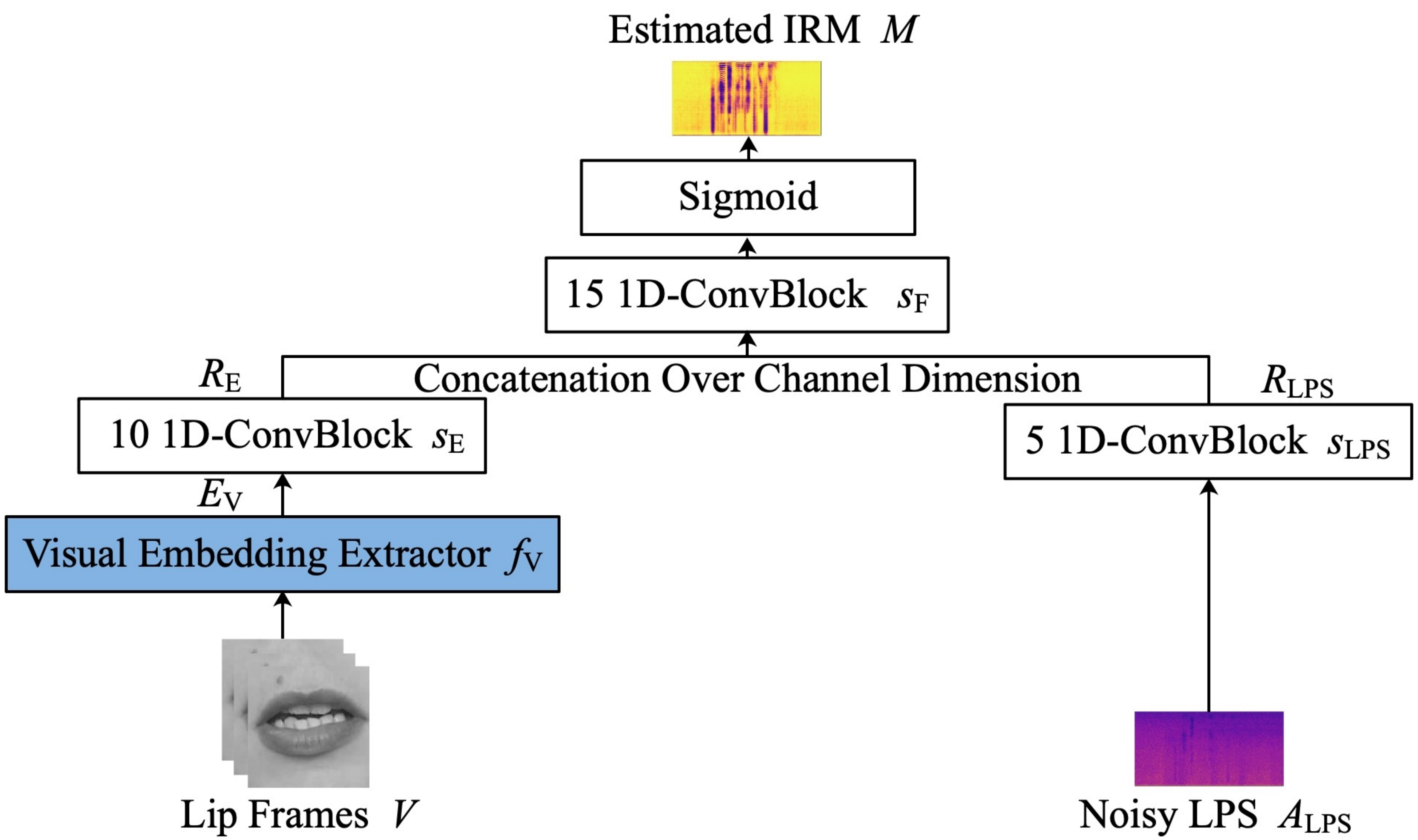


Fig.5. Illustration of the visual embedding aware speech enhancement (VEASE) model

Articulation Place Based Multimodal Embedding Extraction

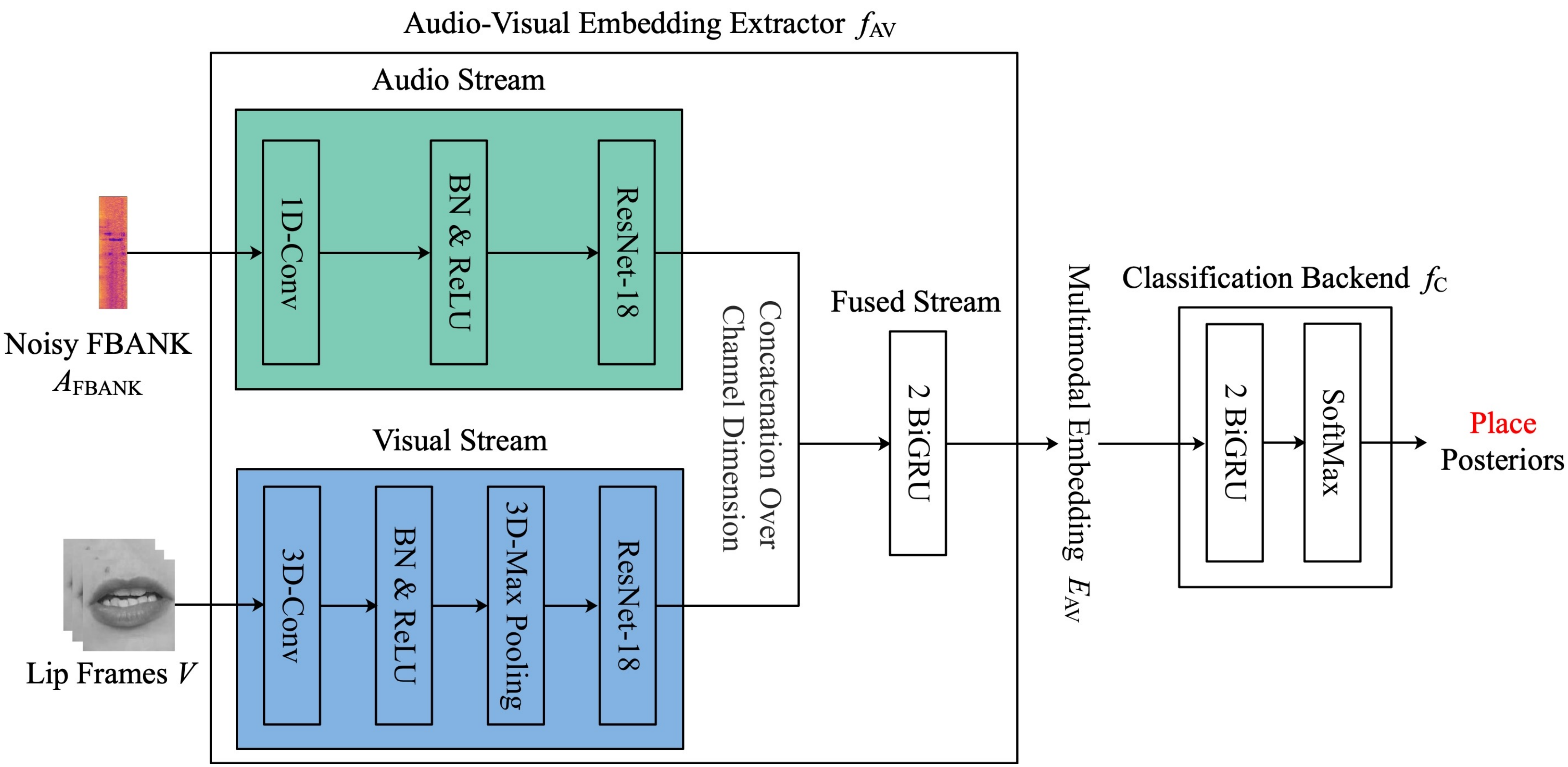


Fig.6. Illustration of the multimodal embedding extractor

The audio-visual embedding extractor takes not only lip frames but also noisy FBANK features as inputs and outputs the multimodal embedding which is learned under the supervision of the articulation place label.

Multimodal Embedding Aware Speech Enhancement

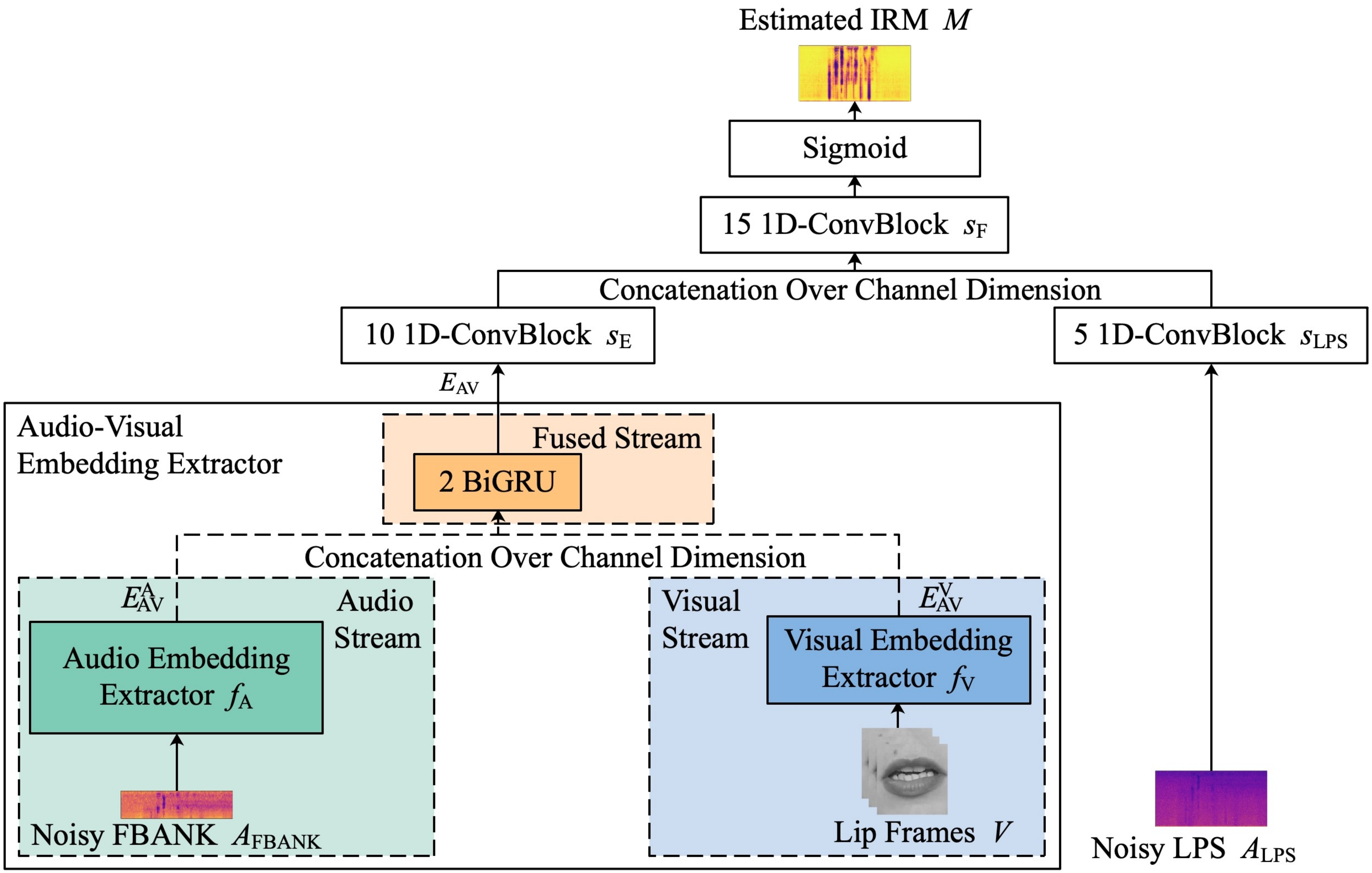


Fig.7. Illustration of the multimodal embedding aware speech enhancement (MEASE) model

Experiment and Result Analyses

Dataset: TCD-TIMIT + noise

- 35-hour train set, 115 noise types, 5 levels of SNRs (15, 10, 5, 0 and −5 dB)
- 8-hour test set, 3 unseen noise types, 5 levels of SNRs (15, 10, 5, 0 and −5 dB)

Model	PESQ					STOI (in %)				
SNR (in dB)	-5	0	5	10	15	-5	0	5	10	15
Noisy	1.70	1.97	2.26	2.56	2.86	54.34	65.11	75.33	84.48	90.88
NoEASE	2.07	2.34	2.64	2.92	3.21	58.79	70.29	80.24	87.83	92.57
VEASE-word	2.16	2.45	2.72	2.99	3.25	66.26	75.11	82.57	88.75	92.98
VEASE-phone	2.14	2.42	2.69	2.96	3.23	66.29	74.89	82.22	88.45	92.79
VEASE-place	2.21	2.47	2.73	3.00	3.26	66.57	75.27	82.64	88.80	92.96

Tab. 2. Average performance comparison of VEASE models with different visual embeddings

VEASE-place not only yields remarkable gains over VEASE-phone but also outperforms VEASE-word (LRW).

The high correlation between the articulation place label and the acoustic information in video is beneficial to the extraction of visual embedding, which is useful for speech enhancement, even if no requirement of additional data.

Experiment and Result Analyses

Model	PESQ					STOI (in %)				
SNR (in dB)	-5	0	5	10	15	-5	0	5	10	15
Noisy	1.70	1.97	2.26	2.56	2.86	54.34	65.11	75.33	84.48	90.88
NoEASE	2.07	2.34	2.64	2.92	3.21	58.79	70.29	80.24	87.83	92.57
VEASE-place	2.21	2.47	2.73	3.00	3.26	66.57	75.27	82.64	88.80	92.96
AEASE	2.09	2.39	2.69	2.98	3.27	60.84	72.24	81.58	88.39	92.76
MEASE	2.29	2.59	2.88	3.16	3.42	68.96	77.64	84.43	89.99	93.64

Tab. 3. Average performance comparison of NoEASE, VEASE, AEASE and MEASE models

MEASE shows significant improvements over VEASE across all evaluation metrics, and larger gains are observed at high SNRs.

But we cannot observe that AEASE performs better than VEASE at high SNRs

Experiment and Result Analyses

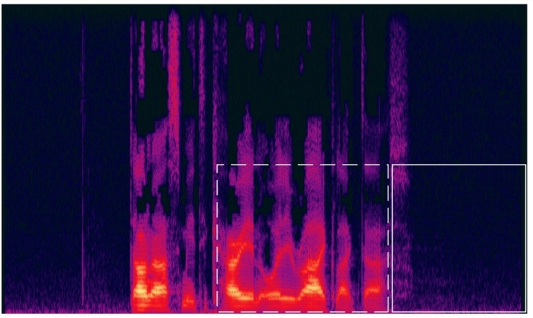
SNR	-5 dB				0 dB				5 dB			
	Model											
Place	NoEASE	AEASE	VEASE	MEASE	NoEASE	AEASE	VEASE	MEASE	NoEASE	AEASE	VEASE	MEASE
Labial	1.28	1.38	1.58	1.76	1.57	1.75	1.81	2.06	2.05	2.18	2.23	2.50
Mid	1.54	1.68	1.86	2.02	2.03	2.21	2.29	2.45	2.58	2.72	2.73	2.96
High	1.38	1.52	1.65	1.81	1.79	1.95	1.99	2.17	2.28	2.39	2.42	2.62
Low	1.63	1.89	2.00	2.29	2.17	2.48	2.46	2.69	2.84	2.99	2.93	3.20
Retroflex	1.46	1.66	1.75	2.00	1.95	2.15	2.12	2.32	2.44	2.57	2.54	2.77
Coronal	1.59	1.74	1.80	1.93	1.92	2.07	2.05	2.23	2.30	2.39	2.35	2.56
Glottal	1.02	1.22	1.36	1.70	1.42	1.71	1.59	1.92	1.95	2.10	2.05	2.30
Velar	1.31	1.44	1.41	1.49	1.48	1.64	1.68	1.86	1.86	2.01	2.00	2.22
Dental	0.94	1.22	1.25	1.64	1.32	1.62	1.36	2.05	1.98	2.21	1.98	2.44

Tab. 4. Average performances of different models on the test set at different SNRs and different articulation places

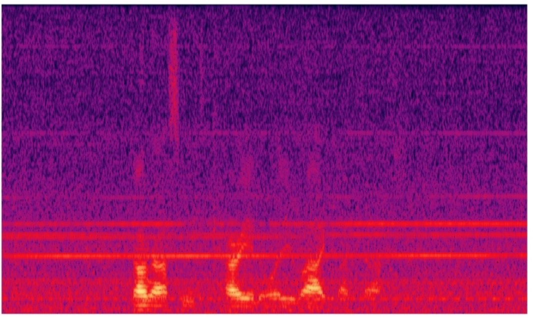
The complementarity between audio and visual embeddings lies in different SNR levels, as well as different articulation places.

More specifically, in the cases where the SNR level is low and the articulation place has high visual correlation, visual embedding performs better.

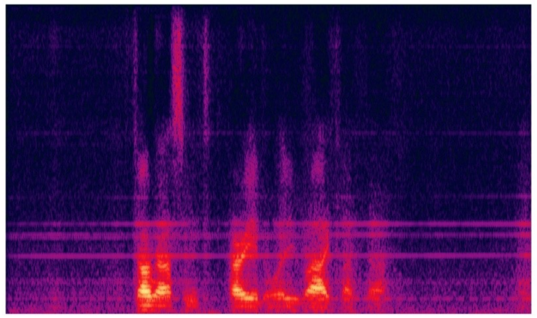
Spectrum Analyses



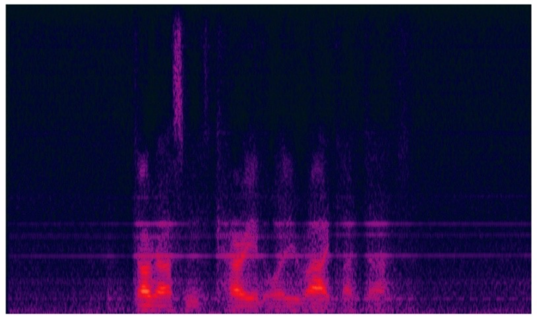
Clean



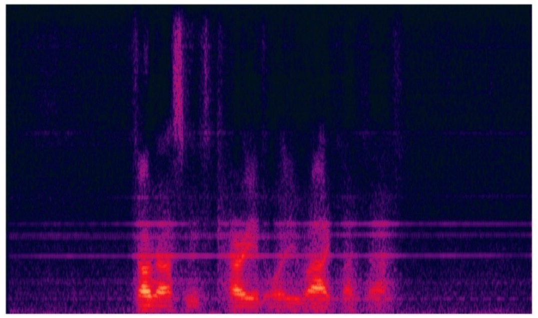
Mixture (5 dB)



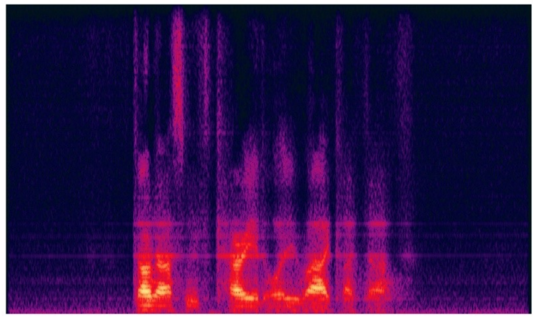
NoEASE (5 dB)



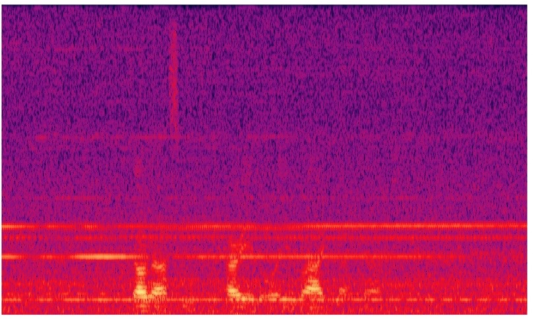
VEASE (5 dB)



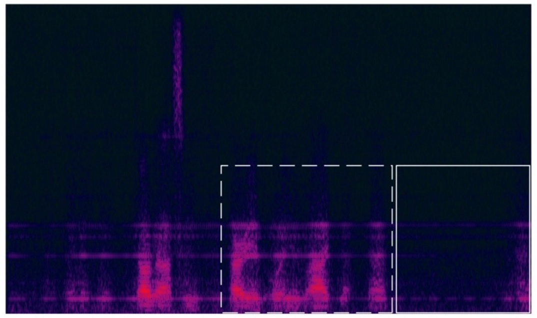
AEASE (5 dB)



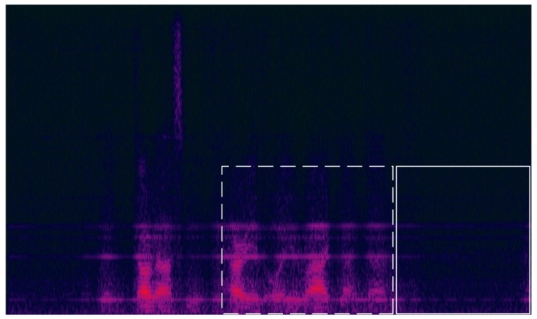
MEASE (5 dB)



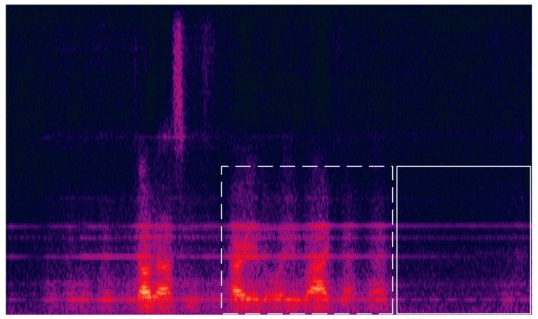
Mixture (0 dB)



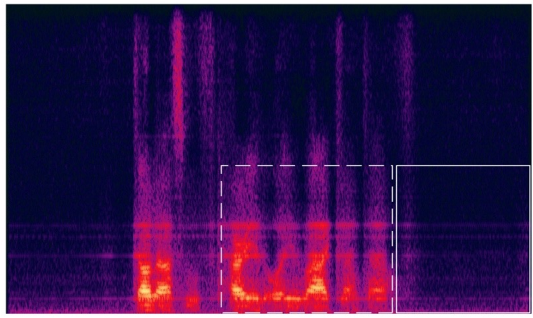
NoEASE (0 dB)



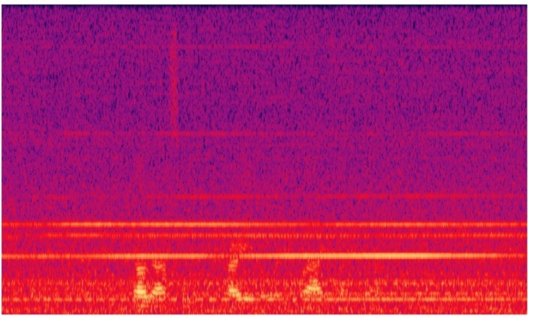
VEASE (0 dB)



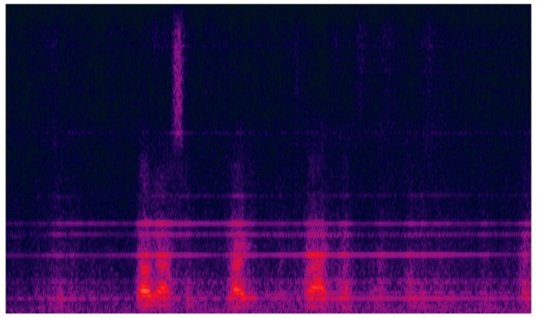
AEASE (0 dB)



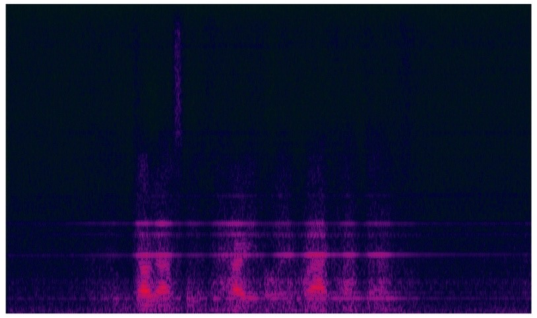
MEASE (0 dB)



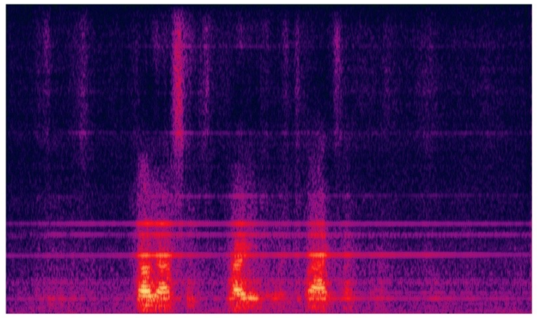
Mixture (-5 dB)



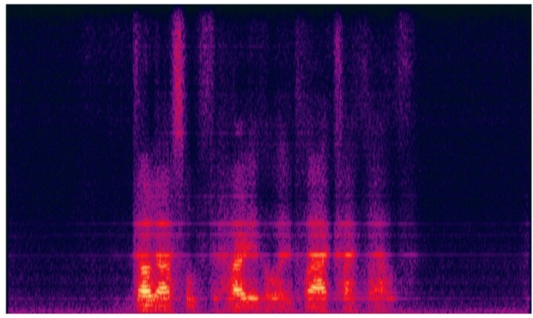
NoEASE (-5 dB)



VEASE (-5 dB)



AEASE (-5 dB)



MEASE (-5 dB)

Recognition Models

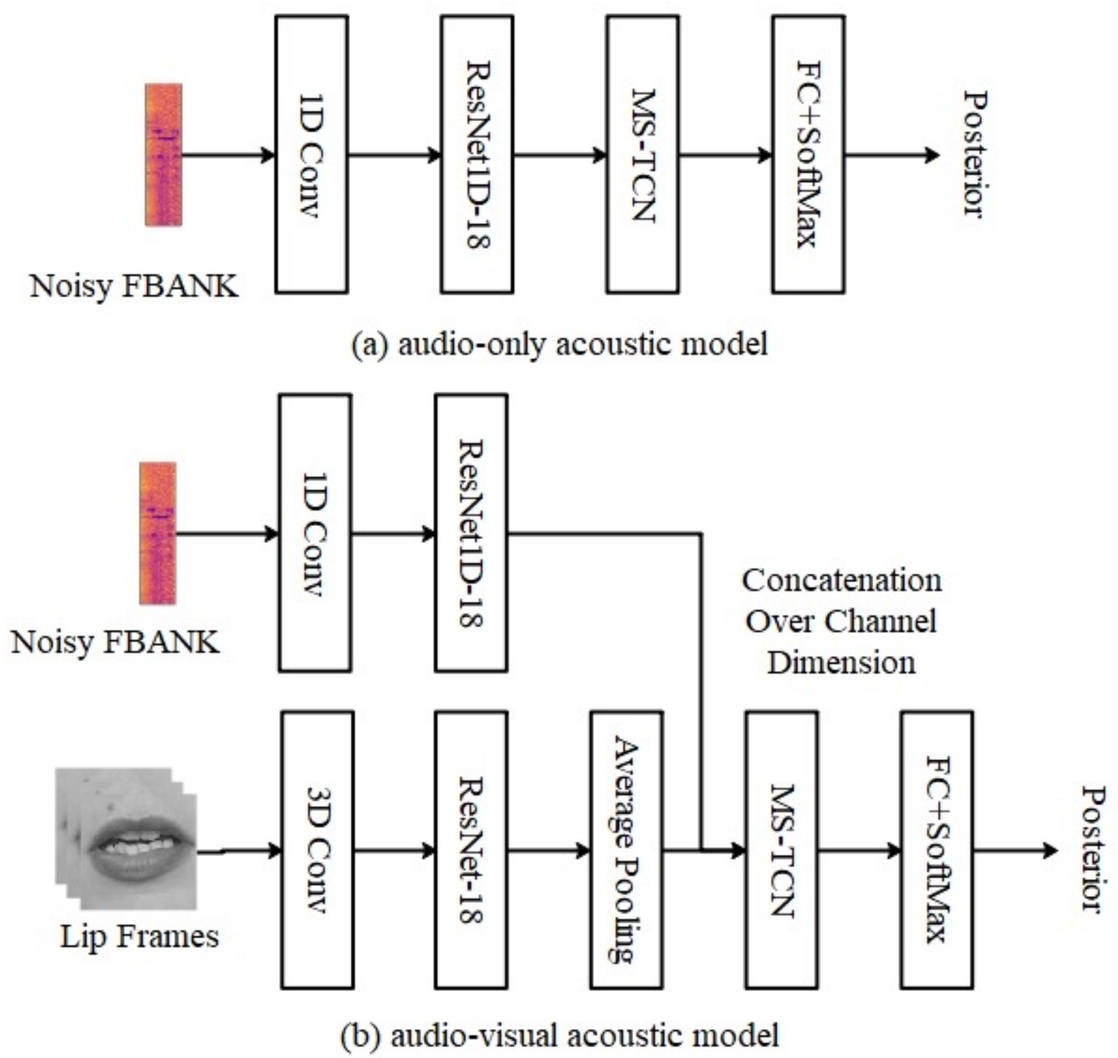


Fig.8. Illustration of Acoustic model

Acoustic model： Resnet18+MS-TCN
Language model： Phone-based bigram
Alignment： 2008 senones
Output： 39 phone sequence
Metric： Phone error rate (PER)

Train processing:
Training and alignment with mono phone
Training and alignment with triphone, using delta and delta feature
Training and alignment with triphone, using LDA and MLLT
Training and alignment with triphone, using SAT
Training NN-based acoustic model and decoding with HMM in 4

Recognition Results

	PER of ASR					PER of AVSR				
SNR	-5	0	5	10	15	-5	0	5	10	15
Raw	74.61	62.75	47.35	36.71	29.72	50.36	41.68	34.39	28.76	25.82
NoEASE	73.45	60.89	46.23	35.01	28.82	54.25	43.59	34.87	28.21	25.69
VEASE-place	64.54	51.69	39.17	31.26	26.69	51.42	42.02	33.57	28.03	25.16
MEASE	56.32	45.33	35.16	29.33	25.60	47.09	38.18	30.78	26.46	24.70

MEASE not only yields remarkable gains in PER of ASR but also in PER of AVSR.

Q&A