Audio-Visual Learning

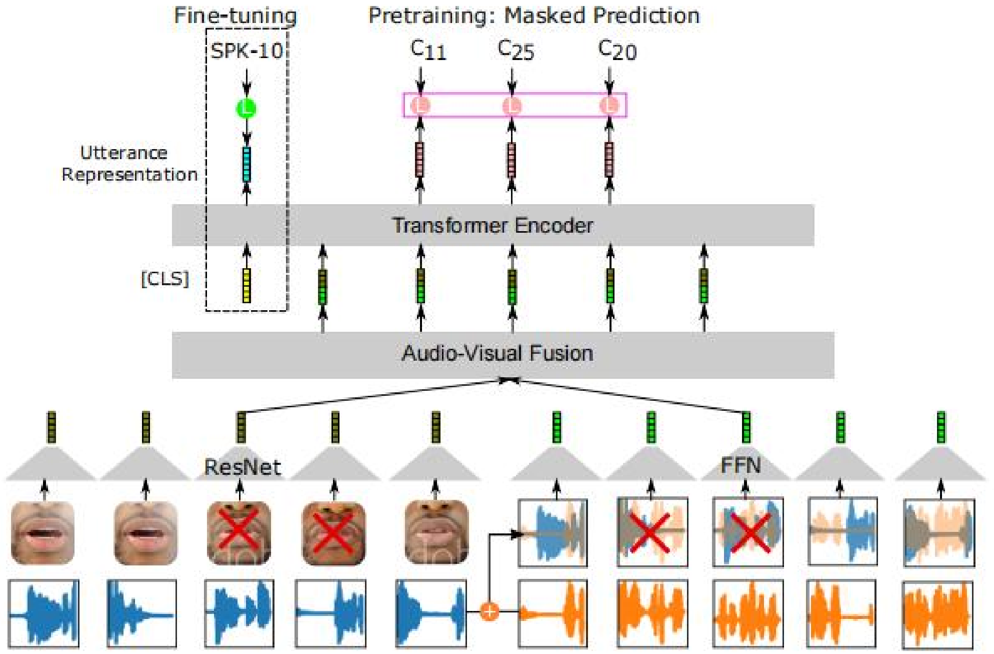
HaoYu Jiang

2022/09/23

Learning Lip-Based Audio-Visual Speaker Embeddings with AV-HuBERT

- Task
  - self-supervised pre-training for audiovisual speaker representation learning
- Motivation
  - several analyses observed that AV-HuBERT still learn rich speaker information especially in earlier layers
- Methods

Figure 1: *AV-HuBERT for learning speaker embedding. Dashed box: added during fine-tuning.*



- Experiments

Table 1: *SV performance on clean and noisy test sets when fine-tuned with various VC2 subsets. The EER averaged over 20 setups (5 SNRs × 4 types) is reported for the noisy test sets.*

PT	FT	Mod.	VC2 EER (%) clean	VC2 EER (%) noisy	VC1 EER (%) clean	VC1 EER (%) noisy
None	VC2-15spk (5h)	A	26.8	39.2	25.1	39.2
None		AV	29.8	35.9	24.6	28.7
VC2+LRS3		A	23.3	33.9	20.0	33.0
VC2+LRS3		AV	22.6	28.0	19.4	21.9
None	VC2-156spk (50h)	A	18.5	34.5	16.1	34.6
None		AV	16.4	24.7	13.1	17.7
VC2+LRS3		A	11.8	28.9	9.4	29.1
VC2+LRS3		AV	9.3	18.8	7.8	12.5
None	VC2-1200spk (485h)	A	11.1	31.6	8.6	30.5
None		AV	9.3	17.6	7.0	9.9
VC2+LRS3		A	7.2	26.1	4.9	25.2
VC2+LRS3		AV	5.7	12.6	3.8	6.1
None	VC2-5h (1740spk)	A	24.4	39.4	21.7	39.5
None		AV	32.8	41.0	30.2	40.3
VC2+LRS3		A	20.1	34.7	17.7	34.5
VC2+LRS3		AV	16.7	28.6	13.9	23.0
None	VC2-50h (5113spk)	A	20.2	35.5	16.1	34.7
None		AV	21.5	26.3	15.7	16.4
VC2+LRS3		A	10.7	29.7	8.0	28.7
VC2+LRS3		AV	7.4	19.8	4.8	11.4
None	VC2-500h (5992spk)	A	10.6	33.1	8.0	31.4
None		AV	6.5	14.5	5.3	7.8
VC2+LRS3		A	4.9	23.7	3.0	22.8
VC2+LRS3		AV	3.7	9.2	1.7	3.9
None	VC2 (5994spk)	A	7.3	29.2	5.1	27.8
None		AV	5.1	11.3	2.9	4.7
VC2+LRS3		A	3.4	20.9	1.9	20.0
VC2+LRS3		AV	2.4	7.8	1.0	2.5

Table 2: *AV-HuBERT fine-tuned on VC2-500h with audio (A) or audio-visual (AV) input, with or without noise augmentation. The abbreviations used are B: Babble, S: Speech, M: Music, and O: Other.*

Noise Aug?	Noise Type	A, VC1 EER (%), SNR (dB)= -10	-5	0	5	10	AV, VC1 EER (%), SNR (dB)= -10	-5	0	5	10
N	B	48.2	36.4	18.5	9.6	6.0	4.4	3.9	3.4	2.6	2.2
	S	48.8	46.5	36.5	18.3	8.5	8.6	6.8	4.8	3.4	2.5
	M	39.3	26.9	14.5	8.3	5.5	6.2	4.3	3.1	2.4	2.0
	O	34.0	23.3	13.7	8.8	5.9	6.0	4.3	3.2	2.6	2.3
Y	B	48.1	27.2	12.7	7.3	5.2	3.4	3.2	2.5	2.2	2.0
	S	24.4	14.9	11.8	12.3	9.6	3.2	2.8	2.6	2.3	2.0
	M	27.3	14.3	8.2	5.6	4.4	3.5	2.8	2.4	2.0	1.8
	O	23.6	13.0	8.0	5.8	4.7	3.1	2.6	2.3	2.1	2.0

Table 3: *Comparing different input. AV-HuBERT is fine-tuned on VC2-500h.*

Model	Input	VC2 EER (%)
AV-HuBERT	audio + face video	2.8
AV-HuBERT	audio + lip video	3.7

- Results

Table 4: *(Top) Comparison with the prior work following the SUPERB fine-tuning protocol. Models are fine-tuned on VC1. (Bottom) Comparison with prior work that does not follow the SUPERB evaluation protocol.*

Method	PT Data	Mod.	VC1 SC-Acc	VC1 SV-EER
FBANK [15]	-	A	8.5E-4	9.56
wav2vec2-B [15]	LS (960 hr)	A	75.18	6.02
HuBERT-B [15]	LS (960 hr)	A	81.42	5.11
WavLM-B [25]	Mix (94k hr)	A	89.42	4.07
wav2vec2-L [15]	LL (60k hr)	A	86.14	5.65
HuBERT-L [15]	LL (60k hr)	A	90.33	5.98
WavLM-L [25]	Mix (94k hr)	A	95.49	3.77
AV-HuBERT-B	VC2+LRS3 (2.8k hr)	A	80.99	5.85
AV-HuBERT-B		AV	93.90	4.85
AV-HuBERT-L		A	91.56	4.42
AV-HuBERT-L		AV	98.06	2.95

Method	PT Data	FT Data	Mod.	VC1 SC-Acc	VC1 SV-EER	VC2 SV-EER
Nagrani et al. [18]	20% VC2	VC1	AV	-	9.43	-
WavLM-L [25]	Mix (94k hr)	VC2	A	-	0.38	-
Shon et al. [32]	-	VC2	AV	-	-	5.29
Unimodal [33]	-	VC2	A	-	2.2	3.5
Multi-view [33]	-	VC2	AV	-	1.8	2.4
Feature Fusion [33]	-	VC2	AV	-	1.4	2.0
Ensemble [33]	-	VC2	AV	-	0.7	1.6
AV-HuBERT-B	VC2+LRS3 (2.8k hr)	VC2	A	-	1.92	3.43
AV-HuBERT-B		VC2	AV	-	1.00	2.41
AV-HuBERT-L		VC2	A	-	1.71	3.11
AV-HuBERT-L		VC2	AV	-	0.84	2.29

- Motivation
  - Unlike works that simply focus on the lip motion, we investigate the contribution of entire visual frames (visual actions, objects, background etc.)
- Methods
  - AudioVisual ASR TrAnsformeR (AVATAR)
    - Audiovisual Encoder： MBT architecture,a transformer based multimodal encoder.
    - Decoder： auto-regressive transformer decoder consisting of 8 layers and 4 attention heads.

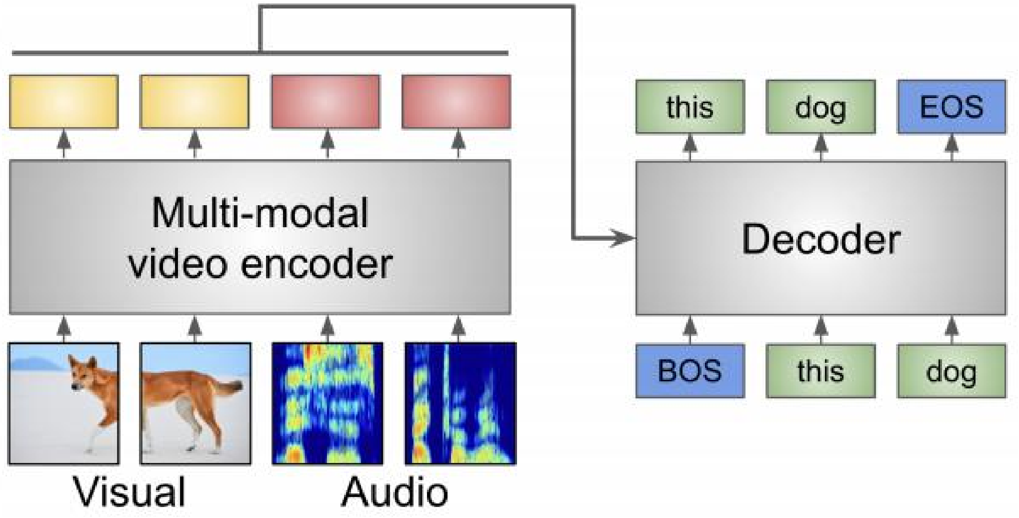


Figure 1: ***AVATAR:*** *We propose a Seq2Seq architecture for audio-visual speech recognition. Our model is trained end-to-end from RGB pixels and spectrograms.*

- VisSpeech Dataset
  - a subset of the publicly released HowTo100M dataset, and is curated using a combination of automatic filtering stages and manual verification.
  - VisSpeech consists of 508 segments from 495 unique videos.
  - with the audio containing background chatter, laughter, music and other environmental sounds.
  - many examples contain speech spoken with challenging English accents from various regions all over the world.
- Results

Table 1: *Audiovisual ASR vs Audio only models under various evaluation noise conditions (Clean, Burst, Environment and Mixed) and with different training masking strategies (Random and Content). Percentage Word Error Rate (%WER) is reported on the How2 test set.* **A:** *Audio-only.* **A+V:** *Audiovisual.* **Rel. Δ:** *Relative improvement of A+V over A.*

Training \ Eval Noise	Clean			Burst Loss			Environment Noise			Mixed Noise		
	A	**A+V**	**Rel. Δ**	**A**	**A+V**	**Rel. Δ**	**A**	**A+V**	**Rel. Δ**	**A**	**A+V**	**Rel. Δ**
No Pretraining	15.72	15.62	0.64%	29.59	28.69	3.05%	50.79	47.70	6.08%	60.51	57.49	5.0%
Vanilla	9.75	9.79	-0.33%	21.97	21.71	1.19%	25.97	25.55	1.61%	39.13	38.96	0.42%
Random Word Masking	9.19	9.11	0.93%	15.60	15.28	2.05%	23.39	22.35	4.45%	32.43	30.64	5.50%
Content Word Masking	9.58	9.25	3.48%	17.26	16.92	1.98%	23.77	22.67	4.65%	33.83	32.26	4.53%

AVATAR: Unconstrained Audiovisual Speech Recognition

- Results

Table 2: ***Comparison to the state-of-the-art on How2.*** *Our model outperforms all previous works when trained from scratch, and pretraining provides a significant boost. We report the best audio-visual numbers for all works.*

Model	%WER
BAS [10]	18.0
VAT [11]	18.0
MultiRes [17]	20.5
LLD [13]	16.7
AVATAR (scratch)	15.6
AVATAR (pretrained)	**9.1**

Table 3: *WERs of AVATAR on our newly introduced test set VisSpeech consisting of real-world noise. The models are trained on automatic ASR from HowTo100M, and finetuned on How2. Note here we do not add any artificial audio degradation at all.*

Training Strategy	A	A+V	Rel. Δ
No pretraining	44.57	43.41	2.61%
Vanilla	12.69	11.91	6.11%
Random Word Masking	12.35	11.86	3.93%
Content Word Masking	12.72	11.28	11.30%

GT: this dessert definitely deserves a happy dance
A: this deserves definitely deserves a happy dance
AV: this dessert definitely deserves a happy dance

GT: the thumb reaches for the coin
A: the thumb reaches for the con
AV: the thumb reaches for the coin

GT: this is a globe eggplant it's a small one
A: this is a glow big plant it's a small one
AV: this is a globe eggplant it's a small one

GT: and repeat the same fold for the opposite side
A: and repeat the simple for the opposite side
AV: and repeat the same fold for the opposite side

Figure 2: ***Qualitative results on the VisSpeech dataset.*** *We show the ground truth (GT), and predictions from our audio only (A) and audio-visual model (A+V). Note how the visual context helps with objects ('desert', 'coin', 'eggplant'), as well as actions ('fold') which may be ambiguous from the audio stream alone. Errors in the predictions compared to the GT are highlighted in red.*

- Motivation
  - Unlike most existing audio-visual methods, our audio-visual model takes audio features (e.g., FBANKs), multi-speaker lip regions of interest (ROIs), and multi-speaker i-vector embbedings as multimodal inputs
- Methods

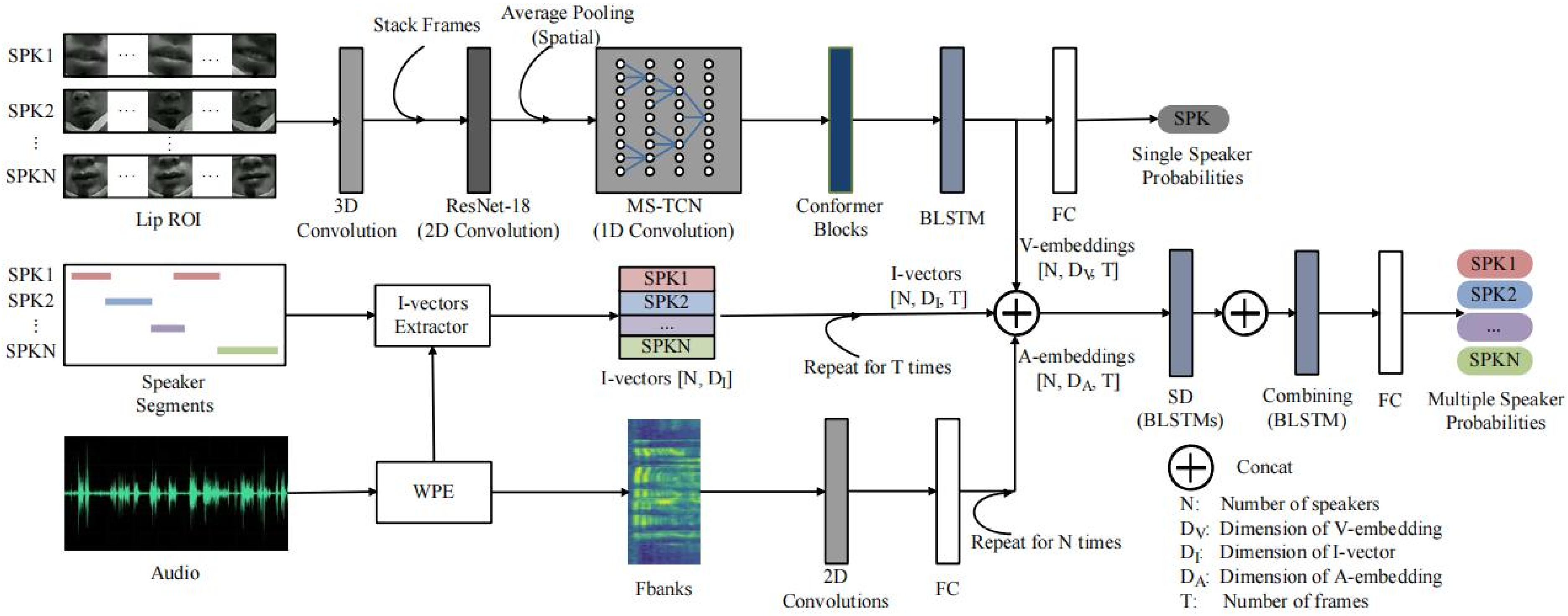


Figure 1: *The illustration of network structure*

End-to-End Audio-Visual Neural Speaker Diarization

- Results

Table 1: *Diarization results on MISP*

Set		DEV					EVAL				
Reference VAD		with				w/o	with				w/o
Modality	**System**	**FA**	**MISS**	**SpkErr**	**DER**	**DER**	**FA**	**MISS**	**SpkErr**	**DER**	**DER**
Audio	VBx	0.00	25.79	7.56	33.35	40.21	0.00	26.25	7.44	33.69	40.82
	TS-VAD	4.30	11.94	11.67	27.91	-	4.27	12.77	11.92	28.95	-
Visual	VSD	4.91	6.74	2.94	14.59	20.63	4.20	6.60	2.28	13.07	19.64
Audio-Visual	[1]AVSD w/o i-vector	3.60	5.06	2.38	11.04	-	2.35	5.90	1.80	10.05	-
	[2]AVSD with i-vector	3.41	5.05	2.10	10.57	-	3.07	5.39	1.56	10.01	-
	[3] + Joint training	3.32	4.67	2.14	10.12	11.68	2.96	4.97	1.56	9.49	10.99
Fusion	DOVER-Lap of 1, 2, 3	2.98	4.68	2.05	9.71	-	2.38	5.09	1.38	8.85	-

diarization error rate (DER) which is calculated as: the summed time of three different errors of false alarm (FA), missed detection (MISS) and speaker errors (SpkErr) divided by the total duration time.

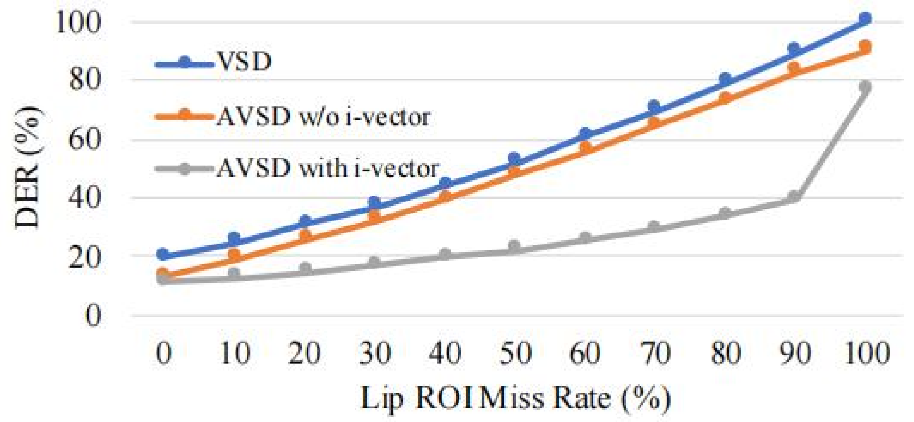


Figure 2: *DER comparison of different lip ROI missing rates without reference VAD on MISP EVAL set.*

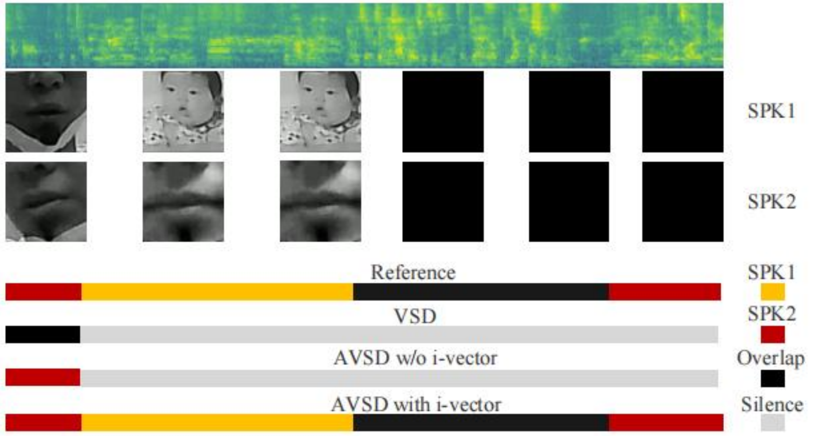


Figure 3: *An example including lip wiggling and lip missing problems in audio-visual recording.*

Thanks