서강대학교 SOGANG UNIVERSITY

Audio-Visual Speech Recognition Based on Joint Training with Audio-Visual Speech Enhancement

2021. 12. 15

황정욱
지능정보처리 연구실

목차

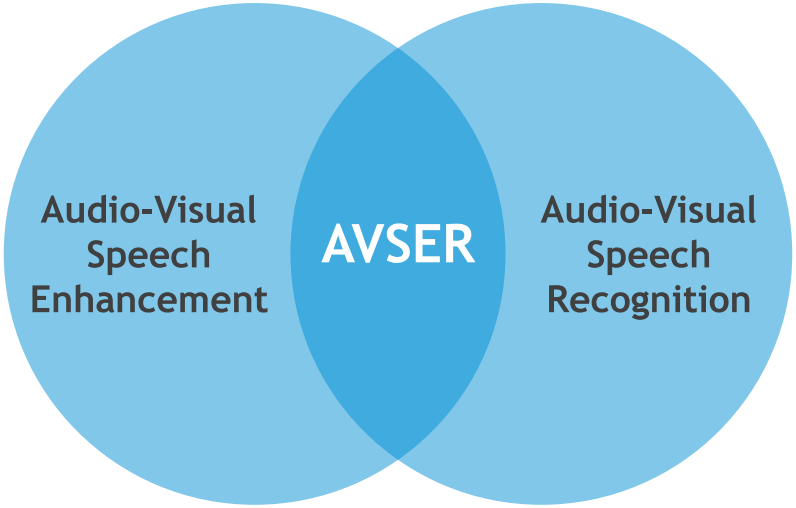
연구배경

- **강인 음성인식을 위한 시청각 음성향상 모델을 포함한 시청각 음성인식**
  - 음성인식을 위한 중요한 정보는 주로 청각 신호에 집중되어 있으며, 시각 정보는 잡음이 많은 환경에서 오디오 신호가 훼손된 경우 인식의 견고성을 높이는 역할임
  - 잡음 환경에서 기존 시청각 음성인식 모델의 인식 성능 향상에는 한계 존재
  - 따라서 시청각 음성향상 모델을 활용해 잡음을 제거하고 음성인식을 수행한다면 잡음 수준이 심한 환경에서도 강인 음성인식을 달성 가능

연구배경

- **효율적인 시청각 음성향상 및 음성인식 융합 모델**
  - 시청각 음성향상 모델과 시청각 음성인식 모델을 융합하여 학습 해야함
  - 따라서 각 모델들과 직접적으로 관련된 목적 함수를 결합하여 새로운 다중 목적 함수를 최적화하도록 공동으로 학습

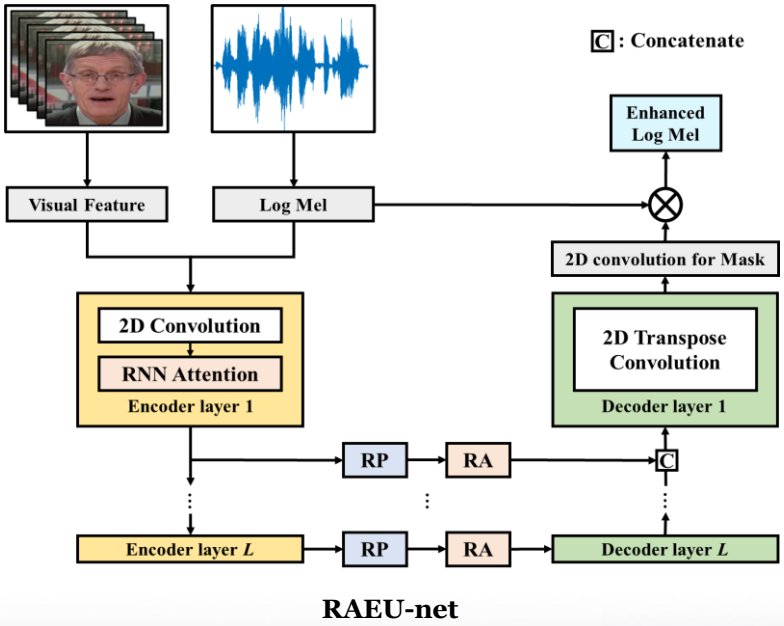
- U-net
- Audio-Visual
- Acoustic Echo Cancellation
- Dereverberation



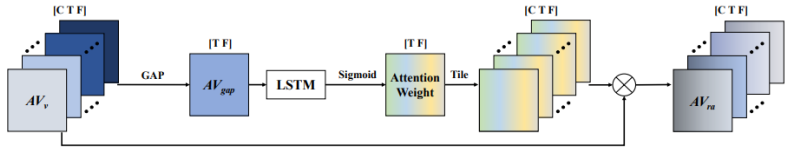
- Seq2seq model, transformer model
- Audio-Visual
- Automatic Speech Recognition
- Lip Reading

관련 연구 내용 소개

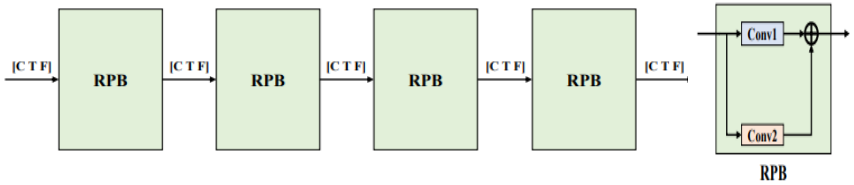
- **AVSE as the First Stage, RAESU-net [15]**
  - 시청각 정보를 모두 활용하여 배경잡음이 섞인 audio로 부터 잡음을 제거하는 모델
    - 시각 정보를 사용함으로써 competing speech noise도 효과적으로 제거
  - U-net 구조를 기반으로 하여, RNN attention과 Res path를 사용한 RAESU-net을 제시
  - 시청각 정보의 early fusion을 통해 하나의 encoder를 사용해 효율적으로 잡음 제거
  - RNN attention는 음성의 고유한 주파수별 특성을 가진 효율적인 표현을 찾는 역할
  - Res path는 encoder-decoder 사이의 semantic gap 문제를 해결하는 역할



RAEU-net



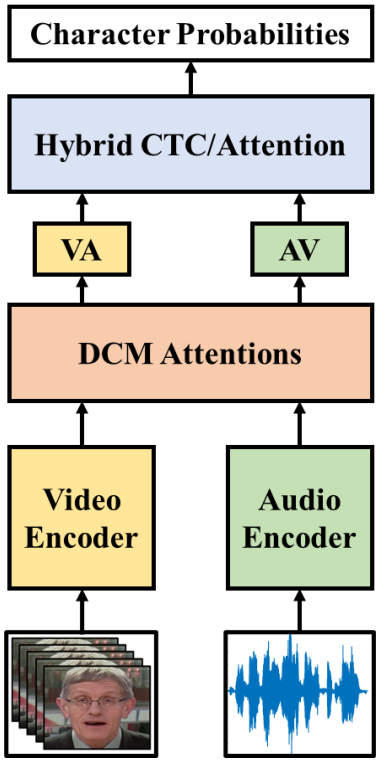
RNN Attention



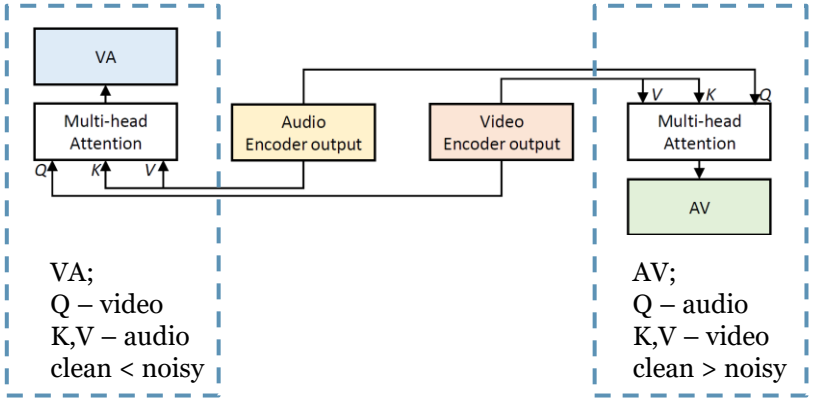
Res path

관련 연구 내용 소개

- **AVSR as the Second Stage, TM-DCM [7]**
  - 시청각 정보를 모두 활용하여 배경잡음이 섞인 청각 정보로부터 음성인식 견고성을 달성하는 모델
  - Transformer 구조를 기반으로 하여, DCM attention 활용 및 Hybrid CTC/attention 구조로 학습
    - DCM attention을 통해 음향 소음 수준에 따라 서로 다른 modality에 가중치 부여



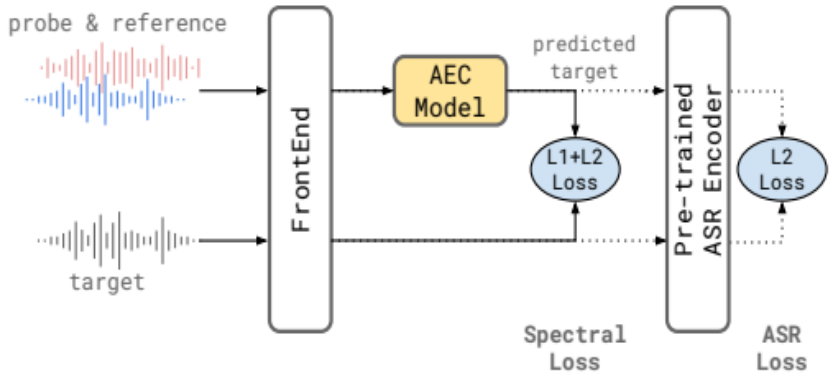
TM-DCM [H]



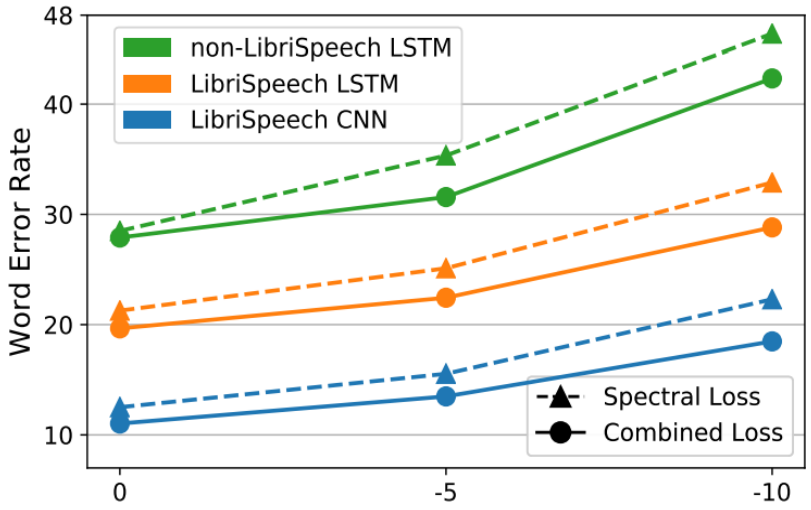
DCM Attention

기존 음성인식을 위한 Two-Stage 모델

- **Acoustic Echo Cancellation(AEC) for ASR [32]**
  - 음성인식 성능을 높이기 위해 AEC 모델과 ASR 모델을 결합함
  - 고정된 ASR 모델을 활용하여 ASR loss를 계산을 하여 다중 목적 함수를 통해 AEC 모델 학습
  - AEC 모델을 사용함으로써 ASR 모델의 인식 정확도가 증가함
  - 그러나, 주 목적이 AEC 모델을 학습하는데 있고, 새로운 ASR 모델을 학습하는 것이 아니라 한계 존재



AEC model with ASR

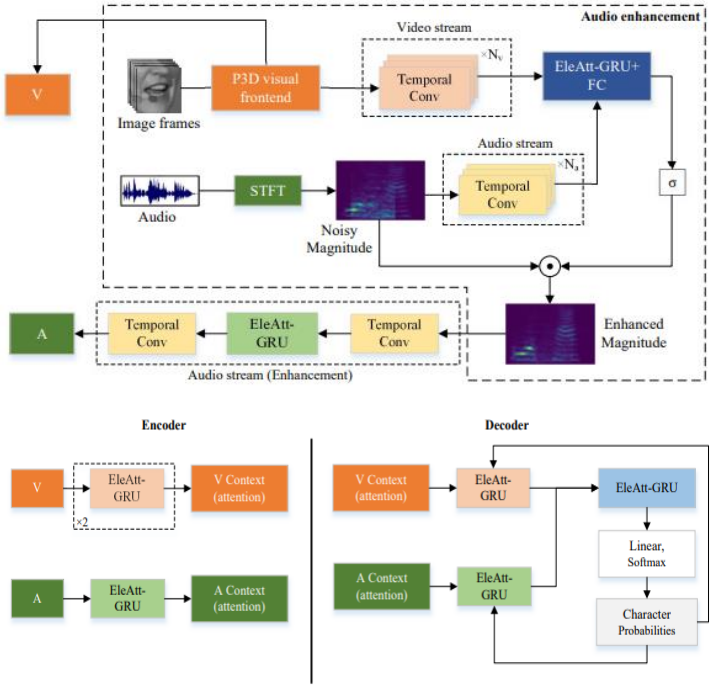


AEC model의 유무에 따른 WER 결과

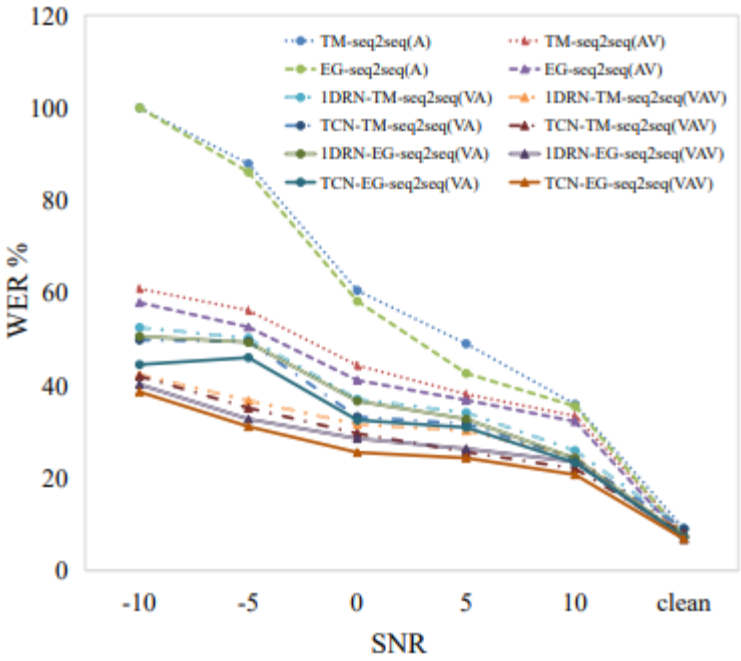
$$loss = loss_{spectral} + \lambda\, loss_{ASR}$$

기존 음성인식을 위한 Two-Stage 모델

- **Non-training AVSE + Training AVSR [33]**
  - 음성인식 성능을 높이기 위해 AVSE 모델과 AVSR 모델을 결합함
  - 고정된 AVSE 모델을 통해 얻어진 향상된 오디오 정보를 AVSR 모델 입력으로 사용하여 학습함
  - AVSE 모델을 사용함으로써, AVSR 모델의 인식 정확도가 크게 증가함
  - 그러나, AVSE 모델이 고정되어 있기에 유연하게 대응되지 않을 수 있음



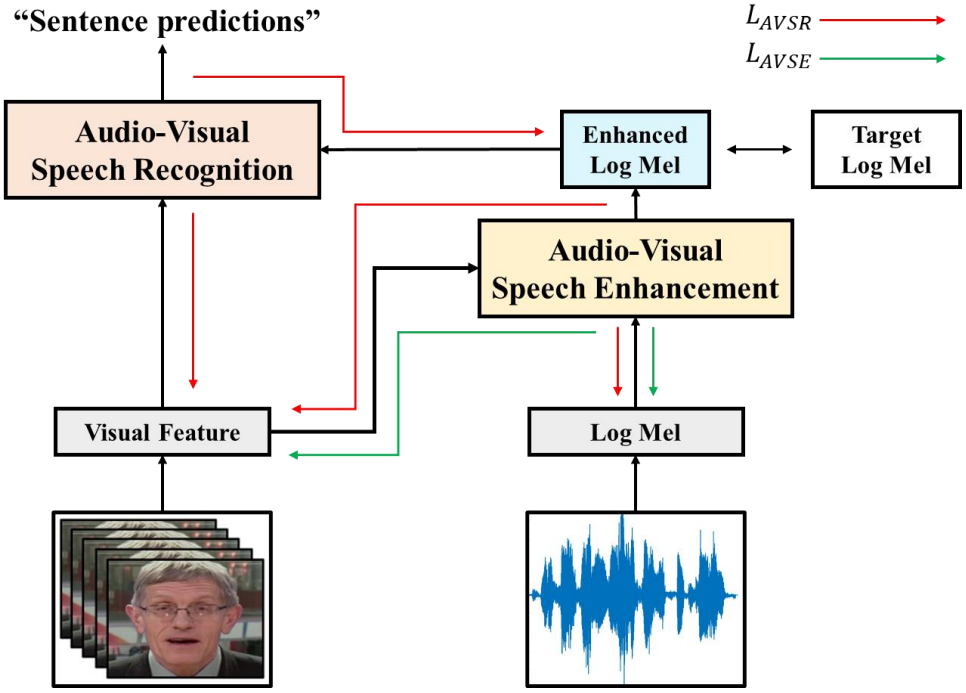
AVSE and AVSR model



AVSE와 AVSR 조합에 따른 실험결과

제안하는 Two-Stage 시청각 음성인식 모델

- **Trainable Two-stage AVSE [15] and AVSR [7] model**
  - AVSE 모델과 AVSR 모델을 모두 학습하는 Two-stage network 제시
  - 시청각 정보를 사용하여 AVSE 모델을 통해 입력 noisy audio의 log-mel feature를 Target log-mel feature와 비교하여 학습
  - AVSE 모델을 통해 향상된 log-mel feature와 시각 정보를 사용하여 AVSR 모델 학습



Proposed Two-stage network

$$loss = \lambda\, loss_{AVSE} + (1-\lambda) loss_{AVSR}$$

실험

- **실험 환경**
  - LRS2-BBC, LRS3-TED datasets
    - Oxford에서 공개한 audio-visual dataset으로 공개된 가장 큰 DB 중 하나
    - 제한된 환경에서 녹화하지 않고 기존 방송 촬영분을 편집한 in the wild dataset

Dataset	Source	Split	Dates	# Spk.	# Utt.	Word inst.	Vocab	# hours
GRID [7]	-	-	-	51	33,000	165k	51	27.5
MODALITY [8]	-	-	-	35	5,880	8,085	182	31
LRW [5]	BBC	Train-val	01/2010 - 12/2015	-	514k	514k	500	165
		Test	01/2016 - 09/2016	-	25k	25k	500	8
LRS2-BBC [6]	BBC	Pre-train	01/2010 - 02/2016	-	96k	2M	41k	195
		Train-val	01/2010 - 02/2016	-	47k	337k	18k	29
		Test	03/2016 - 09/2016	-	1,243	6,663	1,693	0.5
		Text-only	01/2016 - 02/2016	-	8M	26M	60k	-
LRS3-TED	TED & TEDx (YouTube)	Pre-train	-	5,090	119k	3.9M	51k	407
		Train-val	-	4,004	32k	358k	17k	30
		Test	-	451	1,452	11k	2,136	1
		Text-only	-	5,543	1.2M	7.2M	57k	-

- LRS2-BBC & LRS3-TED DB의 training data는 pre-train과 train-val 두 디렉토리로, 합치면 약 650 h 이상
- Test data는 약 1.5 h 으로 2,695개의 발화로 구성

실험

- **실험 환경**
  - Noisy reverberant mixing
    - 잡음환경에 강인한 음성인식을 위해 training data에 cafeteria, restaurant 환경에서 획득된 실잡음 데이터를 signal-to-noise ratios (SNRs) -5 ~ 20 dB로 mixing을 진행
    - 추가적으로 -15 dB와 -10 dB로 mixing하여 data을 test data로 사용

실험

- **학습 전략**
  - Step 1 : AVSE model
    - 1. **noisy** train-val set을 학습
  - Step 2 : AVSR model
    - 1. pre-train set을 3~4 단어로 분절하여 **clean short sentence** 학습
    - 2. pre-train & train-val set을 같이 **clean sentence** 학습
    - 3. clean과 미리 준비한 **noisy** train-val set을 학습
  - Step 3 : AVSE + AVSR joint training
    - 각각 학습된 AVSE, AVSR 모델의 최적의 상태인 pre-trained weight를 사용하여 joint training을 진행
    - λ 를 0, 0,1, 0,5, 0,9, 1로 선택하여 5가지 경우로 비교 학습 진행

$$\boldsymbol{loss = \lambda\, loss_{AVSE} + (1-\lambda) loss_{AVSR}}$$

실험

- **실험 평가 지표**
  - 단어 오류율 (Word Error Rate, WER)
    - $WER = \left(\frac{D+S+I}{N}\right) * 100\ (\%)$
      - 전체 단어의 수 : N, 잘못 인식한 단어 : S, 문장에 없는 단어를 인식 : I, 인식하지 못한 단어 : D

실험

- **실험 결과**
  - 각 λ에 따른 실험결과를 살펴보면 ASR 모델을 제외하고, 모든 AVSR 모델에서 λ = 0.9일 때 인식성능이 제일 좋음
  - 즉, 음성인식 성능을 향상시키기 위해서, AVSR 모델을 향상된 음성에 적응시키는 것보다 잡음이 많은 음성을 더욱 향상시키는 것이 더 중요
  - [32]의 학습 방법 (λ = 1.0, AVSE만 학습) 보다 효과적임을 알 수 있음

AVSR Model	Modality	Criterion	λ	Input SNR (dB)									
				Clean	20	15	10	5	0	-5	-10	-15	Avg.
TM-seq2seq	A	CE	0.0	7.9	7.9	8.7	9.1	10.9	13.4	22.9	42.5	73.5	21.9
			0.1	8.0	8.1	8.4	9.1	10.7	14.1	23.1	42.6	70.9	21.7
			0.5	7.9	8.2	8.3	9.1	10.1	13.5	22.3	42.9	72.3	21.6
			0.9	8.0	8.0	8.5	9.2	10.5	13.3	23.1	43.2	73.9	22.0
			1.0	8.2	8.3	8.5	9.1	10.5	15.4	24.6	47.5	76.8	23.2
TM-seq2seq	AV	CE	0.0	9.6	10.0	10	10.4	11.6	14.6	23.4	40.6	67.8	22.0
			0.1	9.1	9.5	9.9	10.3	11.6	15.1	23.5	40.9	68.3	22.0
			0.5	9.0	9.6	9.8	9.9	11.3	15.1	23.9	41.1	68.0	22.0
			0.9	9.1	9.3	9.4	10.0	11.7	15.3	23.1	40.8	67.4	21.8
			1.0	9.4	9.8	9.9	10.3	11.7	15.9	25.6	43.2	71.9	23.1
TM-DCM	AV	CE	0.0	7.8	7.8	7.8	8.3	9.7	12.8	19.6	37.1	65.4	19.6
			0.1	7.9	8.0	7.9	8.9	9.9	12.9	19.6	36.6	68.0	20.0
			0.5	7.8	7.5	7.8	8.4	10.1	12.6	19.5	37.6	65.7	19.7
			0.9	7.9	8.0	8.0	8.8	9.7	12.3	19.0	36.7	64.4	19.4
			1.0	8.4	8.5	8.4	9.2	10.8	15.7	24.8	45.7	72.3	22.6
TM-DCM	AV	H	0.0	7.4	7.5	7.5	7.9	9.5	12.0	**18.8**	**36.1**	63.6	18.9
			0.1	7.1	7.3	**7.3**	8.1	9.0	12.0	19.2	36.2	63.1	18.8
			0.5	7.1	7.2	7.4	7.6	9.2	12.2	18.9	36.3	63.8	18.9
			0.9	**7.0**	**7.1**	7.4	**7.5**	**8.7**	**11.9**	19.1	36.4	**62.1**	**18.6**
			1.0	7.7	7.6	8.1	8.4	10.3	14.5	24.7	44.7	70.7	21.9

실험

- **실험 결과**
  - AVSE 모델과 AVSR 모델을 서로 다른 4가지의 방법으로 통합한 실험 결과 비교
    - WF : pre-trained AVSR without AVSE
    - FF : pre-trained AVSE, pre-trained AVSR
    - FT : pre-trained AVSE, retraining AVSR
    - TT : retraining AVSE, retraining AVSR (λ = 0.9일 때)
  - 4가지 AVSR 모델 모두 우리가 제시한 TT 학습 방법이 제일 효과적인 것을 알 수 있음
  - [33]의 학습 방법 (FT) 보다 효과적임을 알 수 있음

AVSR Model	Modality	Criterion	Type	Input SNR (dB)									
				Clean	20	15	10	5	0	-5	-10	-15	Avg.
TM-seq2seq	A	CE	WF	8.1	8.4	8.8	10.3	13.4	20.5	36.7	63.3	90.3	28.9
			FF	9.3	9.3	9.9	10.2	12.3	19.0	32.3	58.0	85.4	27.3
			FT	8.7	8.8	9.1	9.6	11.2	14.9	25.2	47.7	79.6	23.9
			TT	8.0	8.0	8.5	9.2	10.5	13.3	23.1	43.2	73.9	22.0
TM-seq2seq	AV	CE	WF	8.7	9.2	9.7	10.8	13.2	20.1	34.7	57.4	80.5	27.1
			FF	10.0	10.1	10.2	11.3	13.4	18.4	30.9	53.9	79.9	26.5
			FT	9.4	9.3	9.7	10.5	11.6	15.3	24.5	44.2	72.7	23.0
			TT	9.1	9.3	9.4	10.0	11.7	15.3	23.1	40.8	67.4	21.8
TM-DCM	AV	CE	WF	8.2	8.3	8.9	9.6	12.0	17.5	30.4	55.0	81.9	25.6
			FF	8.7	8.7	9.3	9.5	11.4	16.4	27.4	52.4	80.8	25.0
			FT	8.1	8.3	8.3	8.8	10.0	13.3	21.2	39.6	70.3	20.9
			TT	7.9	8.0	8.0	8.8	9.7	12.3	19.0	36.7	64.4	19.4
TM-DCM	AV	H	WF	7.5	7.3	8.0	8.7	11.1	16.3	30.9	53.7	78.3	24.6
			FF	8.1	8.1	8.3	8.9	10.5	15.9	28.0	51.5	79.0	24.3
			FT	7.5	7.5	8.0	8.1	9.6	12.8	20.5	40.3	69.1	20.4
			TT	**7.0**	**7.1**	**7.4**	**7.5**	**8.7**	**11.9**	**19.1**	**36.4**	**62.1**	**18.6**

실험

- **실험 결과**
  - Decoding results for -5 dB using TM-DCM model with the hybrid CTC/attention architecture
  - 우리가 제시한 학습 방법 (TT)으로 학습한 모델이 인식 정확도가 제일 높음을 알 수 있음
  - AVSE 모델을 사용하지 않은 AVSR 모델은 음성의 손상으로 인해 정확도가 떨어짐

Type	Transcription
Ground truth	like hundreds of thousands of people do every year
WF	like hundreds of thousands of **being more** do **everything**
FF	like hundreds **and** thousands of people do **over the** year
TF	like hundreds **and** thousands of people do every year
TT	like hundreds of thousands of people do every year
Ground truth	we might say then well let's not worry about this
WF	we might say then **what I love to** worry about this
FF	we might say then well let's **untold what** about this
TF	we might say then well let's not **what** about this
TT	we might say then well let's not worry about this
Ground truth	and then I thought there's got to be a better way
WF	and then I thought **this** got to be a better way
FF	and then I thought **this** got to be a better way
TF	and then I thought **this** got to be a better way
TT	and then I thought there's got to be a better way

결론 및 추후 과제

- **결론**
  - 시청각 정보를 기반으로 AVSE를 수행하여 향상된 청각 정보와 시각 정보를 사용한 end-to-end 모델을 제시
  - 특히, AVSE 모델과 AVSR 모델을 joint training 하는 방법을 제시
  - 다양한 인식 환경에서 기존의 방법보다 우수한 성능을 나타냄을 확인

- **추후 과제**
  - 모델 경량화 알고리즘 적용 모색
  - 실시간 음성인식이 가능하도록 효율적인 인식 모델 모색

감사합니다.

참고문헌

[1] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina et al., "State-of-the-art speech recognition with sequence-to-sequence models," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018, pp. 4774-4778.

[2] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018, pp. 5884–5888.

[3] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," arXiv:1904.03288, 2019.

[4] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," arXiv:2005.03191, 2020.

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu et al., "Conformer: Convolution-augmented transformer for speech recognition," arXiv:2005.08100, 2020.

[6] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.

[7] Y.-H. Lee, D.-W. Jang, J.-B. Kim, R.-H. Park, and H.-M. Park, "Audio–visual speech recognition based on dual cross-modality attentions with the transformer model," Applied Sciences, vol. 10, no. 20, p. 7263, 2020.

[8] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid CTC/attention architecture," in Proc. IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 513–520.

참고문헌

[9] K. Tan and D. Wang, “Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019, pp. 6865–6869.

[10] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, “Phase-aware speech enhancement with deep complex u-net,” arXiv:1903.03107, 2019.

[11] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement,” in Proc. Interspeech, 2020, pp. 2472–2476.

[12] A. Pandey and D. Wang, “A New Framework for Supervised Speech Enhancement in the Time Domain.” in Proc. Interspeech, 2018, pp. 1136–1140.

[13] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019, pp. 6875–6879.

[14] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, “End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks,” IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 26, no. 9, pp. 1570–1584, 2018.

[15] J.-W. Hwang, R.-H. Park, and H.-M. Park, “Efficient Audio-Visual Speech Enhancement Using Deep U-Net With Early Fusion of Audio and Video Information and RNN Attention Blocks,” IEEE Access, vol. 9, pp. 137584–137598, 2021.

[16] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, “Cochleanet: A robust language-independent audio-visual model for speech enhancement,” Information Fusion, vol. 63, pp. 273-285, 2020.

참고문헌

[17] T. Afouras, J. S. Chung, and A. Zisserman, “The conversation: Deep audio-visual speech enhancement,” in Proc. Interspeech, 2018, pp. 3244–3248.

[18] T. Afouras, J. S. Chung, and A. Zisserman, “My lips are concealed: Audio-visual speech enhancement through obstructions,” arXiv:1907.04975, 2019.

[19] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, “Audio-visual speech enhancement based on multimodal deep convolutional neural network,” arXiv:1709.00944, 2017.

[20] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in Proc. IEEE Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3444–3453.

[21] D. Serdyuk, O. Braga, and O. Siohan, “Audio-visual speech recognition is worth 32×32×8 voxels,” arXiv:2109.09536, 2021.

[22] P. Ma, S. Petridis, and M. Pantic, “End-to-end audio-visual speech recognition with conformers,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021, pp. 7613–7617.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv:2010.11929, 2020.

[24] K. Tan, Y. Xu, S.-X. Zhang, M. Yu, and D. Yu, “Audio-visual speech separation and dereverberation with a two-stage multimodal network,” IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 3, pp. 542–553, 2020.

[25] M. Togami, “Joint training of deep neural networks for multi-channel dereverberation and speech source separation,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2020, pp. 3032–3036.

참고문헌

[26] Y. Zhao, Z.-Q. Wang, and D. Wang, “Two-stage deep learning for noisy reverberant speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 1, pp. 53–62, 2018.

[27] C. Fan, J. Tao, B. Liu, J. Yi, and Z. Wen, “Joint training for simultaneous speech denoising and dereverberation with deep embedding representations.” in Proc. Interspeech, 2020, pp. 4536-4540.

[28] J.-Y. Son and J.-H. Chang, “Attention-based joint training of noise suppression and sound event detection for noise-robust classification,” Sensors, vol. 21, no. 20, p. 6718, 2021.

[29] X. Tan and X.-L. Zhang, “Speech enhancement aided end-to-end multi-task learning for voice activity detection,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021, pp. 6823–6827.

[30] Y. Jung, Y. Kim, Y. Choi, and H. Kim, “Joint Learning Using Denoising Variational Autoencoders for Voice Activity Detection.” in Proc. Interspeech, 2018, pp. 1210–1214.

[31] T. Xu, H. Zhang, and X. Zhang, “Joint training ResCNN-based voice activity detection with speech enhancement,” in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 1157–1162.

[32] N. Howard, A. Park, T. Z. Shabestary, A. Gruenstein, and R. Prabhavalkar, “A Neural Acoustic Echo Canceller Optimized Using An Automatic Speech Recognizer and Large Scale Synthetic Data,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2021, pp. 7128–7132.

[33] B. Xu, C. Lu, Y. Guo, and J. Wang, “Discriminative multi-modality speech recognition,” in Proc. IEEE Computer Vision and Pattern Recognition (CVPR), 2020, pp. 14433–14442

참고문헌

[34] T. Stafylakis and G. Tzimiropoulos, “Combining residual networks with LSTMs for lipreading,” in Proc. Interspeech, 2017, pp. 3652–3656.

[35] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[36] N. Ibtehaz and M. S. Rahman, “MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation,” Neural Networks, vol. 121, pp. 74–87, 2020.

[37] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, “RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images,” IEEE Access, vol. 7, pp. 21420–21428, 2019.

[38] G. Sterpu, C. Saam, and N. Harte, “Attention-based audio-visual fusion for robust automatic speech recognition,” in Proc. ACM International Conference on Multimodal Interaction, 2018, pp. 111–115.