MODALITY ATTENTION FOR END-TO-END AUDIO-VISUAL SPEECH RECOGNITION

Pan Zhou[1], Wenwen Yang[2], Wei Chen[2], Yanfeng Wang[2], Jia Jia[1]

[1]Department of Computer Science and Technology, Tsinghua University, China

[2]Voice Interaction Technology Center, Sogou Inc. , Beijing, China

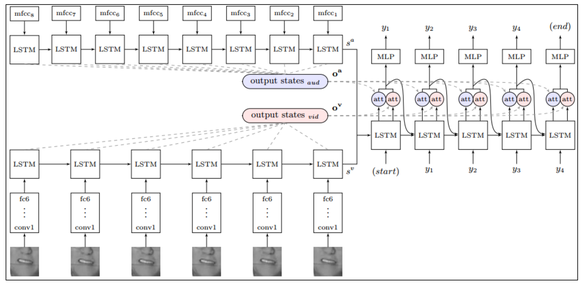
Introduction

- Humans understand speech not only by listening but also by considering visual cues of lips and faces.
- Audio-visual speech recognition (AVSR) is thought to be one of the most promising solutions for robust speech recognition in noisy conditions.
- End-to-end approaches, e.g. CTC, LAS, RNN-T show promising results in ASR.
- Watch, listen, attend and spell (WLAS) propose a framework to fuse information from audio and video.
- **Contribution:** using additional modality attention to learn fused representation of audio and video in sequence-to-sequence architectures for AVSR.
- Experiments show relative improvement from 2% to 36% over auditory modality alone are obtained depending on different SNR, which is better than feature concatenation methods.

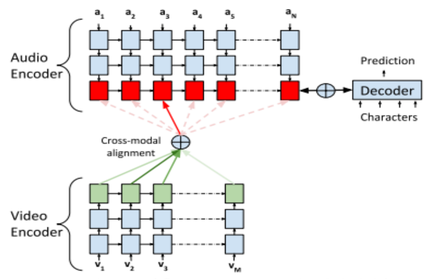
Attention based Encoder Decoder

- Encoder/listener: extract higher level acoustic representation.
- Decoder/speller: RNN net for predicting the output units.
- Attender: compute context vector using encoder output and decoder states for the decoder to output next unit

$$\boldsymbol{h} = Encoder(\boldsymbol{x})$$
$$s_i = DecoderRNN(s_{i-1}, y_{i-1}, c_{i-1})$$
$$e_{i,u} = Energy(s_i, h_u) = V^T tanh(W_h h_u + W_s s_i + b)$$
$$\alpha_{i,u} = \frac{exp(e_{i,u})}{\sum_{u'} exp(e_{i,u'})}$$
$$c_i = \sum_u \alpha_{i,u} h_u$$
$$P(y_i|\boldsymbol{x}, y_{<i}) = DecoderOut(s_i, c_i)$$



WLAS (Chung et al. 2017)



AV_Align (Sterpu et al. 2018)

Modality Attention

- **Scoring function over each modality feature**

$$z_t^m = Z(f_{1\ldots t}^m)$$

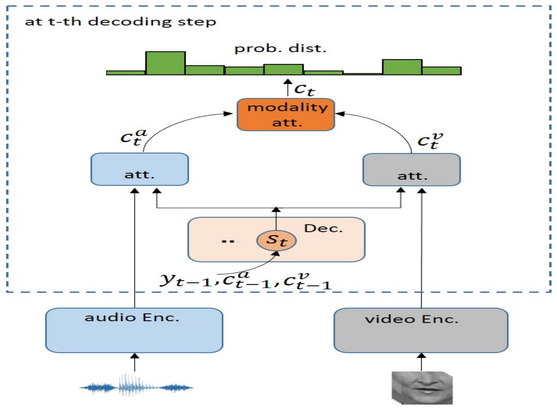
- **Softmax operation over modality scores at time t**

$$\alpha_t^m = \frac{exp(z_t^m)}{\sum_{j=1}^M exp(z_t^j)}$$

- **Weighted sum on modality feature for fused representation**

$$f_t^M = \sum_{m=1}^M \alpha_t^m f_t^m$$

- **Features of each modality supposed to have equal length**



Modality Attention (this work)

Modality Attention for AVSR

- **Encoders: audio encoder & video encoder**
- **Attenders: compute context vector c_t^a for audio and c_t^v for video at each decoding step**
- **Modality attender: attend over c_t^a and c_t^v to get merged vector c_t^{av}**
- **Output: c_t^{av} is used for generating output**

Table 1. Details for video encoder

CNN layer	operation	output size
0	Resize	Tx3x64x80
1-2	Conv-Selu-Conv-Selu-MP-BN	Tx32x32x40
3-4	Conv-Selu-Conv-Selu-MP-BN	Tx48x16x20
5-6	Conv-Selu-Conv-Selu-MP-BN	Tx72x8x10
7-8	Conv-Selu-Conv-Selu-MP-BN	Tx108x4x5
9-10	Conv-Selu-Conv-Selu-MP-BN	Tx128x2x2
11-12	BLSTM-BLSTM	Tx512

Experimental Setup

- **Broadcast TV news audio-visual data, 100 speakers, 150 hours of training set and 42 hours of test set**
- **Add Gaussian noise to audio**
- **71 fbank features extracted every 10ms**
- **Lip region resize to 64x80**
- **Audio encoder: 4 BLSTM with 256 cells**
- **Video encoder: 10CNN + 2 BLSTM**
- **Decoder: 1 LSTM with 512 cells**
- **Output units: 6784 Chinese characters, 26 English characters, SOS, EOS, UNK**

Results

Table 2. CER for different model at different SNR video encoder

	clean	10dB	5dB	0dB
LAS	7.08	10.33	12.93	18.65
WAS	44.62			
WLAS	7.00	9.07	10.23	12.34
AV_align	7.6	10.89	13.69	19.21
MD_ATT	6.95	8.54	9.87	11.93
MD_ATT_MC	6.85	8.12	9.74	13.65

Table 3. Attention weights of MD_ATT_MC for audio and video

test SNR	attention weights	
	α^a	α^v
clean	0.641	0.359
10dB	0.633	0.367
5dB	0.624	0.376
0dB	0.607	0.393

Table 4. Recognition results for comparison

Lab:	不少城市气温连创入冬以来新低	是新疆的风光和美食	与此同时韩国多个市民
LAS:	不少城市气温连创录中以来袭击	新疆的相关关系	与平时韩国政府不明
WAS:	不少人是因为他传入冬冬的拉心地	人心在的疯狂和美石	对此同时韩国多个市民
MD_ATT:	不少城市气温连创入冬以来新低	是新疆的风光和美食	与此同时韩国多个自民