Audiovisual Speech Activity Detection with Advanced Long Short-Term Memory

Fei Tao and Carlos Busso

Erik Jonsson School of Engineering & Computer Science at the University of Texas at Dallas, Richardson, Texas 75080, USA

INTERSPEECH 2018
SEPTEMBER 2-6 | HYDERABAD, INDIA
HYDERABAD INTERNATIONAL CONVENTION CENTRE

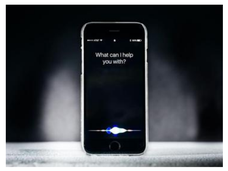
Motivation

Background:

- Speech activity detection (SAD) is an important pre-processing step in speech-based interfaces
- Introducing visual information can improve performance and robustness
- **Longer periodicity** in the acoustic and visual features is crucial to distinguish speech activity
  - Recurrent connections in LSTM only consider previous frame

Our Work:

- This study proposes to explore the **advanced LSTM (A-LSTM)** layers to improve the temporal dependency of our AV-SAD system



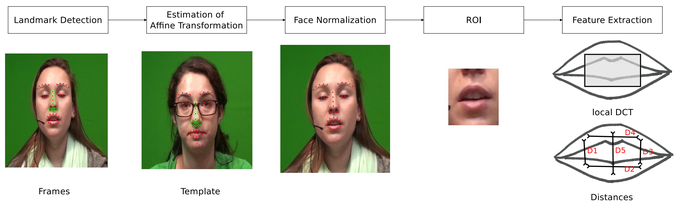
Data and Features

CRSS-4ENGLISH-14 Corpus :

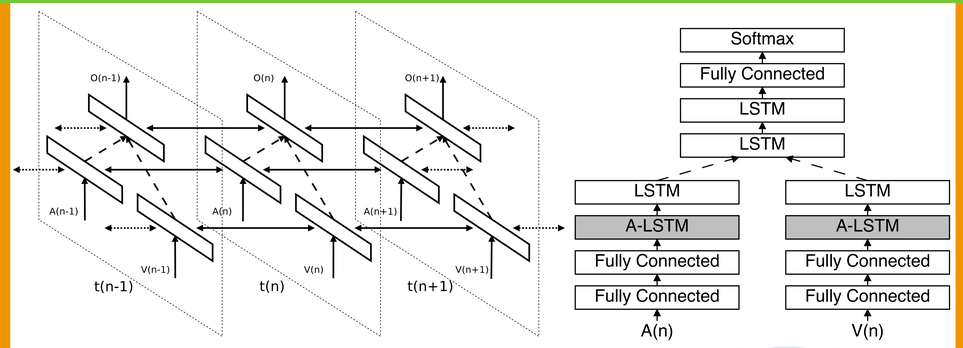
- 442 subjects in total with American, Hispanic, Indian and Australian accent
- This study only uses data from 105 American speakers
  - train (70), validation (10) and test (25)
- Clean and noisy sections with two channels
  - Ideal condition: HD camera and close talk mic (HD)
  - Tablet condition: camera and mic from tablet (TG)

Audiovisual Features:

- 5-D acoustic feature: harmonicity, clarity, prediction gain, periodicity and perceptual spectral flux
- 26-D visual feature following the flowchart:



Advanced Bimodal RNN



Advanced LSTM:

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C}_t$$

$$C' = \sum_T w_{C_T} \times C_T; w_{C_T} = \frac{\exp(W \cdot C_T)}{\sum_T \exp(W \cdot C_T)}$$

$$\widetilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

- The parameter *C'* is the linear combination of the cell values at the selected time points *T* in the past (*t-1* and *t-6* in this study)

Advanced Bimodal RNN:

- ***A branch***: maxout(16) + maxout(16) + A-LSTM(16) + LSTM(16)
- ***V branch***: maxout(64) + maxout(64) + A-LSTM(64) + LSTM(64)
- ***Top network***: LSTM(128) + LSTM(128) + maxout(128) + softmax

Experiment and Results

Speech Activity Detection:

- Baseline: BRNN proposed in [Tao & Busso, 2007] with LSTM
- Run experiment on Nvidia GTX 1070 (8GB)

Env.	Approach	Test Condition	Acc [%]	Pre [%]	Rec [%]	F[%]
C	BRNN	HD	90.1	94.5	84.8	89.4
	BRNN	TG	90.1	91.9	87	89.3
	A-BRNN	HD	90.6	94.5	85.3	89.7
	A-BRNN	TG	90.4	92.3	87.4	89.8*
N	BRNN	HD	93.3	93	94.1	93.5
	BRNN	TG	83.1	77.7	96.6	86.1
	A-BRNN	HD	93.1	92.7	94.8	93.7
	A-BRNN	TG	84.6	79.4	96.8	87.2*

* : statistically significant improvements

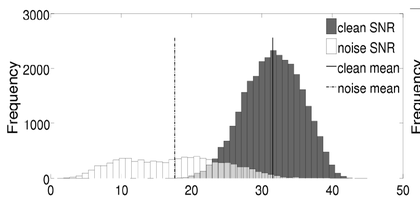
SNR Analysis:

- HD channel is not affected much by noise (mic is close to the mouth)
- TG channel is affected since it is close to the noise source
- Clean session contains spontaneous speech, which is a harder task

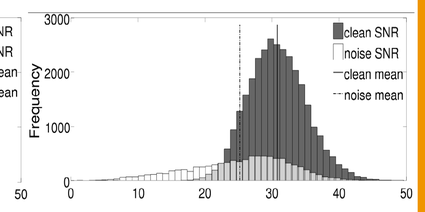
Non-speech Segments with Active Lip Motion:

- Evaluate the robustness to lip movements that are not associated with speech (smiles, lip-smack, deep breath)
- We manually identified 7,397 frames across speakers containing non-speech lip motion
- Report in F-score

Approach	HD		TG	
	C	N	C	N
BRNN	94.2	93.1	94.7	89.0
A-BRNN	94.7	93.4	94.7	89.4



SNR in HD channel



SNR in TG channel

Conclusions

- This study extended BRNN using A-LSTM for AV-SAD
- The proposed framework takes advantage of BRNN
- It has low latency and better time dependency modeling
- It is better in non-speech segments with active lip motion

Future Work

- The current implementation only uses A-LSTM in one layer, which is limited by the hardware requirement
  - A-LSTM can be used in more layers.
- More frames in the past can be consider
- Learn facial and acoustic features with CNN
  - Training the approach as an end-to-end system

This work was funded by NSF awards IIS-1718944 and IIS- 1453781 (CAREER)