Bimodal Recurrent Neural Network for Audiovisual Voice Activity Detection

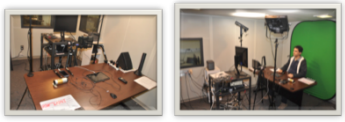
Fei Tao and Carlos Busso
Multimodal Signal Processing (MSP) Lab,
The University of Texas at Dallas, Richardson TX

Motivation

Background:

- VAD is an important preprocessing step in speech-based systems

- Conventional VAD are normally impaired by noise in the environment
- An alternative approach is audiovisual VAD (AV-VAD)
  - Mobile devises have frontal cameras
  - Suitable for human-robot interaction
- This study proposes a bimodal recurrent neural network for AV-VAD, relying on bidirectional LSTMs
- Approach is ideal for real applications

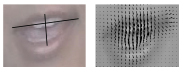


Data Collection

UT-CRSS-4EnglishAccent corpus:

- 442 subjects:
  - **American (115)**
  - Indian (112)
  - Australian (103)
  - Hispanic (112)
- We use American speakers' data (74 speak.)
  - 32.5h training, 4.7h testing
- Conditions:
  - Ideal setup: HD camera and close talk mic
  - Tablet setup: mic and camera from a tablet
- Clean segments + noisy segments
  - Noise played during recordings

Audiovisual Features



- **25D Video features** [Tao et al. 2015]
  - Optical flow: OFx, OFy and OFx+OFy (OFxy)
  - Geo. feat.: height (H), width (W), WxH and H+W

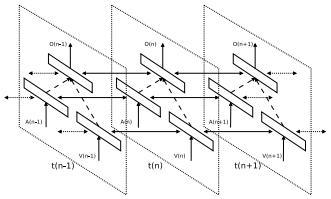
Set	OFx	OFy	OFxy	H	W	W+H	WxH
Temporal Variance	✓	✓	✓	✓	✓	✓	✓
Zero Crossing Rate	✓	✓	✓	✓	✓	✓	✓
Speech Periodic Characteristic	✓	✓	✓	✓	✓	✓	✓
First Order Derivative				✓	✓	✓	✓

- **5D Audio features** [Sadjadi and Hansen, 2013]
  - Harmonicity, clarity, prediction gain, periodicit and perceptual spectral flux

Proposed System

Bimodal Recurrent Neural Network (RNN)

- Hierarchal RNN network that offers powerful feature representation
- Dynamic weighting capability to model relation between modalities
- No assumption about distribution of relationship between modalities
  - The unimodal RNNs capture dependencies within the modalities
  - Second RNN captures the temporal dependency between modalities



B-RNN Architecture

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

Experimental Evaluation

Experimental Setup:

- Train on clean data
- Test under different conditions
  - ideal vs. tablet & clean vs. noisy
- Metrics:
  - Accuracy (ACC)
  - F-score (F)

Baselines:

- audio based VAD (A-VAD)
- video based VAD (V-VAD)
- Bimodal approaches:
  - AV-VAD DNN: only maxout layers without LSTM
  - Concat RNN: concatenated AV features

Recognition Result with Clean Data:

Approach	Condition	Acc[%]	F[%]
A-VAD	Ideal	85.59	85.59
	Tablet	86.88	85.95
V-VAD	Ideal	79.98	81.80
	Tablet	78.62	79.84
AV-VAD DNN	Ideal	87.75	87.88
	Tablet	88.39	88.50
Concat RNN	Ideal	93.19	**93.51**
	Tablet	92.78	93.22
Bimodal RNN	Ideal	93.18	93.48
	Tablet	92.86	**93.27**

Recognition Result with Noisy Data:

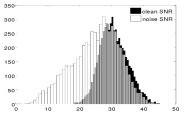
Approach	Condition	Acc[%]	F[%]
A-VAD	Ideal	75.38	78.92
	Tablet	58.75	69.64
V-VAD	Ideal	82.20	84.52
	Tablet	78.93	80.85
AV-VAD DNN	Ideal	84.76	84.76
	Tablet	72.41	78.26
Concat RNN	Ideal	93.18	93.82
	Tablet	84.50	86.94
Bimodal RNN	Ideal	93.82	**94.38**
	Tablet	85.84	**88.16**

Observations:

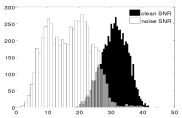
- Audiovisual fusion improves VAD performance under different conditions
- RNN improves audiovisual VAD system
- Bimodal RNN consistently outperforms other bimodal VAD approaches
- Performance of ~90% for the bimodal RNN
  - Solution is an appealing solution for real application

Analysis of Results:

- Tablet condition more affected by noise
- SNR distribution across different channels



SNR in Ideal channel



SNR in tablet channel

- SNR differences due to location of mics
  - Tablet (2m), close-talk microphone (few cm)

Conclusion and Future Work:

- We propose a novel deep learning architecture for multimodal learning
  - Bimodal RNN for AV-VAD
- Robust performance under different conditions
- Future work
  - Build end-to-end system
  - Architectures that deal with missing features
  - Use framework in other multimodal problems

This work was funded by the NSF under grant IIS-1450349