RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Children's
National Medical Center

Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment

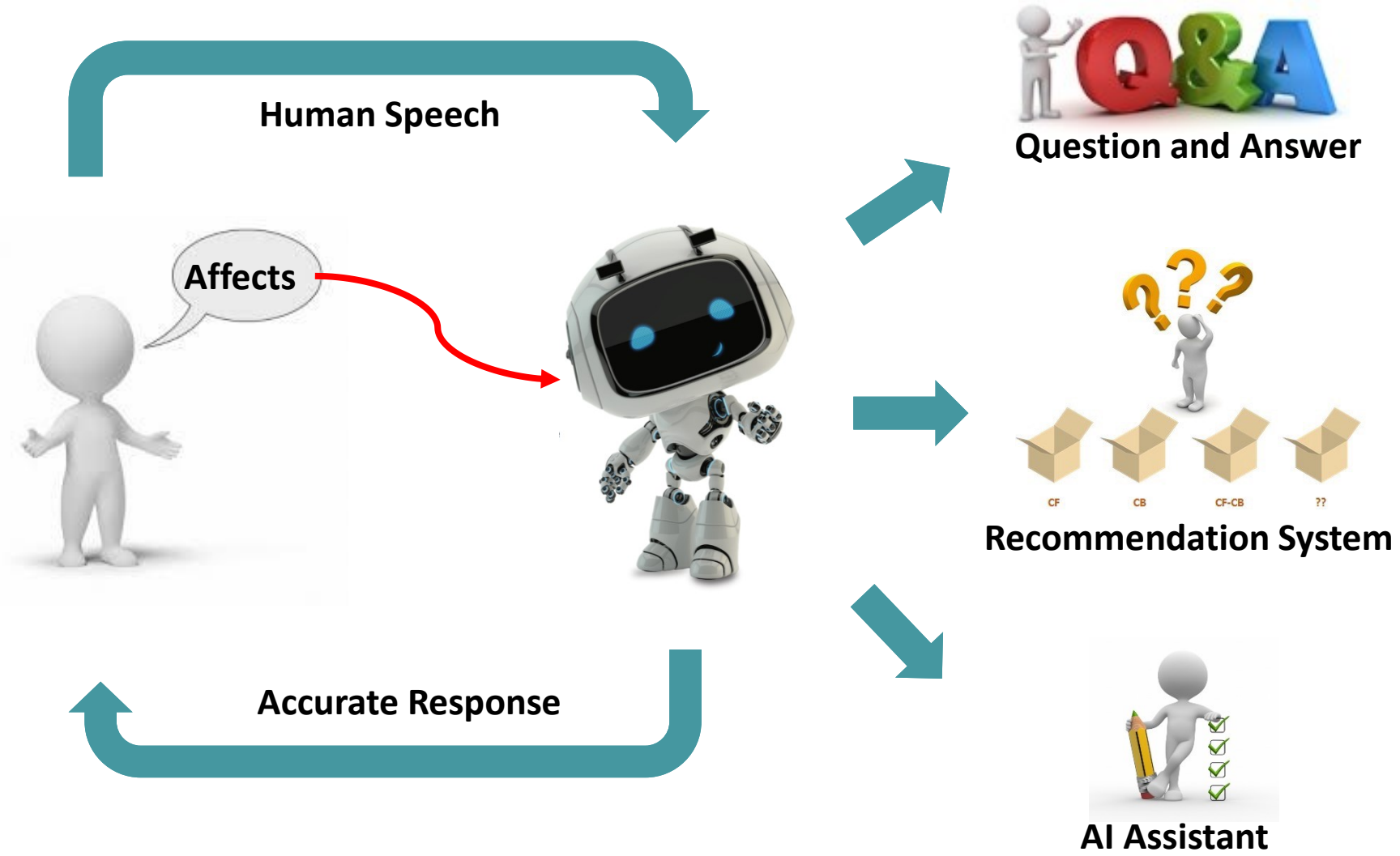
Yue Gu*, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, Ivan Marsic

Multimedia Image Processing Lab

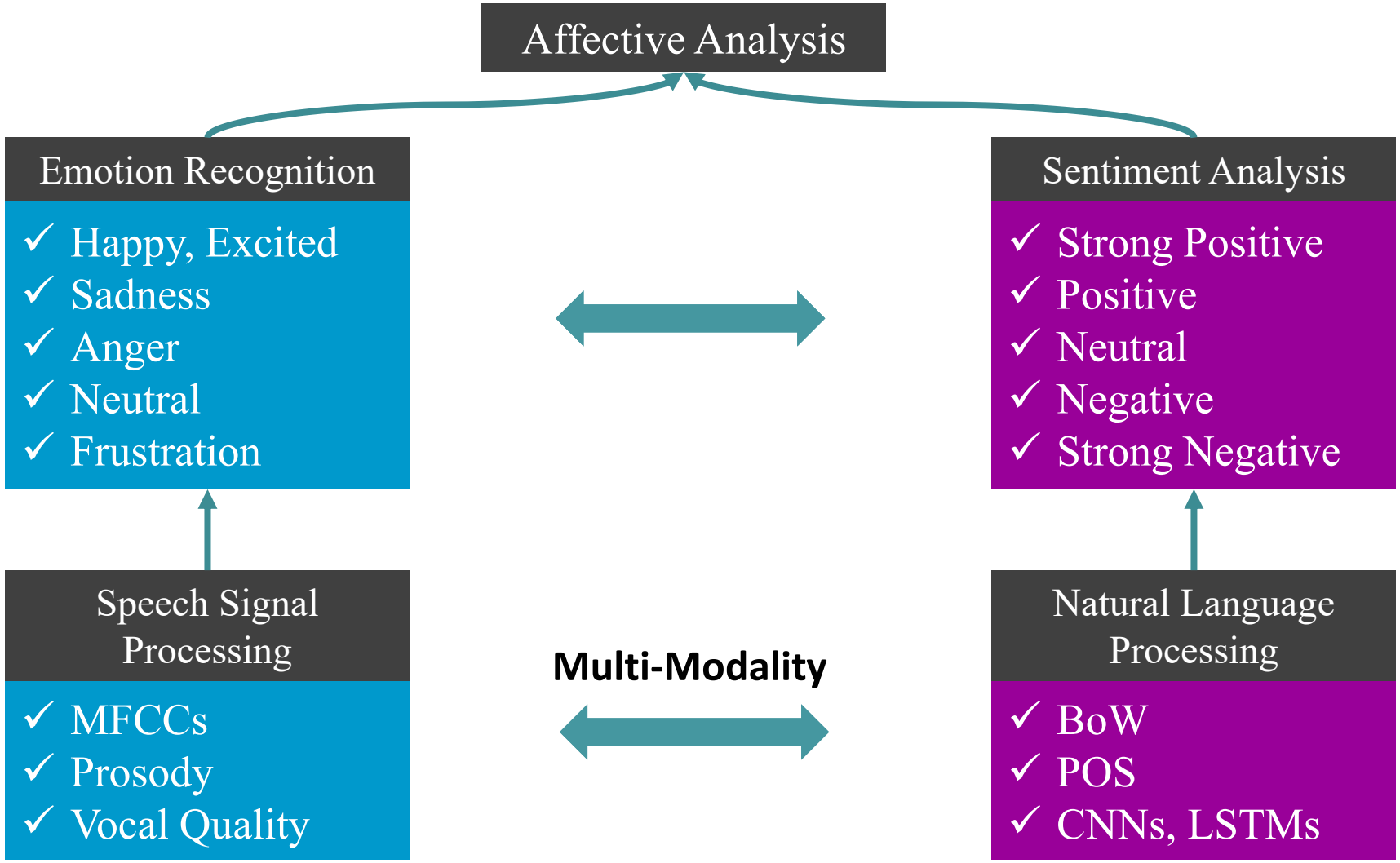
Electrical and Computer Engineering Department

Rutgers, The State University of New Jersey

Why the affective analysis is necessary?

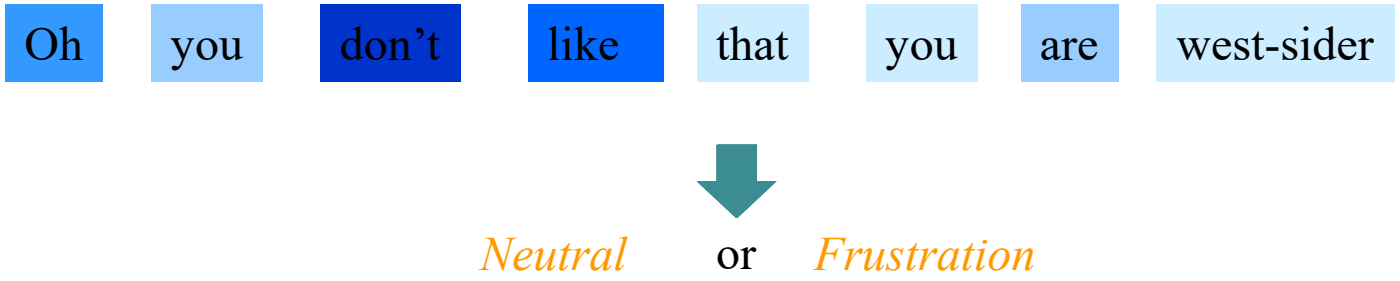


Progress of Affective Computing



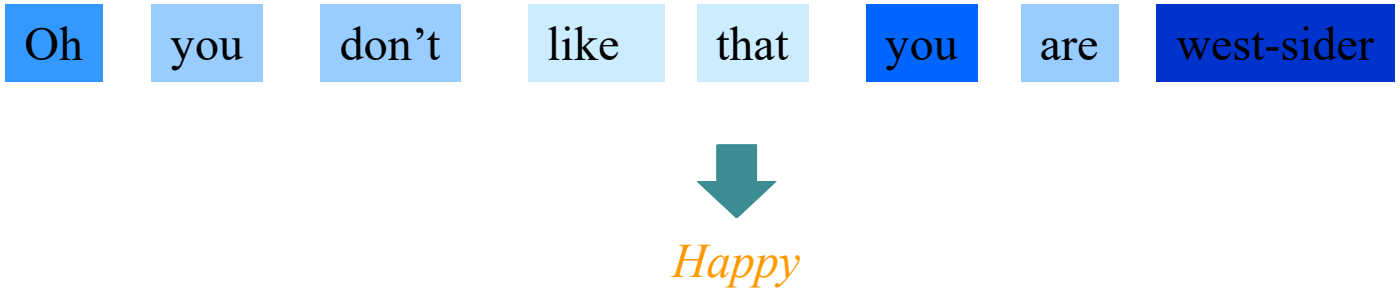
Is multi-modality needed?

- Vocal signal prominence



Is multi-modality needed?

- Vocal signal prominence



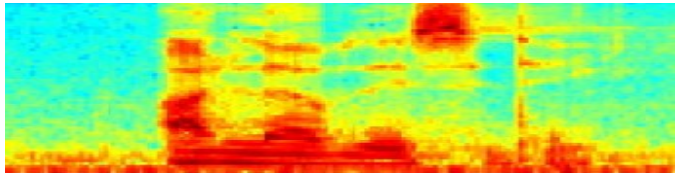
Is multi-modality needed?

- Vocal signal prominence

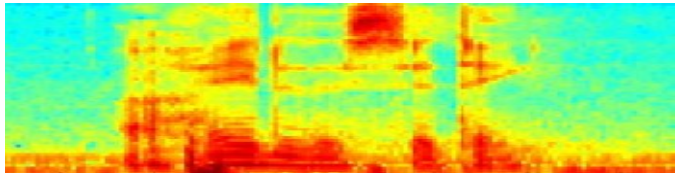
Oh you don't like that you are west-sider

Happy

- Acoustic ambiguity



"I love this city!"



"I hate this city!"

Challenges: Feature Extraction

- Gap between features and actual affective states
  - Lack of high-level associations
  - Not all parts contribute equally

Challenges: Modality Fusion

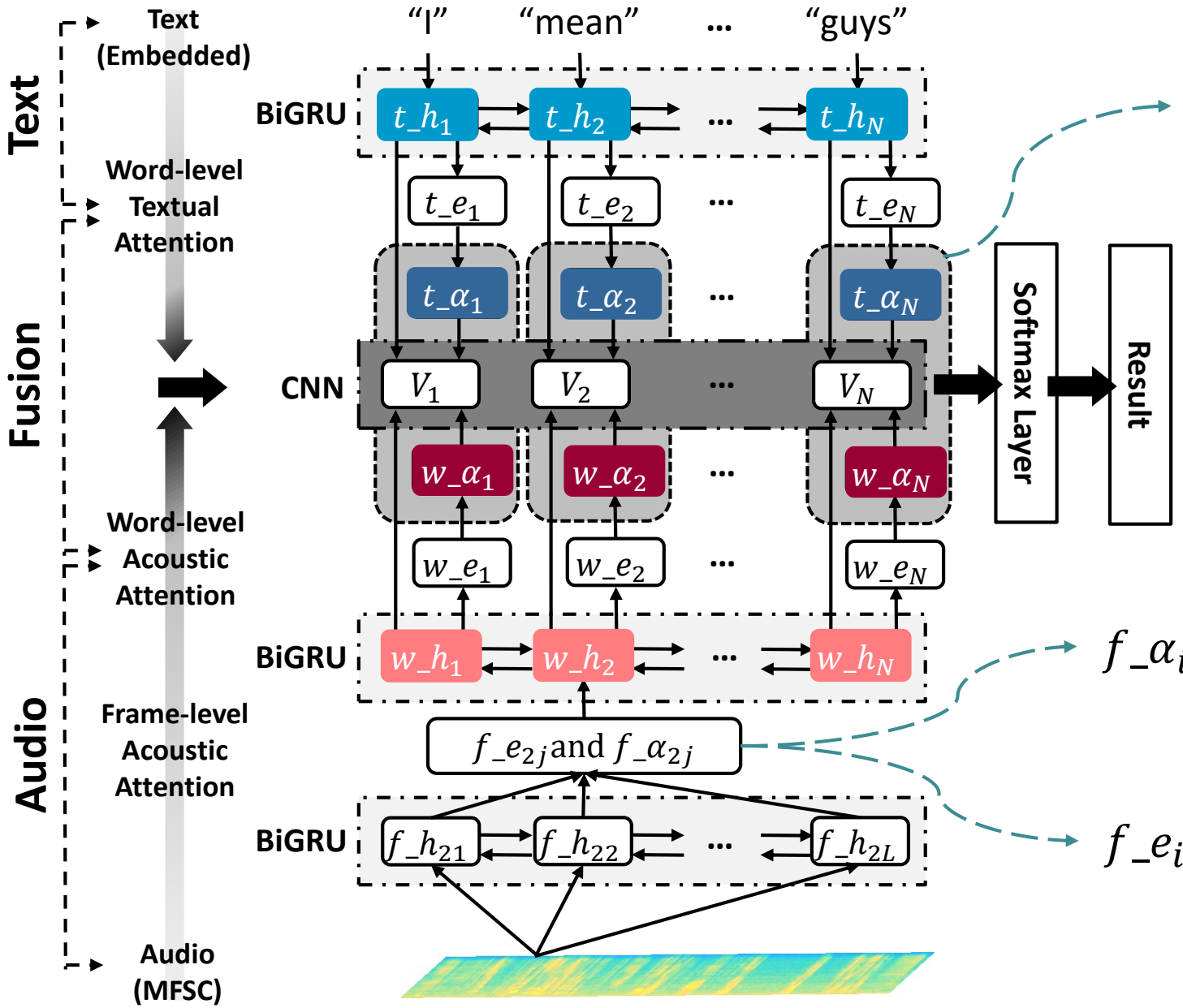
- Decision-level Fusion
  - Lack of mutual association learning
- Feature-level Fusion
  - Fail to learn time-dependent interactions
  - Lack of consistency

Proposed Solutions

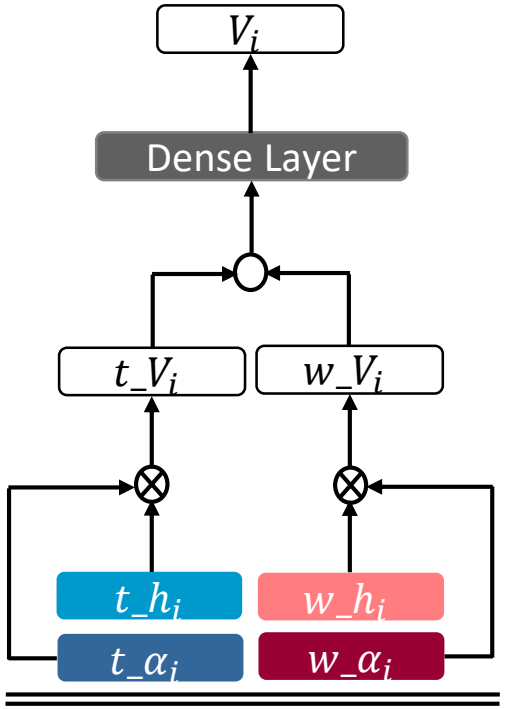
- Feature Extraction
  - Hierarchical attention based bidirectional GRUs
- Modality Fusion
  - Word-level fusion with attention
  - An End-to-End multimodal network

Data Pre-processing

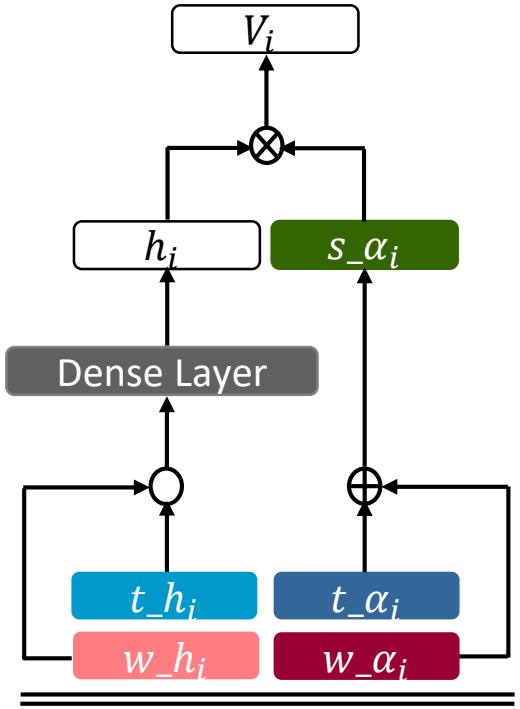
- Text Branch
  - Word Embedding: *word2vec*
- Audio Branch
  - Mel-frequency spectral coefficients (MFSCs)
- Synchronization
  - Word-level forced alignment


Text
(Embedded)
Text
Word-level
Textual
Attention
Fusion
Word-level
Acoustic
Attention
Audio
Frame-level
Acoustic
Attention
Audio
(MFSC)
"I"
"mean"
...
"guys"
BiGRU
t_h_1
t_h_2
t_h_N
t_e_1
t_e_2
t_e_N
t_α_1
t_α_2
t_α_N
CNN
V_1
V_2
V_N
w_α_1
w_α_2
w_α_N
w_e_1
w_e_2
w_e_N
BiGRU
w_h_1
w_h_2
w_h_N
$f_e_{2j}$and f_α_{2j}
BiGRU
f_h_{21}
f_h_{22}
f_h_{2L}
Softmax Layer
Result
Word-level
Fusion
$f\_\alpha_{ij} = \frac{exp(f\_e_{ij}{}^{\top} v_f)}{\sum_{k=1}^{L} exp(f\_e_{ik}{}^{\top} v_f)}$
$f\_e_{ij} = tanh(W_f f\_h_{ij} + b_f)$

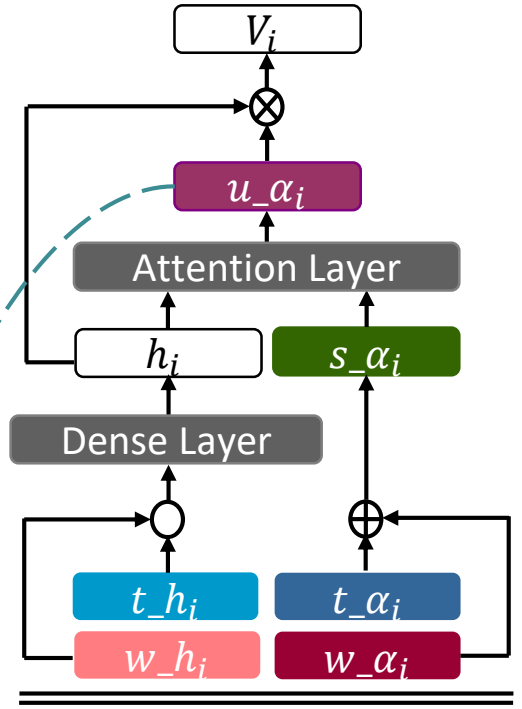
Word-level Fusion



(a) Horizontal Fusion



(b) Vertical Fusion



(c) Fine-tuning Attention Fusion

w_α_i **Word-level acoustic attention distribution**

t_α_i **Word-level textual attention distribution**

w_h_i **Word-level acoustic contextual state**

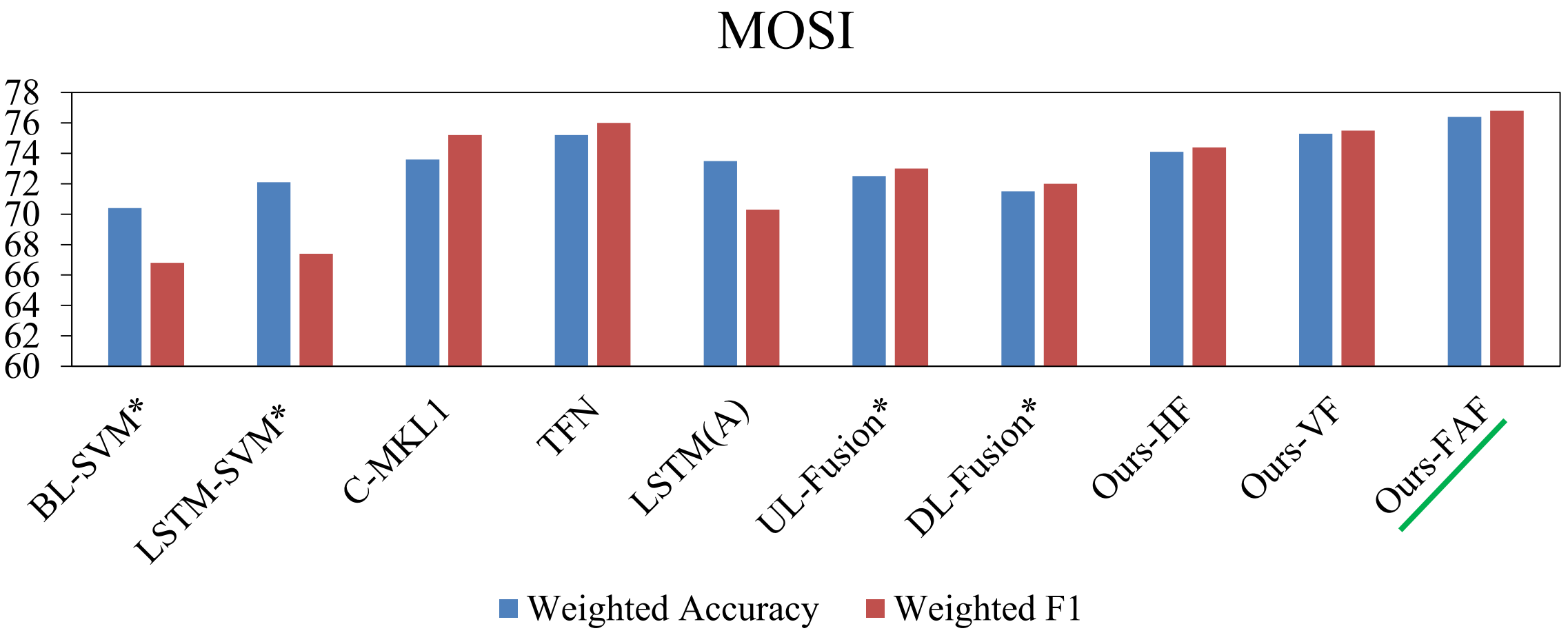
t_h_i **Word-level textual contextual state**

$$u\_\alpha_i = \frac{exp(u\_e_i^{\top} v_u)}{\sum_{k=1}^{N} exp(u\_e_k^{\top} v_u)} + s\_\alpha_i$$

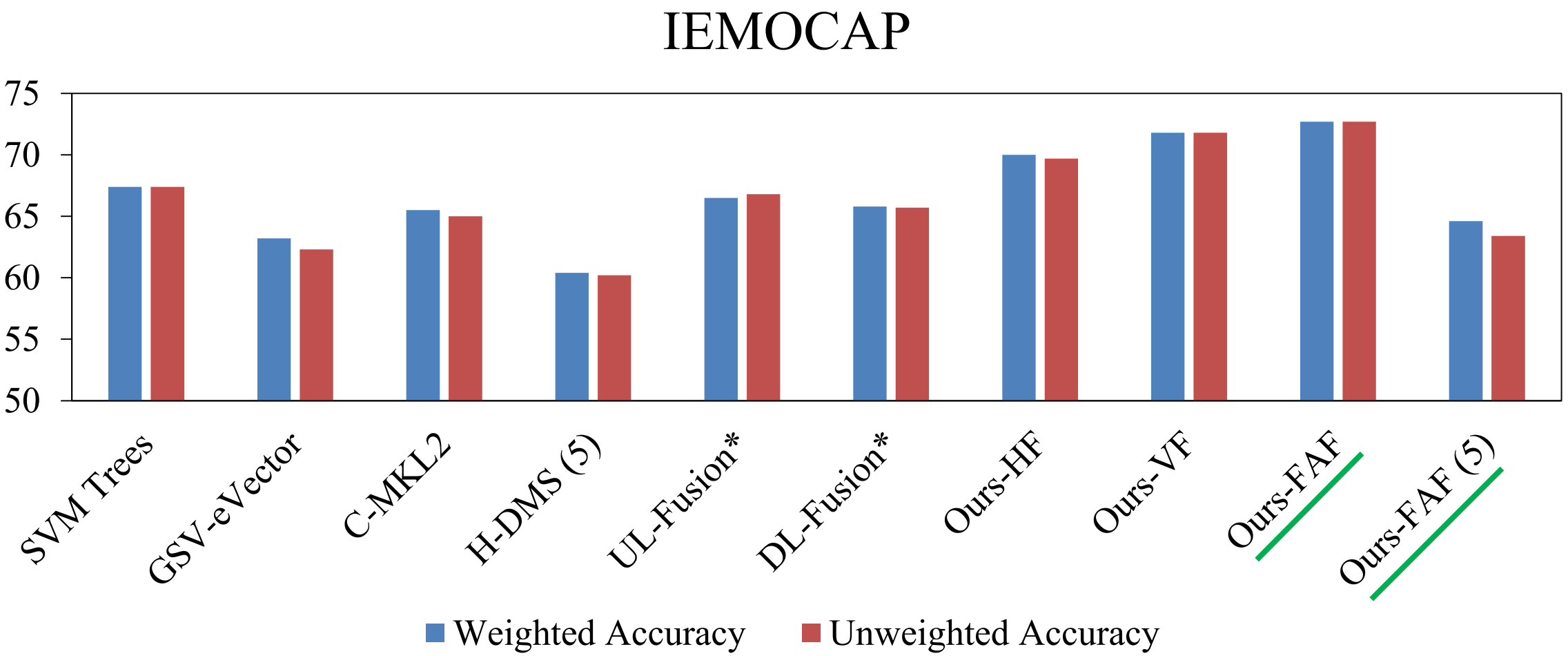
Baselines

- Sentiment Analysis
  - BL-SVM, LSTM-SVM
  - C-MKL, TFN, LSTM(A)
- Emotion Recognition
  - SVM Trees, GSV-eVector
  - C-MKL, H-DMS
- Fusion
  - Decision-level, Feature-level (utterance-level)

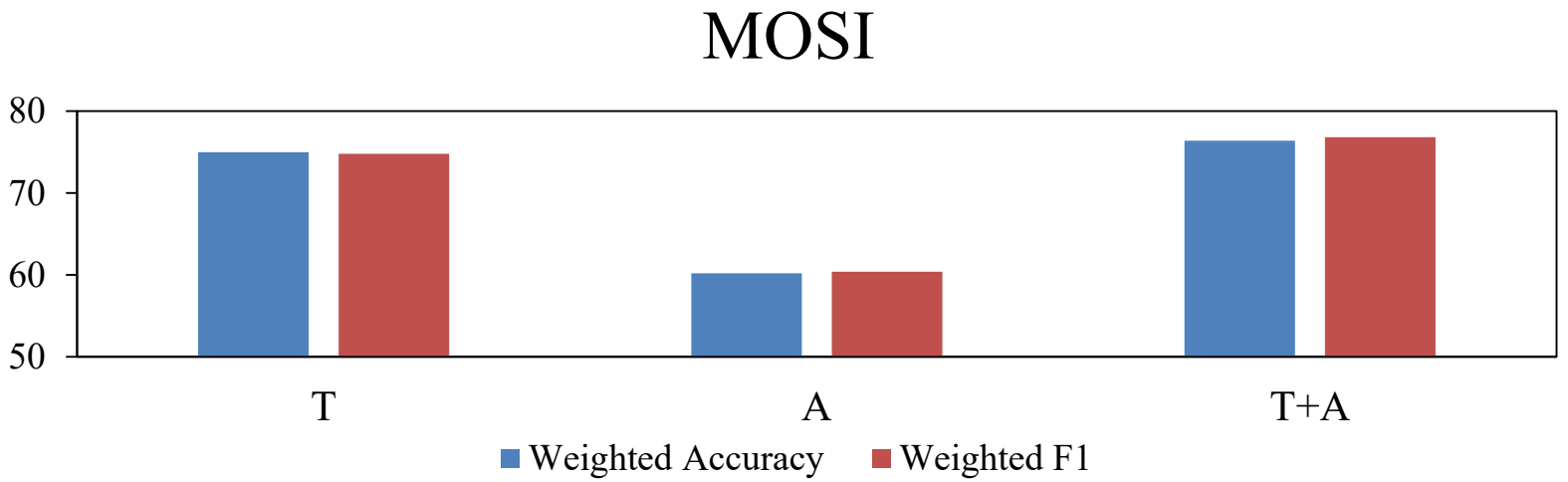
Sentiment Analysis Result

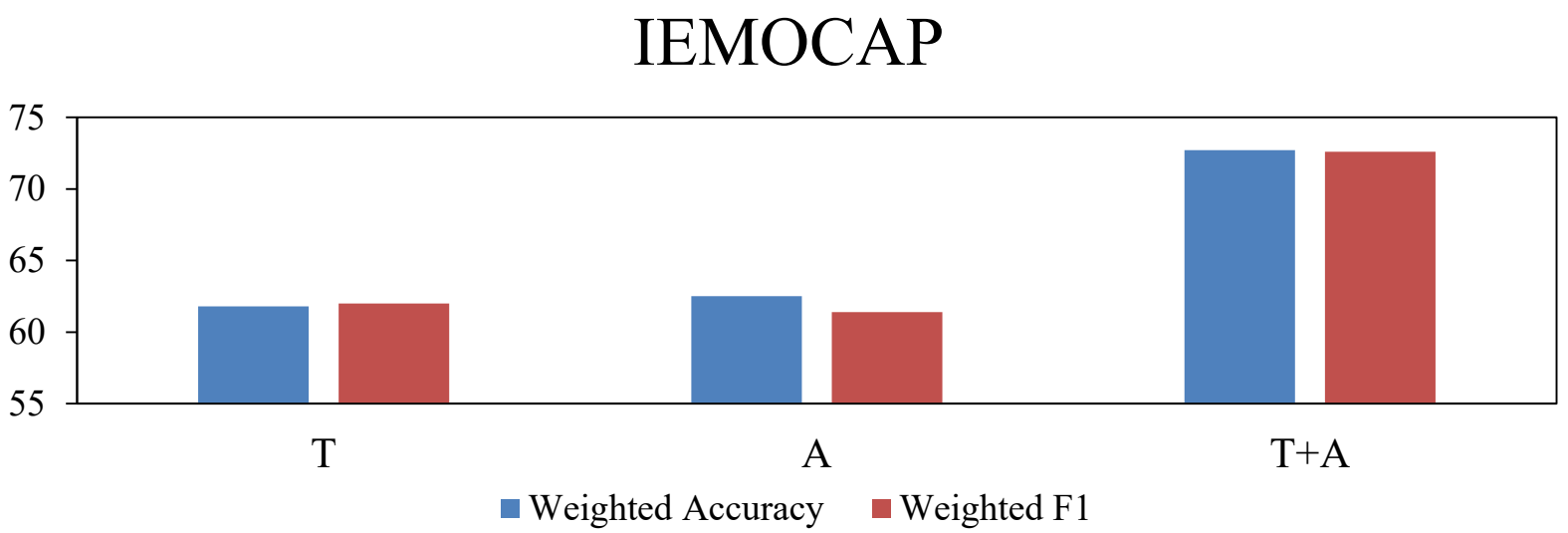


Emotion Recognition Result

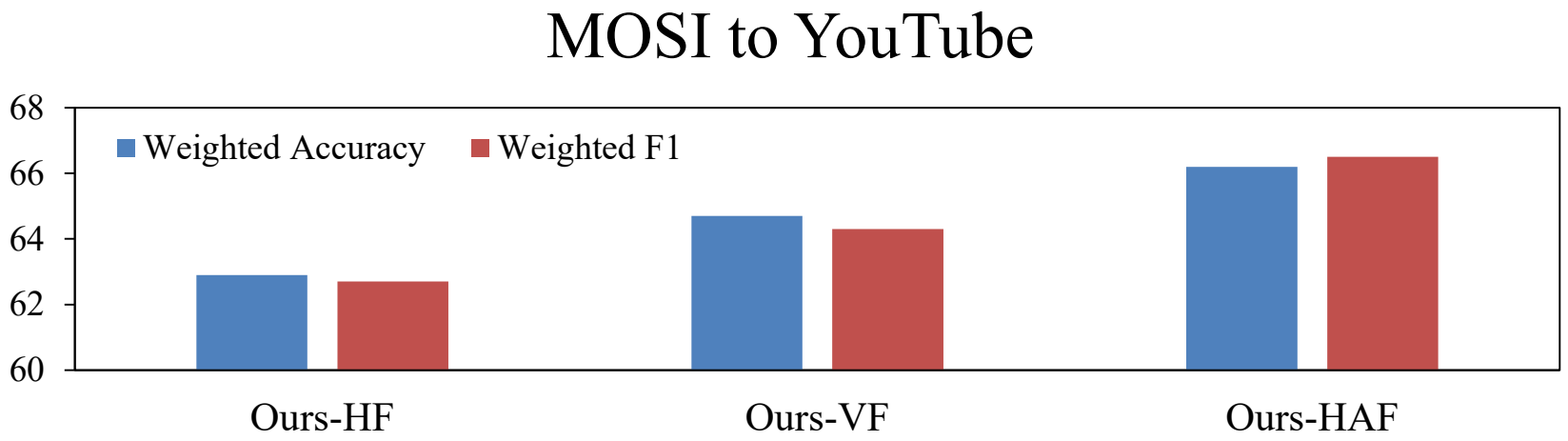


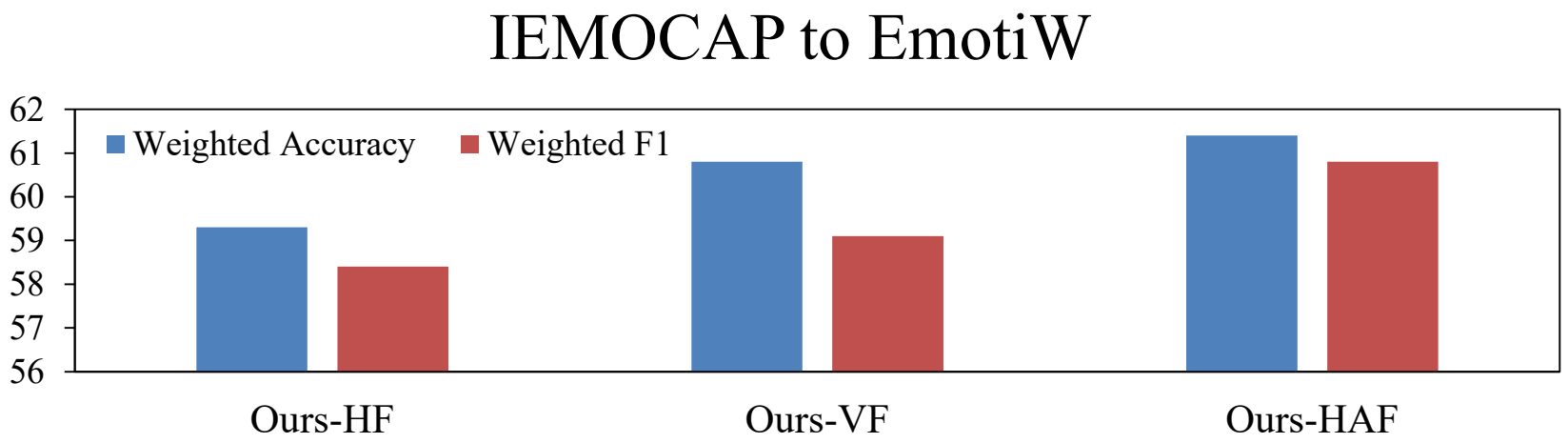
Multimodal architecture is needed





Generalization

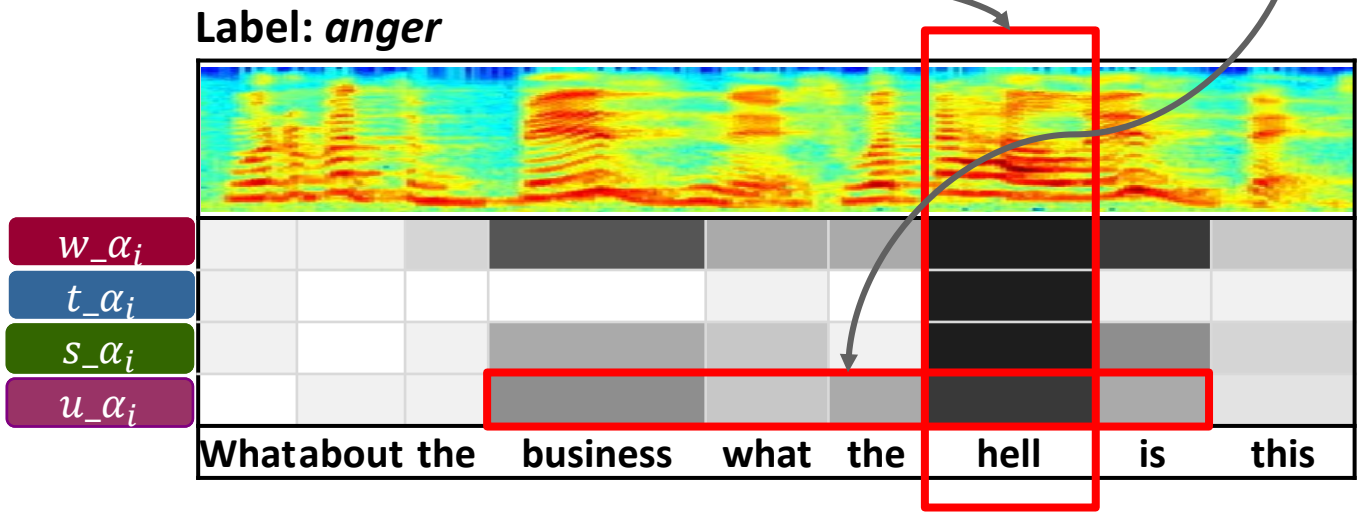




Attention Visualization

Carry representative information in both text and audio

Successfully combine both textual and acoustic attentions



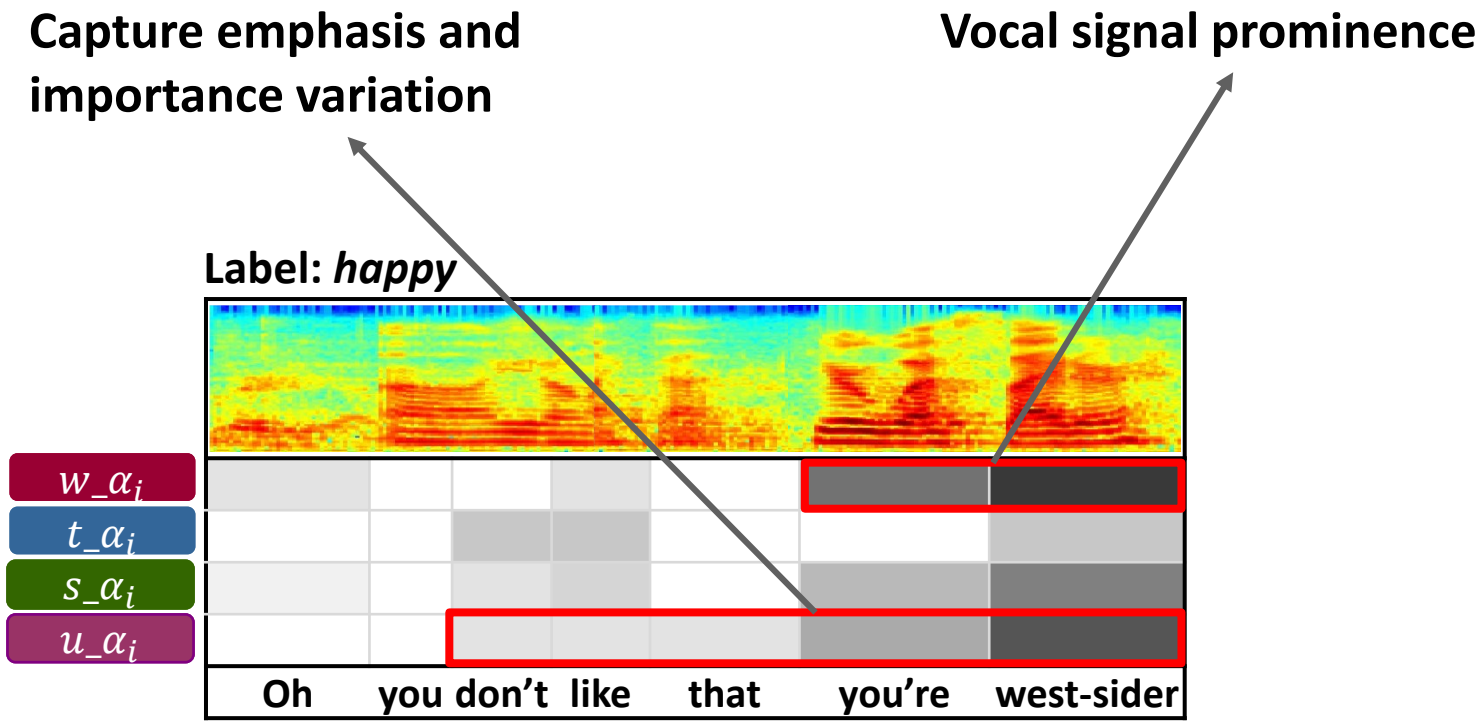
w_α_i Word-level acoustic attention distribution

t_α_i Word-level textual attention distribution

s_α_i Shared attention distribution

u_α_i Fine-tuning attention distribution

Attention Visualization



w_α_i Word-level acoustic attention distribution

t_α_i Word-level textual attention distribution

s_α_i Shared attention distribution

u_α_i Fine-tuning attention distribution

Summary

- A hierarchical attention based multimodal structure
- The word-level fusion strategies
- Word-level attention visualization

Thank you !

Email: yg202@scarletmail.rutgers.edu

Homepage: www.ieyuegu.com