Enhancing Resilience to Missing Data in Audio-Text Emotion Recognition with Multi-Scale Chunk Regularization

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Wei-Cheng Lin, Lucas Goncalves, Carlos Busso

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

Motivation

Background:

- Audio-Text Multimodal Emotion Recognition
  - Unclear role of temporal synchronization (i.e., alignment) between the input audio and text sequences
  - Current model-level and feature-level fusion techniques cannot investigate this research question
  - Multimodal modeling can effectively improve recognition performance but also reduces the model robustness against missing data scenario

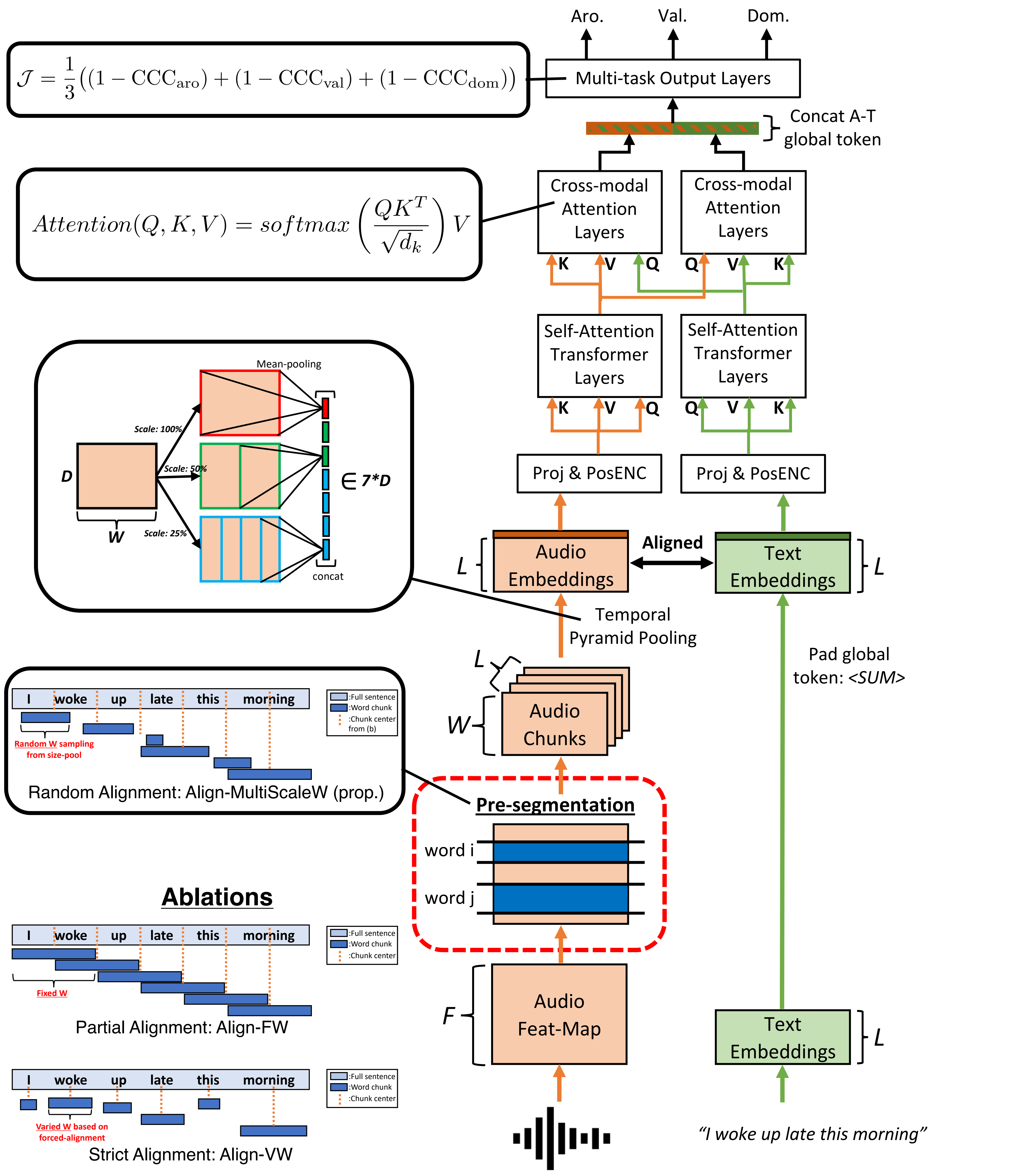
Our Work:

- Proposes a novel word-chunk modeling concept for audio-text emotion recognition
- Confirms that the self-attention mechanism is powerful enough to capture temporal alignment across audio-text
- Propose to leverage multi-scale chunk regularization to improve model's robustness against missing data

Proposed Framework

Setups:

- Input Features
  - Wav2vec2-large-robust (Audio: 1,024D)
  - RoBERTa-base (Text: 768D)
- Model Architecture
  - [3 X Self-Attention Blocks] + [3 X Cross-Attention Blocks]
  - Hidden dim: 256D



$$\mathcal{J} = \frac{1}{3}((1-\text{CCC}_{\text{aro}}) + (1-\text{CCC}_{\text{val}}) + (1-\text{CCC}_{\text{dom}}))$$

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Ablations

Random Alignment: Align-MultiScaleW (prop.)

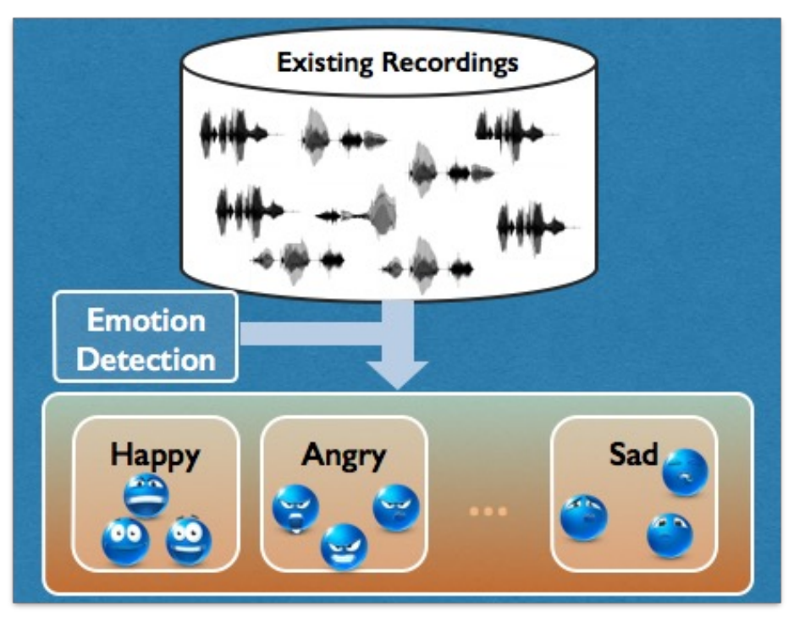
Partial Alignment: Align-FW

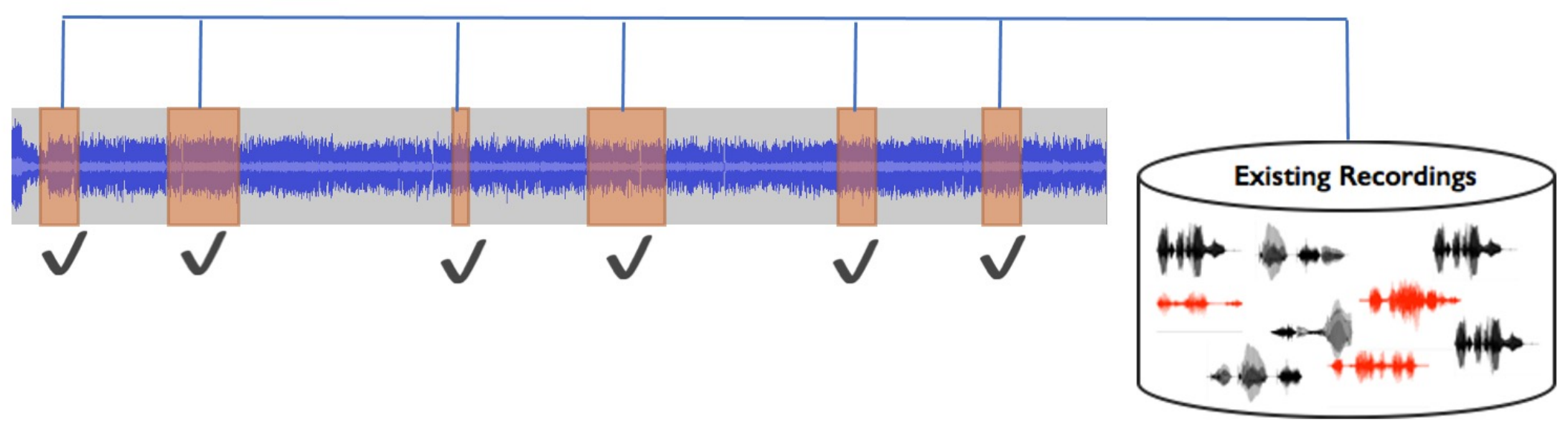
Strict Alignment: Align-VW

Resources

Dataset:

- The MSP-PODCAST v1.10 corpus
  - Largest spontaneous speech emotion corpus collected from existing podcast recordings
  - Annotated using crowdsourcing framework (AMT)
  - Includes train/dev/test: 63,076/10,999/16,903 clips (~166hrs)
  - Regression problem: arousal, dominance, and valence
  - Montreal-forced-aligner (MFA)
  - Divide into speaker turns (2.75 - 11 secs)





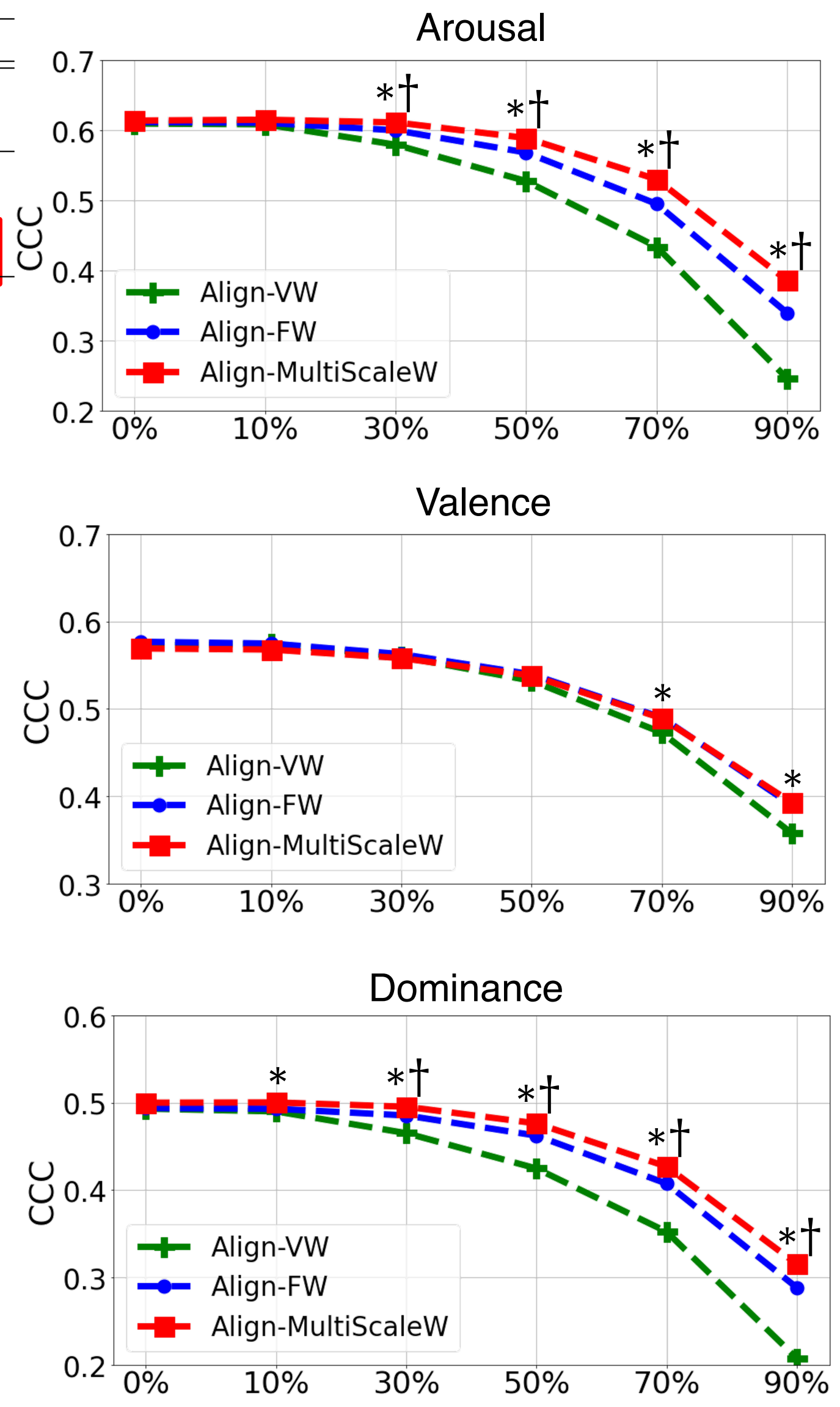
Experimental Results

SER Performance Comparison

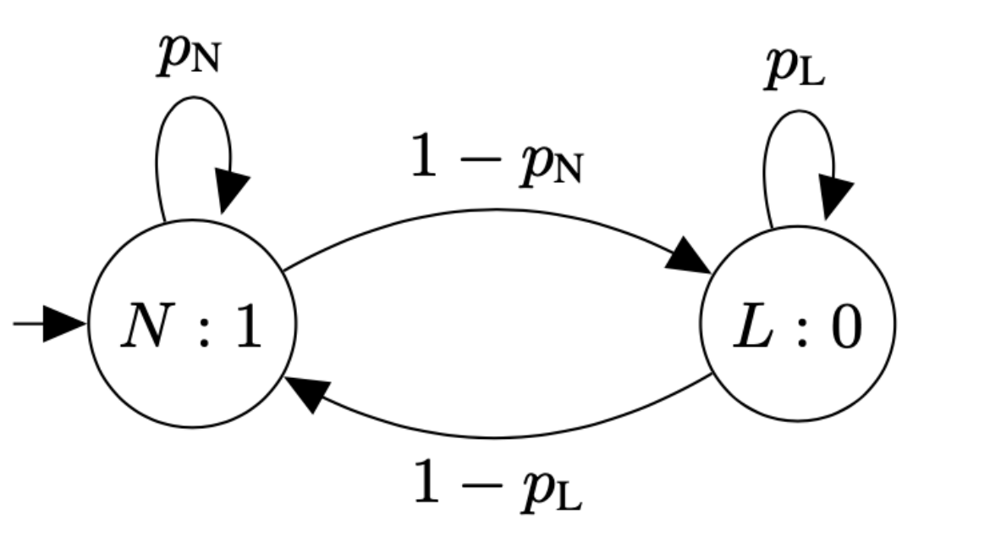
Approach	Arousal	Valence	Dominance
MDRE-GRU [52]	0.6090	0.5592	0.4842
MDREA-GRU [52]	0.6029	0.5603	0.4792
Align-VW	0.6094	$0.5723^{*\dagger}$	$0.4932^{*\dagger}$
Align-FW	$0.6117^{\dagger}$	$\mathbf{0.5767}^{*\dagger}$	$0.4939^{*\dagger}$
Align-MultiScaleW	$\mathbf{0.6137}^{\dagger}$	$0.5691^{*\dagger}$	$\mathbf{0.5000}^{*\dagger}$

Improved and robust performance for all three emotions under different missing data testing conditions

Missing Data Robustness for Audio-Text: Random Drop Words



Missing Data Robustness for Audio-Only: Packet-Loss



Obtain the same robust performance trend as well as for Audio-Only SER

Conclusions

- Word-chunk concept can explicitly model and control the alignment level between audio-text sequences
- Multi-scale chunk regularization can effectively improve model robustness against missing data conditions, which is valid for both the audio-text and audio-only scenarios

Github link: https://github.com/winston-lin-wei-cheng/MultiScale-Chunk-Regularization

Future Work

- Extending word-chunk concept to more modalities (e.g., audio-text-video) for temporal synchronization
- Apply multi-scale chunk regularization based on different modalities to improve model performance

This work was supported by NSF under Grant CNS-2016719