Role of Lexical Boundary Information in Chunk-Level Segmentation for Speech Emotion Recognition

Wei-Cheng Lin and Carlos Busso

UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

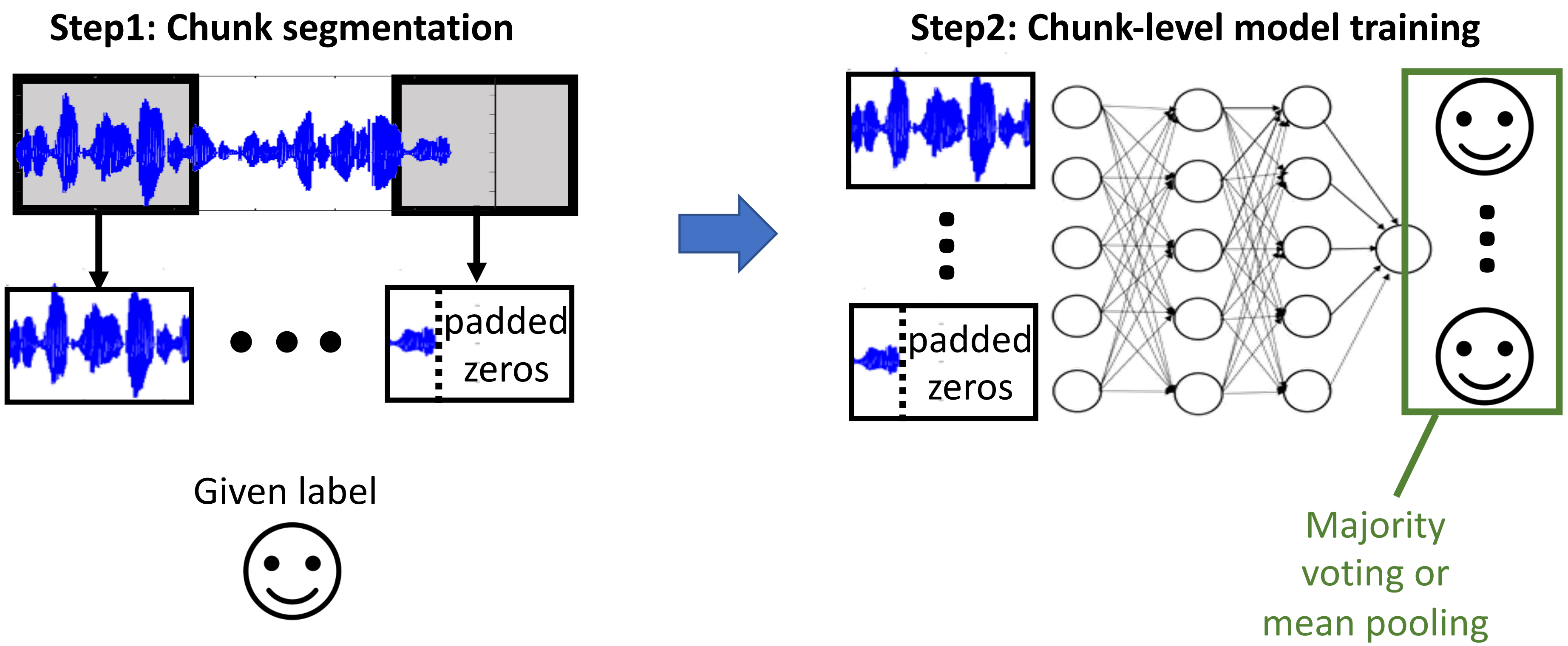
UT Dallas MSP Multimodal Signal Processing Laboratory

Outline

1. **Motivation**
2. **Lexical-based Segmentation**
3. **Experimental Results and Analysis**
4. **Conclusions & Future Works**

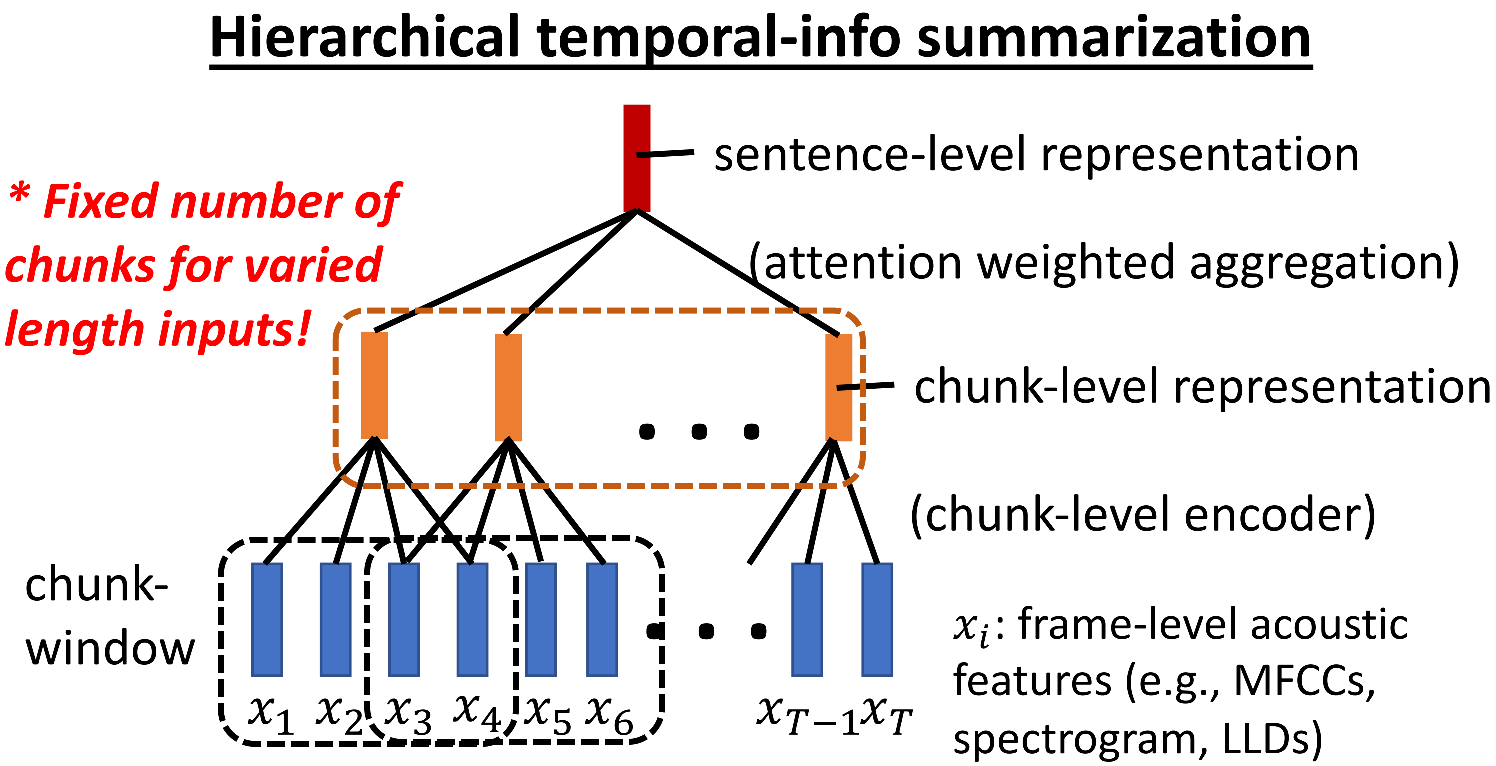
Motivation

- **Chunk-level modeling for *speech emotion recognition* (SER)**



Motivation

- **Dynamic chunk segmentation [1]**

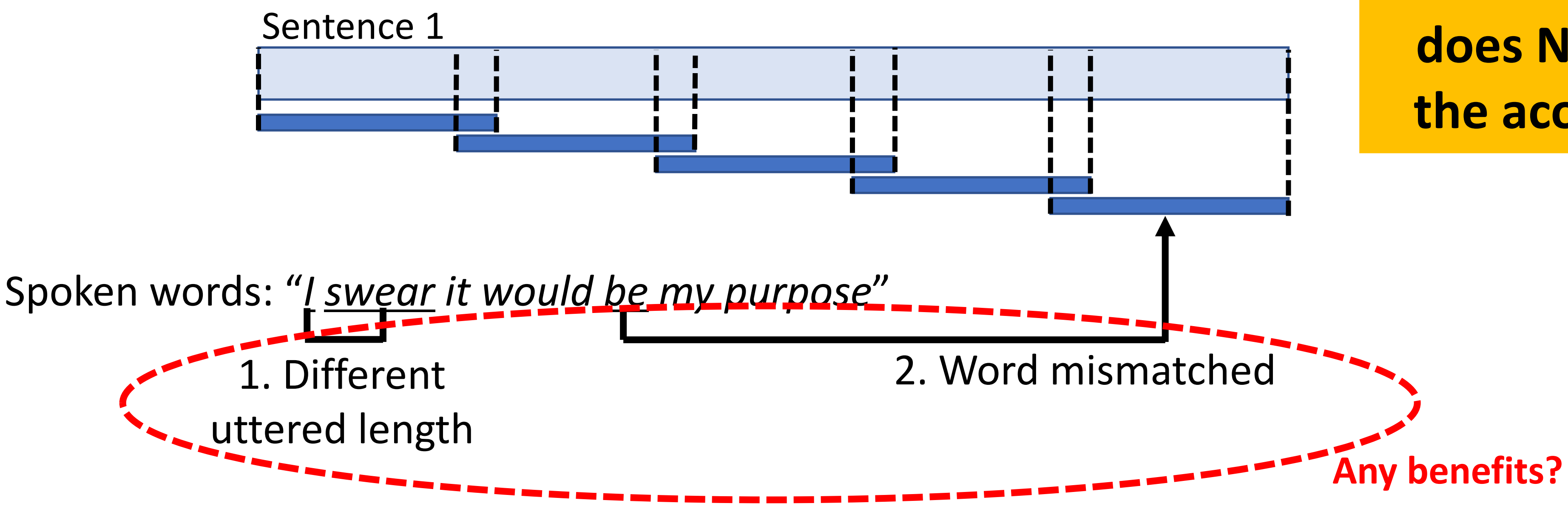


[1] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," IEEE Transactions on Affective Computing, vol. Early Access, 2022.

Motivation

- **Time-based chunk segmentation**
  - Split data chunks without considering their actual lexical boundaries
  - E.g., dynamic chunk segmentation

Lexical information does NOT align with the acoustic frames!



Lexical-based Segmentation

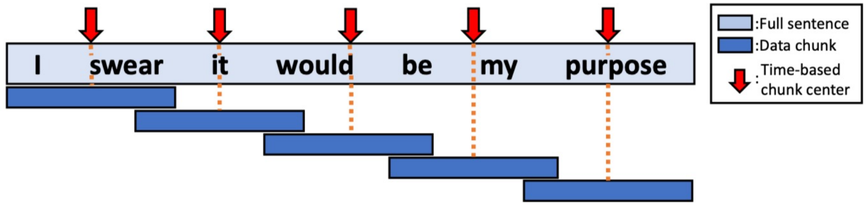
- **Lexical chunk segmentation**
  - Factor *C*: number of chunks to split
    => fixed or number of spoken words
  - Factor *W*: chunk window size
    => fixed or word-level alignment boundary
  - Result in different combinations, e.g., FixedW-VariedC (FW-VC)

The model does NOT consider the semantic meaning of words, still an SER task!

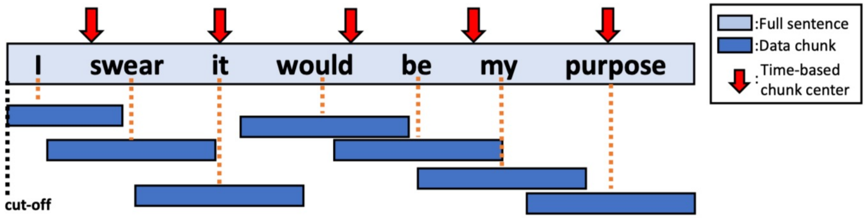
Lexical-based Segmentation

- **Visualization of different options**

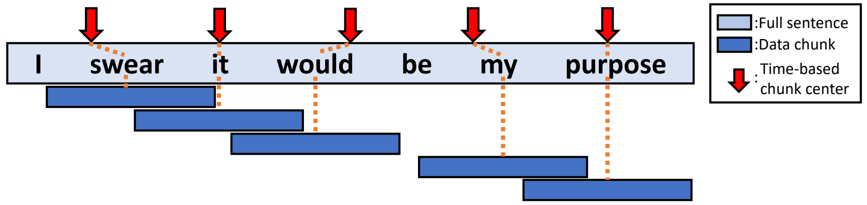
Time-based FixedW-FixedC (tFW-FC)



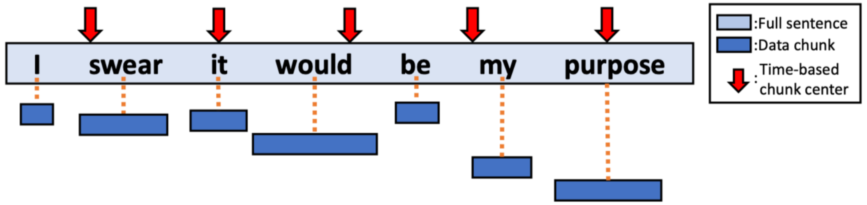
Lexical FixedW-VariedC (FW-VC)



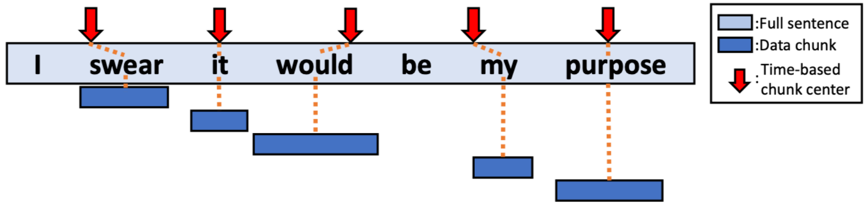
Lexical FixedW-FixedC (FW-FC)



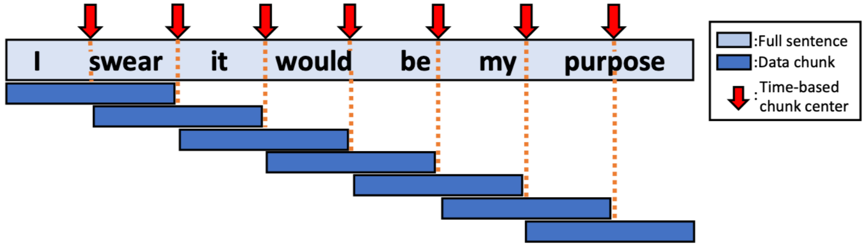
Lexical VariedW-VariedC (VW-VC)



Lexical VariedW-FixedC (VW-FC)



Combine FixedW-VariedC (cFW-VC)



Experimental Results and Analysis

UT Dallas MSP Multimodal Signal Processing Laboratory

- **Experimental Setting**
  - Datasets: *IEMOCAP* and *MSP-Podcast v1.10* [2]
  - Acoustic features: *LLDs* and *wav2vec2* [3]
  - Model arch: *LSTM* chunk-level encoder + *Multi-Heads Self-Attention* aggregation
  - Emotion: *arousal*, *dominance,* and *valence*
  - Evaluation metric: *concordance correlation coefficient* (CCC)

The only difference is how we segment the data chunks!

[2] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing, vol. 10, no. 4, pp. 471–483, October-December 2019.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems (NeurIPS 2020), vol. 33, Virtual, December 2020, pp. 12 449–12 460.

Experimental Results and Analysis

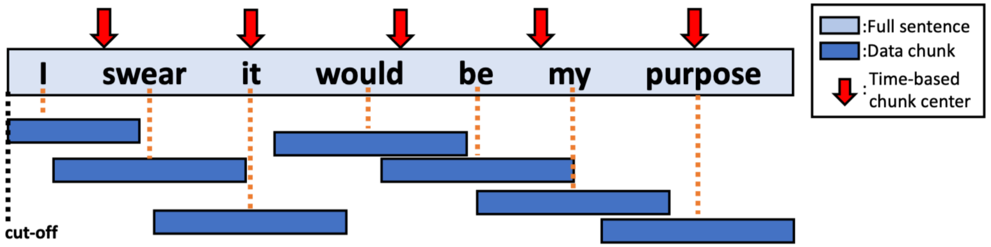
Performance comparison summary

Method	CovrR/OvrR [%]	LLDs (CCC) Aro.	LLDs (CCC) Val.	LLDs (CCC) Dom.	Wav2Vec2 (CCC) Aro.	Wav2Vec2 (CCC) Val.	Wav2Vec2 (CCC) Dom.
MSP-Podcast v1.10							
tFW-FC	99 / 50	0.528	**0.216***	0.430	0.604	**0.352**	0.478
FW-FC	90 / 56	0.529	0.191	0.423	0.598	0.344	0.475
VW-FC	57 / 35	0.534	0.170	0.427	0.595	0.352	0.460
FW-VC	94 / 68	0.544*	0.141	0.455*	**0.620***	0.349	0.497*
VW-VC	82 / 0	0.546*	0.118	0.459*	0.613*	0.336	0.492*
cFW-VC	99 / 65	**0.562**†	0.207*	**0.468**†	0.616*	0.343	**0.499***
IEMOCAP							
tFW-FC	99 / 79	0.614	0.353	0.406	0.709*	0.554	0.531
FW-FC	82 / 83	0.593	0.257	0.411	0.700	0.537	0.539
VW-FC	47 / 71	0.595	0.279	0.409	0.688	0.532	0.526
FW-VC	81 / 67	0.633*	0.378	0.451*	0.713*	0.582*	0.538
VW-VC	61 / 0	0.626*	0.395*	0.433	**0.719***	0.577*	0.558*
cFW-VC	99 / 60	**0.636***	**0.404**†	**0.463**†	0.718*	**0.584***	0.549*

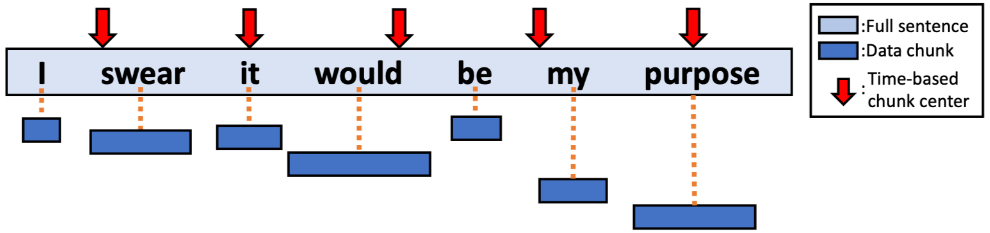
* means statistically significant better performance over other approaches without a marker

† means the results are statistically significant better than all other approaches

Lexical FixedW-VariedC (FW-VC)



Lexical VariedW-VariedC (VW-VC)



Knowing the precise word boundary (i.e., VariedW) does NOT bring significant performance benefits!

Experimental Results and Analysis

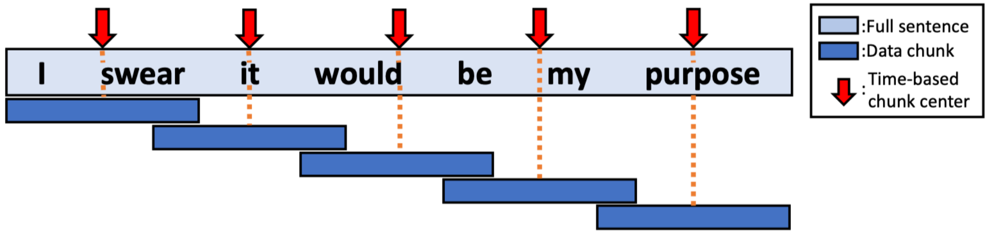
Performance comparison summary

Method	CovrR/OvrR [%]	LLDs (CCC) Aro.	LLDs (CCC) Val.	LLDs (CCC) Dom.	Wav2Vec2 (CCC) Aro.	Wav2Vec2 (CCC) Val.	Wav2Vec2 (CCC) Dom.
MSP-Podcast v1.10							
tFW-FC	99 / 50	0.528	**0.216***	0.430	0.604	**0.352**	0.478
FW-FC	90 / 56	0.529	0.191	0.423	0.598	0.344	0.475
VW-FC	57 / 35	0.534	0.170	0.427	0.595	0.352	0.460
FW-VC	94 / 68	0.544*	0.141	0.455*	**0.620***	0.349	0.497*
VW-VC	82 / 0	0.546*	0.118	0.459*	0.613*	0.336	0.492*
cFW-VC	99 / 65	**0.562†**	0.207*	**0.468†**	0.616*	0.343	**0.499***
IEMOCAP							
tFW-FC	99 / 79	0.614	0.353	0.406	0.709*	0.554	0.531
FW-FC	82 / 83	0.593	0.257	0.411	0.700	0.537	0.539
VW-FC	47 / 71	0.595	0.279	0.409	0.688	0.532	0.526
FW-VC	81 / 67	0.633*	0.378	0.451*	0.713*	0.582*	0.538
VW-VC	61 / 0	0.626*	0.395*	0.433	**0.719***	0.577*	**0.558***
cFW-VC	99 / 60	0.636*	**0.404†**	**0.463†**	0.718*	**0.584***	0.549*

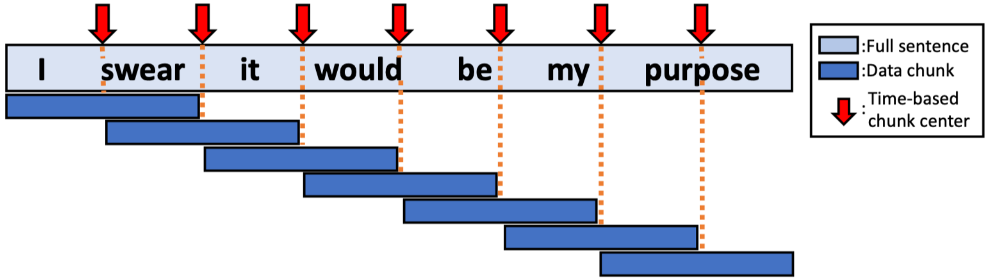
* means statistically significant better performance over other approaches without a marker

† means the results are statistically significant better than all other approaches

Time-based FixedW-FixedC (tFW-FC)



Combine FixedW-VariedC (cFW-VC)



Knowing how many chunks use to split (i.e., depending on the number of words) is crucial!

Conclusions & Future Works

- **Conclusion**
  - We found a minor role of word-level timing boundaries for chunk-level SER
  - We found that splitting data chunks according to number of words is the key leads to better chunk-level SER

- **Future Work**
  - Multimodal segmentation and benefits for modeling (e.g., video-speech-text)

Welcome to join our poster session in ICASSP 2023
Paper ID: 4354
Session Date/Time: 6/8/2023 14:00:00 (EEST)
Session Name: Speech Emotion Recognition: Multimodality

Thank you for your attention !

This work was funded by NSF (CNS-2016719; IIS-1453781)

Questions or Contact:
wei-cheng.lin@utdallas.edu
busso@utdallas.edu