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

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

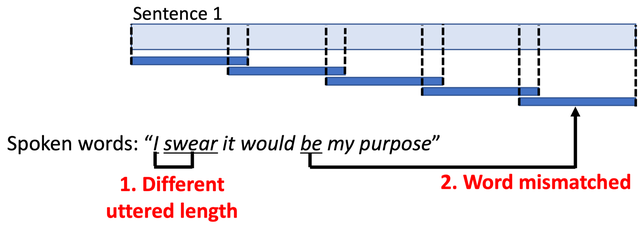
Wei-Cheng Lin and Carlos Busso

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

Motivation

Background:

- Chunk-level speech emotion recognition (SER):
  - Conventional time-based segmentation does NOT consider the role of lexical boundaries
  - Is there any benefit in providing precise lexical boundary information to segment the speech into chunks
    - e.g., word-level alignments?



Our Work:

- Investigating the role lexical boundary information plays in data chunks segmentation for chunk-level SER

Resources

Datasets:

- The MSP-PODCAST v1.10 corpus
  - Largest spontaneous speech emotion corpus collecting from existing podcast recordings
  - Includes 63,076 (train), 10,999 (dev), 16,903 (test) clips (~166hrs)
  - Regression problem: arousal, dominance, and valence
- The USC-IEMOCAP corpus
  - Contains 10,039 clips (~12hrs)
  - Leave-one-session-out cross-validation
  - Regression task: arousal, dominance, and valence

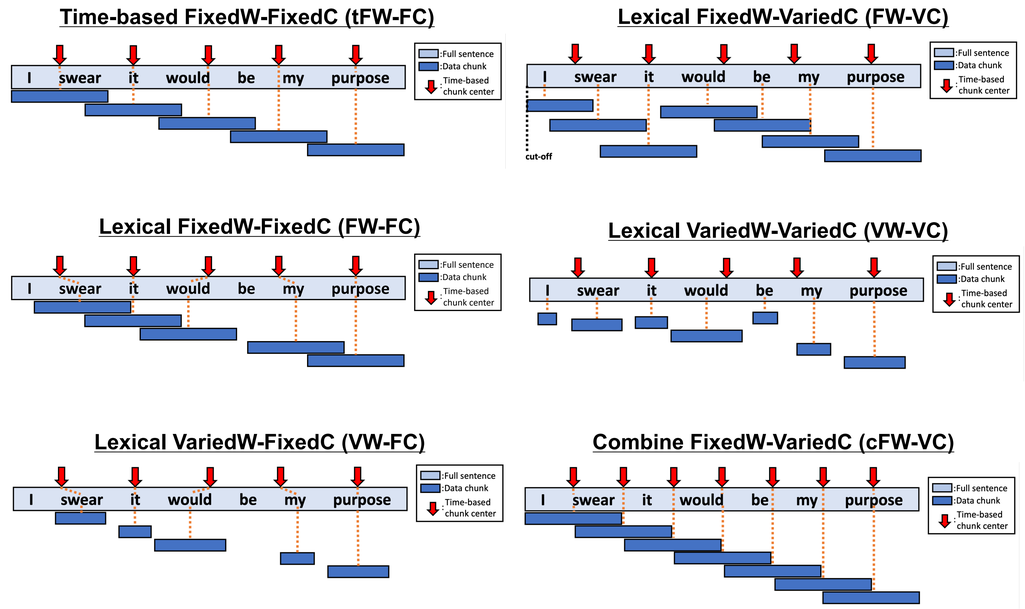
Acoustic Features:

- Low-level descriptors (LLDs, 130D)
- Wav2vec2.0 (pre-trained model, 1,024D)

Word-level Alignment:

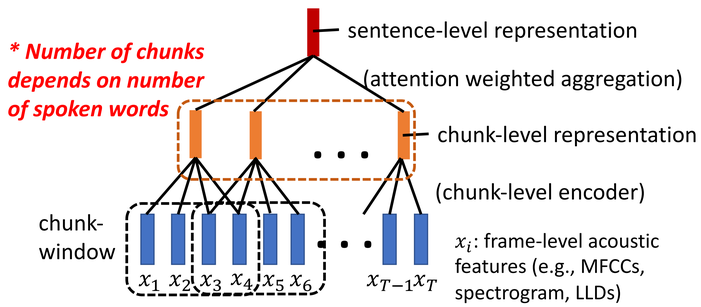
- Both datasets provide transcriptions
- Word boundary with Montreal forced aligner (MFA)

Different Segmentation Options



Chunk-Level SER Modeling

Hierarchical temporal-info summarization



Model Setups [1]:

- LSTM chunk-level encoder
- Multi-heads self-attention aggregation
- Loss/evaluation metric:
  - Concordance correlation coefficient (CCC)

Experimental Results and Analysis

Method	*CovrR/OvrR* [%]	*LLDs* (CCC) Aro.	Val.	Dom.	*Wav2Vec2* (CCC) Aro.	Val.	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							
Method	*CovrR/OvrR* [%]	*LLDs* (CCC) Aro.	Val.	Dom.	*Wav2Vec2* (CCC) Aro.	Val.	Dom.
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*

Knowing the precise word boundary does NOT bring significant performance benefits!

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

* means statistically significant better performance over other approaches without a marker
† means the results are statistically significant better than all other approaches

Conclusions

- We found a minor performance role of using word-level timing boundaries for chunk-level SER
- The key benefit provided by lexical information in the chunk segmentation process is the number of words
  - It can determine the number of chunks to segment a sentence

Future Work

- Explore benefit of multimodal lexical segmentation
  - e.g., video-speech-text

References:
[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.

This work was supported by NSF under Grant CNS-2016719