INTERSPEECH 2020

OCTOBER 25-29/ SHANGHAI , CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

# An Efficient Temporal Modeling Approach for Speech Emotion Recognition by Mapping Varied Duration Sentences into Fixed Number of Chunks

Wei-Cheng Lin and Carlos Busso

UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

UT Dallas MSP
Multimodal Signal Processing Laboratory

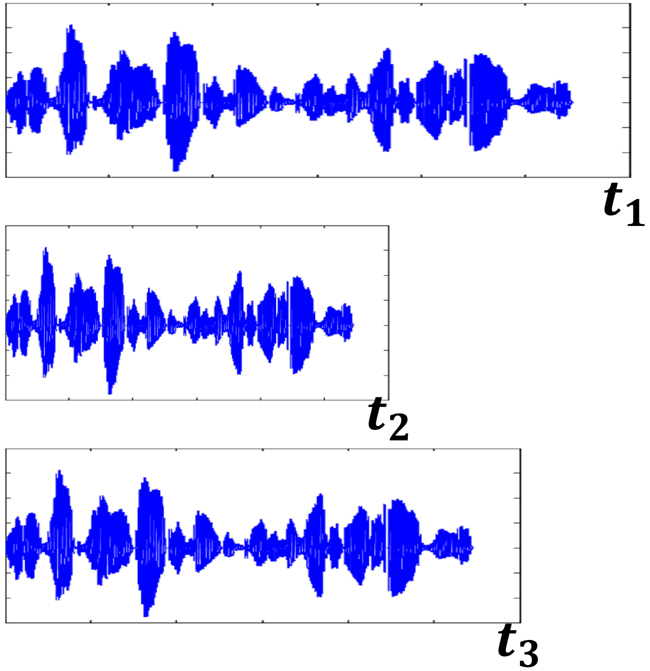
# Outline

1. Background
2. Proposed Methodology
3. Experimental Results and Analysis
4. Conclusions & Future Works

# Speech Emotion Recognition (SER)

- **Sequence-to-One Problem**

**<u>Varied Duration Inputs</u>**




**Temporal Modeling**

**Sentence-level Representation**

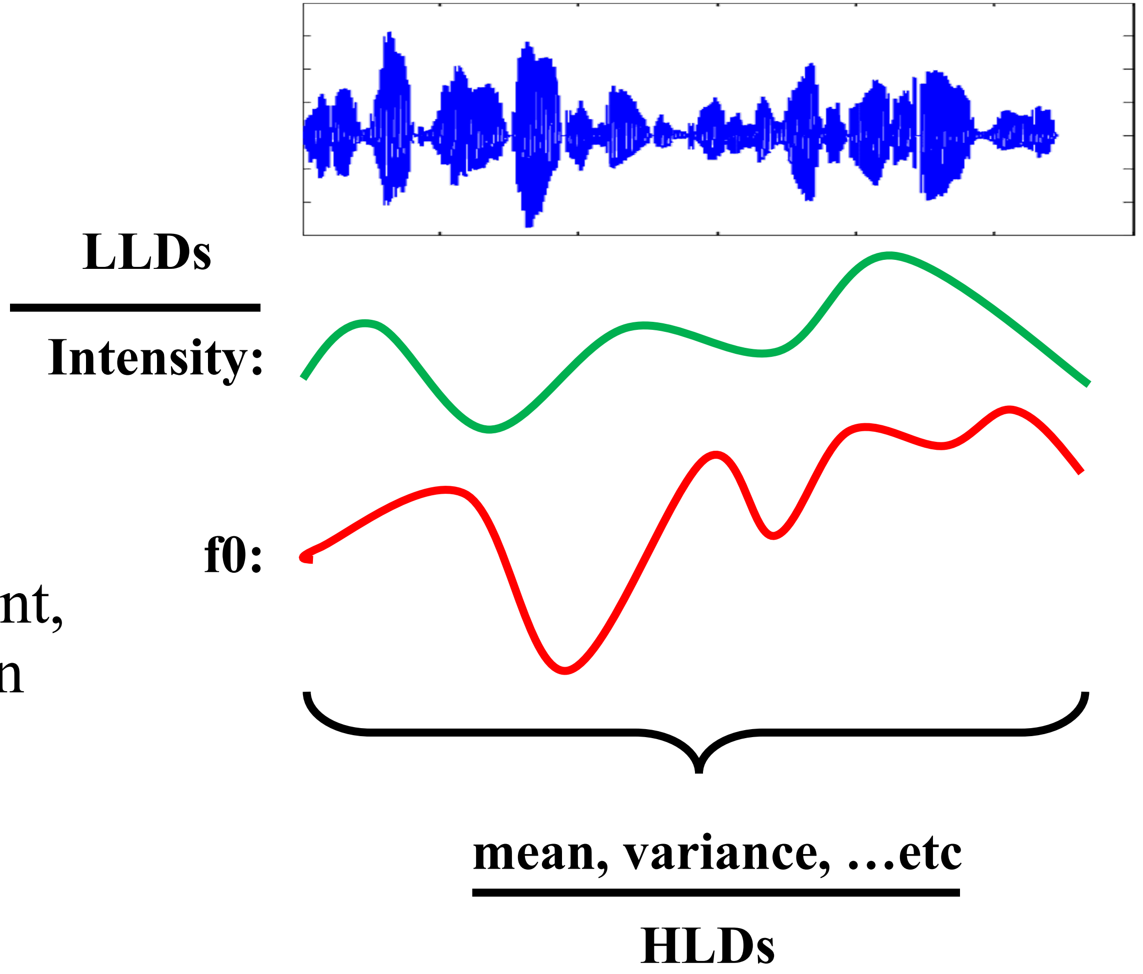
*But How?*

**<u>Sentence-level Emotions</u>**

# Temporal Modeling



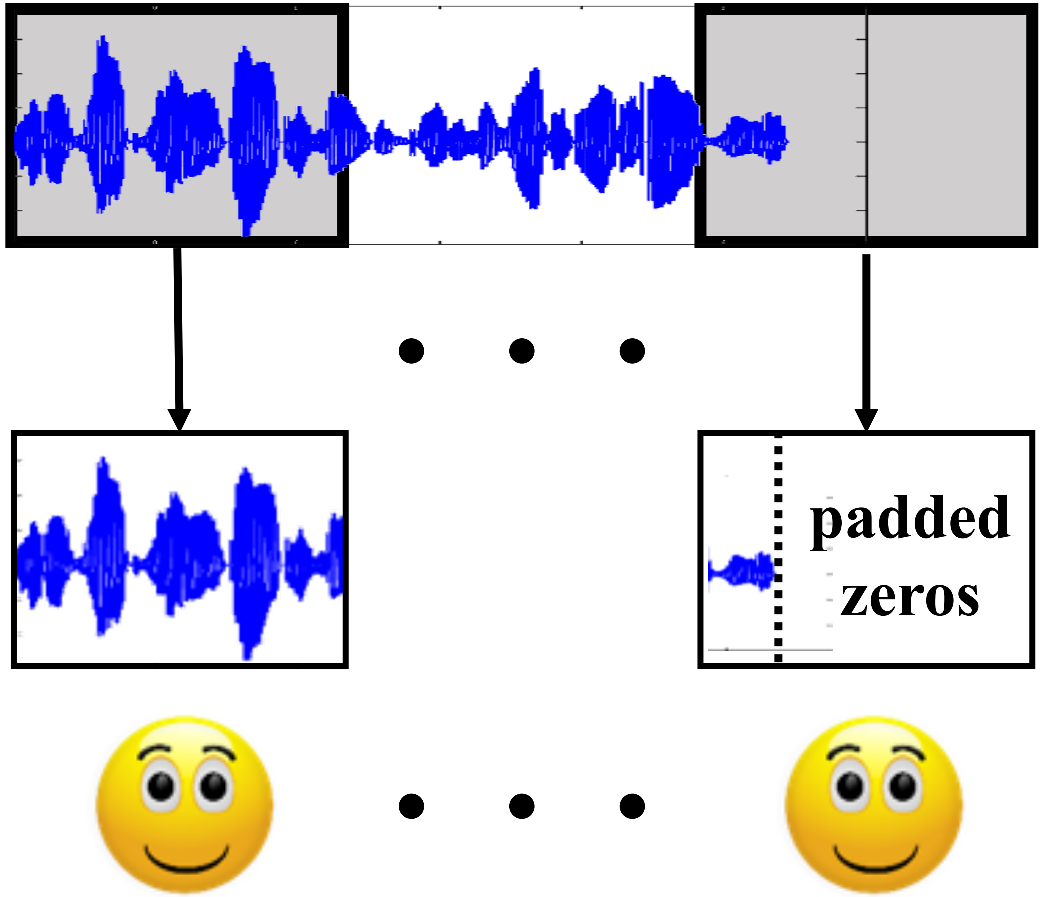
- **Traditional Approach**
  - Frame-level LLDs (e.g., f0, MFCCs, energy)
  - Sentence-level HLDs (e.g., mean, variance)
  - Learning models (e.g., SVM, FCNN)




- **Issues**
  - HLDs assume all the frames are equally important, ignoring non-uniform externalization of emotion
  - Static-encoding vector
  - It is not able to dynamically reflect emotional changes over time

# Temporal Modeling

- **Deep Learning Approach**
  - Jointly trained feature extractor with discriminator
    - Powerful feature representation
  - Cropping and Zero-Padding to deal with sentences with different length
  - CNN, LSTM or hybrid CNN-LSTM
    - Temporal mean pooling
    - Majority voting or averaging outputs
- **Issues**
  - Truncate original temporal information
  - Mean pooling/majority voting/averaging outputs still treat every segment the same (i.e., **non-dynamic temporal modeling**)



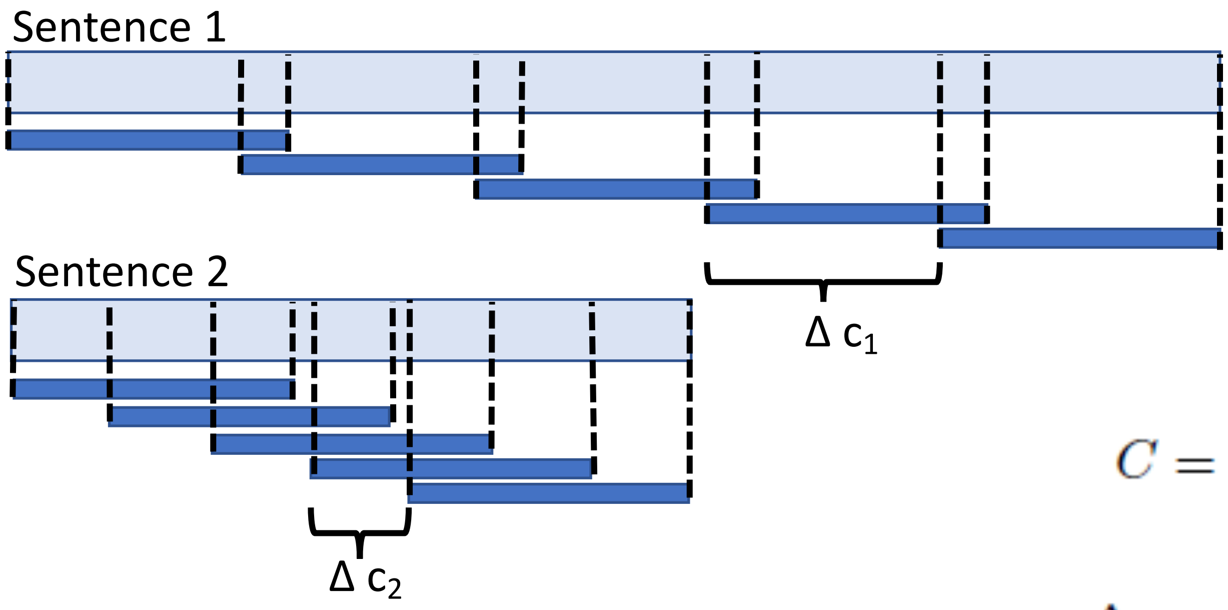
# Proposed Methodology

- **Goal:** dynamic temporal modeling
- **Key Problem:** sentences have varied number of segments/frames
- **Proposed:** Novel Chunk Segmentation Process

**Fixed size and fixed number of chunks**
**for different duration sentences**

# Proposed Methodology

- Chunk Segmentation Process




**No zero-padding is required!**

<u>Pre-defined Parameters:</u>

1. $T_{max}$ (sec): maximum sentence duration in the corpus
2. $w_c$ (sec): desired chunk window length

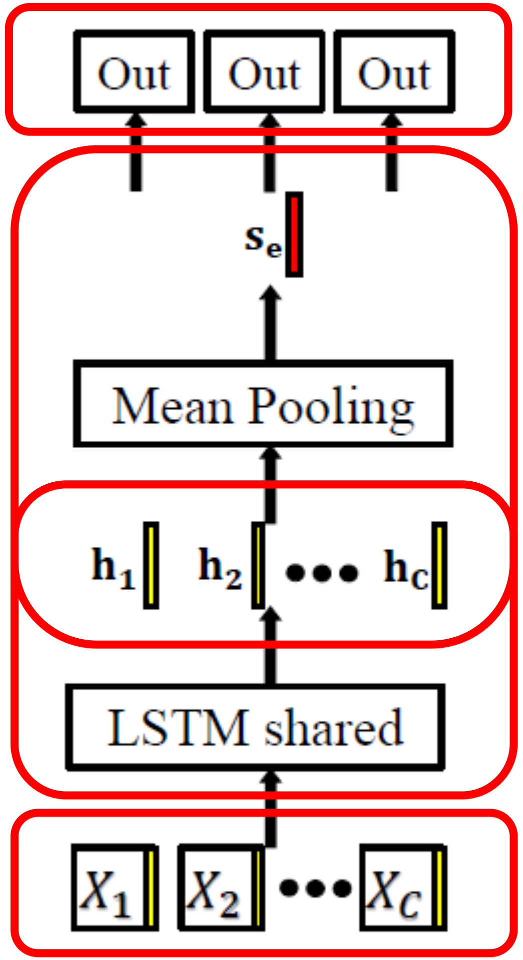
$$C = \left\lceil \frac{T_{\max}}{w_c} \right\rceil$$ : number of chunks per sentence

$$\Delta c_i = \frac{T_i - w_c}{C - 1}$$ (sec): chunk step size depends on sentence duration

# Proposed Methodology

: Chunk-level feature representation

: Sentence-level feature representation




**Step4: multi-task learning outputs**

**Step3: sentence-level feature representation**

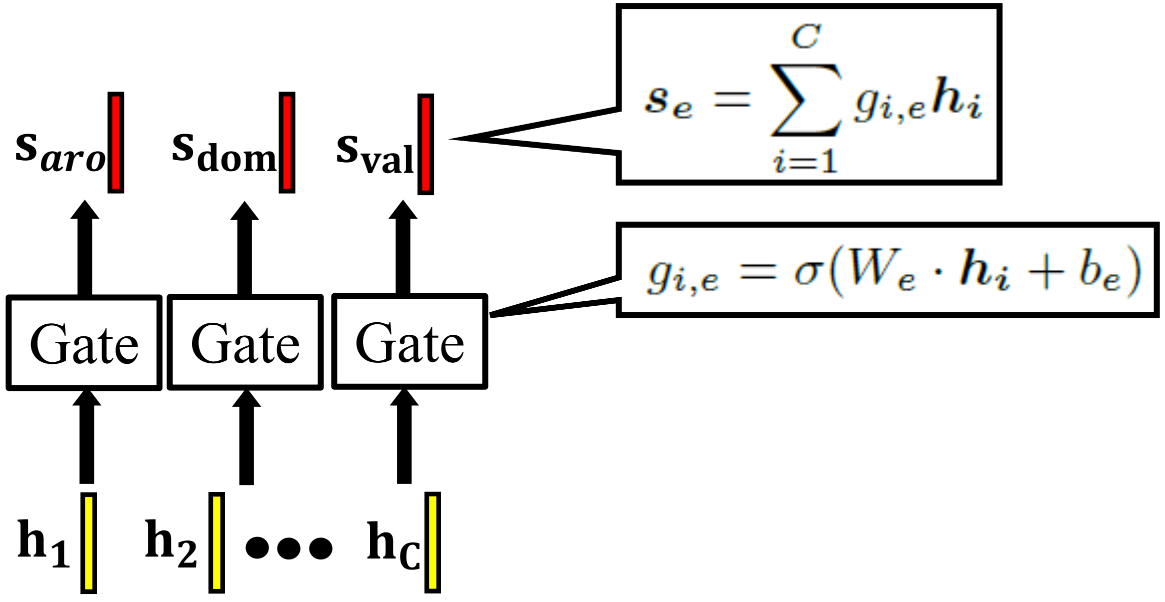
**Step2: chunk-level feature representation**

**Step1: chunk segmentation**

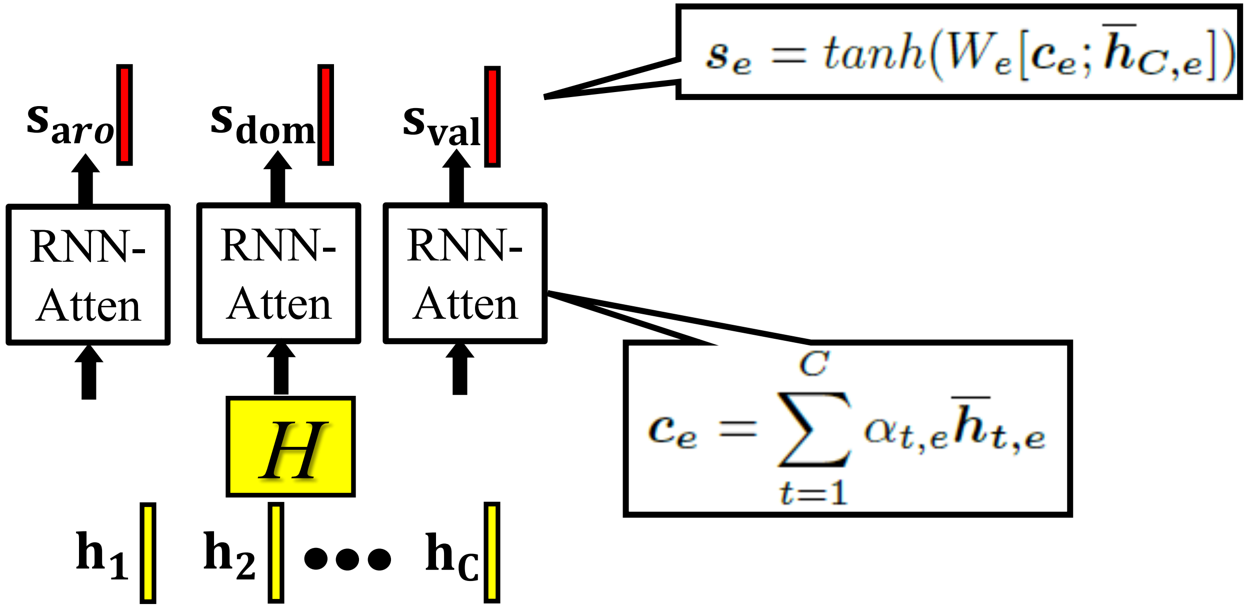
**NonAtten**

# Proposed Methodology

- Dynamically Combining Chunk-Based Feature Representations



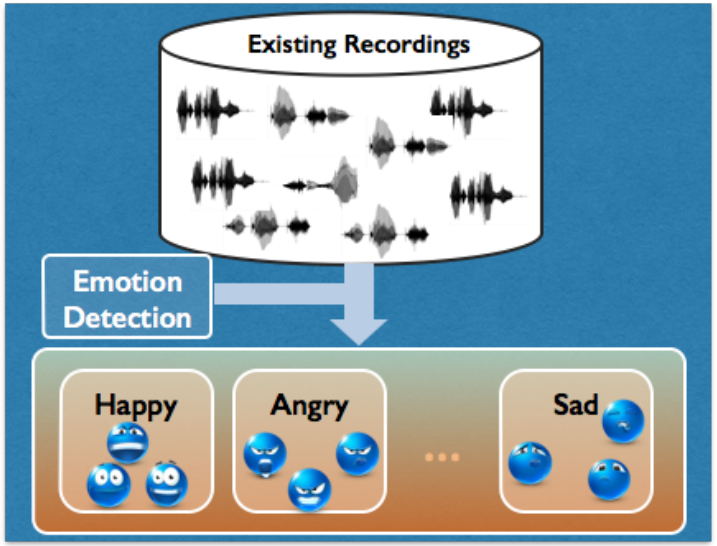

**GatedVec**

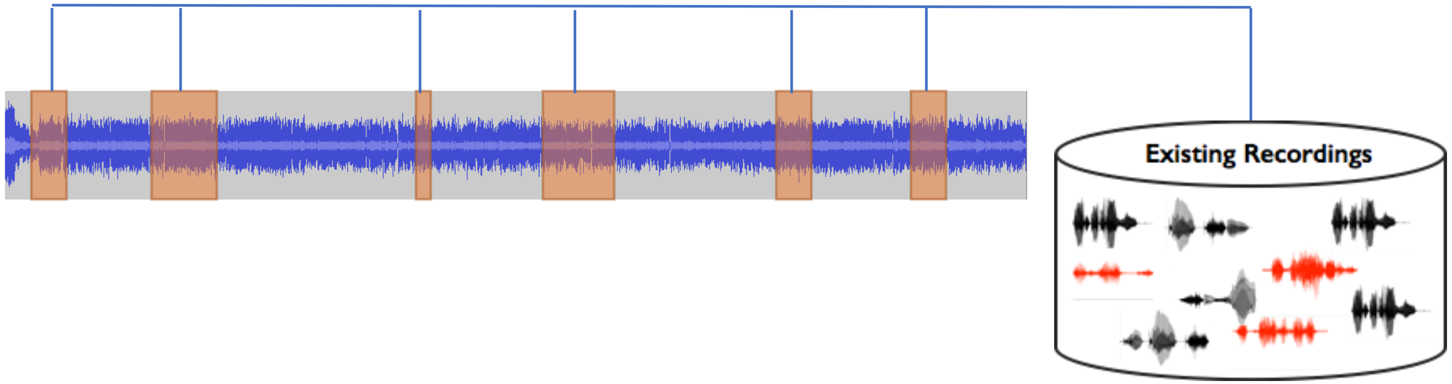



**AttenVec**

# Experimental Settings

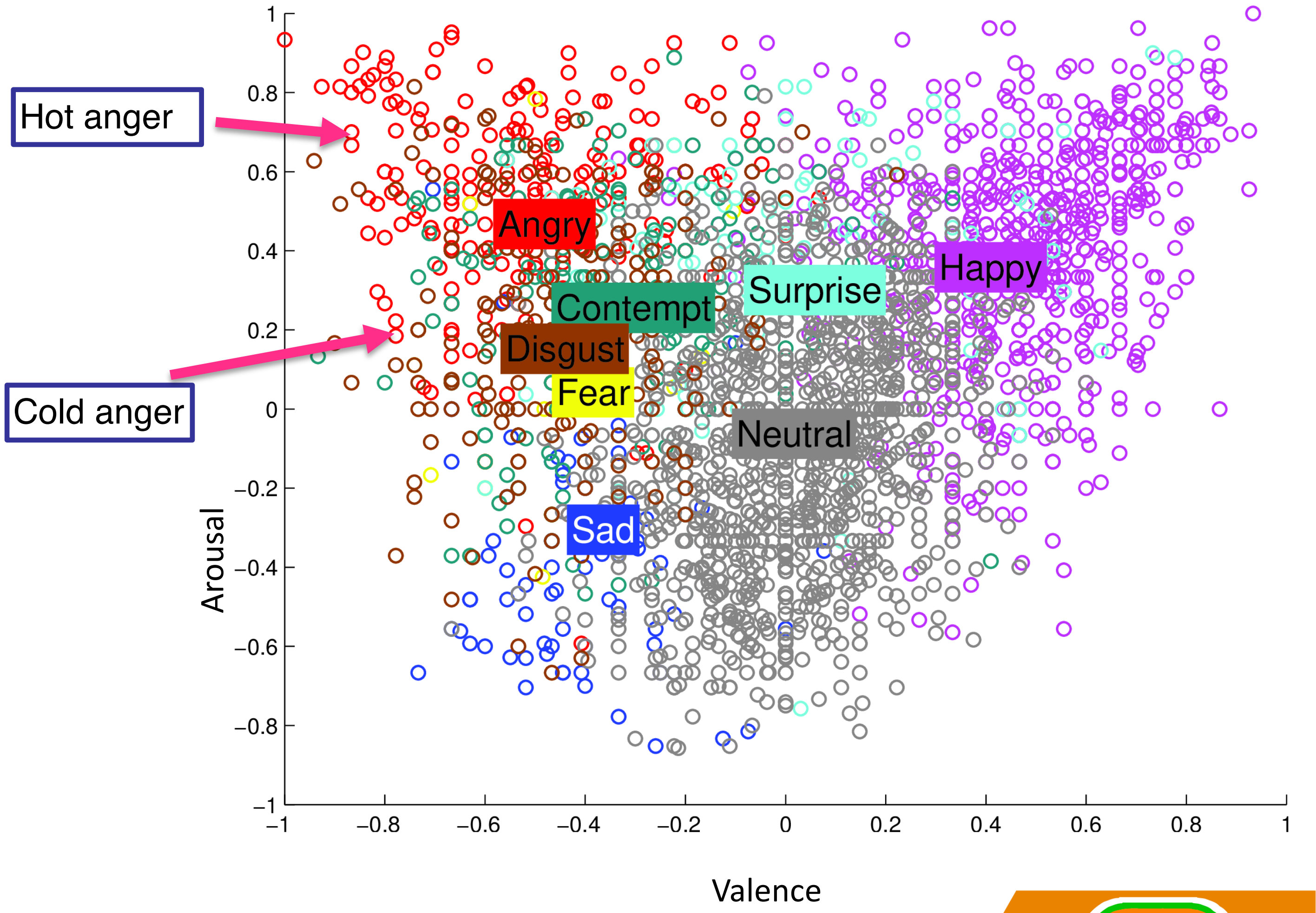
- Corpus: The MSP-Podcast v1.6
  - Use existing podcast recordings
  - Divide into speaker turns
  - Emotion retrieval to balance the emotional content
  - Annotate using crowdsourcing framework



# Experimental Settings

- The MSP-Podcast v1.6
  - 50,362 (83h,29m)
  - duration range: 2.75 ~ 11 secs
- Corpus partition with minimal speaker overlap sets:
  - Test data
    - 10,124 samples from 50 speakers (25 males, 25 females)
  - Validation data
    - 5,958 samples from 40 speakers (20 males, 20 females)
  - Train data
    - Remaining 34,280 samples



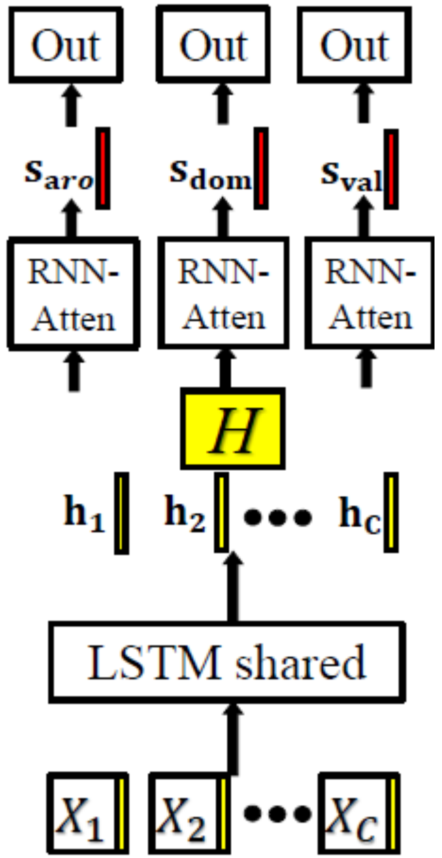
# Experimental Settings



- Acoustic Features
  - Opensmile
  - Interspeech 2013 computational paralinguistics challenge (IS13ComparE)
  - 65 low level descriptors (LLDs) + its delta value = 130-dimensions in total
    - mel frequency cepstral coefficients (MFCCs)
    - Fundamental frequency (f0)
    - Intensity (energy)
    - …

# Experimental Settings

- **Parameters Settings:**
  - MTL tasks: arousal (Aro.), dominance (Dom.) and valence (Val.)
  - $w_c$ : 1 (sec)
  - $T_{max}$ : 11 (secs)
  - C = 11 (chunks/per sentence)
  - Network nodes: 130 for all layers (the same as input-dim)
  - Adam optimizer
  - 128 batch size
  - Loss function: concordance correlation coefficient (CCC)
    - Same as the evaluation metric



# Experimental Results

- **Baseline Models:**
  - Padding zeros to the max length (i.e., 11 secs) for all sentences
  - *LSTM(130),* number of nodes in the LSTM shared layers is 130 number of nodes
  - *LSTM(260),* number of nodes in the LSTM shared layers is 130 number of nodes
- ***Best performance for all emotional attributes***

| Model | Aro [CCC] | Dom [CCC] | Val [CCC] |
|---|---|---|---|
| *LSTM (130)* | 0.6520 | 0.5711 | 0.2031 |
| *LSTM (260)* | 0.6875 | 0.6045 | 0.2847 |
| *NonAtten* | 0.6781 | 0.6019 | **0.2925** |
| *GatedVec* | 0.6747 | 0.5944 | **0.3199** |
| *AttenVec* | **0.6947** | **0.6132** | **0.3072** |

# Experimental Results

- **Improvement of model efficiency**
  - *Chunks are parallel processed by GPU*
  - *Significant reduction in MFLOPs (i.e., roughly C=11 times faster)*

| Model | # of Par. [$10^6$] | MFLOPs [MFLOPS] | Train [sec/epoch] | Online [ms/uttr] |
|---|---|---|---|---|
| *LSTM (130)* | 0.323 | 5.67 | 437.1 | 547.5 |
| *LSTM (260)* | 1.052 | 18.49 | 439.4 | 598.1 |
| *NonAtten* | 0.323 | **0.49** | **74.9** | **42.2** |
| *GatedVec* | 0.324 | **0.49** | 246.6 | **44.6** |
| *AttenVec* | 0.577 | **1.50** | 353.1 | **45.6** |

# Experimental Results

- Robustness for different duration
  - Short (≤ 5 secs): 4,280 sentences
  - Middle (5-8 secs): 3,684 sentences
  - Long (≥ 8 secs): 2,160 sentences
- ***Robust to different duration inputs (especially for long sequences)***

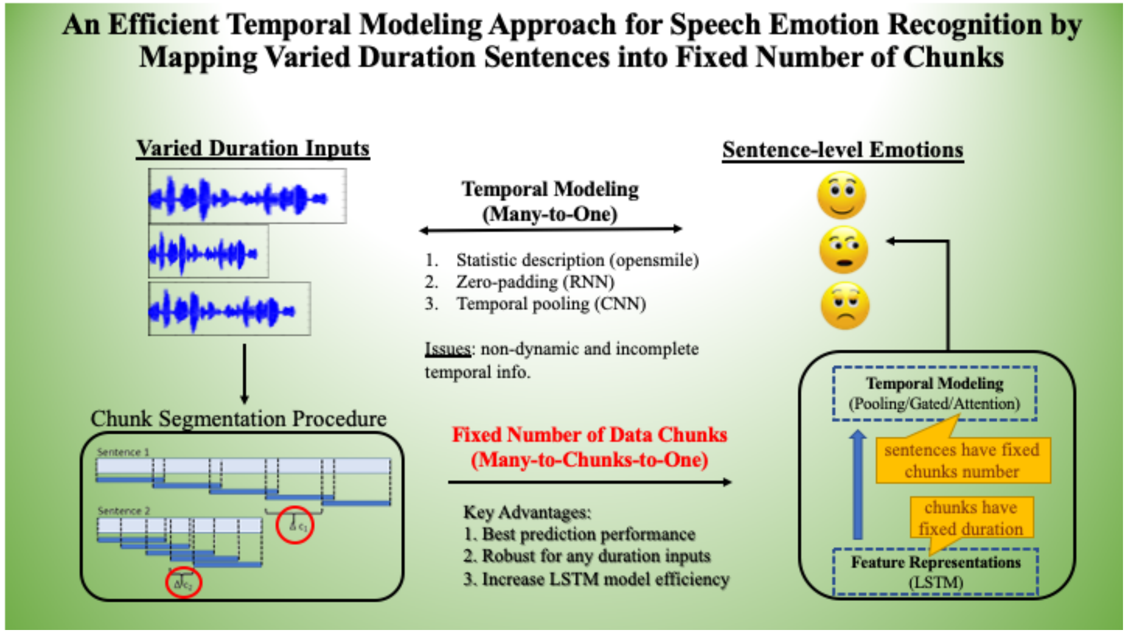
| Short ( < 5sec) | Aro-CCC | Dom-CCC | Val-CCC |
|---|---|---|---|
| LSTM(130) | 0.6636 | 0.5812 | 0.2389 |
| NonAtten | 0.6761 | 0.6077 | 0.3129 |
| GatedVec | 0.6621 | 0.5865 | 0.3263 |
| AttenVec | 0.7003 | 0.6192 | 0.3363 |

| Long ( > 8sec) | Aro-CCC | Dom-CCC | Val-CCC |
|---|---|---|---|
| LSTM(130) | 0.6314 | 0.5559 | 0.1737 |
| NonAtten | 0.6811 | 0.5822 | 0.2331 |
| GatedVec | 0.6933 | 0.6030 | 0.2835 |
| AttenVec | 0.6912 | 0.5989 | 0.2539 |

| Middle ( 5~8sec) | Aro-CCC | Dom-CCC | Val-CCC |
|---|---|---|---|
| LSTM(130) | 0.6484 | 0.5642 | 0.1735 |
| NonAtten | 0.6779 | 0.6071 | 0.2839 |
| GatedVec | 0.6807 | 0.6042 | 0.3279 |
| AttenVec | 0.6880 | 0.6129 | 0.2978 |

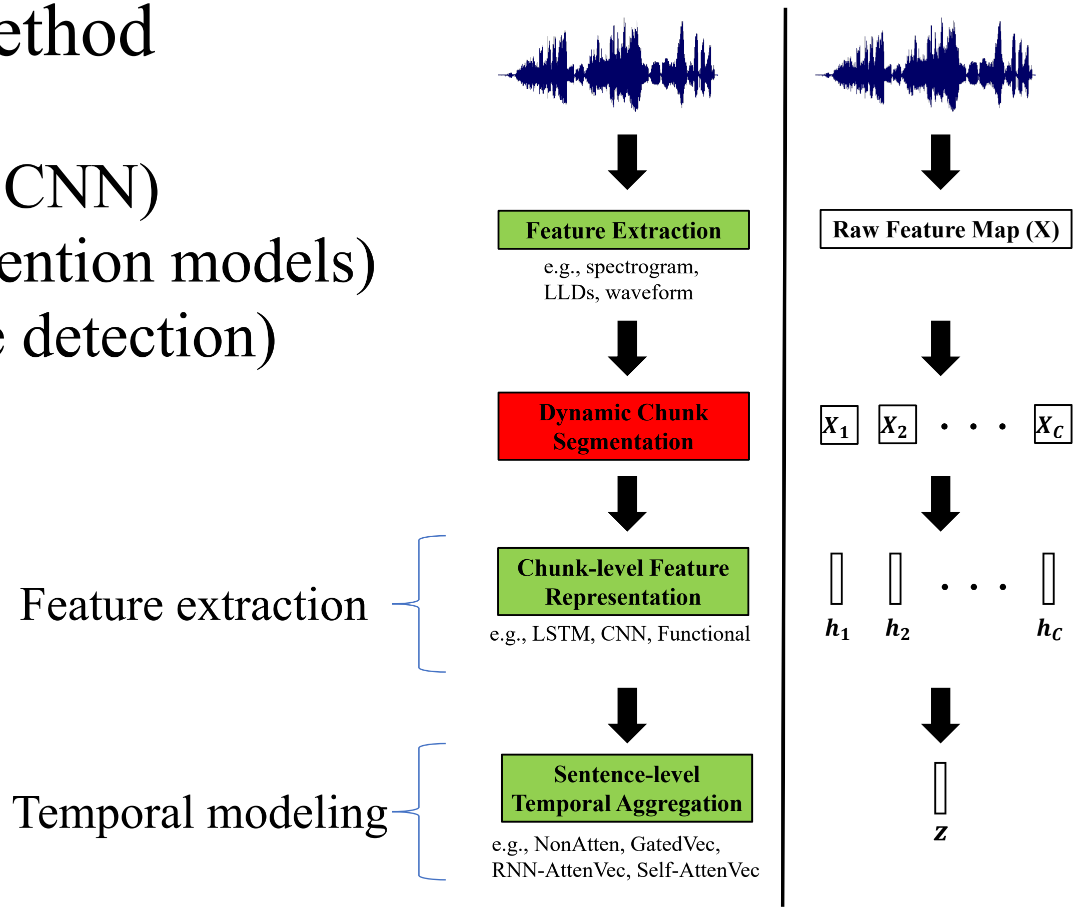
# Conclusions

- Novel segmentation approach that can split a sentence into a **fixed number** of chunks, which have **the same** duration
- **Flexibly** and **dynamically** combine temporal information
- Best prediction **performance**
- Improve model **efficiency**
- **Robust** for different duration sentences



# Future Works

- General framework of the proposed method
  - Multiple datasets
  - Different feature extraction models (e.g., CNN)
  - Different temporal modeling (LSTM, Attention models)
  - Different sequence-to-one tasks (e.g., age detection)



# Release of the MSP-Podcast Corpus

- **Academic license**
  - Federal Demonstration Partnership (FDP) Data Transfer and Use Agreement
  - Free access to the corpus
- **Commercial license**
  - Commercial license through UT Dallas


https://msp.utdallas.edu

# Thank you for your attention !

This work was funded by NSF
(CNS-1823166; IIS-1453781)

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