
# DeepEmoCluster: A Semi-Supervised Framework for Latent Cluster Representation of Speech Emotions

Wei-Cheng Lin, Kusha Sridhar and Carlos Busso

UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

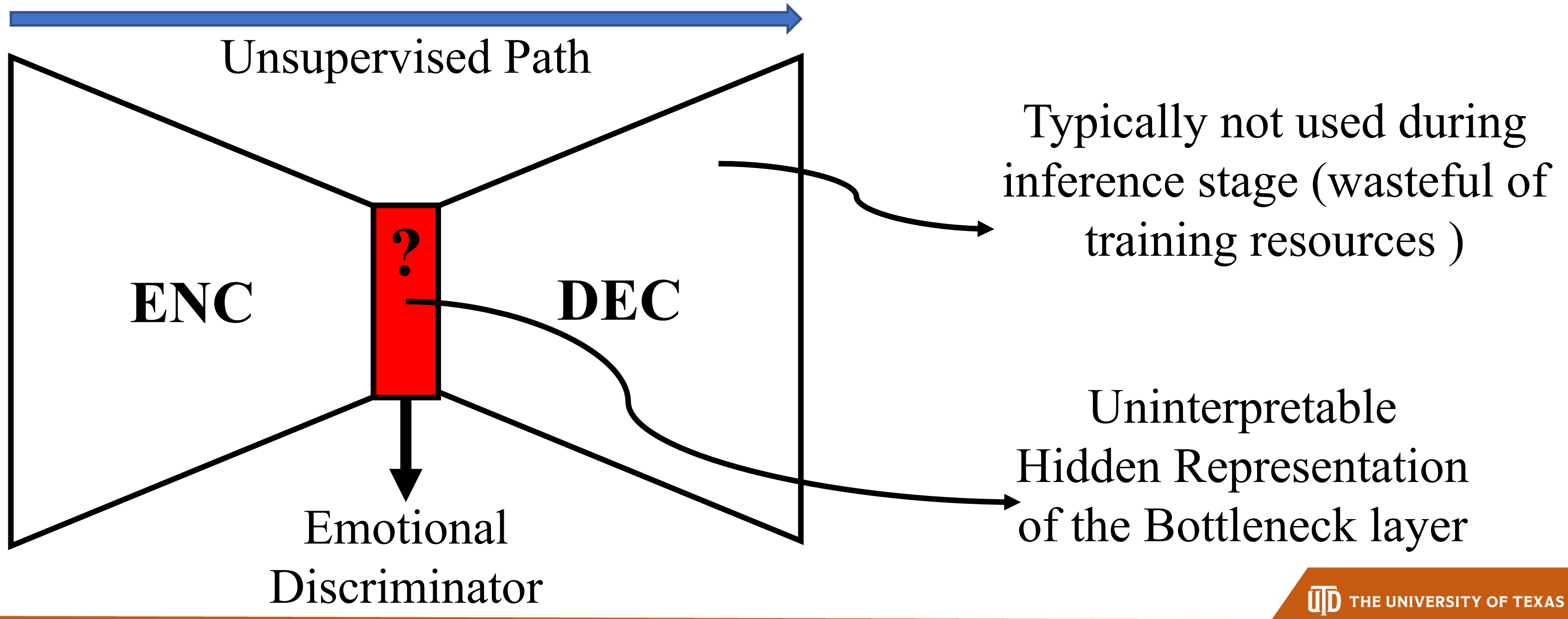
UT Dallas MSP Multimodal Signal Processing Laboratory

# Outline

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

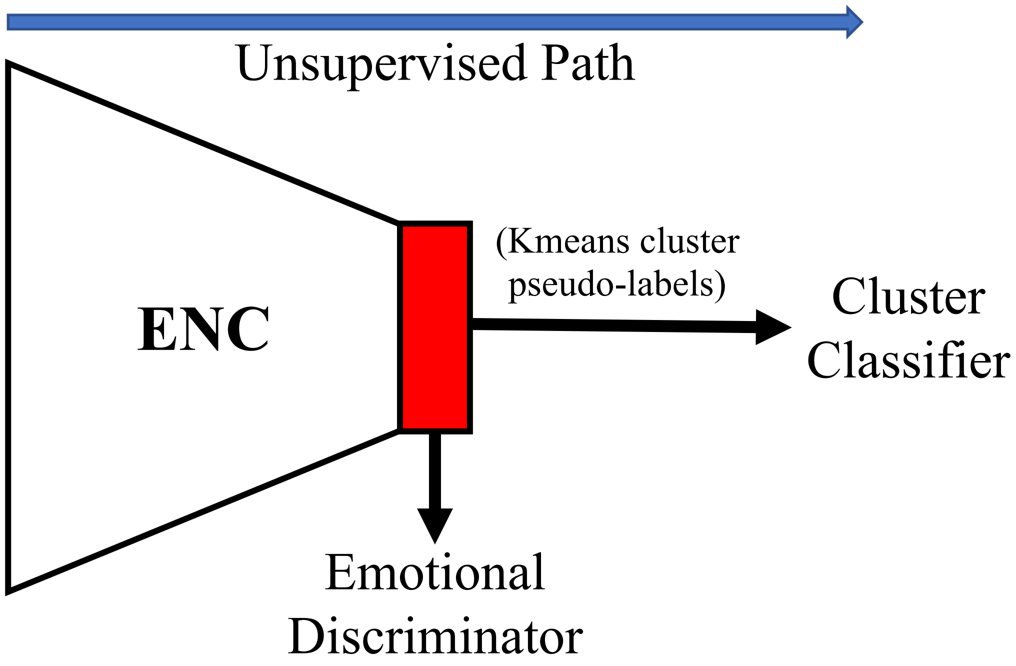
# SSL in SER

- **Semi-Supervised Learning (SSL)**
  - Leverage large amounts of unlabeled data to improve recognition generalization ability
- **SSL in Speech Emotion Recognition (SER)**
  - Reconstruction-based architecture (i.e., AE [1], VAE [2] and LadderNet [3])



# Proposed Methodology

- **DeepEmoCluster framework**
  - Inductive SSL scheme (i.e., pseudo-labeling by K-means clustering)

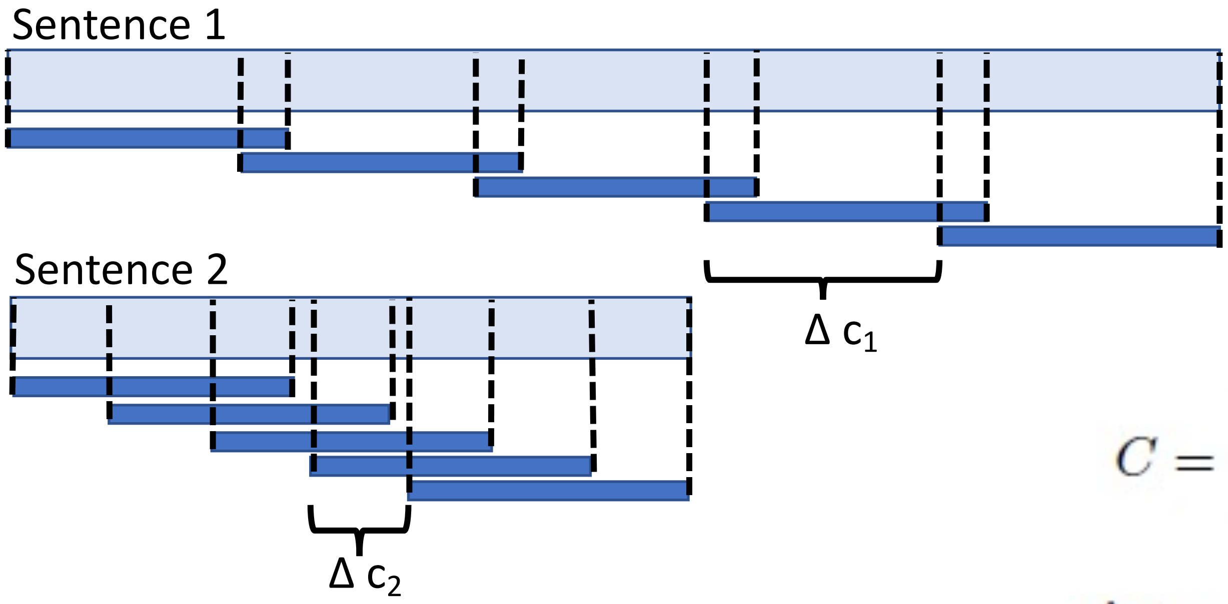



1. No **Decoder** in the framework
2. Meaningful and interpretable hidden representations, resulting in **emotional clusters**!

# DeepEmoCluster



- **End-to-End SSL framework**
  - Input with 128-Mel spectrogram (32ms window size and 16ms overlaps)
  - **Step 1:** chunk-segmentation pre-processing [Lin and Busso, 2020]




No zero-padding is required!

Pre-defined Parameters:

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

Create equation in ppt

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

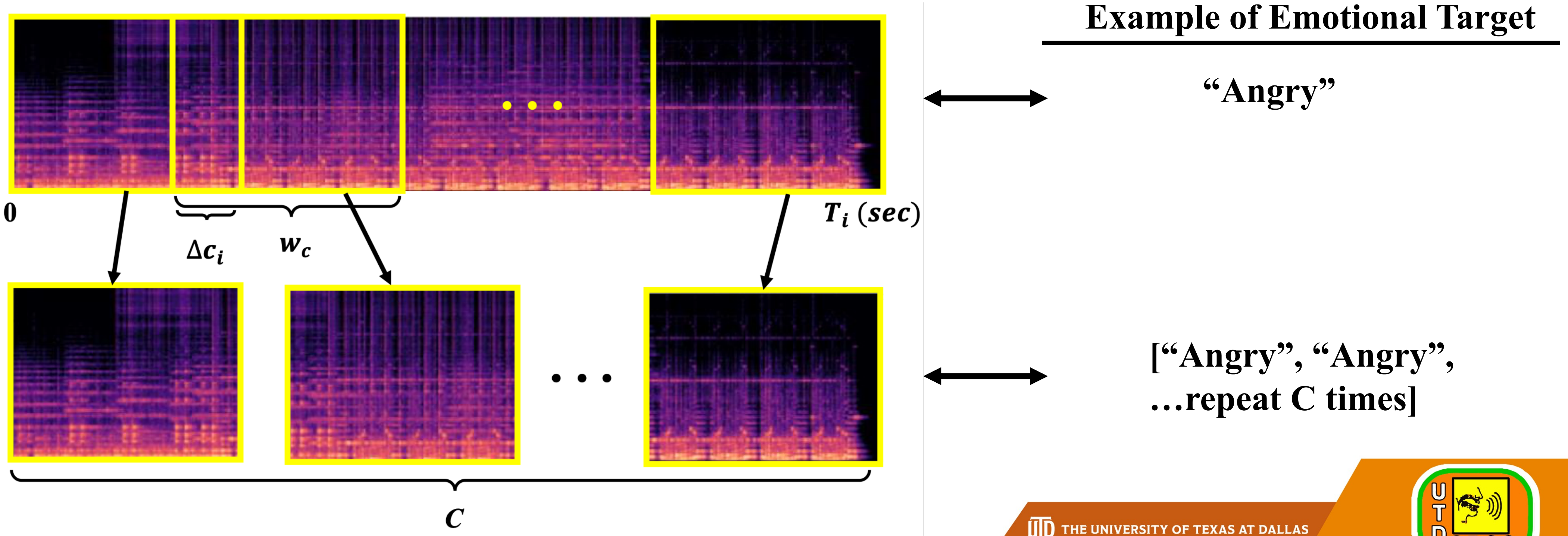
$$\Delta c_i = \frac{T_i - w_c}{C - 1}$$ (sec): chunk step size depends on sentence duration



Wei-Cheng Lin and Carlos Busso, "An Efficient Temporal Modeling Approach for Speech Emotion Recognition by Mapping Varied Duration Sentences into Fixed Number of Chunks," in Interspeech 2020, Shanghai, China, October 2020, pp. 2322-2326.

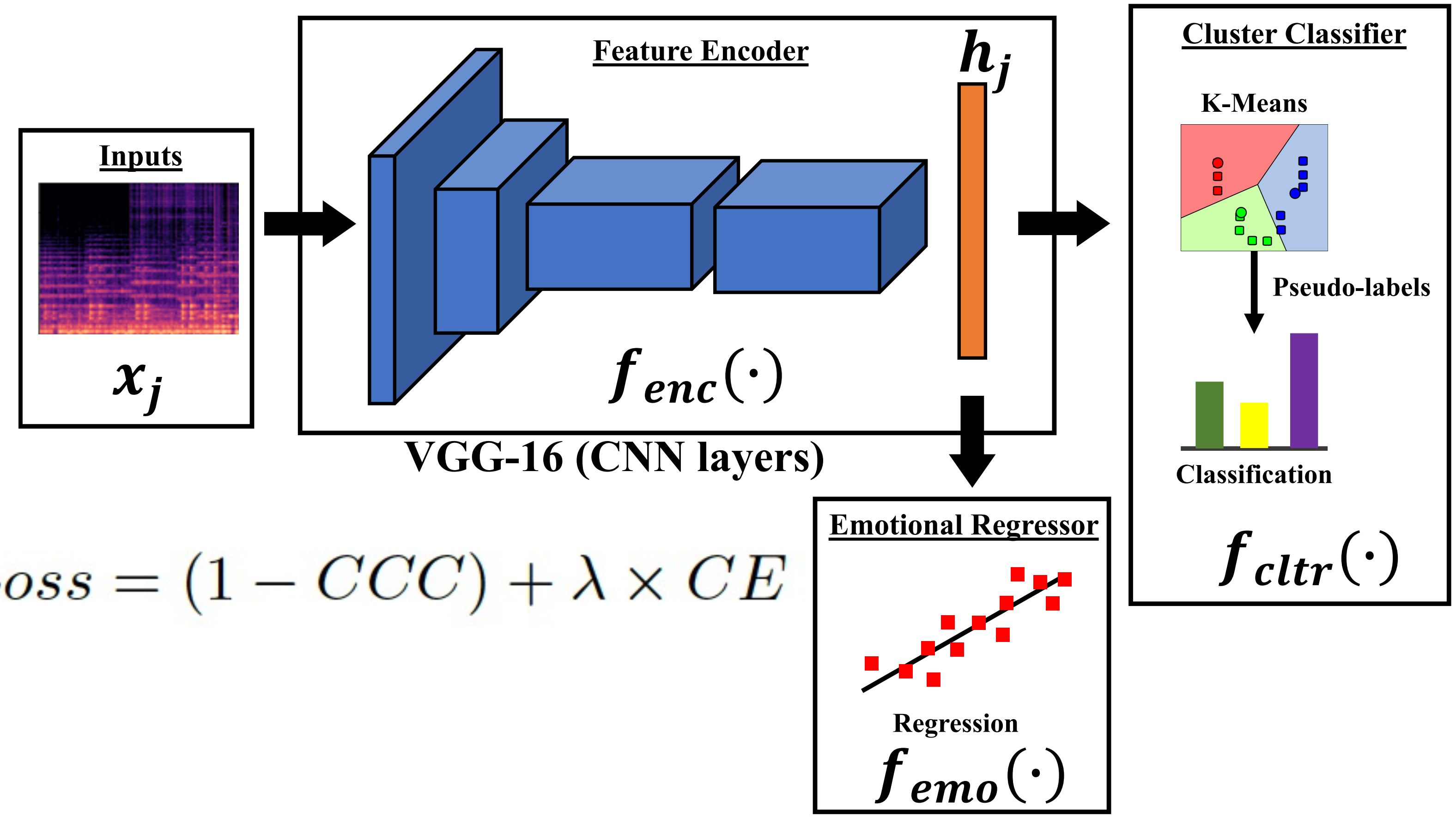
# DeepEmoCluster

- **Visualization of 128-Mel spectrogram data chunks**
  - Originally arbitrary length of audios are mapped into fixed size and fixed number of **“small spec-images”** as the inputs.
  - These images are shared the same sentence-level emotional target during training procedure.



# DeepEmoCluster

- **Step 2:** unsupervised (Stage I) + joint-optimization (Stage II)



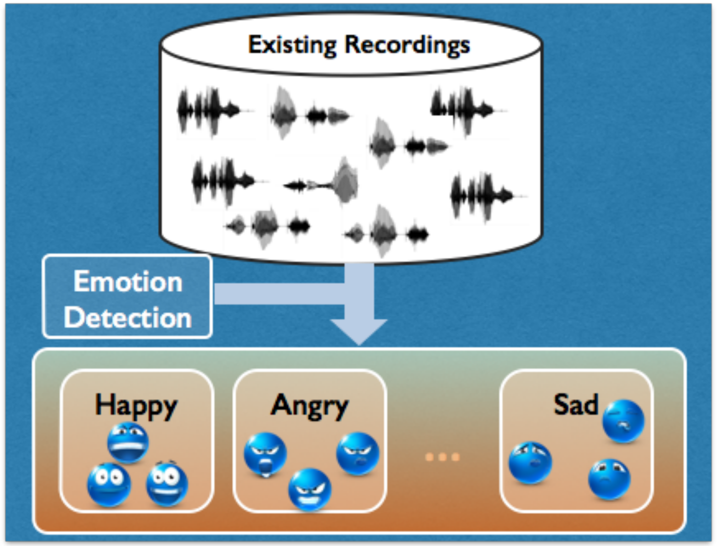

$$Loss = (1 - CCC) + \lambda \times CE$$

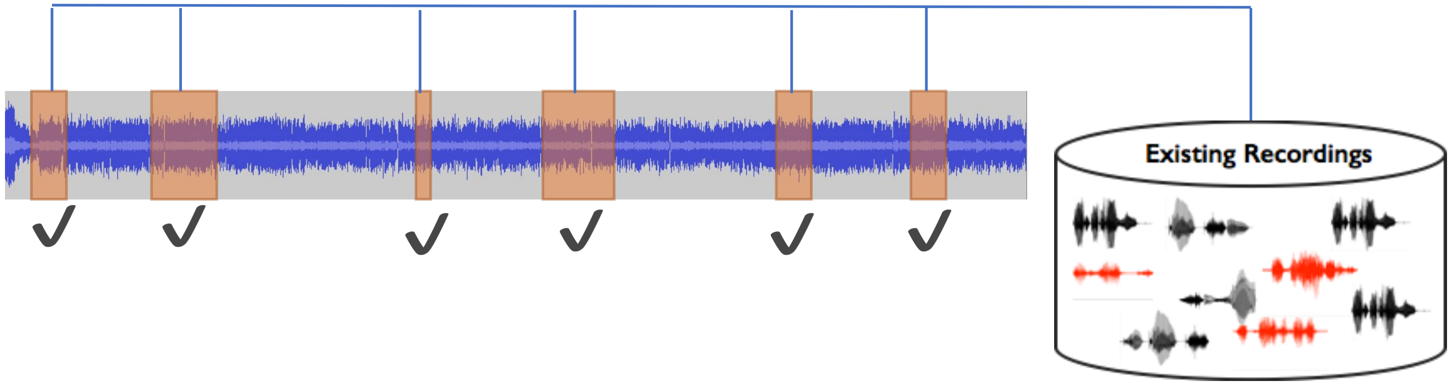
**Some details**

1. Stage I is for the unlabeled data
2. Stage II is for the labeled data
3. We reassign the K-means clustering pseudo-labels on every new epoch

# Experimental Settings

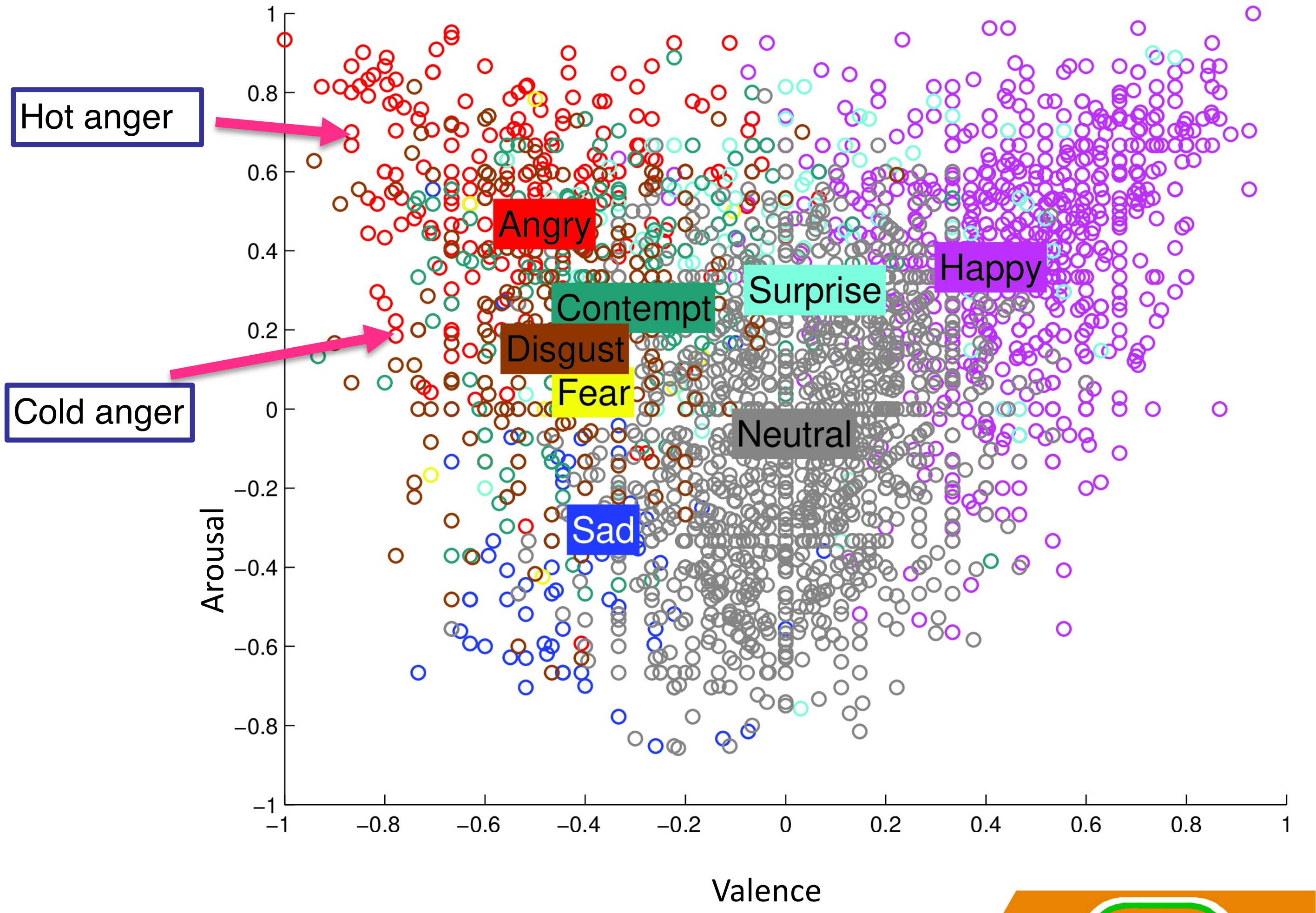
- Corpus: The MSP-Podcast v1.6
  - Use existing podcast recordings
  - Divide into speaker turns
  - Emotion retrieval to balance the emotional content
  - Annotate using crowdsourcing framework



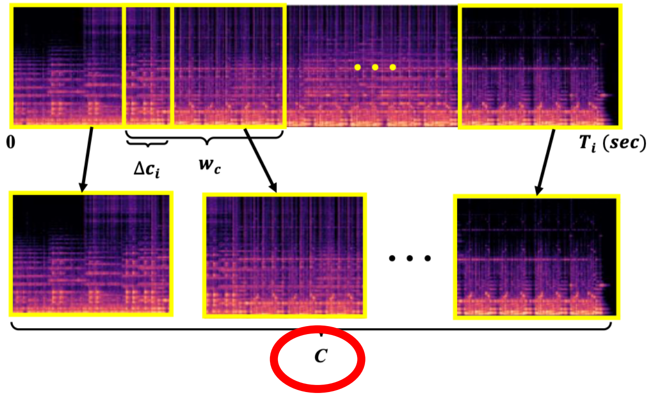
# Experimental Settings

- 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
    - 50 speakers (25 males, 25 females)
  - Development data: 5,958 samples
    - 40 speakers (20 males, 20 females)
  - Train data: 34,280 samples
    - from remaining speakers
  - Unlabeled data
    - Totally around 500,000 samples



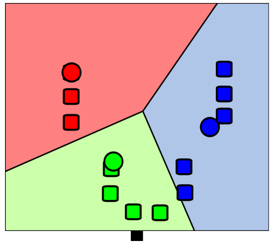
# Experimental Settings

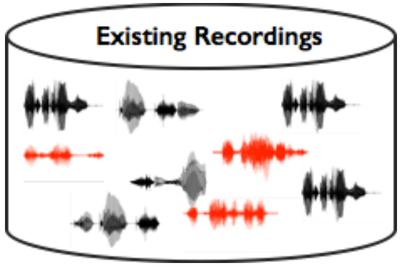
- **Parameters Settings:**
  - $w_c$:1 (sec), $T_{max}$:11 (secs) ➔ **C=11** (sub-images/per sentence)
  - Joint-optimization weighting factor of the loss function ***λ*=1**
  - 64 batch size, early stopping criteria
    (for saving the best model based on the min. development loss)
  - # of K-means cluster = [10, 20, 30]
    (finetuned parameter depending on the size of unlabeled set)
  - Size of unlabeled set = [0, 15K, 40K]
    (random sample from the unlabeled data pool)




$$Loss = (1 - CCC) + \lambda \times CE$$

K-Means





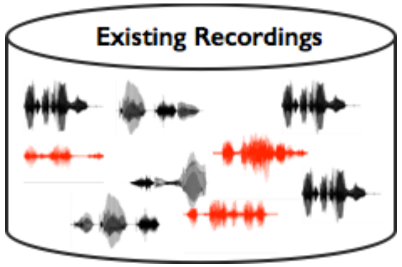
# Experimental Results

- **Recognition performance under *fully supervised learning* (FSL)**
- **Baseline Models (all use VGG-16 structure):**
  - *CNN-regressor*
  - *CNN-AE (autoencoder)*
  - *CNN-VAE (variational autoencoder)*
- ***Statistically significant outperforms baseline models***

| Model | Aro [CCC] | Dom [CCC] | Val [CCC] |
|---|---|---|---|
| *CNN-regressor* | 0.6177 | 0.4928 | 0.1696 |
| *CNN-AE* | 0.6338 | 0.5111 | 0.1354 |
| *CNN-VAE* | 0.5586 | 0.4800 | **0.1826** |
| *DeepEmoCluster (10-clusters)* | **0.6502** | **0.5426** | 0.1510 |

# Experimental Results

- **Further improved recognition performance while adopting *semi-supervised learning* (SSL)**
  - ADD KEY RESULTS



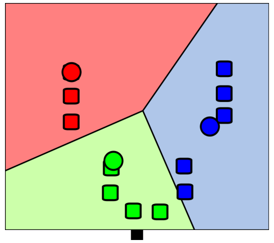

**DeepEmoCluster (10-clusters)**

| *Size of Unlabeled Data set* | Aro [CCC] | Dom [CCC] | Val [CCC] |
|---|---|---|---|
| *0 (FSL)* | 0.6502 | **0.5426** | 0.1510 |
| *15K (SSL)* | 0.6504 | 0.5400 | **0.1714** |
| *40K (SSL)* | **0.6611** | 0.5400 | 0.1572 |

# Experimental Results

- **Finetuned parameter of # K-means clusters**
  - ADD KEY RESULTS

**K-Means**

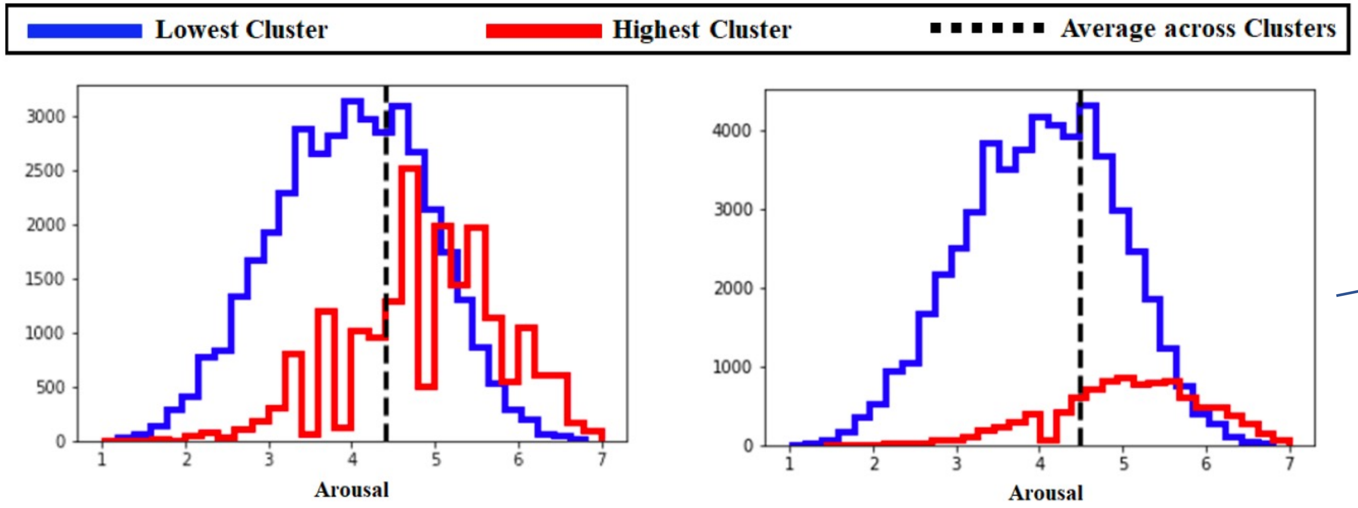


**DeepEmoCluster (SSL-40K)**

| *# of clusters* | Aro [CCC] | Dom [CCC] | Val [CCC] |
|---|---|---|---|
| *10-clusters* | **0.6611** | 0.5400 | 0.1572 |
| *20-clusters* | 0.6491 | 0.5459 | **0.1756** |
| *30-clusters* | 0.6416 | **0.5490** | **0.1752** |

# Experimental Results

- **Emotional clusters reflecting by the ground-truth emotion distributions under each cluster**
  - ADD KEY RESULTS




(a) Unsupervised Clusters (b) Supervised Clusters

**Fig. 3**. Emotional distributions of the clusters with the highest and lowest average level of arousal. The distributions are farther apart with the addition of the supervised SER task.

Add original images

# Conclusions

- We introduced a new SSL framework in SER field
- DeepEmoCluster achieved the best and competitive recognition performances comparing to other existing SSL frameworks in SER
- DeepEmoCluster could result in meaningful hidden representations
- We discussed and determined the important parameter
  - The number of K-means clusters is a finetuned parameter depending on the size of unlabeled dataset

# Future Works

- Extension of the framework from a single modality (speech) to a multimodal system (speech, language and visual)
  - Forming a comprehensive behavioral emotional clusters

- Strengthen the connections between the latent clusters and the target emotions by utilizing information theory based metric

# 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

# Reference

**[1]** J. Deng, X. Xu, Z. Zhang, S. Fr¨uhholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 31–43, January 2018.

**[2]** S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion," in Interspeech 2018, Hyderabad, India, September 2018, pp. 3107–3111.

**[3]** S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2697–2709, September 2020.

**[4]** W.-C. Lin and C. Busso, "An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks," in Interspeech 2020, Shanghai, China, October 2020, pp. 2322–2326.

# Thank you for your attention !

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

Github link:
https://github.com/winston-lin-wei-cheng/DeepEmoClusters
Questions or Contact:
**wei-cheng.lin@utdallas.edu**