Unsupervised domain adaptation for preference learning based speech emotion recognition

- Abinay Reddy Naini, Mary A. Kohler, Carlos Busso

Presentation by Abinay Reddy Naini

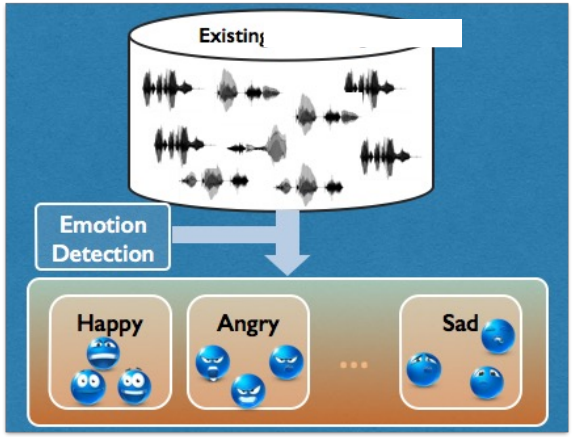
UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

UT Dallas MSP Multimodal Signal Processing Laboratory

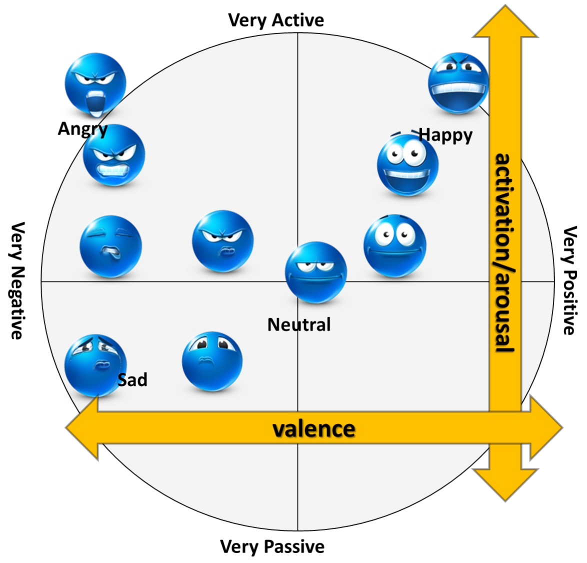
Role of Emotion Recognition

- **Emotion recognition is critical for the Intelligence Community (IC)**
  - Analyze massive amount of information available through media domains
  - Identify and preselect segments with potentially threatening behaviors



Expression of Emotion

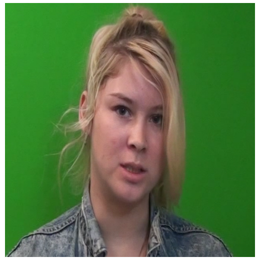
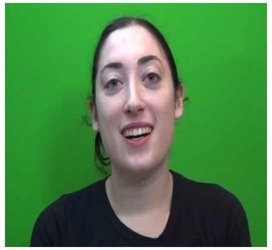
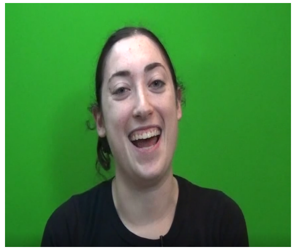
- **Categorical labels**
  - Anger, happiness, sadness, neutral
- **Dimensional or attribute-based labels**
  - Valence (negative versus positive)
  - Arousal (calm versus active)
  - Dominance (weak versus strong)
  - More accurate emotion descriptors (intensity)



Ordinal Representation of Emotion

- **Thesis: emotions are intrinsically ordinal (relative)**
  - The benefits of representing them that way are many!
  - This thesis is supported by theoretical arguments across disciplines and empirical evidence in Affective Computing

How positive
is this image?

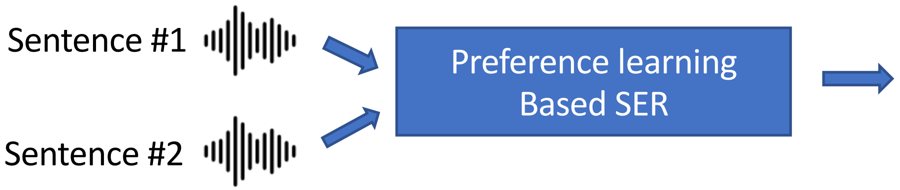

<

<


Georgios N. Yannakakis, Roddy Cowie, and Carlos Busso, "The ordinal nature of emotions: An emerging approach," IEEE Transactions on Affective Computing, vol. To appear, 2019

Georgios N. Yannakakis, Roddy Cowie, and Carlos Busso, "The ordinal nature of emotions," in International Conference on Affective Computing and Intelligent Interaction (ACII 2017), San Antonio, TX, USA, October 2017, pp. 248-255.

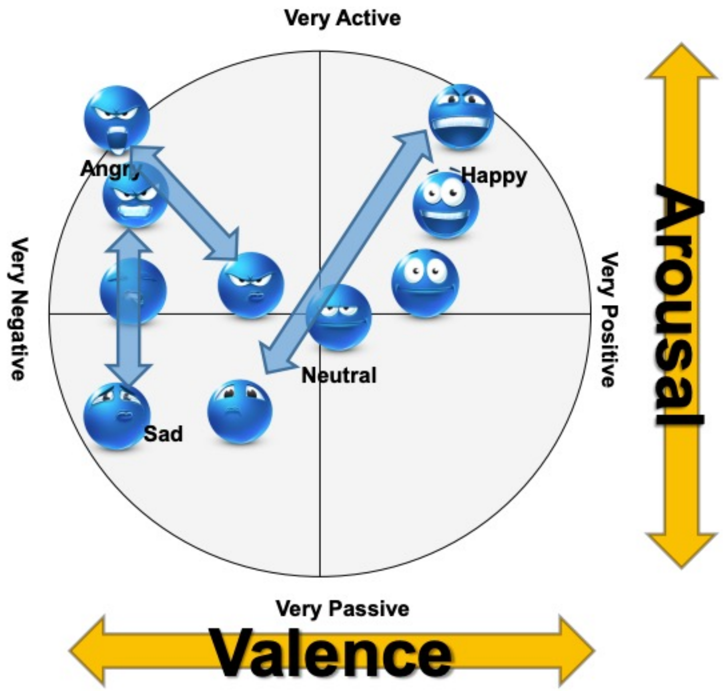
Preference learning formulation

- **Preference learning**



Arousal (Sentence #1 >> Sentence #2)
Valence (Sentence #2 >> Sentence #1)
Dominance (Sentence #2 >> Sentence #1)

- **Why preference learning?**
  - Humans are better at relative comparisons than absolute values
  - Appealing to Emotional Retrieval tasks
  - Better use of training data
    - N(N-1)/2 potential pairs



Getting preference labels ?

Qualitative Agreement

	Sentence 1	Sentence 2
Rater 1	4.0	2.0
Rater 2	3.0	3.0
Rater 3	5.0	3.0
Rater 4	3.0	3.0
Rater 5	-	2.0
Rater 6	-	4.0

Sentence 2

Sentence 1	R1	R2	R3	R4	R5	R6
R1	↑	↑	↑	↑	↑	=
R2	↑	=	=	=	↑	↓
R3	↑	↑	↑	↑	↑	↑
R4	↑	=	=	=	↑	↓

QA-based labels for sentence-level annotations

- The goal is to define trends in the evaluations
- In the above example, there are 15 preferences for sentence one, 2 preferences for sentence two, and 7 draws

S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in Interspeech 2018, Hyderabad, India, September 2018, pp. 252–256.

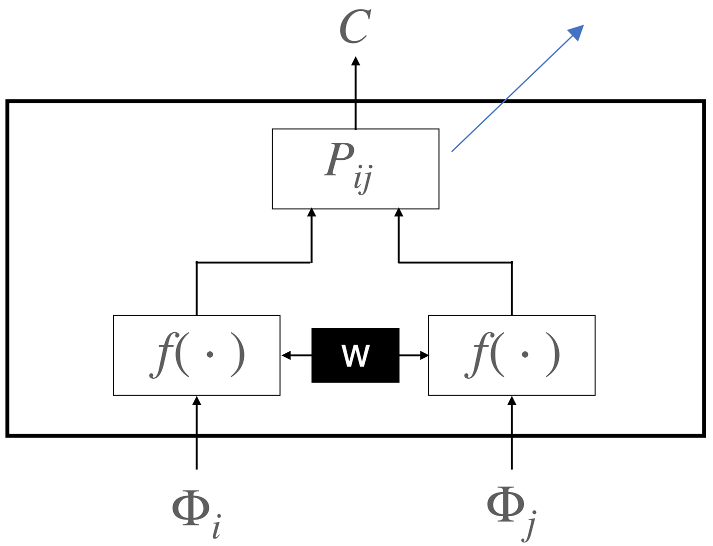
RankNet (Previous works)

Ideal probabilities $\overline{P}_{ij}$ is set according to the preference in pairs of samples.

- $\overline{P}_{ij} = 0$ if $j \gg i$
- $\overline{P}_{ij} = 1$ if $i \gg j$

$$\boxed{C = -\bar{P}_{ij} log P_{ij} - (1 - \bar{P}_{ij}) log(1 - P_{ij})}$$

$$P_{ij} \equiv \frac{1}{1+e^{-\sigma(f(\Phi_i)-f(\Phi_j))}}$$

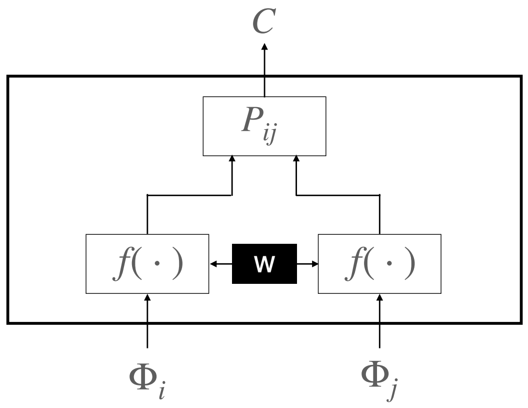


Sample sentences *i*,*j*, with features $\boldsymbol{\Phi}_i$, $\boldsymbol{\Phi}_j$

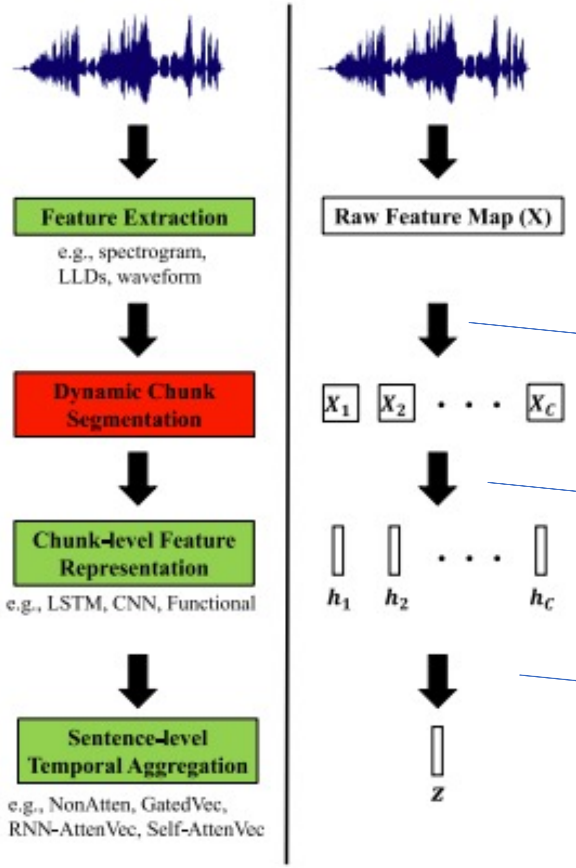
C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, “Learning to rank using gradient descent,” in International conference on Machine learning (ICML 2005), Bonn, Germany, August 2005, pp. 89–96.

Speech Emotion Recognition Formulation

- **We can implement function *f()* with arbitrary architectures**
- **Focus on generalization**
  - Train on one domain and test on another
- **We consider two alternative and complementary domain adaptation schemes**
  - Ladder networks (feature reconstruction)
  - Adversarial domain adaptation (feature representation)



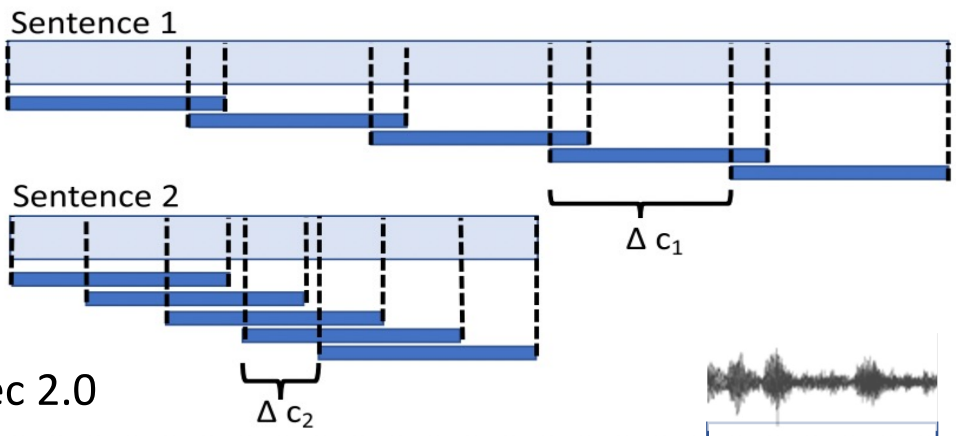
Chunk based segmentation

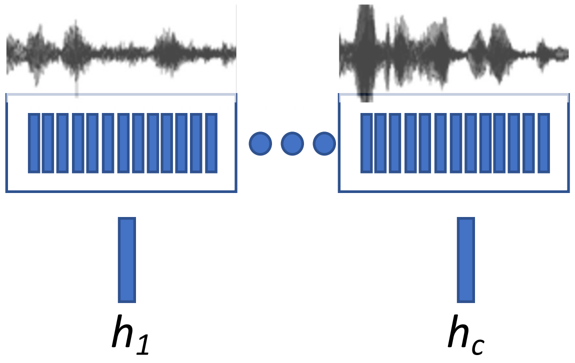


Wav2Vec 2.0

2-LSTM layers

RNN-AttenVec

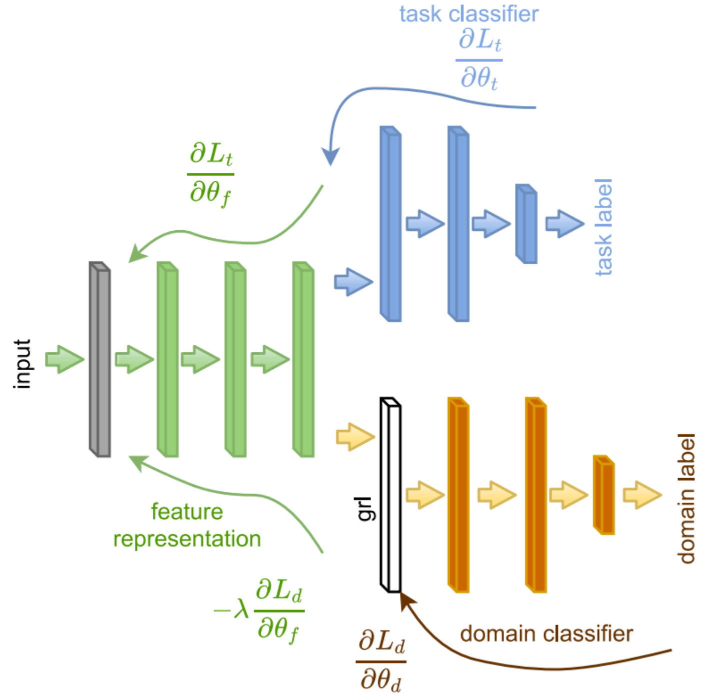


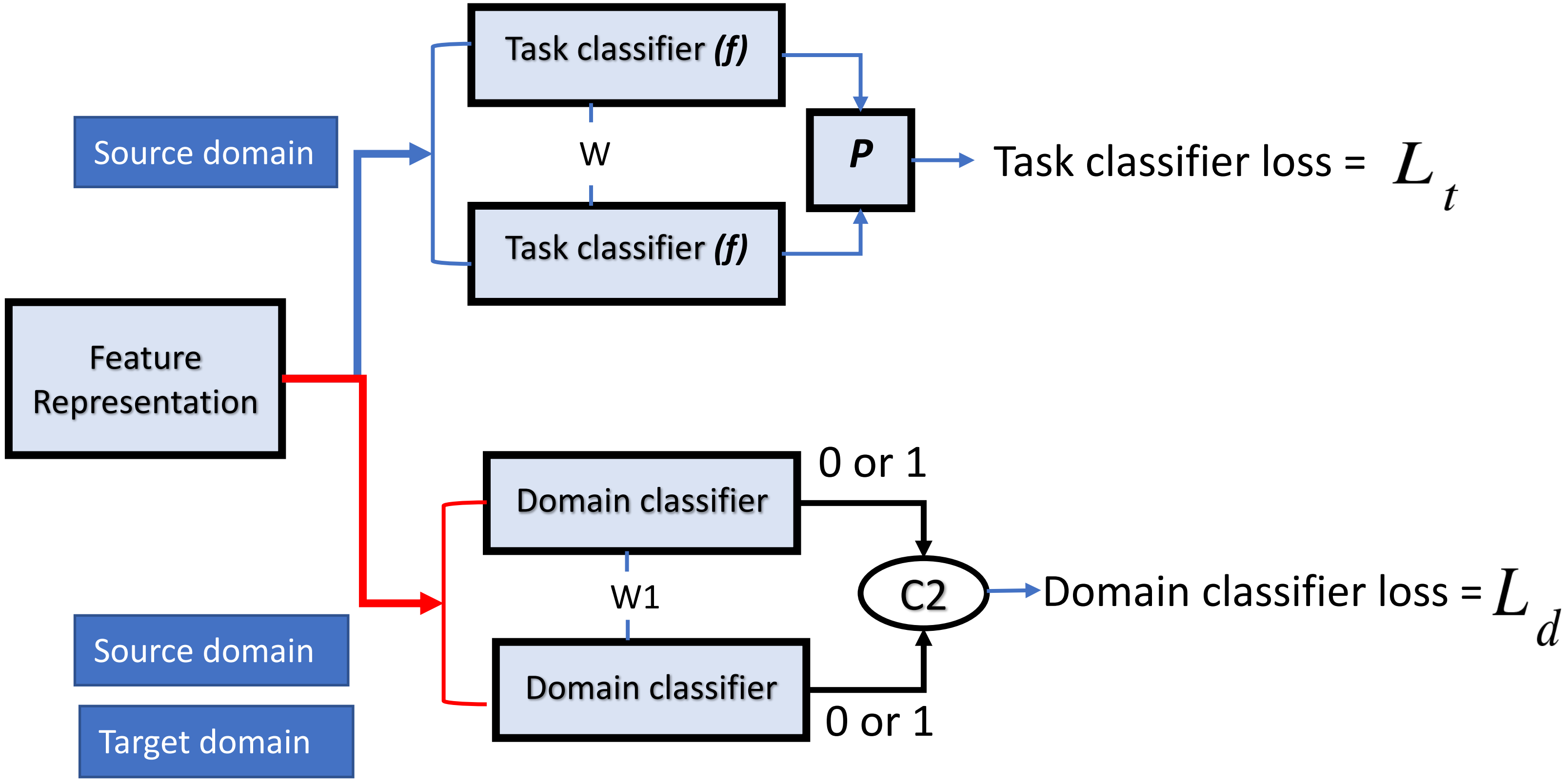


$$v = \sum_{t=1}^{C} \alpha_t \bar{h}_t$$

$$z = tanh(W[v; \bar{h}_C]).$$

Adversarial Domain Adaptation Model

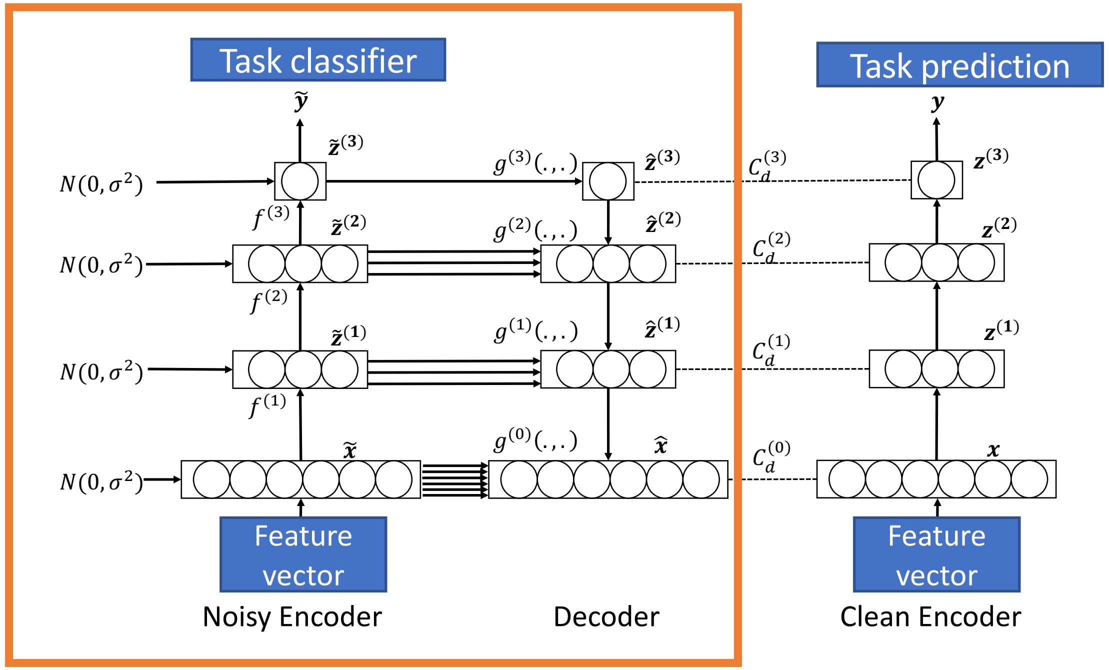




M.Abdelwahaband, C.Busso,"Domain adversarial for acoustic emotion recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 12, pp. 2423–2435, December 2018

Ladder network for domain adaptation (LN)

- The goal is to recover a clean version of the encoder while obtaining task-specific encoded features
- Source domain: Both task classifier loss along with reconstruction loss
- Target domain: Only reconstruction loss.

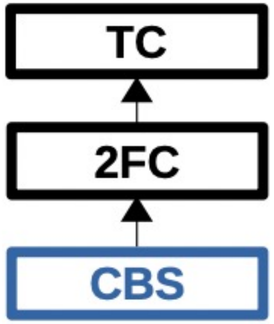


Total cost =

$$C = C_c + \lambda_l \sum_l C_d^{(l)}$$

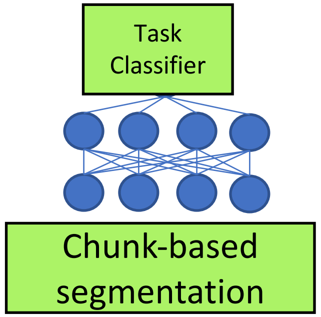
S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2697–2709, September 2020.

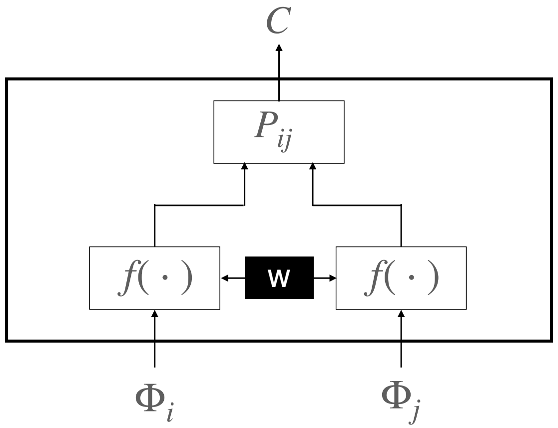
Proposed Architectures



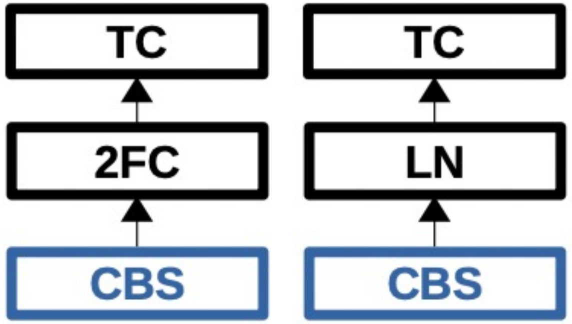
CBS + 2FC

CBS: Chunk-based segmentation, TC: Task classifier, FC: Fully connected layers, LN: Ladder Network, DA: Adversarial Domain Adaptation



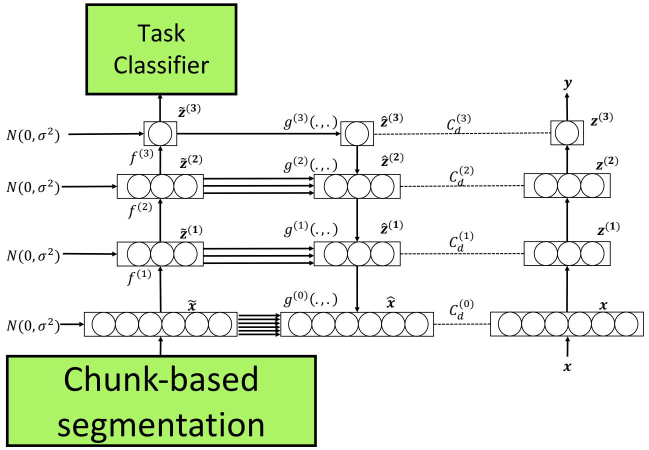


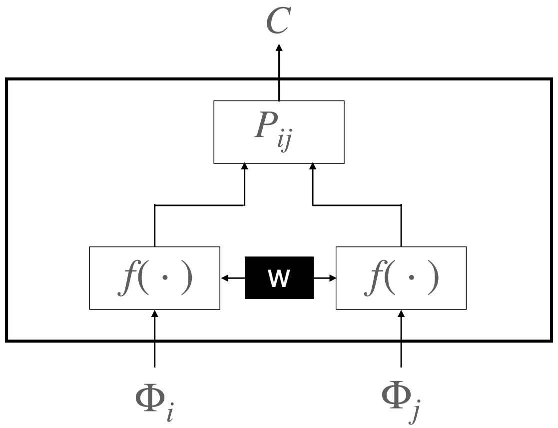
Proposed Architectures



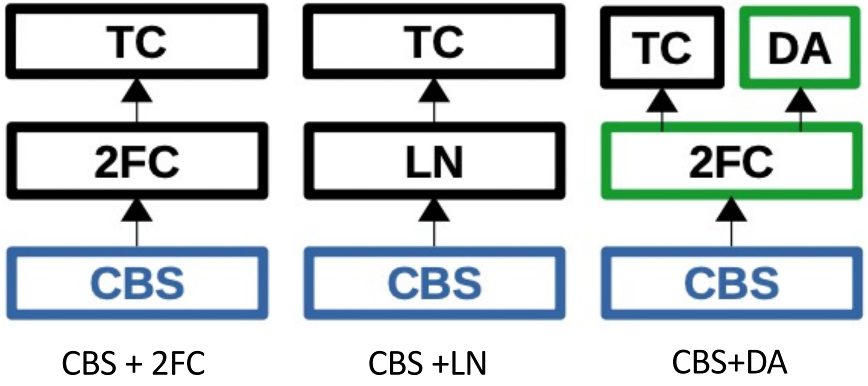
CBS + 2FC   CBS +LN

CBS: Chunk-based segmentation, TC: Task classifier, FC: Fully connected layers, LN: Ladder Network, DA: Adversarial Domain Adaptation

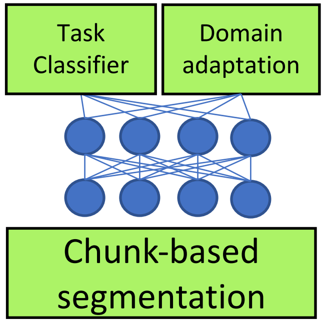


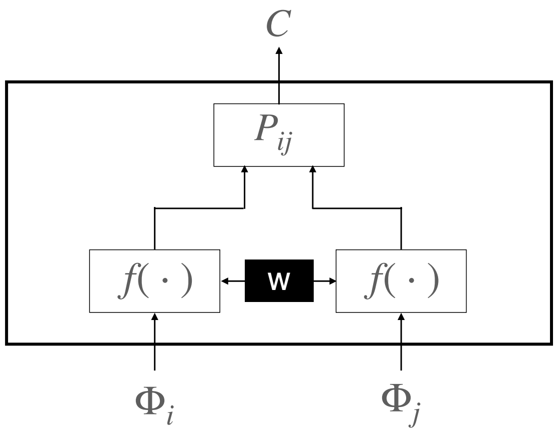


Proposed Architectures

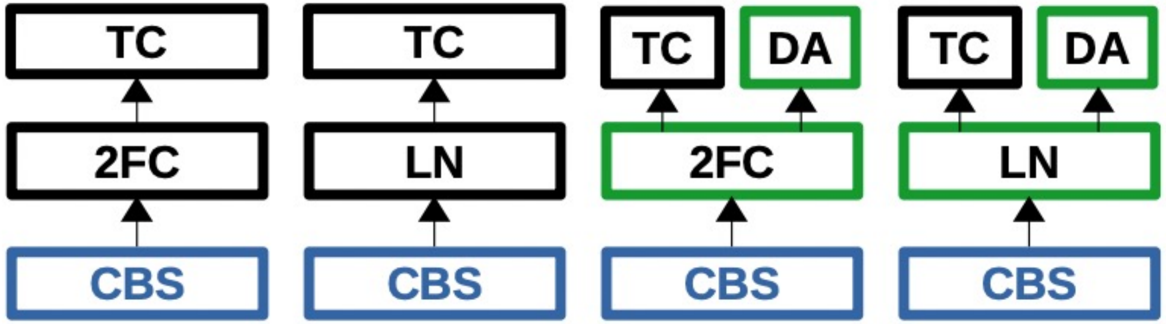


CBS: Chunk-based segmentation, TC: Task classifier, FC: Fully connected layers, LN: Ladder Network, DA: Adversarial Domain Adaptation



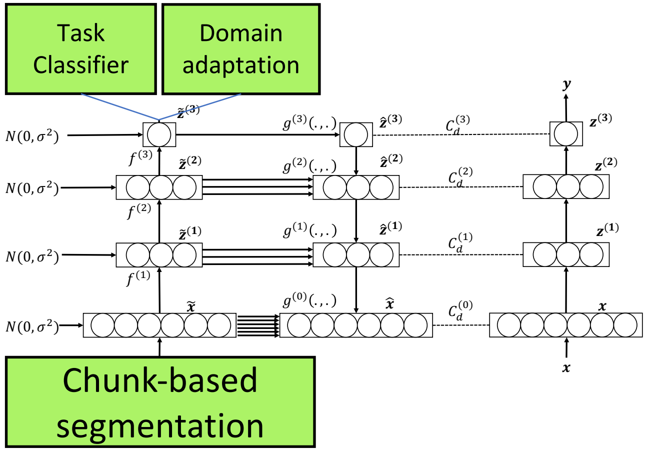


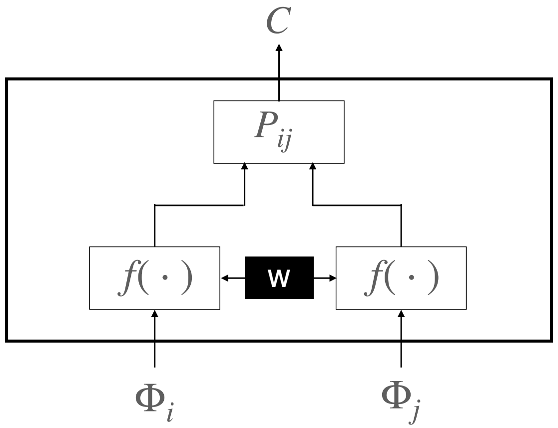
Proposed Architectures



CBS + 2FC | CBS +LN | CBS+DA | CBS+LN+DA

CBS: Chunk-based segmentation, TC: Task classifier, FC: Fully connected layers, LN: Ladder Network, DA: Adversarial Domain Adaptation





Experiment setting

- **Data preparation**
  - MSP-Podcast v1.10 (Source domain)
    - Recordings are annotated for emotional attribute labels (arousal, valence, dominance)
    - We have used ~420k pairs of samples from MSP-Podcast training set.
  - MSP-IMPROV (Target domain)
    - Recordings are annotated for emotional attribute labels (arousal, valence, dominance) like MSP-Podcast.
    - Data from the first three sessions are used as the test set, remaining sessions are reserved for adaptation.
    - We samples equal number of pairs (~420k) for adaptation.

Experiment setting

- **Feature extraction**
  - Wav2vec2-large-robust[1]
    - wav2vec2-large feature representation (1024) using pre-trained Wav2vec2.0 large model from the HuggingFace library.
    - Then, we prune the top 12 transformer blocks, and fine-tuned the model using the MSP-Podcast corpus.
  - extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)[2]
    - We also present a baseline using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which includes 88 acoustic features.

1. 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), Virtual, December 2020, vol. 33, pp. 12449–12460.
2. F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, April-June 2016.

Kendall's Tau correlation coefficient (KT)

- If (x_1,y_1) …. (x_n,y_n) be a set of observations
- (x_i,x_j), (y_i,y_j) are said to be concordant if the sort order agrees.

$$KT = \frac{(\text{Number of concordant pairs}) - (\text{Number of discordant pairs})}{\binom{n}{2}}$$

- For testing: 200 utterance are sampled randomly. This process is repeated 20 times, then mean and SD of the result are reported.

Results

KT: Kendall's Tau correlation coefficient

Accuracy: Overall test accuracy

Method	Labeled	Un-labeled	Arousal		Valence		Dominance	
			KT	Accuracy	KT	Accuracy	KT	Accuracy
eGeMAPS+2FC	MSP-PODCAST	-	0.372	64.3	0.208	58.7	0.316	62.7
CBS+2FC	MSP-PODCAST	-	0.417	71.9	0.258	61.8	0.392	69.2
CBS+LN	MSP-PODCAST	MSP-IMPROV	0.484	79.2	0.313	63.9	0.447	72.3
CBS+DA	MSP-PODCAST	MSP-IMPROV	0.462	78.5	0.301	63.7	0.442	71.5
CBS+LN+DA	MSP-PODCAST	MSP-IMPROV	**0.506**	**80.6**	**0.312**	**64.2**	**0.461**	**74.7**

- **Observations:**
  - Lower performance without domain adaptation

Results

KT: Kendall's Tau correlation coefficient

Accuracy: Overall test accuracy

Method	Labeled	Un-labeled	Arousal		Valence		Dominance	
			KT	Accuracy	KT	Accuracy	KT	Accuracy
eGeMAPS+2FC	MSP-PODCAST	-	0.372	64.3	0.208	58.7	0.316	62.7
CBS+2FC	MSP-PODCAST	-	0.417	71.9	0.258	61.8	0.392	69.2
CBS+LN	MSP-PODCAST	MSP-IMPROV	0.484	79.2	0.313	63.9	0.447	72.3
CBS+DA	MSP-PODCAST	MSP-IMPROV	0.462	78.5	0.301	63.7	0.442	71.5
CBS+LN+DA	MSP-PODCAST	MSP-IMPROV	**0.506**	**80.6**	**0.312**	**64.2**	**0.461**	**74.7**

- **Observations:**
  - Both adaptation methods lead to improvements
  - Best performance achieved by combining ladder network and domain adaptation

Results

KT: Kendall's Tau correlation coefficient

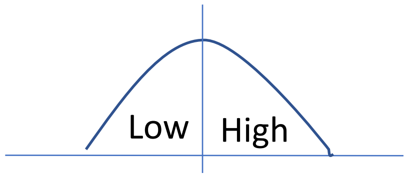
Accuracy: Overall test accuracy

Method	Labeled	Un-labeled	Arousal		Valence		Dominance	
			KT	Accuracy	KT	Accuracy	KT	Accuracy
eGeMAPS+2FC	MSP-PODCAST	-	0.372	64.3	0.208	58.7	0.316	62.7
CBS+2FC	MSP-PODCAST	-	0.417	71.9	0.258	61.8	0.392	69.2
CBS+LN	MSP-PODCAST	MSP-IMPROV	0.484	79.2	0.313	63.9	0.447	72.3
CBS+DA	MSP-PODCAST	MSP-IMPROV	0.462	78.5	0.301	63.7	0.442	71.5
CBS+LN+DA	MSP-PODCAST	MSP-IMPROV	**0.506**	**80.6**	**0.312**	**64.2**	**0.461**	**74.7**
CBS+LN*	MSP-IMPROV	MSP-IMPROV	0.617	84.7	0.375	66.6	0.558	81.2

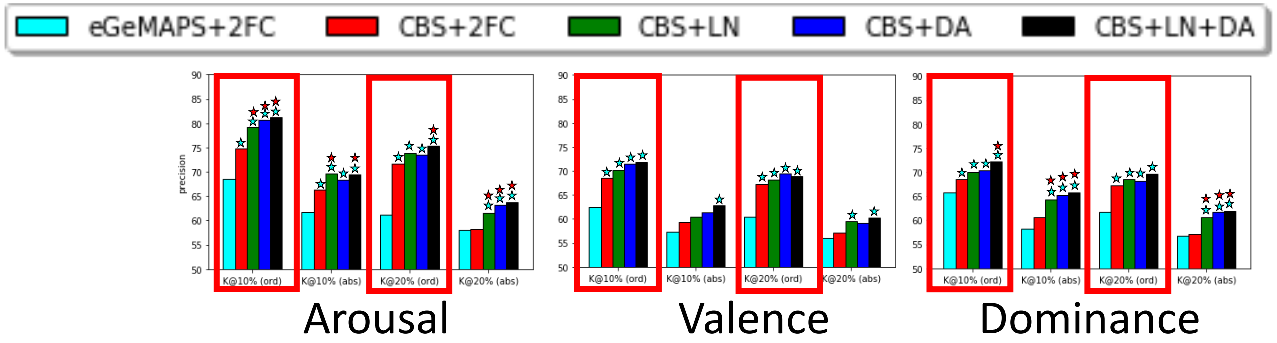
- **Observations:**
  - Both adaptation methods lead to improvements
  - Best performance achieved by combining ladder network and domain adaptation

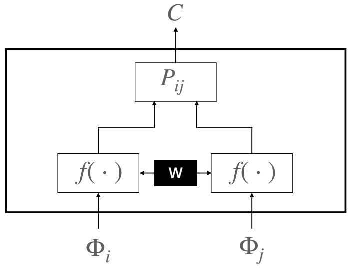
Precision at K

- **Precision as the number of retrieved samples increases**
  - We evaluate 10% and 20% of the data
  - We retrieve samples with low and high values of an attribute
    - Arousal, valence, and dominance
  - Success: retrieved samples belong to the correct class created with a median split



  - Baselines: models using *f()*, trained to predict absolute scores





Summary

- In this work we explored different preference learning based architectures for SER.
- We observed ladder network and Adversarial Domain Adaptation are complementary while adapting SER model to new domain.

This study is supported by

LAS **Laboratory for Analytic Sciences**

Thank You