Contrastive Unsupervised Learning for Speech Emotion Recognition

Mao Li[2], Bo Yang[1], Joshua Levy[1], Andreas Stolcke[1], Viktor Rozgic[1], Spyros Matsoukas[1], Constantinos Papayiannis[1], Daniel Bone[1], Chao Wang[1]

Amazon Alexa[1] and University of Illinois at Chicago[2]

Motivation

- Modern speech systems are mainly designed for speech content understanding, while speech emotion recognition (SER) becomes a key technology to enable natural human-machine communication.

- Application scenarios of SER system:
  - voice assistant
  - human health assistant
  - chat-bots & social robot

[Image credit: Patrick J. Kiger's blog]

amazon alexa

Motivation

- Modern speech systems are mainly designed for speech content understanding, while speech emotion recognition (SER) becomes a key technology to enable natural human-machine communication.

- The **deficiency of emotion annotated data** is the bottleneck for development of SER system.

    - labeling is expensive

    - labeling emotion data is challenging due to annotator disagreements

amazon alexa

Motivation

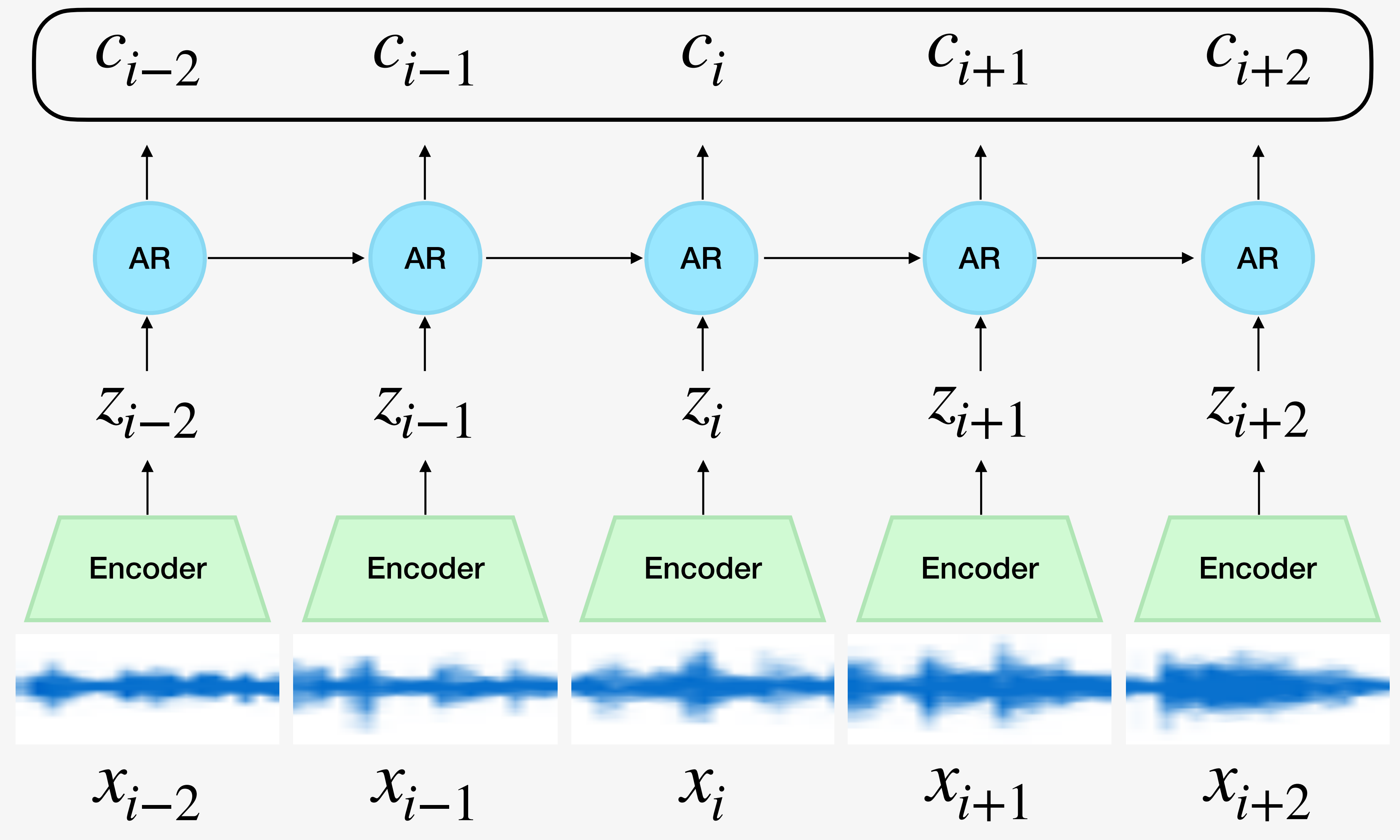
- Modern speech systems are mainly designed for speech content understanding, while speech emotion recognition (SER) becomes a key technology to enable natural human-machine communication.
- The **deficiency of emotion annotated data** is the bottleneck for development of SER system.
- Two solutions:
  - Transfer learning from a related speech task
  - Unsupervised representation learning

amazon alexa

Proposed Method

Contrastive Predictive Coding (CPC)

- sequence of audio frames: $\{x_1, x_2, \ldots, x_n\}$
- non-linear encoder f
- frame-level representation: $z_i = f(x_i)$
- autoregressive model g
- contextual representation: $c_i = g(z_{\leq i})$

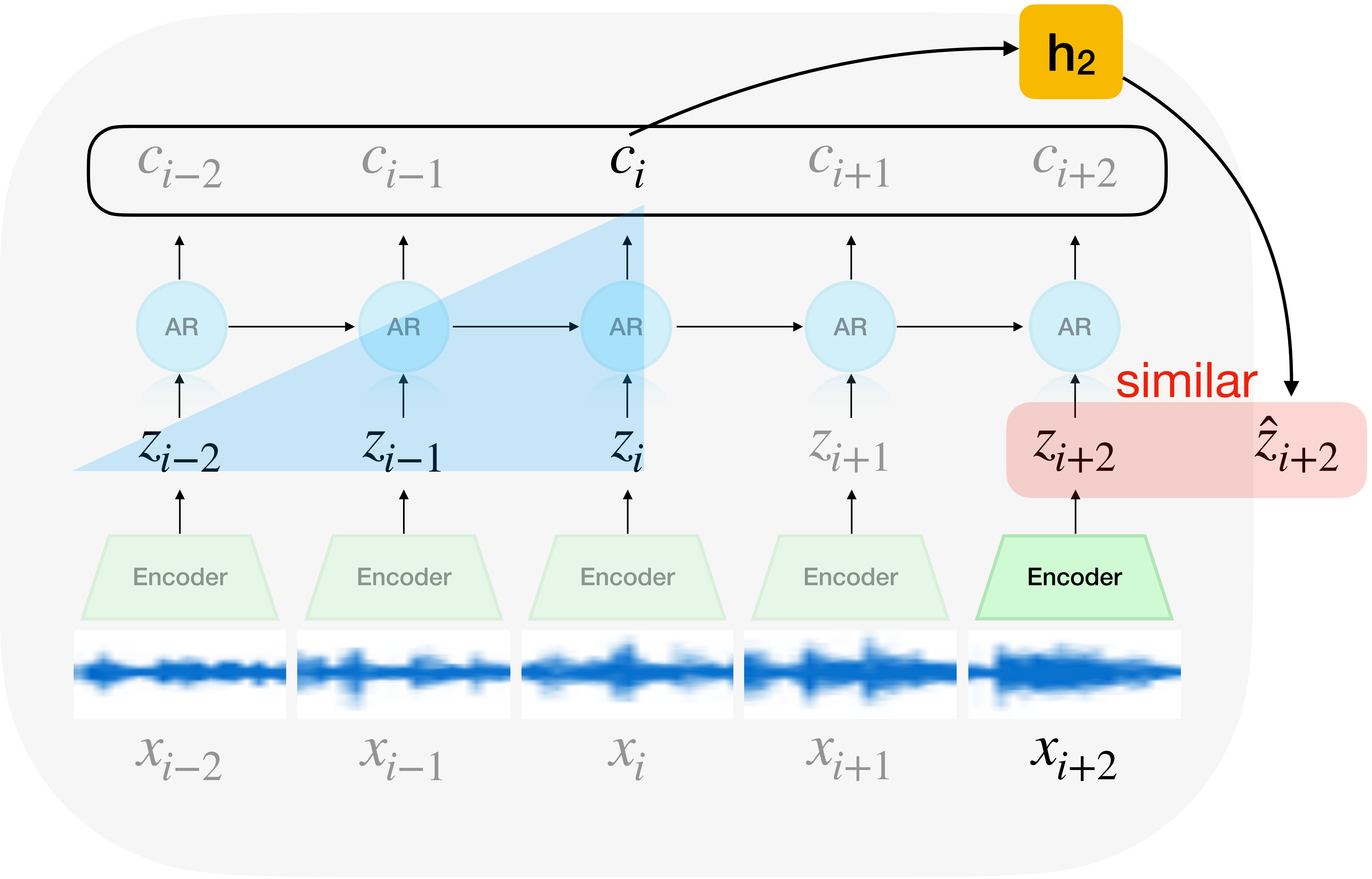


Proposed Method

Contrastive Predictive Coding (CPC)

- frame-level representation: $z_i = f(x_i)$
- contextual representation: $c_i = g(z_{\leq i})$
- prediction function for a specific k: h_k
- predict future : $\hat{z}_{i+k} = h_k(c_i) = h_k(g(z_{\leq i}))$
- InfoNCE Loss:

$$\mathscr{L} = -\sum_{m=1}^{k}\left[\log \frac{\exp\left(\hat{z}_{i+m}^{\top} z_{i+m}\right)}{\exp\left(\hat{z}_{i+m}^{\top} z_{i+m}\right) + \sum_{i=1}^{N-1} \exp\left(\hat{z}_{i+m}^{\top} z_i\right)}\right]$$



amazon alexa

Proposed Method

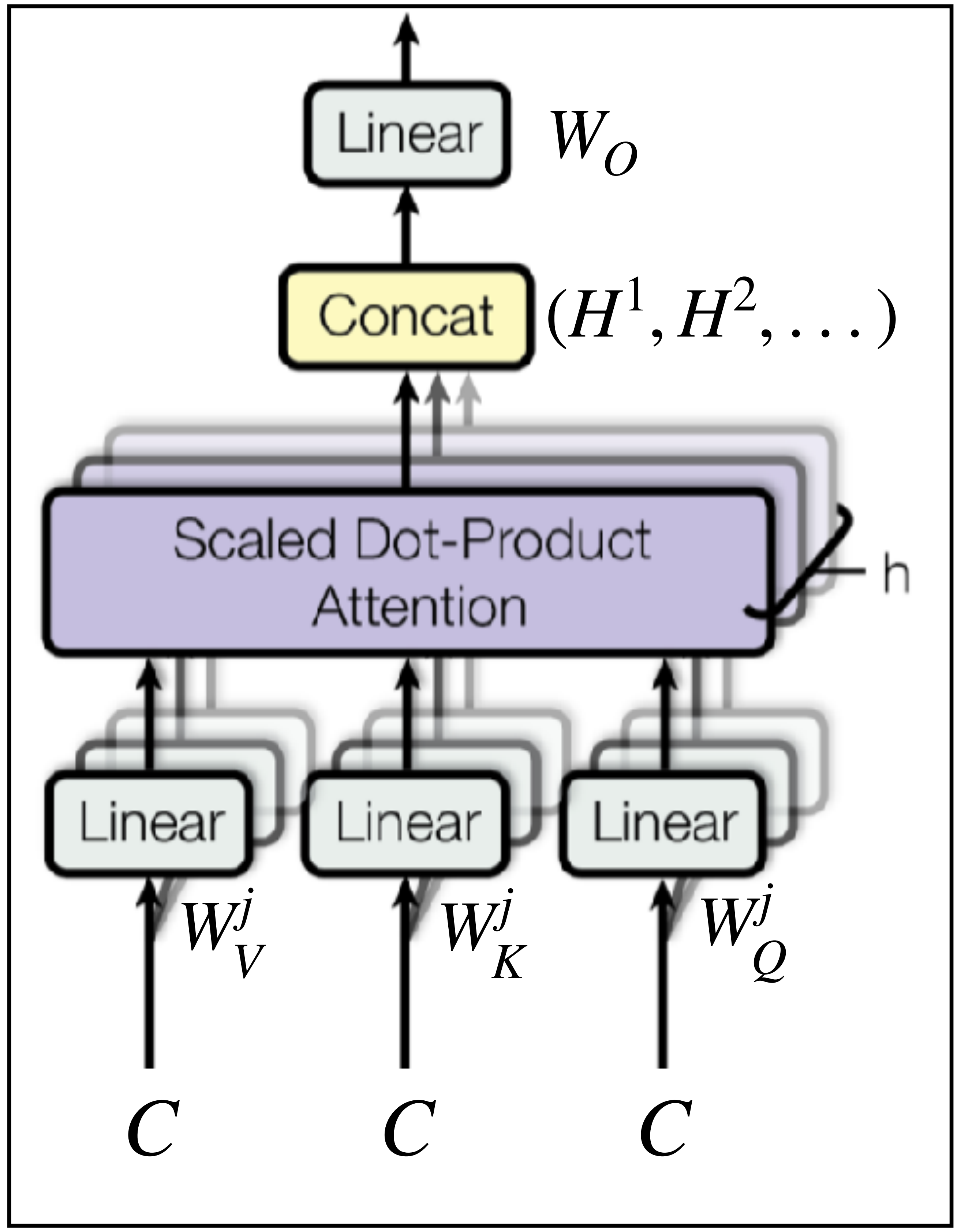
Attention-based Emotion Recognizer

- Multi-head attention
  - C is the output of CPC
  - W_*^j, W_O are trainable weights
  - H^j is attention score of a single head
  - h is number of heads, d_K is the dimension

$$H^j = \mathrm{softmax}\left(W_Q^j C \left(W_K^j C\right)^\top / \sqrt{d_K}\right) W_V^j C$$

$$U = \mathrm{Concat}\left(H^1, H^2, \ldots, H^n\right) W_O$$

Multi-head Attention



amazon alexa

Proposed Method

Attention-based Emotion Recognizer

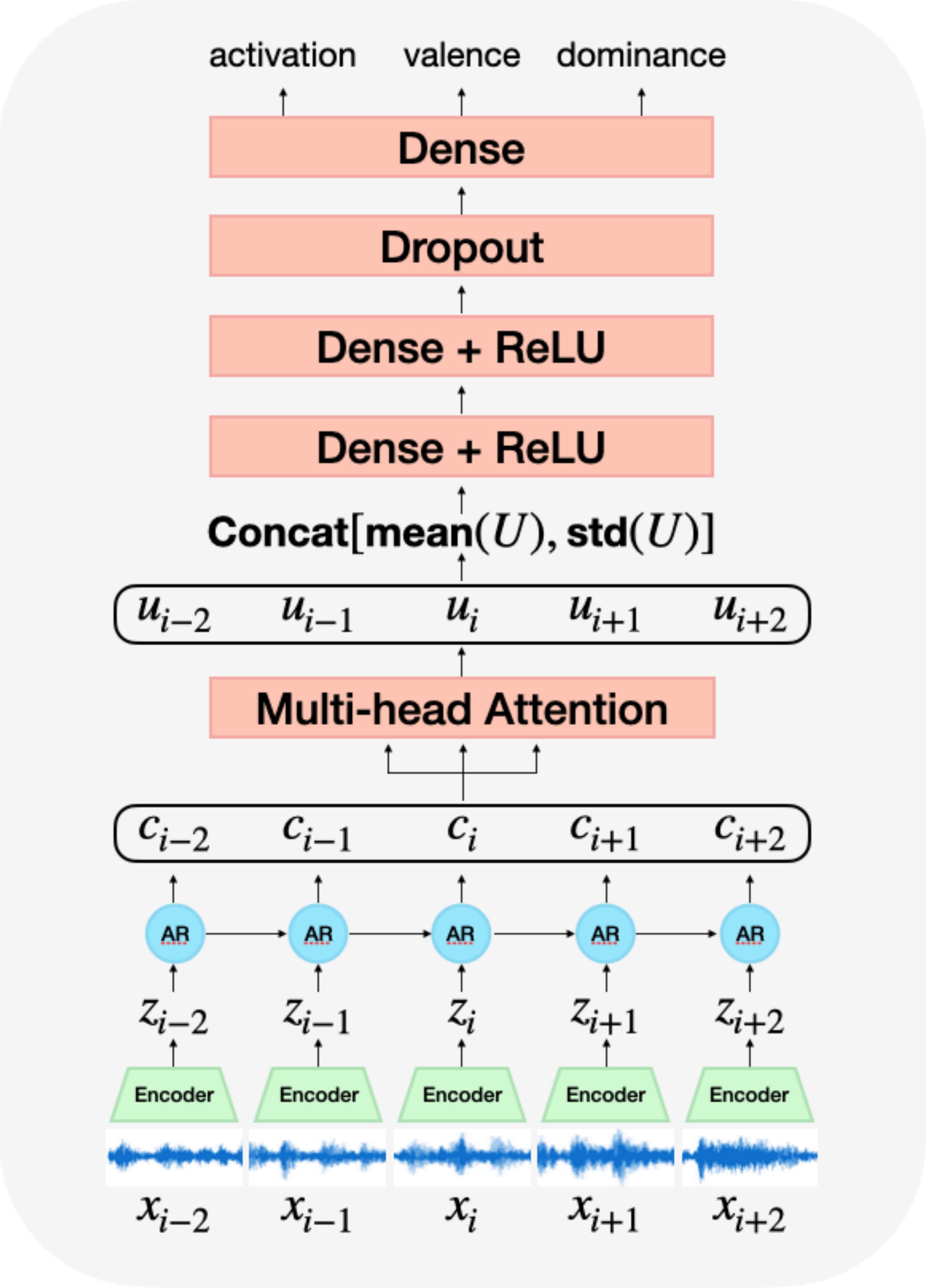
- Utterance embedding $u = [mean(U); std(U)]$
- Concordance Correlation Coefficient (CCC) measures alignment of two random variables:

$$\mathrm{CCC}(X, Y) = \rho \frac{2\sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2 + \left(\mu_X - \mu_Y\right)^2}, \rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

  - X： ground truth score; Y: predicted score

- Loss function:

$$\mathscr{L} = 1 - \alpha \mathbf{CCC}_{act} - \beta \mathbf{CCC}_{val} - (1 - \alpha - \beta)\mathbf{CCC}_{dom}$$



amazon alexa

Experiments

Datasets

Dataset	Label	Duration	Usage
LibriSpeech	N/A	100 hours of audio	Unsupervised pre-training
IEMOCAP	Primitive labels; Categorical labels	12 hours of audio	Model evaluation; Visualization
MSP-Podcast	Primitive labels; Categorical labels	84 hours of audio	Model evaluation

Experiments

Experimental Setups

- preCPC: pre-trained CPC (LibriSpeech) + supervised (IEMOCAP/MSP-Podcast)

- Sup: supervised only (IEMOCAP/MSP-Podcast)

handcrafted V.S learned feature

Experiments

Experimental Setups

- preCPC: pre-trained CPC + supervised
- Sup: supervised only

Table 1: CCC scores (mean/std) on the IEMOCAP dataset

Methods	CCC avg	CCC act	CCC val	CCC dom
Sup	.664 ± .007	.638 ± .017	.718 ± .004	.635 ± .009
jointCPC	.562 ± .012	.549 ± .032	.642 ± .013	.491 ± .016
miniCPC	.660 ± .005	.673 ± .028	.702 ± .009	.606 ± .019
preCPC	.731 ± .003	.752 ± .014	.752 ± .009	.691 ± .009

Table 2: CCC scores (mean/std) on the MSP-Podcast dataset

Methods	CCC avg	CCC act	CCC val	CCC dom
Sup	.458 ± .005	.596 ± .007	.266 ± .004	.501 ± .013
jointCPC	.491 ± .008	.628 ± .006	.280 ± .006	.568 ± .007
miniCPC	.549 ± .006	.688 ± .009	.345 ± .005	.615 ± .011
preCPC	.571 ± .004	.706 ± .006	.377 ± .008	.639 ± .012

Hypothesis:

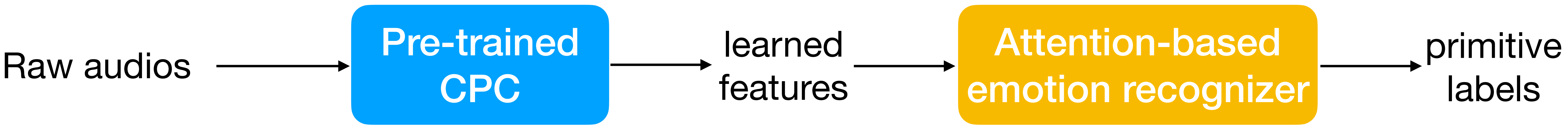
- Representations learned by CPC are superior to handcrafted features for speech emotion recognition task

amazon alexa

Experiments

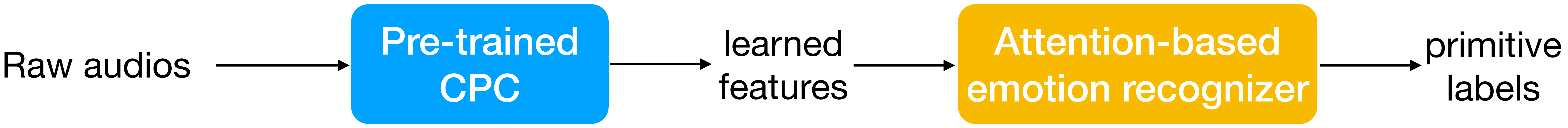
Experimental Setups

- preCPC: pre-trained CPC (LibriSpeech) + supervised (IEMOCAP/MSP-Podcast)



LibriSpeech: large-scale, independent with downstream task

- Sup: supervised only (IEMOCAP/MSP-Podcast)
- miniCPC: pre-trained CPC (IEMOCAP/MSP-Podcast)+ supervised (IEMOCAP/MSP-Podcast)



IEMOCAP/MSP-Podcast: small size, dataset for downstream task

amazon alexa

Experiments

Experimental Setups

- preCPC: pre-trained CPC + supervised
- Sup: supervised only
- miniCPC: pre-trained CPC + supervised

Table 1: CCC scores (mean/std) on the IEMOCAP dataset

Methods	CCC_{avg}	CCC_{act}	CCC_{val}	CCC_{dom}
Sup	$.664 \pm .007$	$.638 \pm .017$	$.718 \pm .004$	$.635 \pm .009$
jointCPC	$.562 \pm .012$	$.549 \pm .032$	$.642 \pm .013$	$.491 \pm .016$
miniCPC	$.660 \pm .005$	$.673 \pm .028$	$.702 \pm .009$	$.606 \pm .019$
preCPC	$.731 \pm .003$	$.752 \pm .014$	$.752 \pm .009$	$.691 \pm .009$

Table 2: CCC scores (mean/std) on the MSP-Podcast dataset

Methods	CCC_{avg}	CCC_{act}	CCC_{val}	CCC_{dom}
Sup	$.458 \pm .005$	$.596 \pm .007$	$.266 \pm .004$	$.501 \pm .013$
jointCPC	$.491 \pm .008$	$.628 \pm .006$	$.280 \pm .006$	$.568 \pm .007$
miniCPC	$.549 \pm .006$	$.688 \pm .009$	$.345 \pm .005$	$.615 \pm .011$
preCPC	$.571 \pm .004$	$.706 \pm .006$	$.377 \pm .008$	$.639 \pm .012$

Hypothesis:

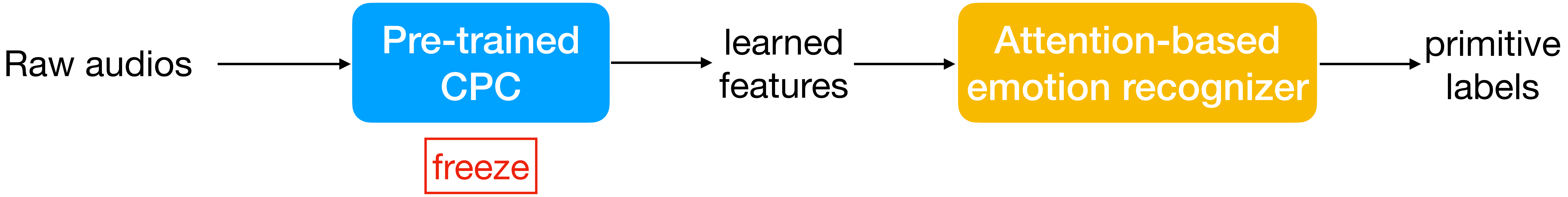
- Exposing the model to more diverse acoustic conditions and speaker variations (LibriSpeech) is beneficial for learning robust features.

amazon alexa

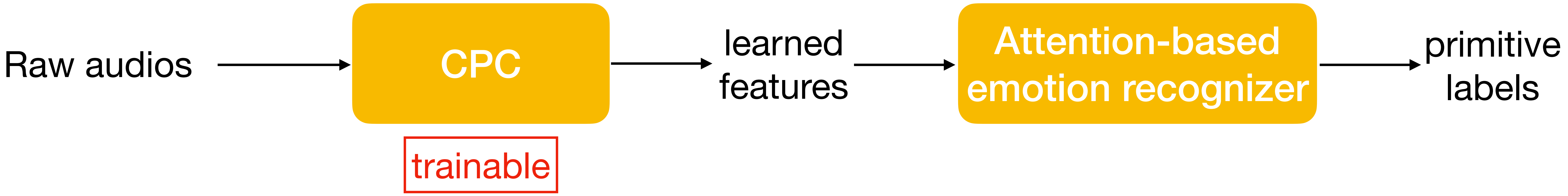
Experiments

Experimental Setups

- preCPC: pre-trained CPC (LibriSpeech) + supervised (IEMOCAP/MSP-Podcast)
- Sup: supervised only (IEMOCAP/MSP-Podcast)
- miniCPC: pre-trained CPC (IEMOCAP/MSP-Podcast)+ supervised (IEMOCAP/MSP-Podcast)



- jointCPC: pre-trained CPC (IEMOCAP/MSP-Podcast)+ supervised (IEMOCAP/MSP-Podcast)



Experiments

Experimental Setups

- preCPC: pre-trained CPC + supervised
- Sup: supervised only
- miniCPC: pre-trained CPC + supervised
- jointCPC: pre-trained CPC + supervised

Hypothesis:

- Unsupervised pre-training produces representations with better generalization which facilitate various downstream tasks

Table 1: CCC scores (mean/std) on the IEMOCAP dataset

Methods	CCC $_{avg}$	CCC $_{act}$	CCC $_{val}$	CCC $_{dom}$
Sup	.664 $\pm$ .007	.638 $\pm$ .017	.718 $\pm$ .004	.635 $\pm$ .009
jointCPC	.562 $\pm$ .012	.549 $\pm$ .032	.642 $\pm$ .013	.491 $\pm$ .016
miniCPC	.660 $\pm$ .005	.673 $\pm$ .028	.702 $\pm$ .009	.606 $\pm$ .019
preCPC	.731 $\pm$ .003	.752 $\pm$ .014	.752 $\pm$ .009	.691 $\pm$ .009

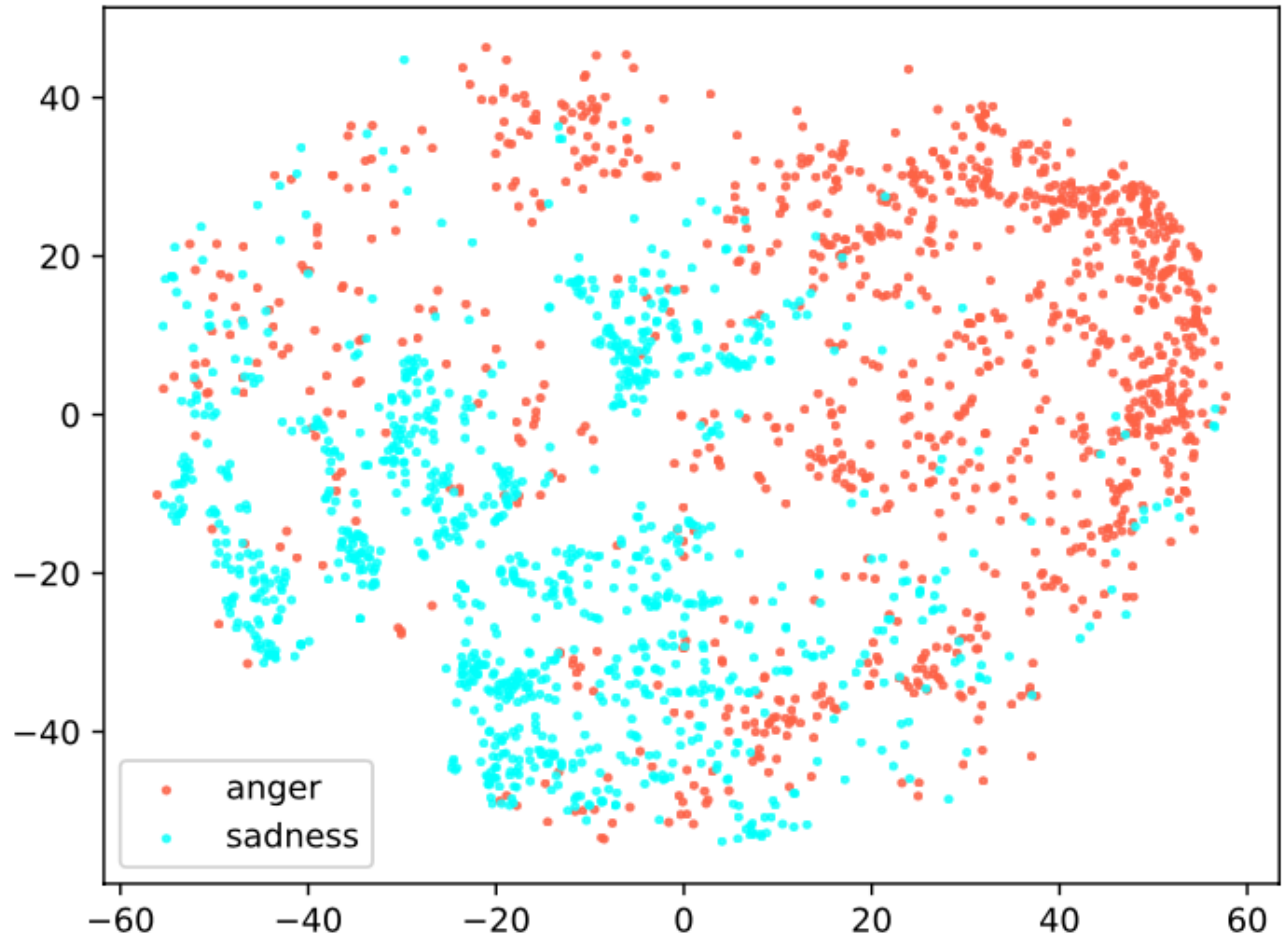
Table 2: CCC scores (mean/std) on the MSP-Podcast dataset

Methods	CCC $_{avg}$	CCC $_{act}$	CCC $_{val}$	CCC $_{dom}$
Sup	.458 $\pm$ .005	.596 $\pm$ .007	.266 $\pm$ .004	.501 $\pm$ .013
jointCPC	.491 $\pm$ .008	.628 $\pm$ .006	.280 $\pm$ .006	.568 $\pm$ .007
miniCPC	.549 $\pm$ .006	.688 $\pm$ .009	.345 $\pm$ .005	.615 $\pm$ .011
preCPC	.571 $\pm$ .004	.706 $\pm$ .006	.377 $\pm$ .008	.639 $\pm$ .012

Experiments

Representation Visualization:

- t-SNE plot of representations that learned from CPC



Data points are well separated, even though trained without emotion labels

Conclusion & Future Work

Conclusion

- CPC can learn salient features from unlabeled speech corpora that benefits emotion recognition task

- Obtained competitive performance on public benchmarks

Future work

- Investigate the impact unsupervised representation learning data on emotion recognition performance (e.g., replace Libri speech with TED data)

- Still to do: end-to-end modeling

amazon alexa

You are very welcome to our poster session!

Speech Emotion 3: Emotion Recognition-Representations, Data Augmentation

Wednesday, 9 June, 15:30 - 16:15

Thank you!