香 港 中 文 大 學
The Chinese University of Hong Kong

Microsoft
Research Asia

A Hierarchical Regression Chain Framework for Affective Vocal Burst Recognition

Jinchao Li[1], Xixin Wu[1], Kaitao Song[2], Dongsheng Li[2], Xunying Liu[1], Helen Meng[1]
[1]The Chinese University of Hong Kong, Hong Kong SAR, China; [2]Microsoft Research Asia, Shanghai, China

ICASSP 2023

Poster #: 6131

Introduction

➢ACII AVB 2022 Challenge tasks:

- TWO (regression): predict values of arousal and valence
- TYPE (classification): classify the type of VB from 8 classes (Gasp, Laugh, Cry, Scream, Grunt, Groan, Pant, Other)
- HIGH (regression): predict the intensity of 10 emotions
- CULTURE (regression): predict the intensity of 40 emotions

➢Motivation:

- Vocal bursts (VB) play a crucial role in conveying emotion
- Emotions are complex and have various labeling methods
- Modeling inner and cross relationships among multiple emotional labels help understanding the emotion better

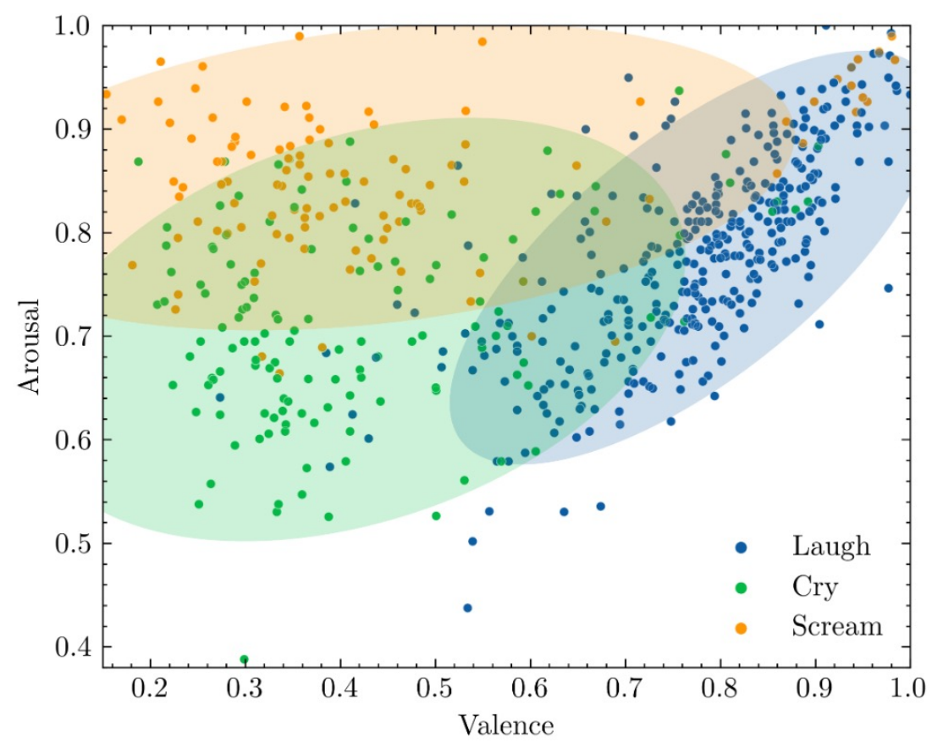


Fig. *Use labels in TWO to predict labels of TYPE.*

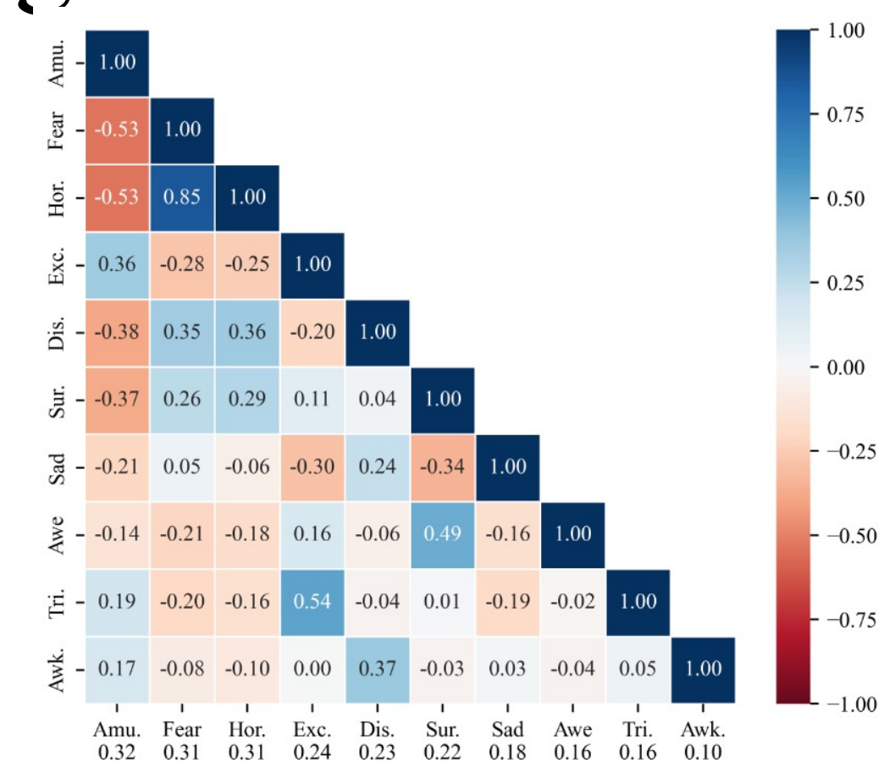


Fig. *Correlation among labels in the HIGH task.*

➢Objective & Contributions:

- Propose a hierarchical multitask model with chain regressors to explicitly learn the label dependency

Experimental Setup

➢The HUME-VB competition data:

	Train	Val.	Test	∑
HH: MM: SS	12 :19 :06	12 :05 :45	12 :22 :12	36 :47 :04
No.	19 990	19 396	19 815	59 201
Speakers	571	568	563	1 702
USA	206	206	—	—
China	79	76	—	—
South Africa	244	244	—	—
Venezuela	42	42	—	—

➢Loss function & training details:

- Cross-entropy loss for TYPE, COUNTRY; CCC loss $(1 - CCC(y, y'))$ for TWO, HIGH, CULTURE; AdamW optimizer

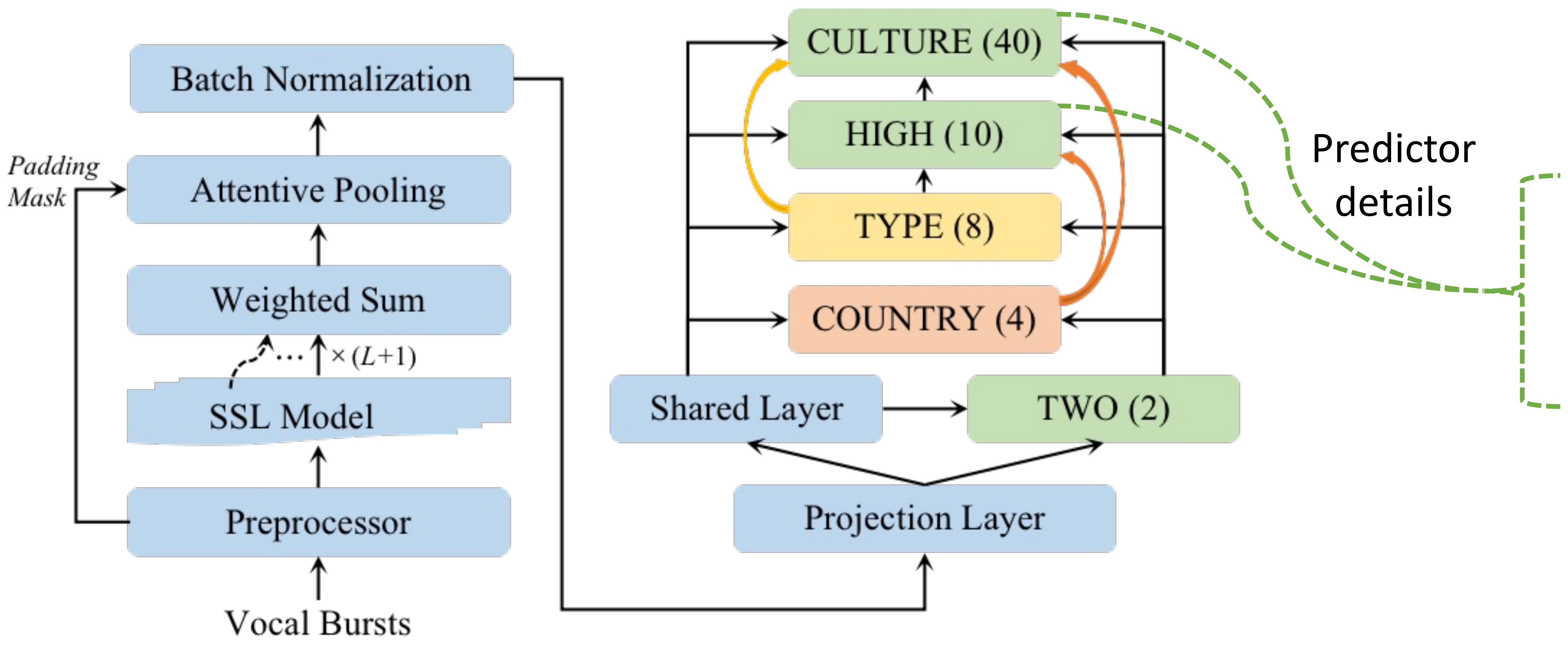
➢Evaluation Protocols:

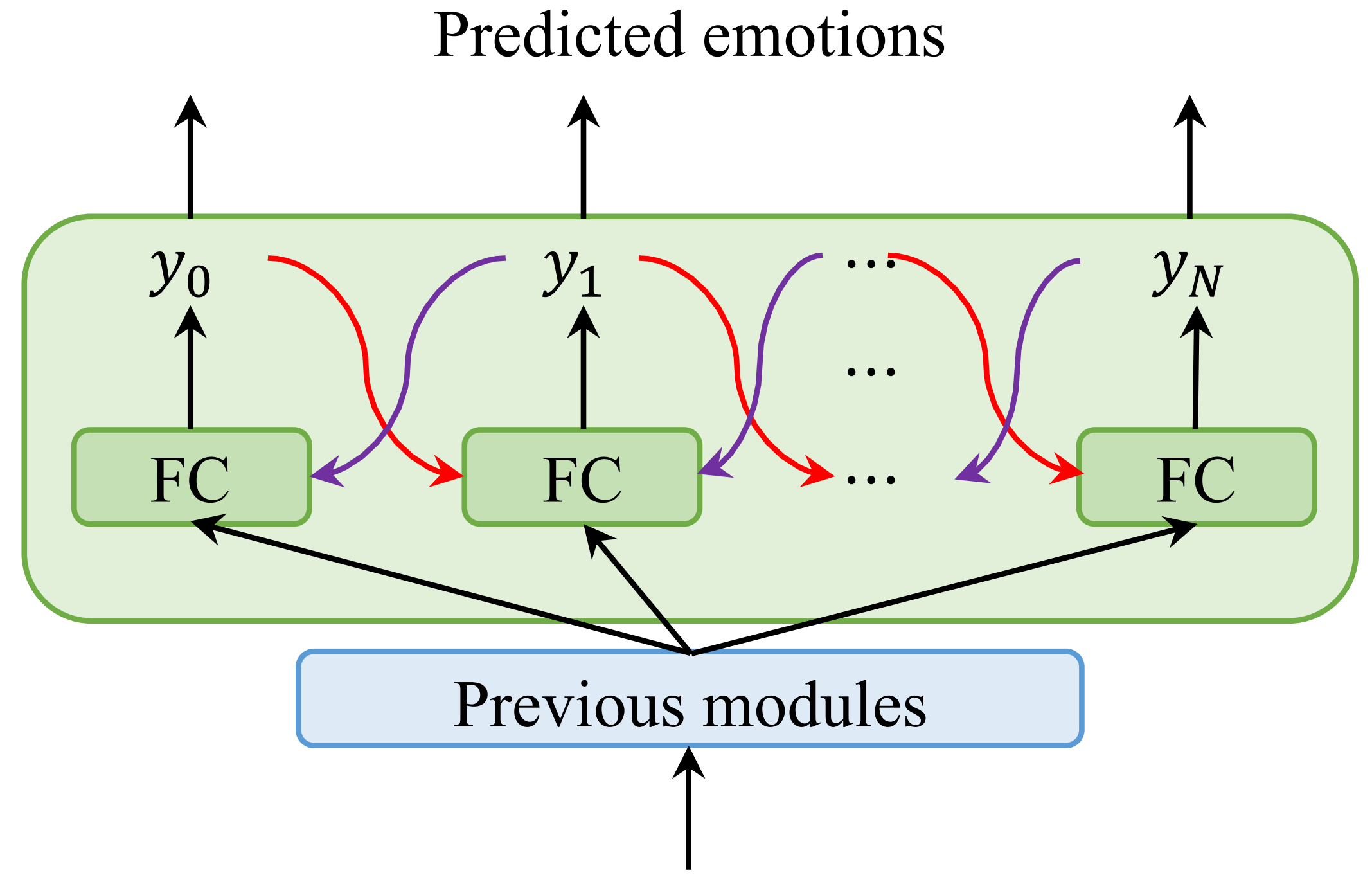
- Concordance correlation coefficient (CCC):

$$CCC(x_i, y_i) = \frac{1}{N}\sum \frac{2 * cov(x_i, y_i)}{\sigma_{x_i}^2 + \sigma_{y_i}^2 + (\mu_{x_i} + \mu_{y_i})^2}$$

Approach

➢Overview of hierarchical multitask learning framework:





The bi-directional chain regressor

➢**Multitask learn multilabel dependency:** lower-dimensional to higher-dimensional labels

➢**Bi-chain regressor learn inner-label dependency:** High to low correlation + low to high correlation

➢Model details:

- SSL Model: Wav2vec 2.0-Large XLSR
- Projector layer, shared layer: 128, 64-dimensional fully-connected layers
- Classifiers: Fully-connected layers for TWO, TYPE, COUNTRY, Bi-Chain Regressors for HIGH, CULTURE

Results

➢Results on different tasks

Approach	TWO		HIGH		CULTURE	
	Val.	Test	Val.	Test	Val.	Test
ComParE [1]	.4942	.4986	.5154	.5214	.3867	.3887
eGeMAPS [2]	.4114	.4143	.4484	.4496	.3229	.3214
END2YOU [3]	.4988	.5084	.5638	.5686	.4359	.4401
Ours	**.6966**	**.6854**	**.7351**	**.7237**	**.6464**	**.6017**

[1] B.Schuller'16, [2] F.Eyben'15, [3] P.Tzirakis'18

➢Ablation study

Approach	Averaged CCC
ComParE [1]	.5154
eGeMAPS [2]	.4484
END2YOU [3]	.5638
Ours	.7351
- Finetune	.6103
- Regression Chain	.6513
- Finetune & Regression Chain	.5540

Conclusion

➢Effectively extract features from pretrained models with finetuning and layer-wise, temporal aggregation

➢Modeling multi-label dependency by hierarchical multitask learning

➢Modeling intra-label dependency by Bi-directional chain regressor

➢**Winner** on the TWO and CULTURE tasks, Second on the HIGH task