DYNAMIC SPEECH EMOTION RECOGNITION USING A CONDITIONAL NEURAL PROCESS

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Luz Martinez-Lucas & Carlos Busso

Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas, Richardson, Texas - 75080, USA

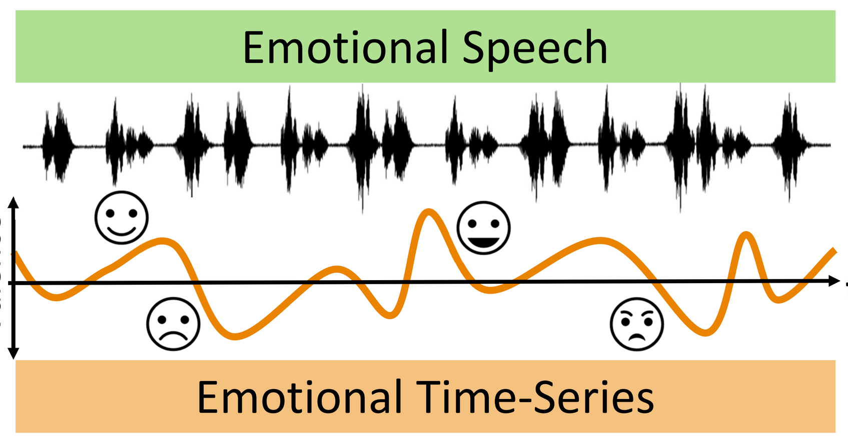
2024 IEEE International Conference on Acoustics, Speech and Signal Processing
Seoul, South Korea

ICASSP 2024 KOREA

Introduction

Background:

- *Speech emotion recognition* (SER)
  - Often predict one emotional value for a short speaking turn
  - Natural and nuanced emotions are dynamic throughout time
- ***Dynamic speech emotion recognition*** **(DSER)**
  - Treat the SER problem as a time-series problem
  - Previous work learns a single emotional distribution



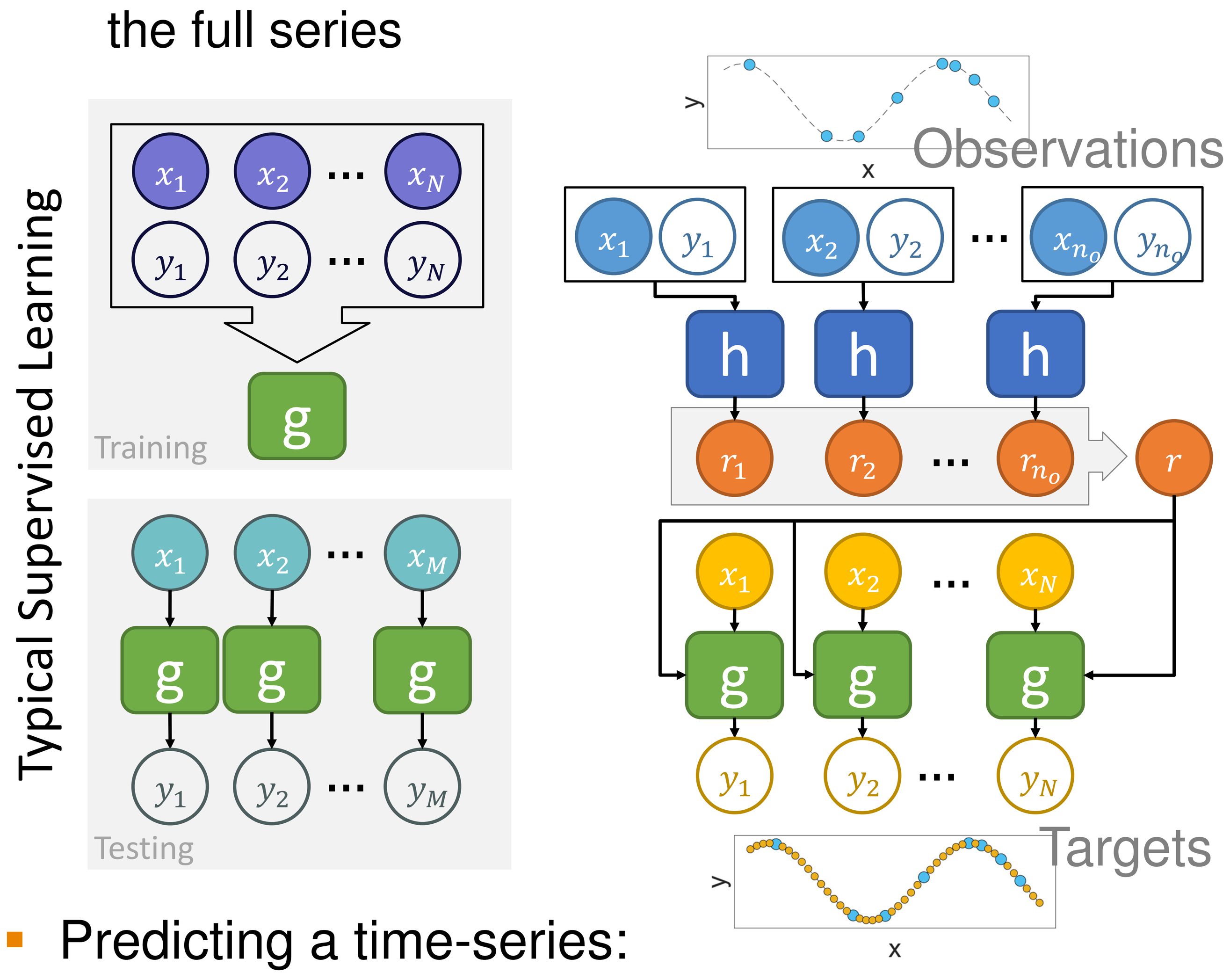
Our Work:

- Use *conditional neural process* (CNP) models for DSER
  - Allows the model to select the emotional distribution

Methodology

Conditional Neural Process:

- Conditional stochastic process that conditions predictions on observations
  - An observation is a time-step chosen to represent the full series

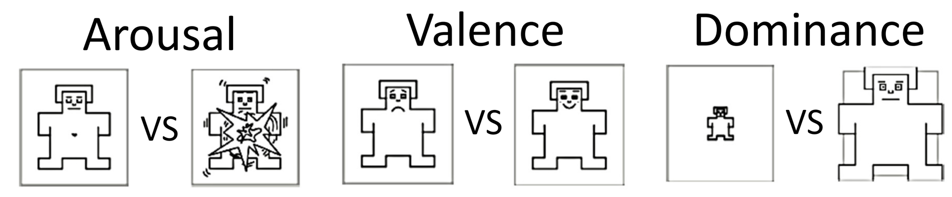


- Predicting a time-series:
  - Step 1: Embed each observation feature-label pair using a neural network
  - Step 2: Aggregate the embeddings using a commutative operation
  - Step 3: Predict each time-step label using the aggregated embedding using a neural network

Databases

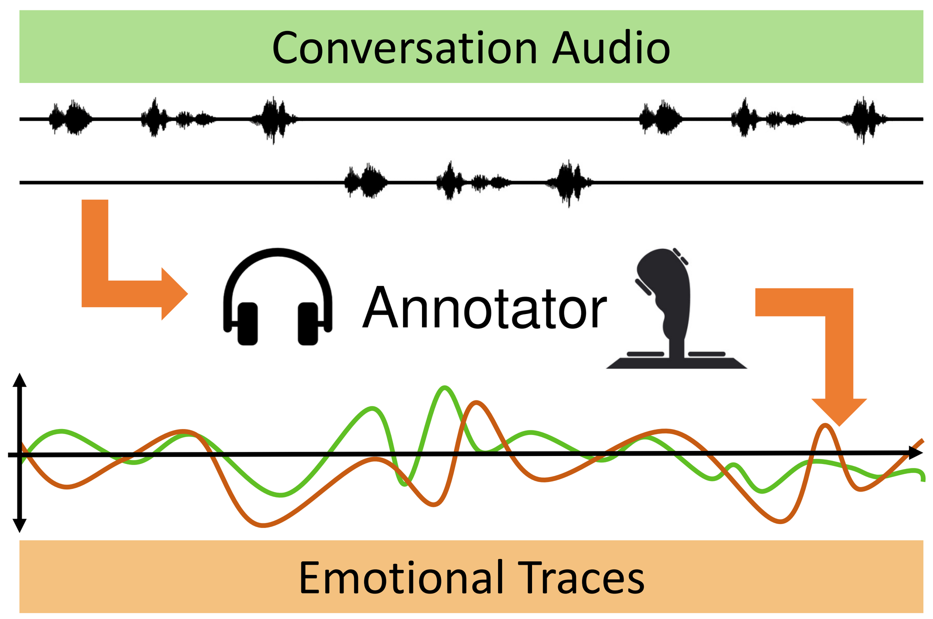
MSP-Podcast Corpus (Segments)

- Used to train the SER model that predicts the observation pseudo-labels
- Speech sentences obtained from publicly available audio sources
- Annotated with single values of:



MSP-Conversation Corpus (Conversations)

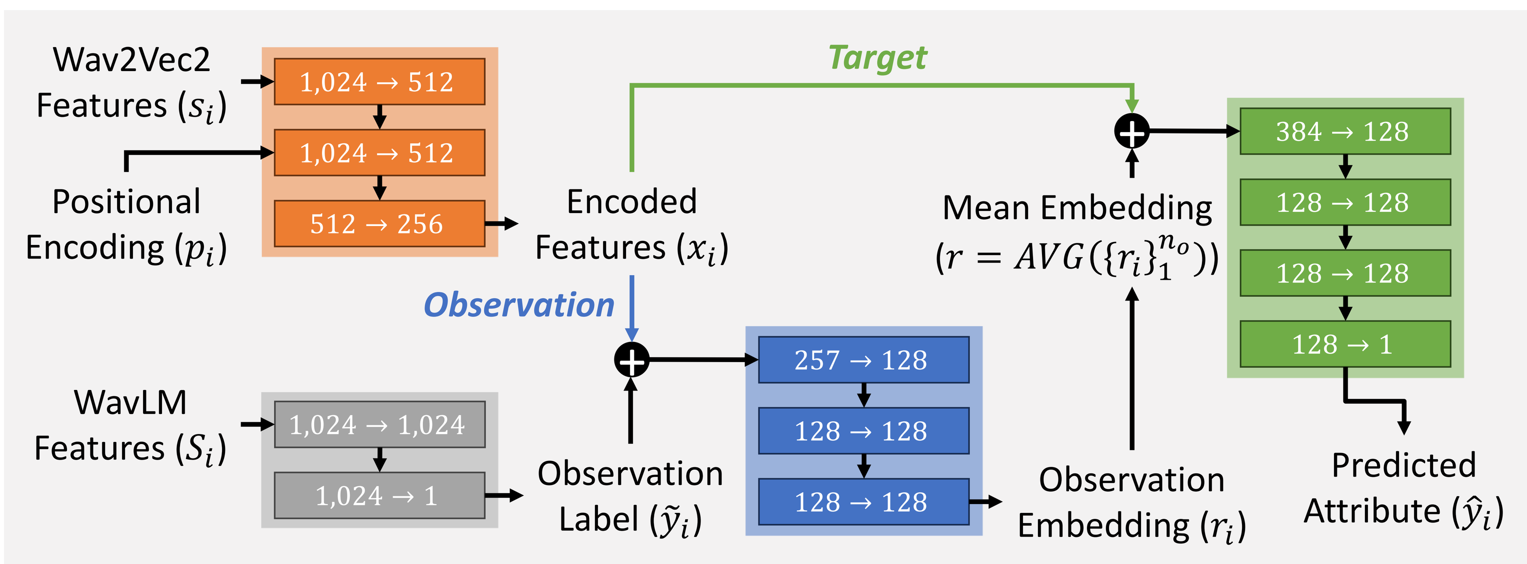
- Used to train the CNP main model
- Audio-only conversations from podcasts in the MSP-Podcast corpus
- Each conversation is annotated with emotional traces of the attributes



Results

- CNP model (ground-truth labels for observations)
  - Add Gaussian noise $\sim \mathcal{N}(0, \sigma_N^2)$ to observation labels
- SER+CNP model (predicted pseudo-labels for observations)

Attribute	Model	# Obs.	CCC ↑	ρ ↑	MSE ↓
Arousal	BiLSTM		0.594	0.602	112
	CNP	20	0.766	0.767	74.3
	SER+CNP	20	0.560	0.574	118
Valence	BiLSTM		0.390	0.397	498
	CNP	35	0.802	0.803	149
	SER+CNP	35	0.474	0.478	441
Dominance	BiLSTM		0.435	0.440	124
	CNP	30	0.801	0.802	150
	SER+CNP	30	0.445	0.455	113

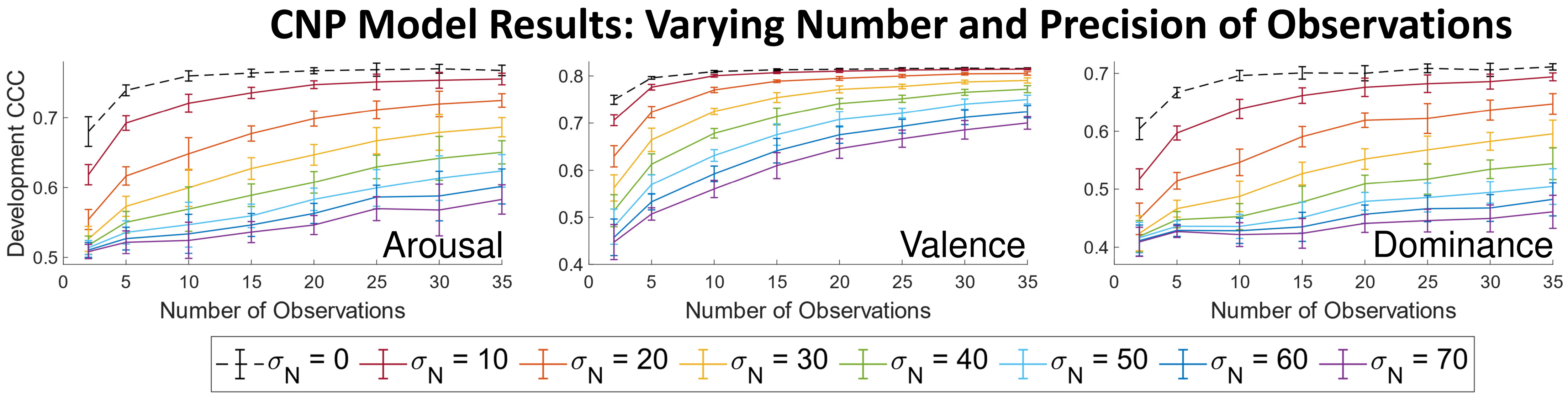


The CNP model (ground-truth observations) performs the best

SER+CNP valence and dominance models perform better than BiLSTM baselines

21% increase for valence

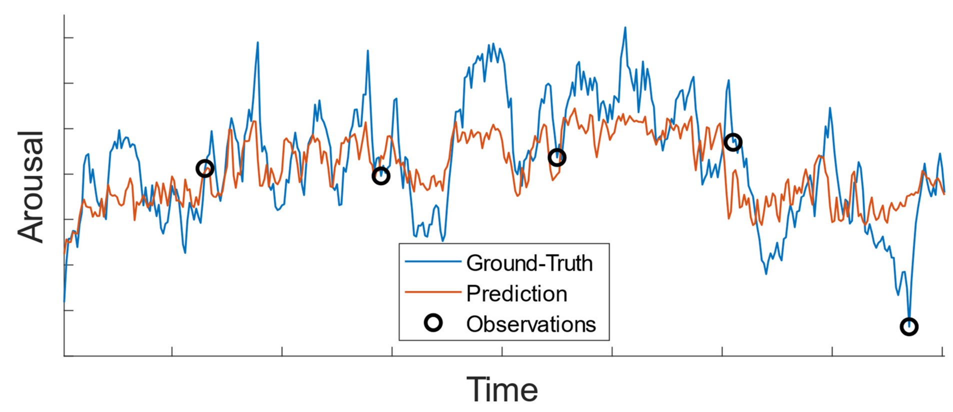
CNP Model Results: Varying Number and Precision of Observations



Conclusions

- This work proposes a DSER model based on a CNP method

Prediction of CNP Model using the Ground-Truth Observation Labels



- Tested our method using two types of observation labels:
  - Ground-truth labels
  - Predicted pseudo-labels
- Future Work:
  - Improve SER model used to predict the pseudo-labels
  - Improve CNP performance by using more sophisticated model

This work was supported by NSF (CNS-2016719)