Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science

Qualification Exam Presentation – Fall 2018

Role of Regularization in the Prediction of Valence from Speech

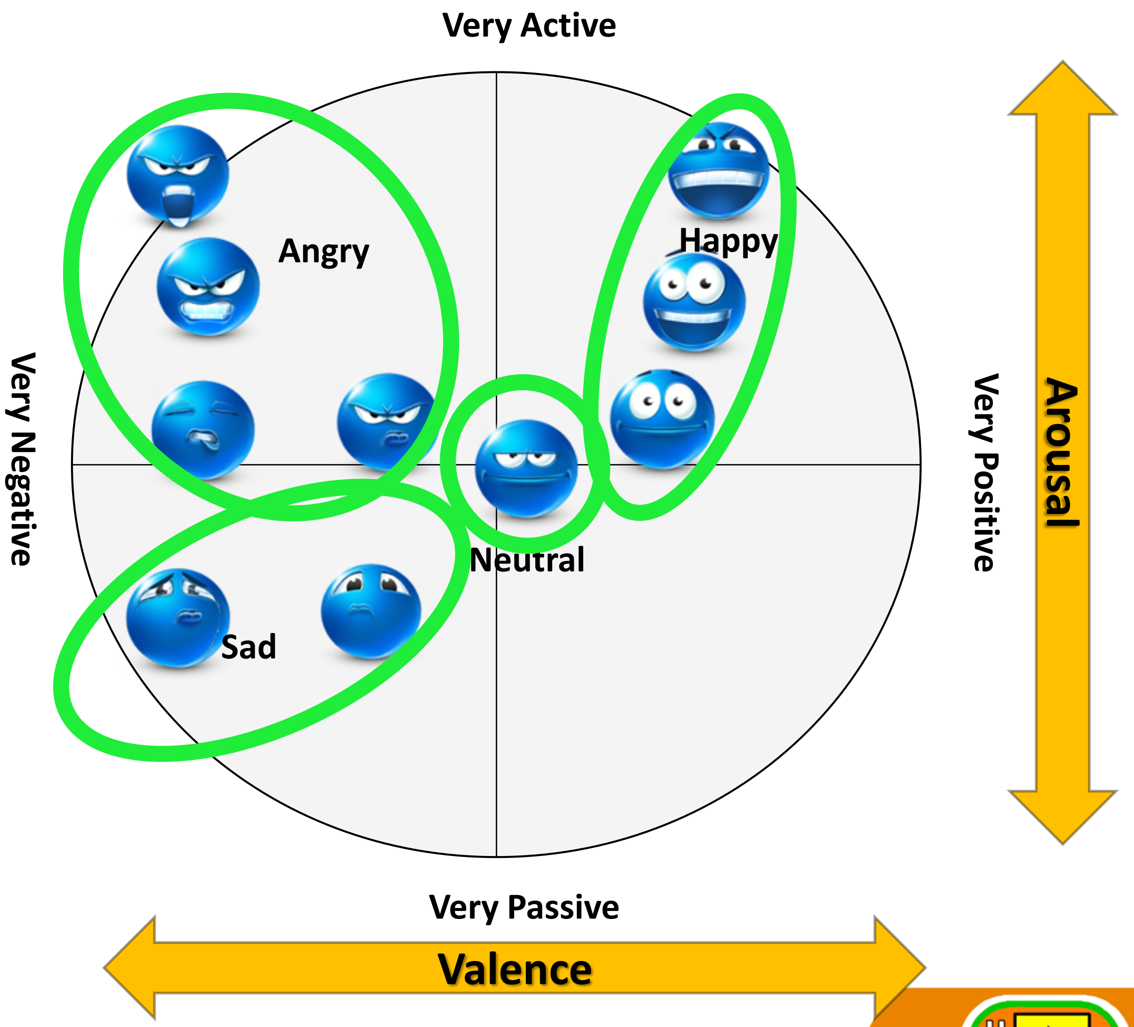
Kusha Sridhar

(presented at Interspeech 2018)

Motivation

<u>Attribute Descriptors</u>

- Supported by core affect theory
- Human interaction consists of mixed emotional content – hard to classify into few distinct classes
- Emotional attributes can describe differences within emotional categories – Appealing !!!



Motivation – From Psychology

Emotional attributes are more suitable to describe complex human behaviors in everyday interactions.

Characteristic behaviors in the expression of valence

- **People express pleasure or displeasure in varied manners**
  - Appraisal of situation dictates behaviors
  - Two people in the same situation often externalize valence differently
  - In self-reported mood, the spread for valence scores is higher than arousal [Feldman 1995]

$$\sigma^2(\text{Valence}) = 2 \,.\, \sigma^2(\text{Arousal})$$

Study of Valence Emotion

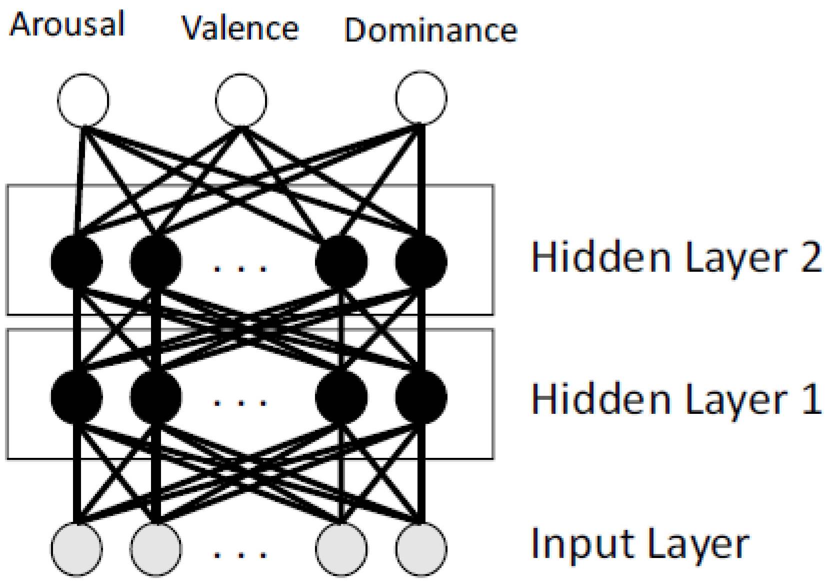
- Valence attribute (negative vs. positive) is key in many applications
  - Mental health, costumer service, security and defense
- Speech-based classifiers often lead to lower performance for valence, compared to other emotional attributes (e.g., arousal and dominance)

Studies	Arousal	Valence
Trigeorgis 2016 (convolutional RNN)	0.686 (CCC)	0.261 (CCC)
Parthasarathy 2016 (rank-based classifier)	89.7% (Accuracy)	65.7% (Accuracy)
Lotfian 2016 (preference learning)	75.1% (Accuracy)	66.8% (Accuracy)

It is important to explore options to improve the performance in detecting valence from speech

Motivation – From speech

- **Previous observations for detecting valence from speech**
  - Few acoustic features are more discriminative for valence alone [Busso&Rahman, 2012]
  - Temporal context can help improve valence prediction [Lee et al., 2009]
  - Improvements when jointly predicting valence with arousal and dominance under a multitask learning framework [Parthasarathy and Busso, 2017,2018]



This paper explores the role of regularization in DNNs as one of the aspects that can lead to better prediction of valence from speech

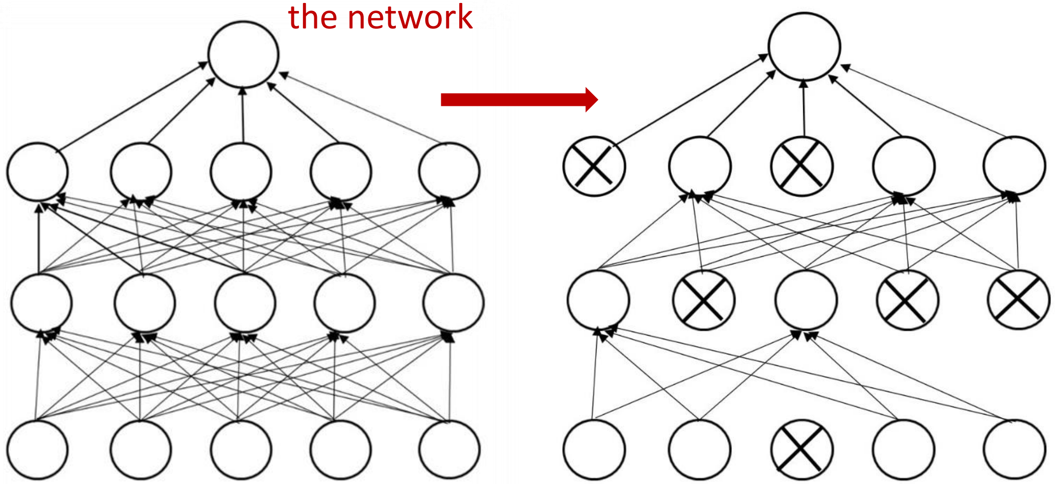
Improving Valence Predictions

- **Role of regularization**
  - **Hypothesis**: Higher regularization leads to better prediction for valence
  - Allows DNN to find consistent trends across speakers
  - Focus is on the role of dropout in the prediction of valence
- **Methodology**
  - Analyzing the model performance as a function of dropout probability
  - Analyzing performance for different DNN configurations (# layers, # nodes, emotional attributes)

Regularization in DNNs

- **Regularization is very important in DNNs to avoid overfitting**
  - Learn general patterns rather than specific trends in training set
- **Different approaches for regularization:**
  - **Dropout**, early stopping, data augmentation, weighted penalties on the training data, multitask learning
- **Dropout**
  - Randomly ignores nodes in the network
  - Essentially, it trains a smaller network on each iteration
  - Prevents learning of interdependent feature weights
  - Prevent co-dependencies across neighbor nodes

Randomly shut down parts of the network



Standard network

Network with dropout

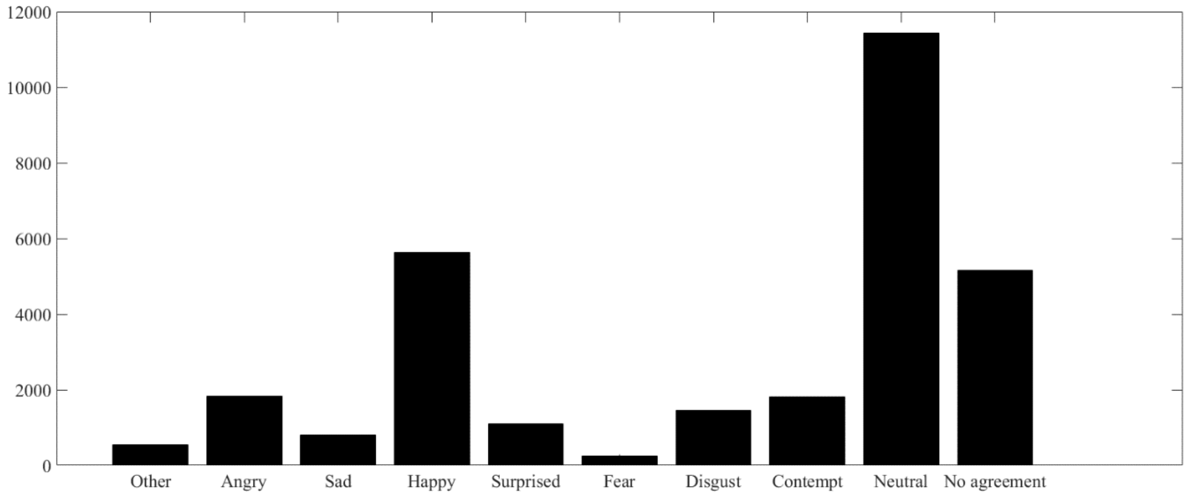
p = Dropout rate

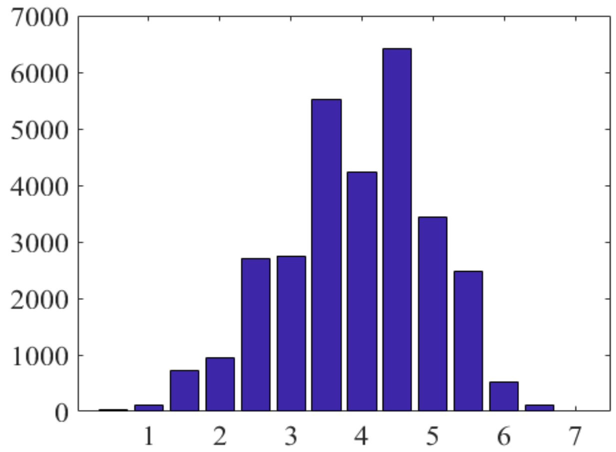
MSP-Podcast database

Ongoing effort

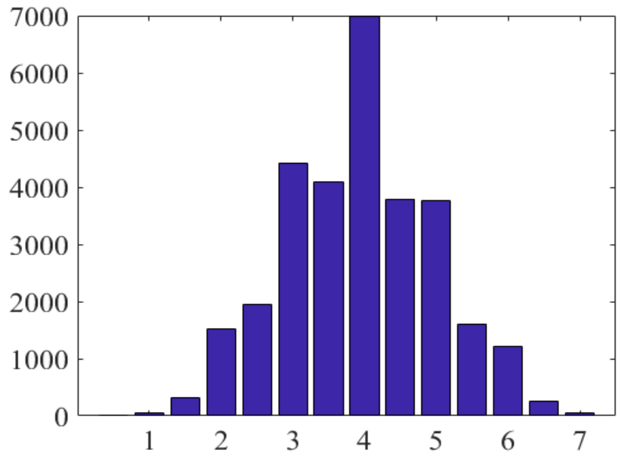
With emotion labels:
30,681 sentences
(50h, 09m)

Segmented turns
244,477 sentences from 1500 podcasts

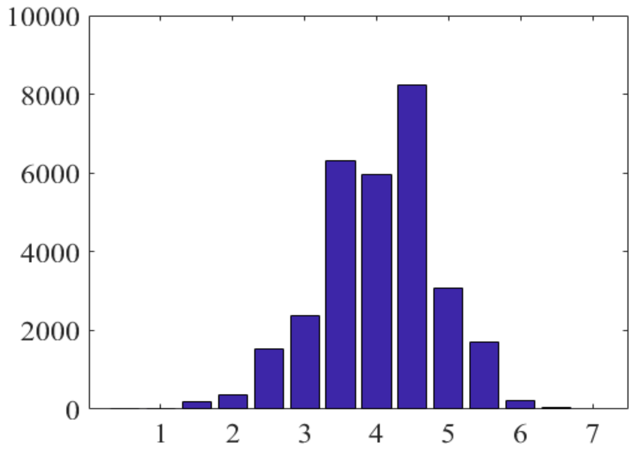




Arousal



Valence

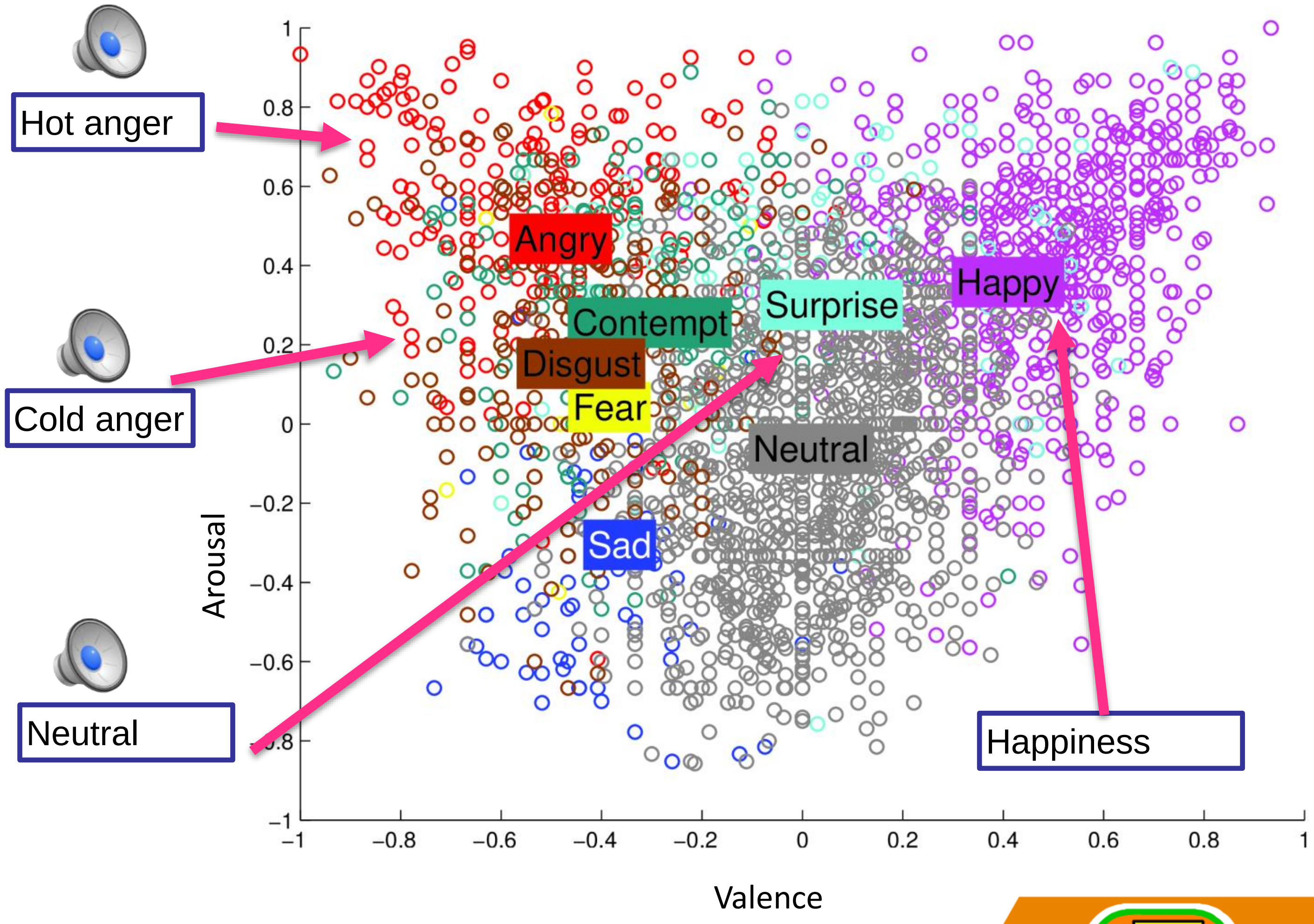


Dominance

THE UNIVERSITY OF TEXAS AT DALLAS

MSP-Podcast database

- Version 1.0 of the **MSP-Podcast** corpus
  - 20,045 (30h43m)
- Corpus partition with minimal speaker overlap sets:
  - Training data: 11,750 samples
  - Test data: 6,069 samples
  - Validation data: 2,226 samples



Experimental Framework

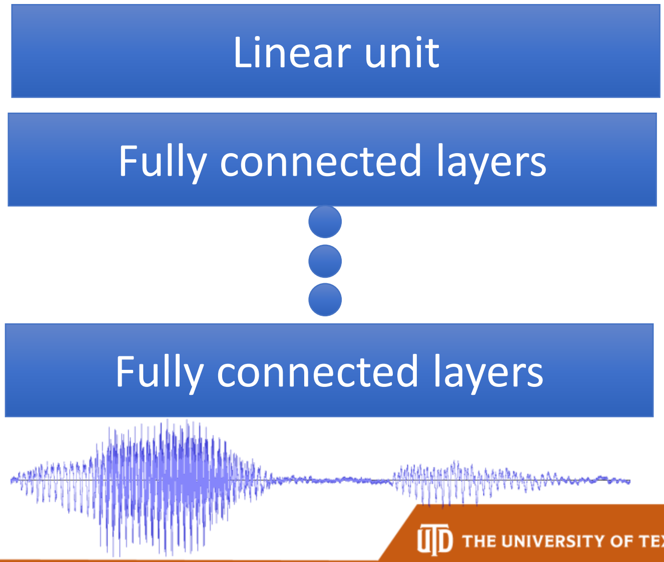
- **The features correspond to the IS2013 ComparE feature set (6,373 features)**

Architecture of the network

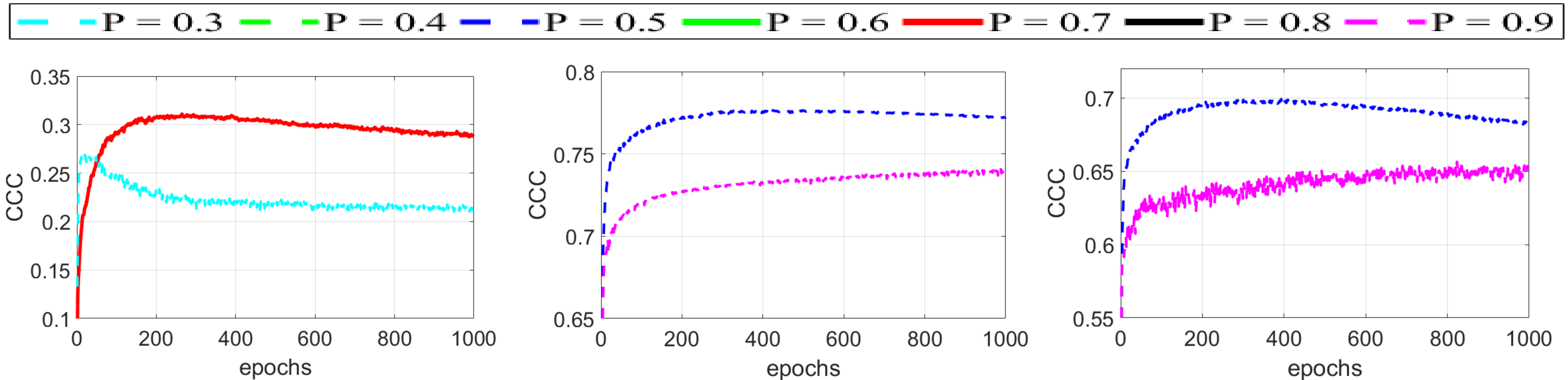
  - 2, 4 or 6 layers
  - 256, 512 and 1024 nodes per layer

- Output of DNN is a prediction score for arousal, valence or dominance
- Batch normalization to normalize the output of each layer
- Trained for 1,000 epochs with early stopping
  - Concordance Correlation Coefficient (CCC) achieved on the validation set

Network parameters	Values
Activation	ReLU
Optimizer	SGD with momentum of 0.9
Learning rate	0.001
Evaluation metric & cost function	Concordance Correlation Coefficient (CCC)



Analysis: Performance in Terms of Dropout



Valence – Test set

Arousal – Test set

Dominance – Test set

- Two layers with 256 nodes
- **Results:**
  - The optimum dropout rate:
    - Valence is in the range {0.7,0.8}
    - Arousal and dominance is in the range {0.4,0.5}

Analysis: Performance in Term of Nodes

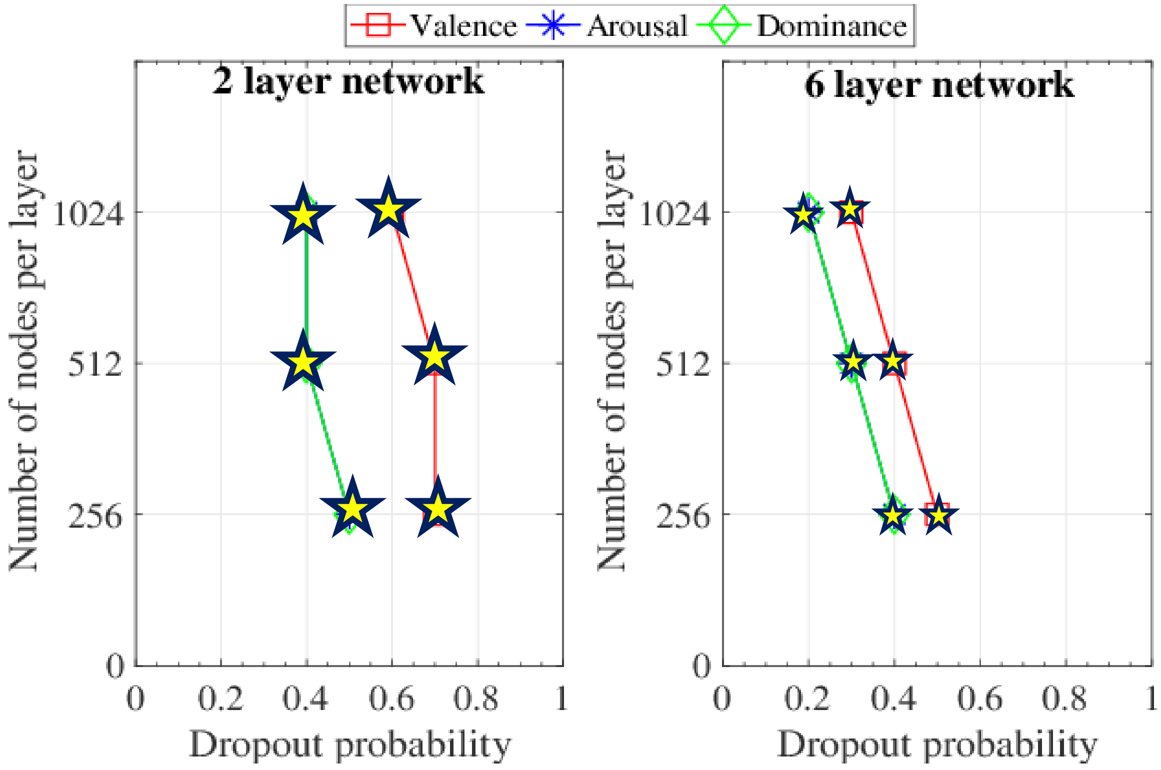
- **DNN with two layers (256, 512, 1,024 nodes)**
- **Average CCC values for p = 0.5 and p=0.7 over 10 trials**
- *** indicate significant differences between both dropout rates (one-tailed t-test)**

- **Results**
  - Better performance for valence with p=0.7
  - Better performance for arousal and dominance with p=0.5

		Test set	
Attributes	Nodes	P = 0.5	P = 0.7
Valence	256	0.2903	0.3102*
	512	0.2870	0.3080*
	1024	0.2841	0.3009*
Arousal	256	0.7733*	0.7577
	512	0.7717*	0.7525
	1024	0.7691*	0.7472
Dominance	256	0.6936*	0.6733
	512	0.6902*	0.6617
	1024	0.6888*	0.6523

Analysis: Optimal Dropout Rate (# Nodes)

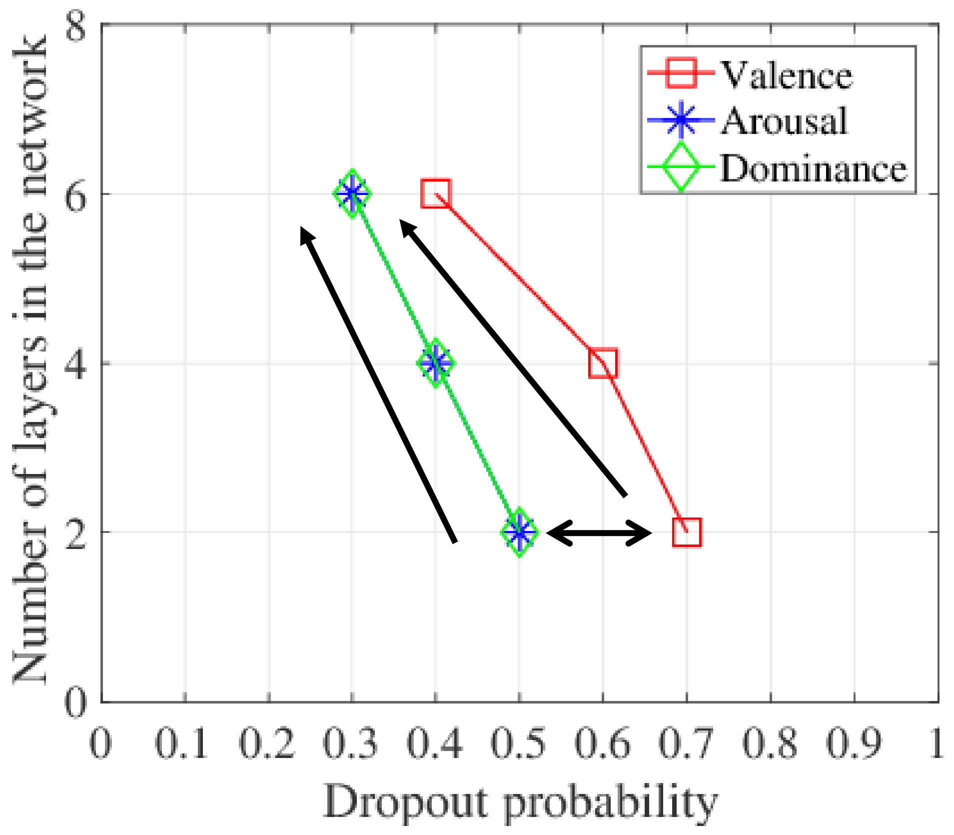
- **DNNs with two or six layers**
  - 256, 512 or 1,024 nodes
  - Dropout on all layers
- **Results:**
  - Optimal dropout rate for arousal and dominance are the same across conditions
  - Optimal dropout rate decreases as the network is implemented with more nodes
  - Gap between optimal dropout rate for valence and arousal/dominance is consistent



Analysis: Optimal Dropout Rate (# Layers)

- **DNNs with two, four or six layers**
  - 256 nodes

- **Results:**
  - Optimal dropout rate decreases as the network is implemented with more layers
  - Gap between optimal dropout rate for valence and arousal/dominance is consistent

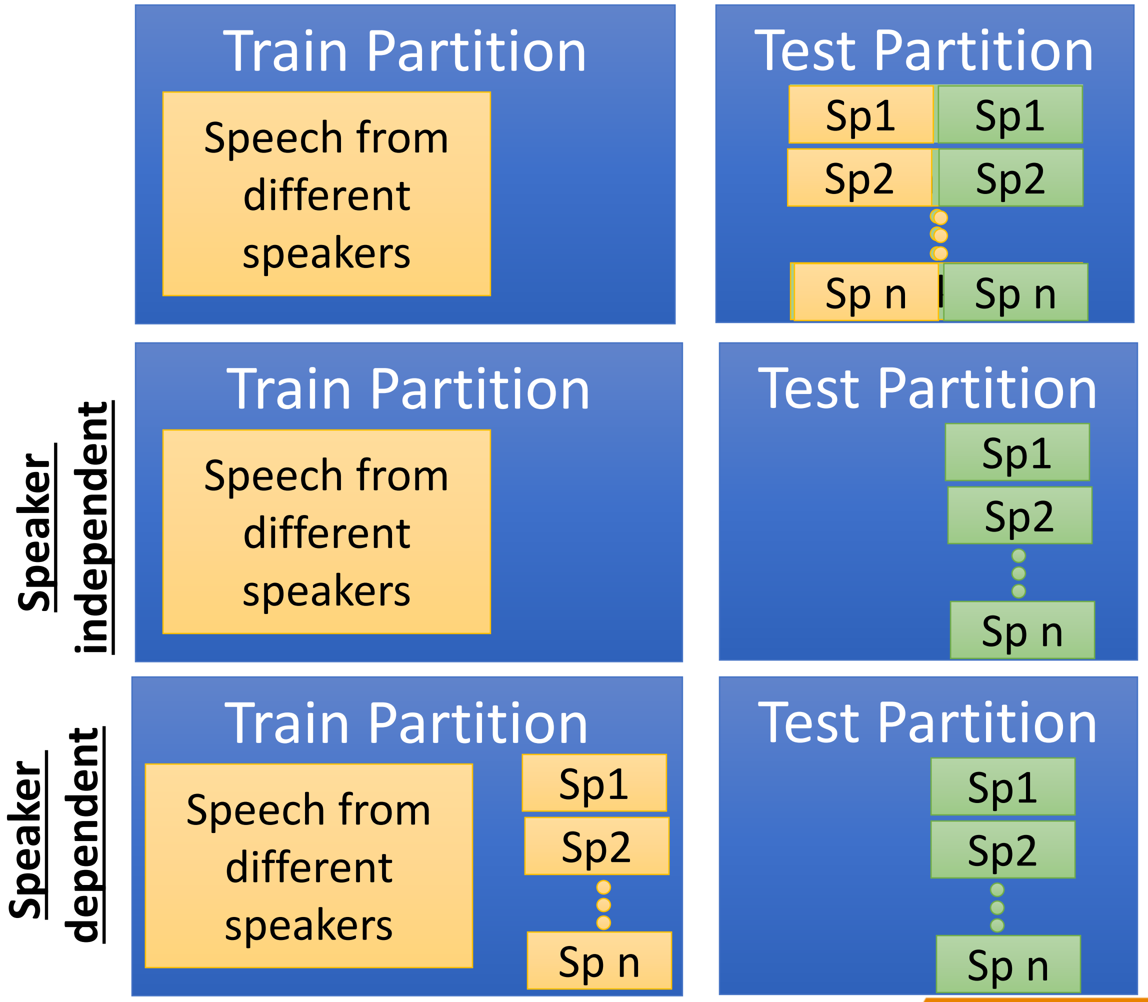


Why Does Valence Need Higher Dropout?

Hypothesis:

- Speaker dependent nature of emotional cues
  - When heavily regularized, the network learns features that are consistent across all speakers
  - It places less emphasis on speaker dependent traits

- **Experiment to validate this hypothesis**
  - Compare DNNs trained on speaker dependent and independent train-test partitions
  - Speaker dependent predictors should lead to higher performance gain for valence
    - They learn patterns from target speaker



Speaker dependent vs independent conditions

		Speaker Independent	Speaker Dependent	Gain (%)
Attributes	Nodes	Test	Test	Test
Valence	256 512 1024	0.2906 0.2835 0.2880	0.3761 0.3686 0.3600	29.42 30.01 28.57
Arousal	256 512 1024	0.7712 0.7720 0.7688	0.7885 0.7813 0.7800	2.24 1.20 1.45
Dominance	256 512 1024	0.6901 0.6837 0.6782	0.7051 0.7052 0.7005	2.17 3.14 3.28

- **DNNs with two layers**
- **Results:**
  - Important performance gain for valence in speaker dependent condition (~30%)
  - Performance gain is not as high for arousal and dominance
  - Significant gap in performance validates our hypothesis that valence is expressed with more speaker dependent cues

Final Remarks

- Predicting valence from speech requires a higher dropout rate than arousal or dominance
  - Optimal dropout rate is consistently higher for valence across different network configuration
- Discriminative acoustic features for detecting valence vary across speakers
  - Dropout regularizes the network to learn consistent patterns across speakers
- Take home message:
  - Valence imposes challenges that should be carefully considered
  - Optimal parameters are not necessarily the same as the ones for arousal or dominance

Future Directions

- Evidence from Speaker dependent experiments – leveraging information learned from train speakers to personalize to target speakers.
  - Using techniques like model adaptation or weighting to achieve personalization.

Questions ?

This work was funded by NSF CAREER award IIS-1453781

[MSP Lab UT Dallas](msp.utdallas.edu)