Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning

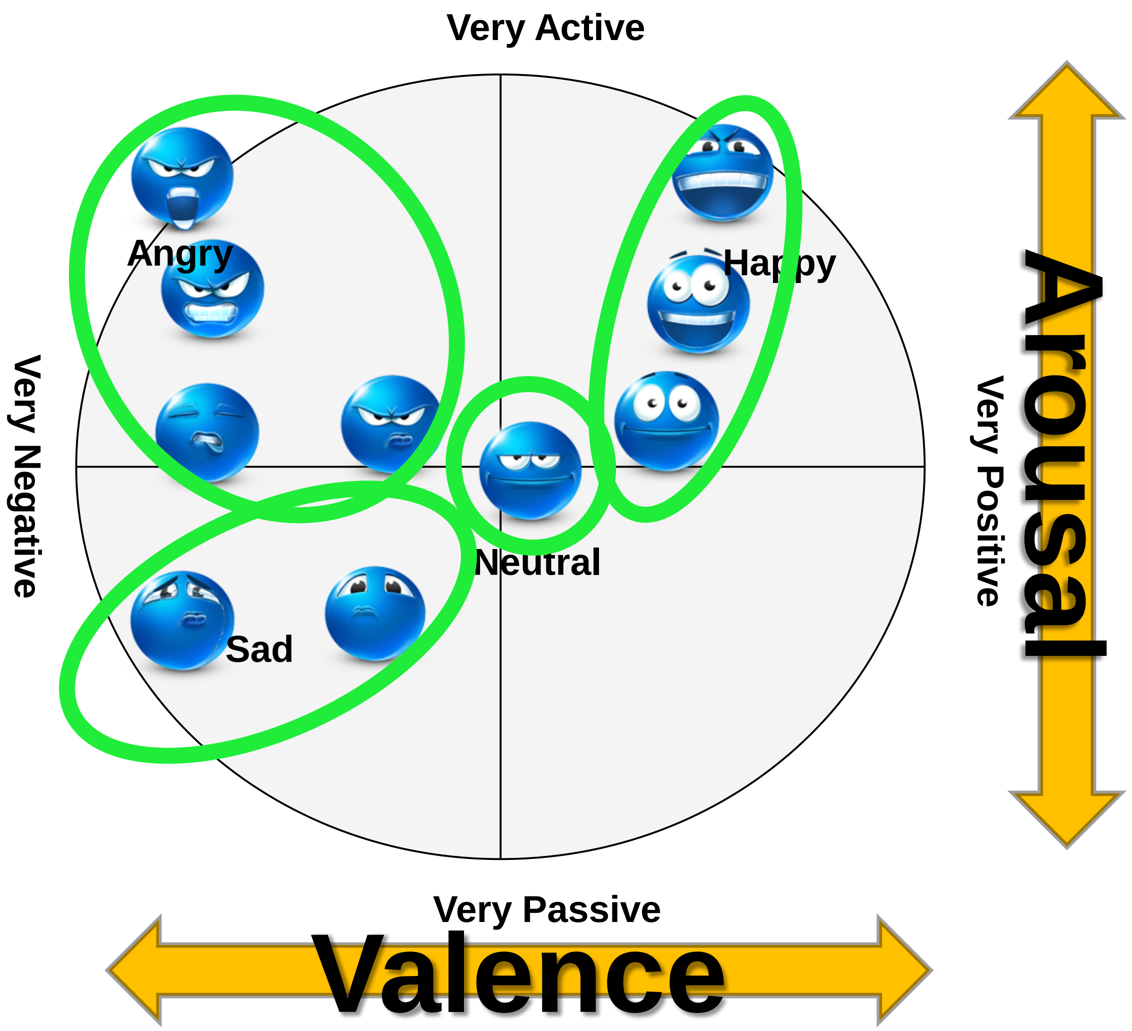
Srinivas Parthasarathy and Carlos Busso

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

August 22, 2017

Motivation

- Emotions represented using emotional attributes
  - Arousal – passive vs. active
  - Valence – negative vs. positive
  - Dominance – weak vs. strong



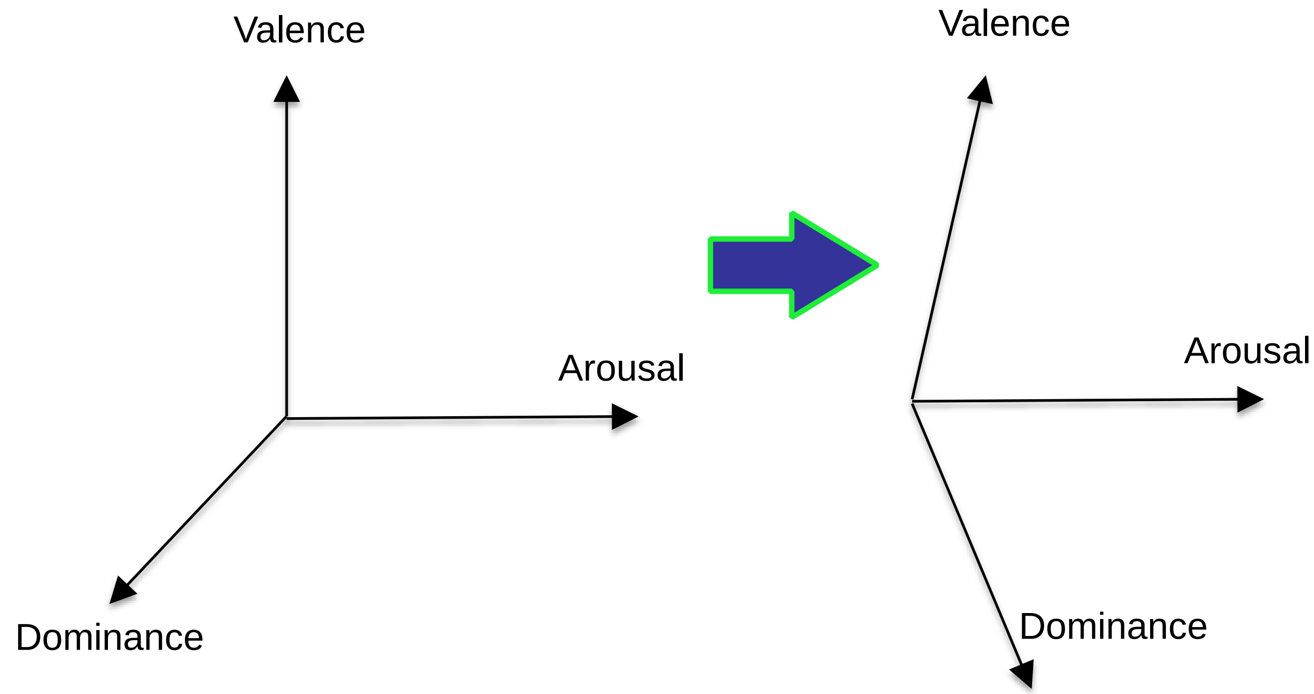
- Attributes are very appealing
  - Some emotions are ambiguous and are hard to label with emotional categories
  - Attributes provide finer granularity to represent emotion

Limitations

- Systems that predict emotional attributes have some limitations
  - Systems predict each emotional attribute independently – ignore dependencies between attributes [1], [2], [3] related work

Correlation between attributes MSP-IMPROV

	Valence	Dominance
Arousal	0.2217	0.6503
Valence		0.3129



[1] P. Lewis, H.Critchley, P.Rotshtein, and R.Dolan, “Neural correlates of processing valence and arousal in affective words,”

[2] J. Russell, “Evidence of convergent validity on the dimensions of affect,”

[3] M. Nicolaou, H. Gunes, and M. Pantic, “Continuous prediction of spontaneous affect from multiple cues and modalities in valence- arousal space,”

Solution

- Appealing solution – Jointly learn multiple emotional attributes
- Leverage dependencies between the attributes
- Multitask Learning (MTL) framework for learning attributes
  - Learning secondary attribute helps primary attribute
  - Regularizes learning
- Learn feature representations that benefit predicting all three attributes while optimizing for target attribute

MTL - Related Work

- Xiu and Liu[1] proposed multi-task learning framework
  - Primary task – emotional categories
  - Secondary task – prediction/classification attribute scores
- Zhang et al.[2] proposed multi-task framework with shared hidden layers
  - Jointly classify emotions with different representations (e.g., varied number of classes, quadrants in arousal-valence space)
- Chang and Scherer[3] proposed jointly learning valence and arousal
  - Valence primary task
  - Three, five-class classification problem

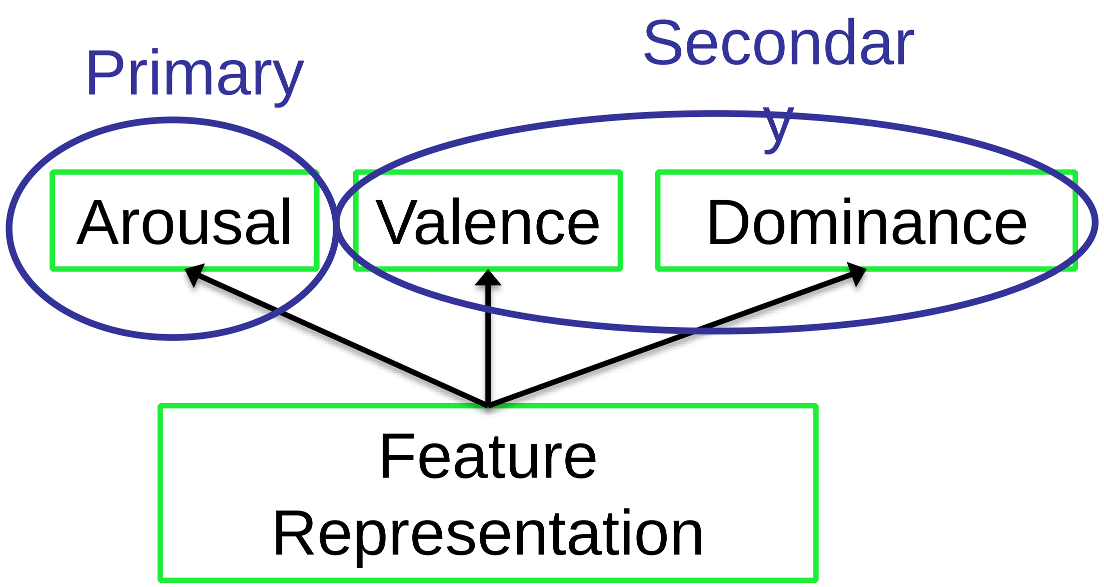
[1] R.Xia and Y.Liu,"A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*

[2] Y.Zhang, Y.Liu, F.Weninger, and B.Schuller,"Multi-taskdeep neural network with shared hidden layers: Breaking down the wall between emotion representations," *(ICASSP 2017)*

[3] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," *(ICASSP 2017)*

Contributions

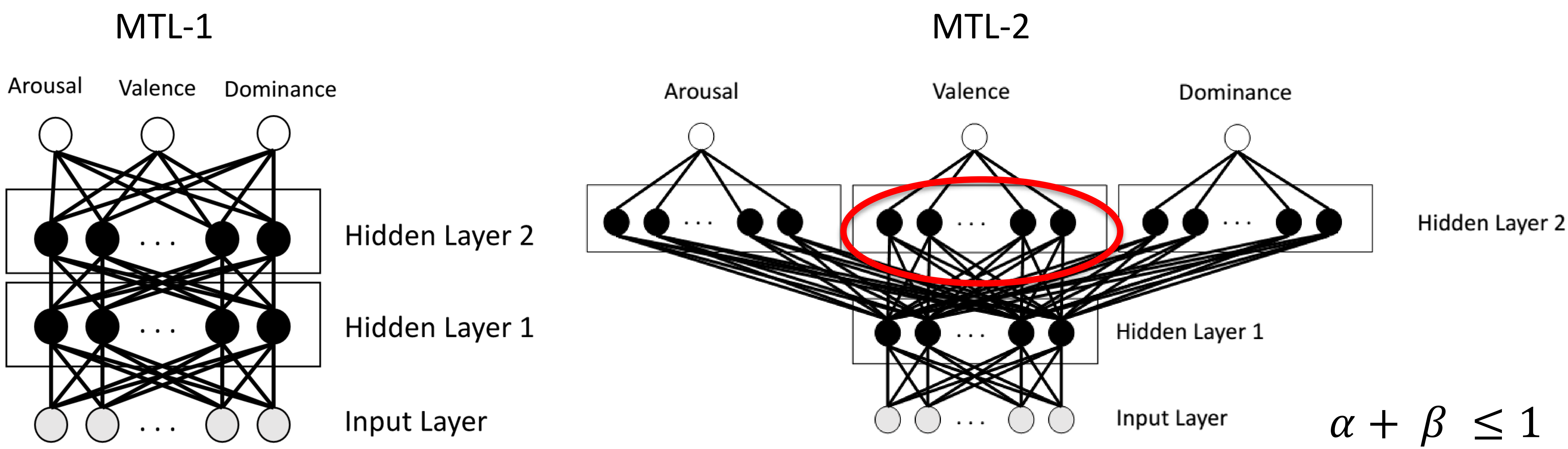
- Use MTL where the primary task is target attribute (e.g., arousal) secondary tasks – other two attributes (e.g., valence, dominance)



- Explore attribute-dependent layers on top of shared hidden layers
- Extensive within corpus and cross corpus evaluations

MTL Framework

- Goal: predicting emotional attributes with a unified framework
  - MTL implemented with deep neural networks (DNN)
  - Loss – Mean squared error

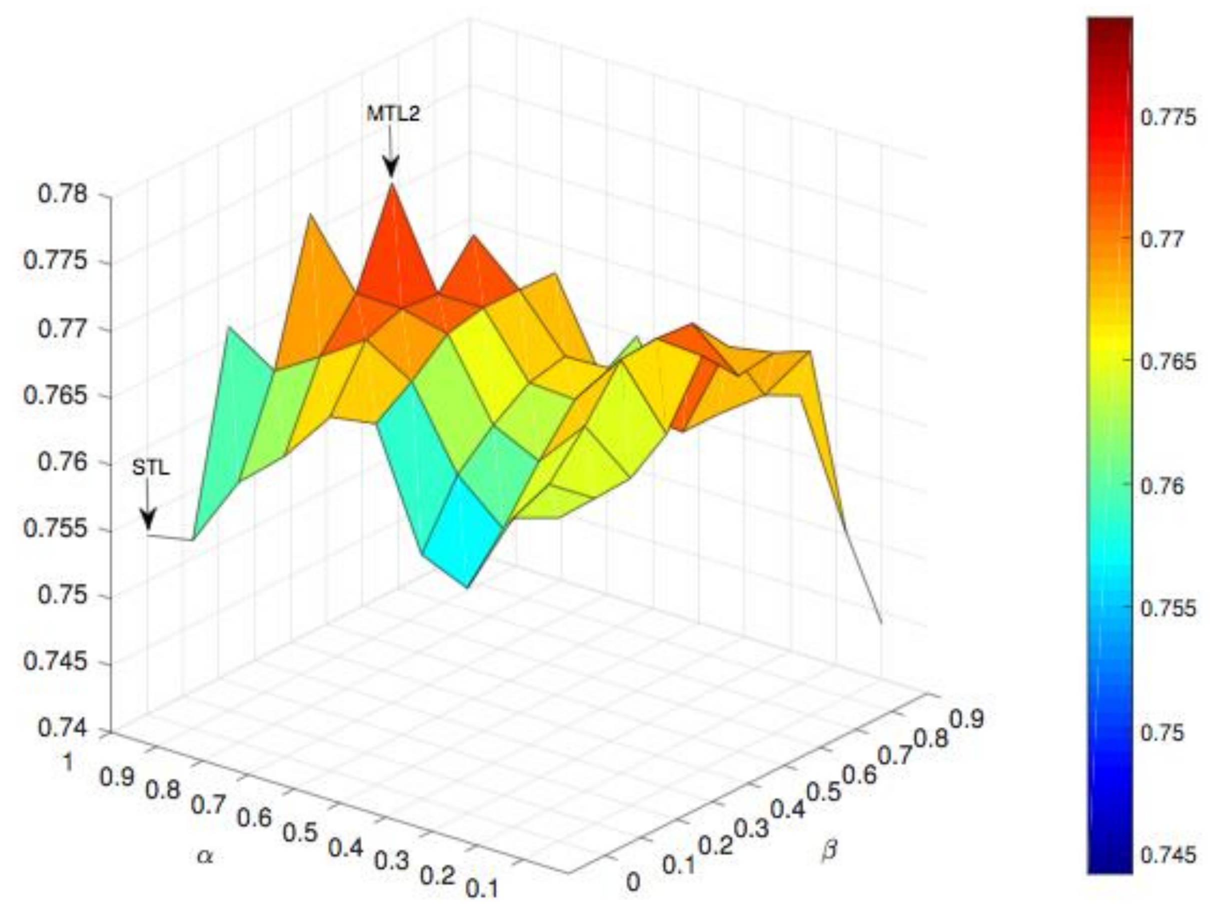


$$\alpha + \beta \leq 1$$

$$MSE_{ov} = \alpha \times MSE_{aro} + \beta \times MSE_{val} + (1 - \alpha - \beta) \times MSE_{dom}$$

MTL Framework

- Weights learned using the development set



STL

$\alpha = 1, \beta = 0$ Arousal

$\alpha = 0, \beta = 1$ Valence

$\alpha = 0, \beta = 0$ Dominance

$$MSE_{ov} = \alpha \times MSE_{aro} + \beta \times MSE_{val} + (1 - \alpha - \beta) \times MSE_{dom}$$

Experimental Evaluation

- Acoustic features
  - Interspeech 2013 feature-set for paralinguistic challenge – 6,373 features
- Implementation
  - 2 hidden layers, 256/ 512/ 1024 nodes with ReLU activation
  - SGD – momentum 0.9, mini-batch 256, dropout 0.5
  - Evaluated on concordance correlation coefficient

$\uparrow$ correlation

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

$\downarrow$ MSE

Experimental Results

- Baseline : Single Task Learning (STL)
  - Individually predict value of arousal, valence, dominance

$$MSE_{ov} = \alpha \times MSE_{aro} + \beta \times MSE_{val} + (1 - \alpha - \beta) \times MSE_{dom}$$

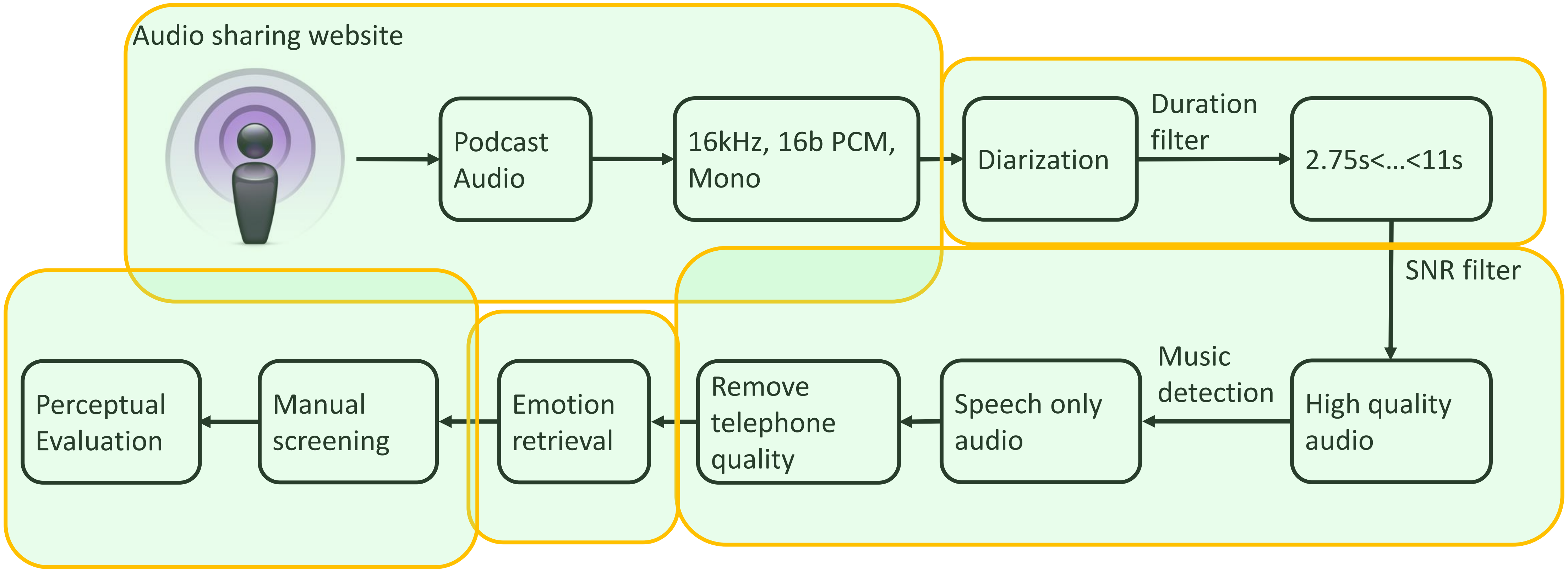
  - Can be formulated setting

    $\alpha = 1, \beta = 0$ for arousal

    $\alpha = 0, \beta = 1$ for valence

    $\alpha = 0, \beta = 0$ for dominance

MSP-PODCAST



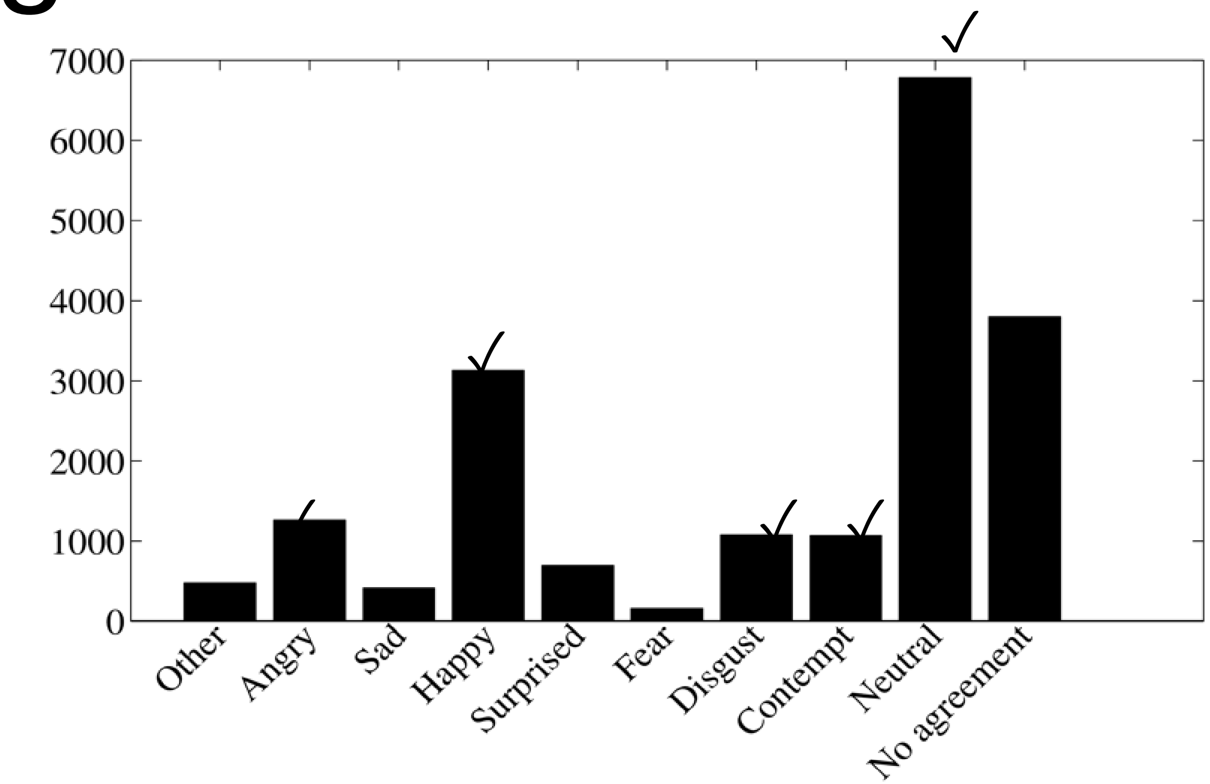
- Collection of audio recordings[1] (Podcasts)
  - Naturalness and the diversity of emotions
  - Creative Commons copyright licenses
  - Duration between 2.75s – 11s
  - Perceptive evaluation of emotional content

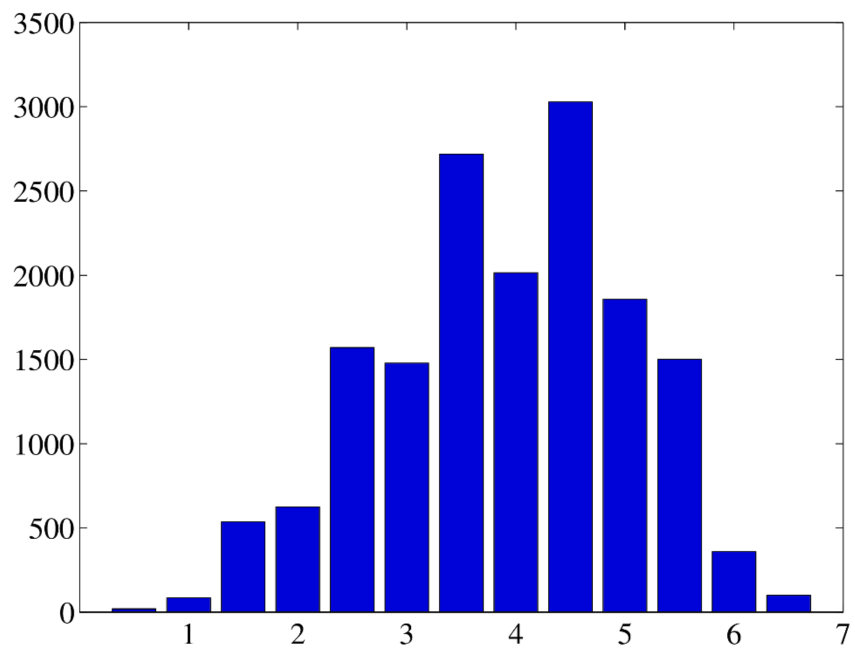
[1] Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing

Status of the MSP-PODCAST: Ongoing work

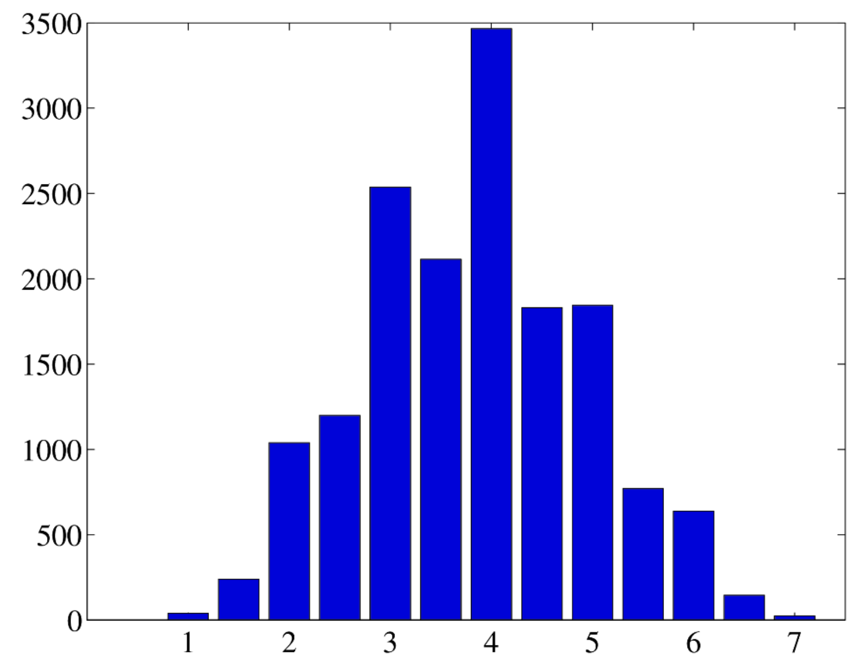
With emotion labels:
19,670 sentences
(29h, 37m)

Segmented turns
190,872 sentences from 952 podcasts

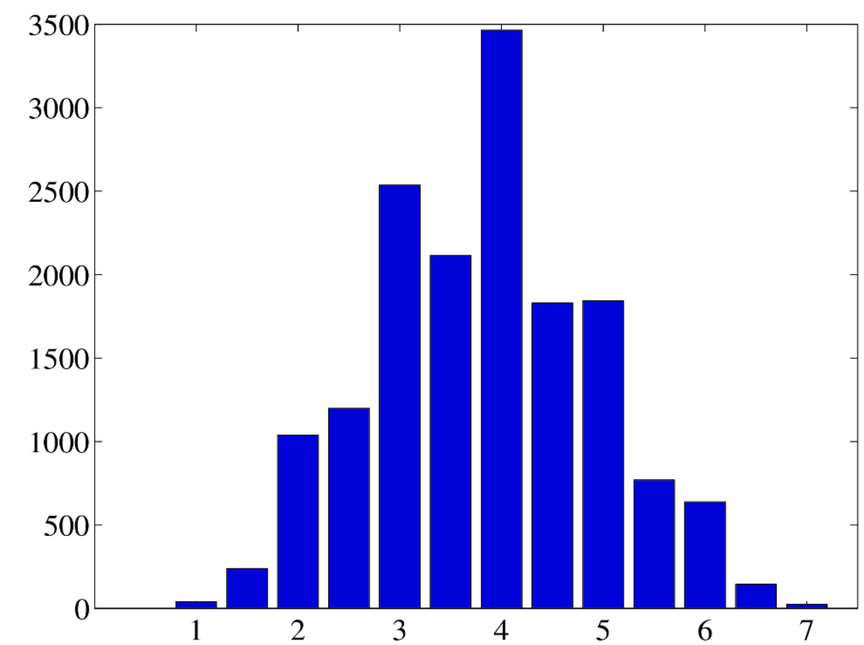




Arousal

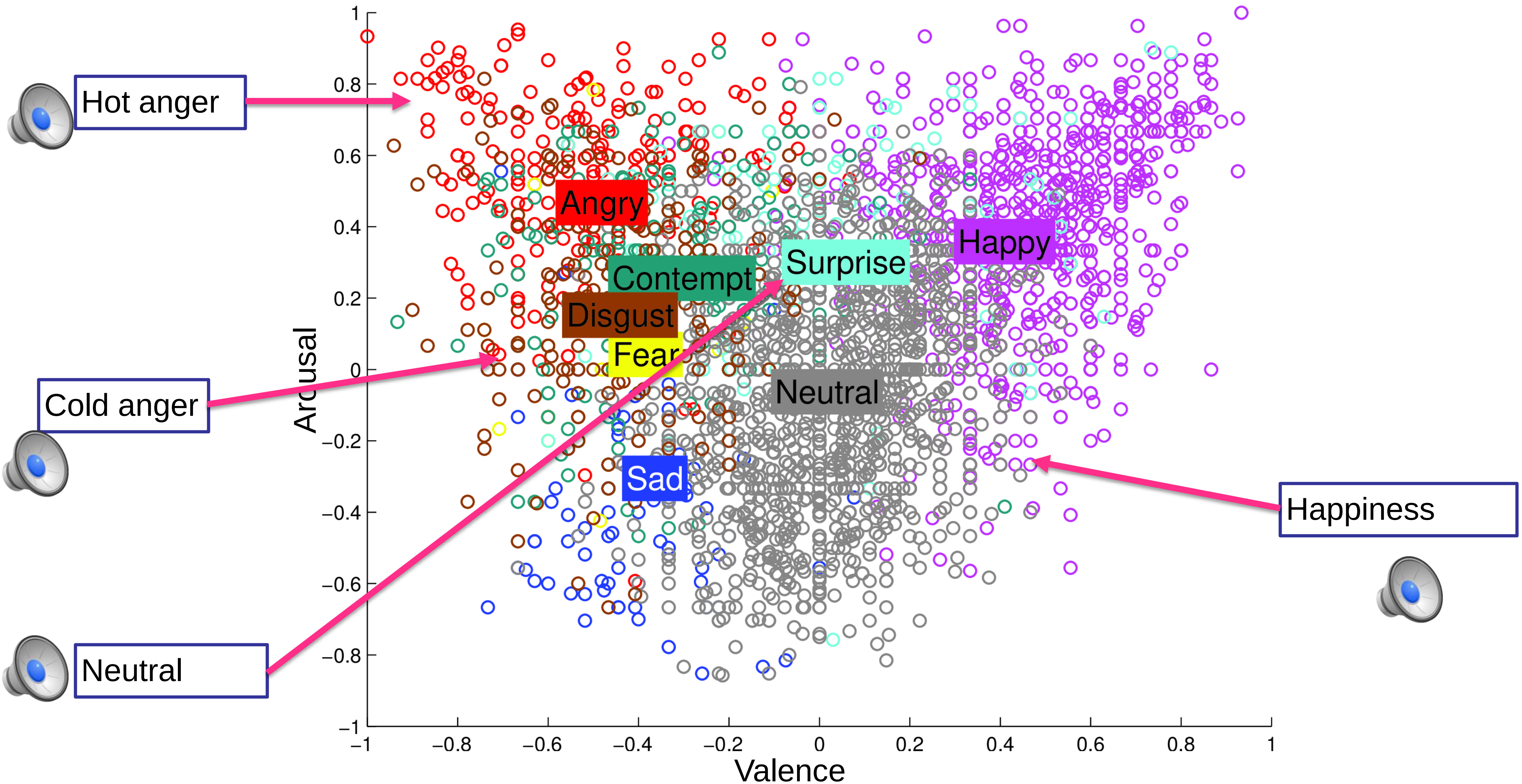


Valence



Dominance

Status of the MSP-PODCAST: Ongoing Work

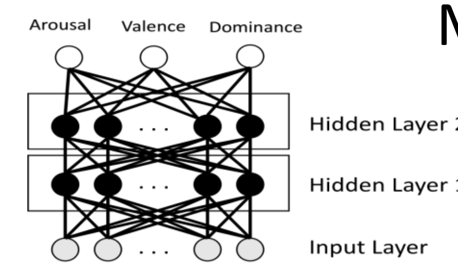


- ✓ Natural recordings
- ✓ The largest database
- ✓ Multiple speakers
- ✓ Rich emotional content

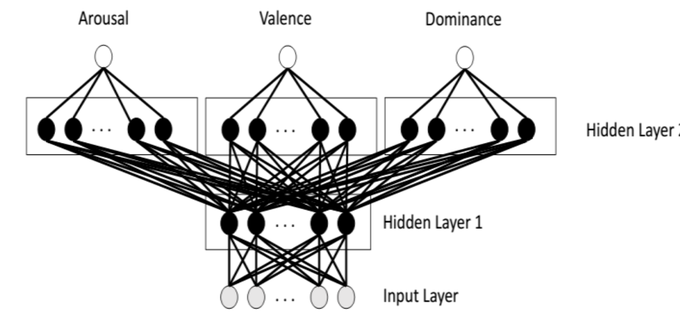
Experimental Results

- Within-corpus evaluation
  - Multi-task learning (MTL) always better than single task learning (STL)
  - Performance increase as we increase number of nodes

MTL-1



MTL-2



Nodes / Layers	*Type of task*	*Concordance Correlation Coefficient*		
		Arousal	*Valence*	*Dominance*
256 / 2	*STL*	*0.7401*	*0.2421*	*0.6697*
	MTL-1	*0.7340*	*0.2721*	*0.6842*
	MTL-2	*0.7496*	*0.2687*	*0.7059*
512 / 2	*STL*	*0.7380*	*0.2702*	*0.6622*
	MTL-1	*0.7489*	*0.2877*	*0.6994*
	MTL-2	*0.7508*	*0.2889*	*0.7097*
1024 / 2	*STL*	*0.7200*	*0.2607*	*0.6796*
	MTL-1	*0.7430*	*0.2826*	*0.6963*
	MTL-2	*0.7635*	*0.2894*	*0.7130*

Within-Corpus Evaluation

Validation Set

10 speakers
887 sent.

Testing Set

50 speakers
5,024 sent.

Training Set

Rest of corpus
6,710 sentences

Other datasets

- USC-IEMOCAP
  - 12 hours of conversational recordings from 10 actors in dyadic sessions
  - Sessions consists of emotional scripts as well as improvised interactions
  - All speaking turns annotated for emotional attributes by two raters on a scale of 1-5

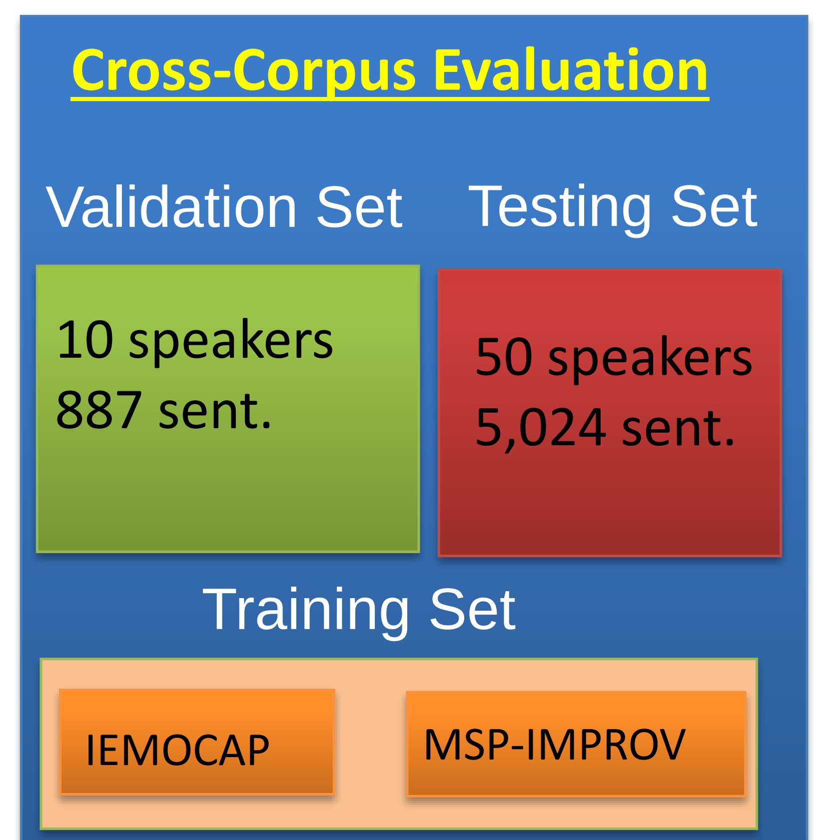
- MSP-IMPROV
  - Improvisation between actors (12 actors)
  - Contains 8,438 speaking turns
  - Annotated by novel crowdsourcing methods on a scale of 1-5 by at least 5 raters

- Emotional values in all databases scaled between [-1,1]

Experimental results

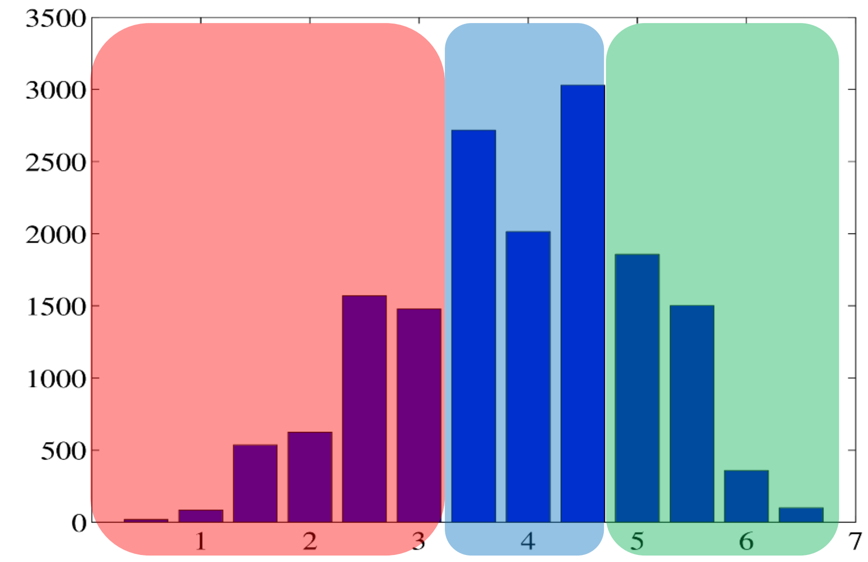
- Cross-corpus evaluation
  - Performance drops with respect to within-corpus evaluations
  - Benefit of multi-task increases – 0.14
  - Best performance with lower number of nodes per layer

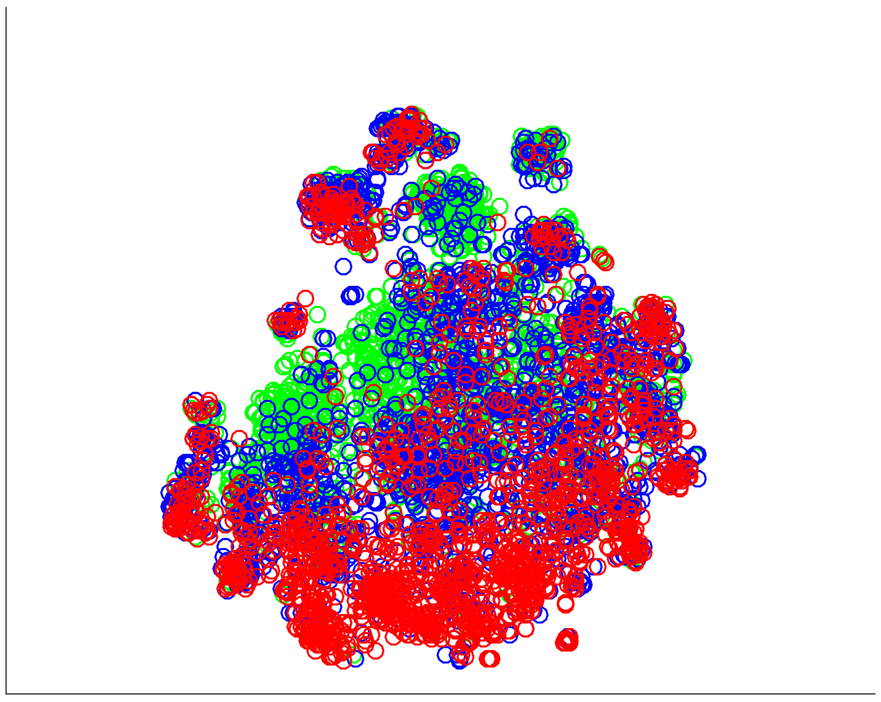
Nodes / Layers	Type of task	Concordance Correlation Coefficient		
		Arousal	Valence	Dominance
256 / 2	STL	0.4052	0.1519	0.3109
	MTL-1	0.4329	0.1519	0.4408
	MTL-2	0.4642	0.1674	0.4512
512 / 2	STL	0.3877	0.1308	0.3006
	MTL-1	0.3985	0.1745	0.4381
	MTL-2	0.4242	0.1843	0.4398
1024 / 2	STL	0.3726	0.1426	0.3131
	MTL-1	0.3908	0.1607	0.4364
	MTL-2	0.4616	0.1697	0.4384



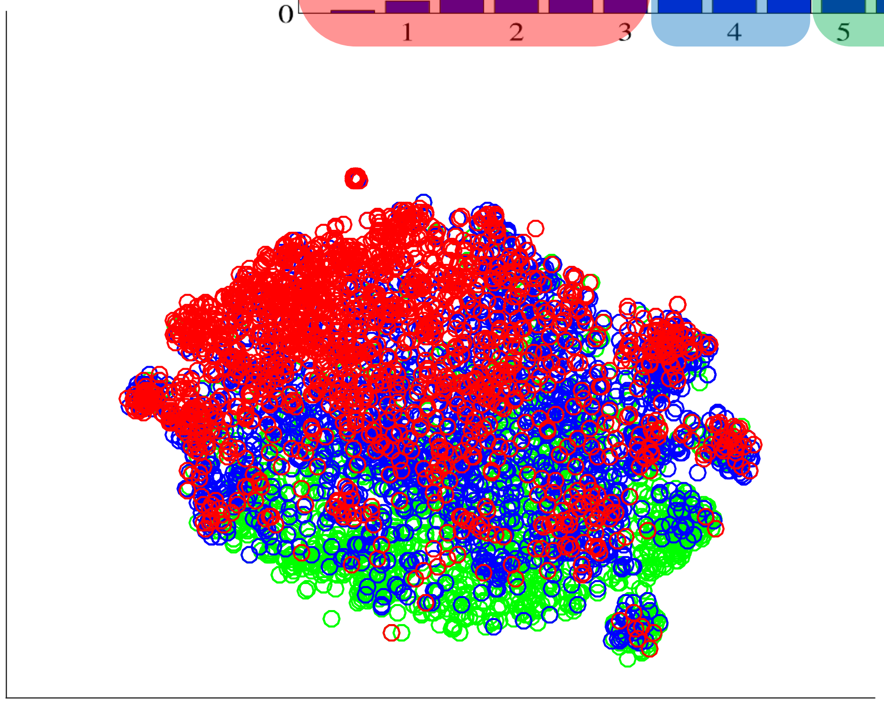
Feature Representation

- Feature representation of best models illustrated with t-SNE for arousal
- Values divided into three classes
- Classes better separated in MTL2





STL



MTL2

Conclusions

- Recognizing emotional attributes is appealing
- Some dependencies exist between various emotional attributes
- Dependencies can be learnt with MTL
- MTL’s with shared hidden layers and attribute dependent layers perform better than STL
- Improvement in concordance correlation coefficient for within corpus and cross corpus tests

Questions ?

This work was funded by NSF CAREER award IIS-1453781

msp.utdallas.edu