Formulating Emotion Perception as a Probabilistic Model with Application to Categorical Emotion Classification

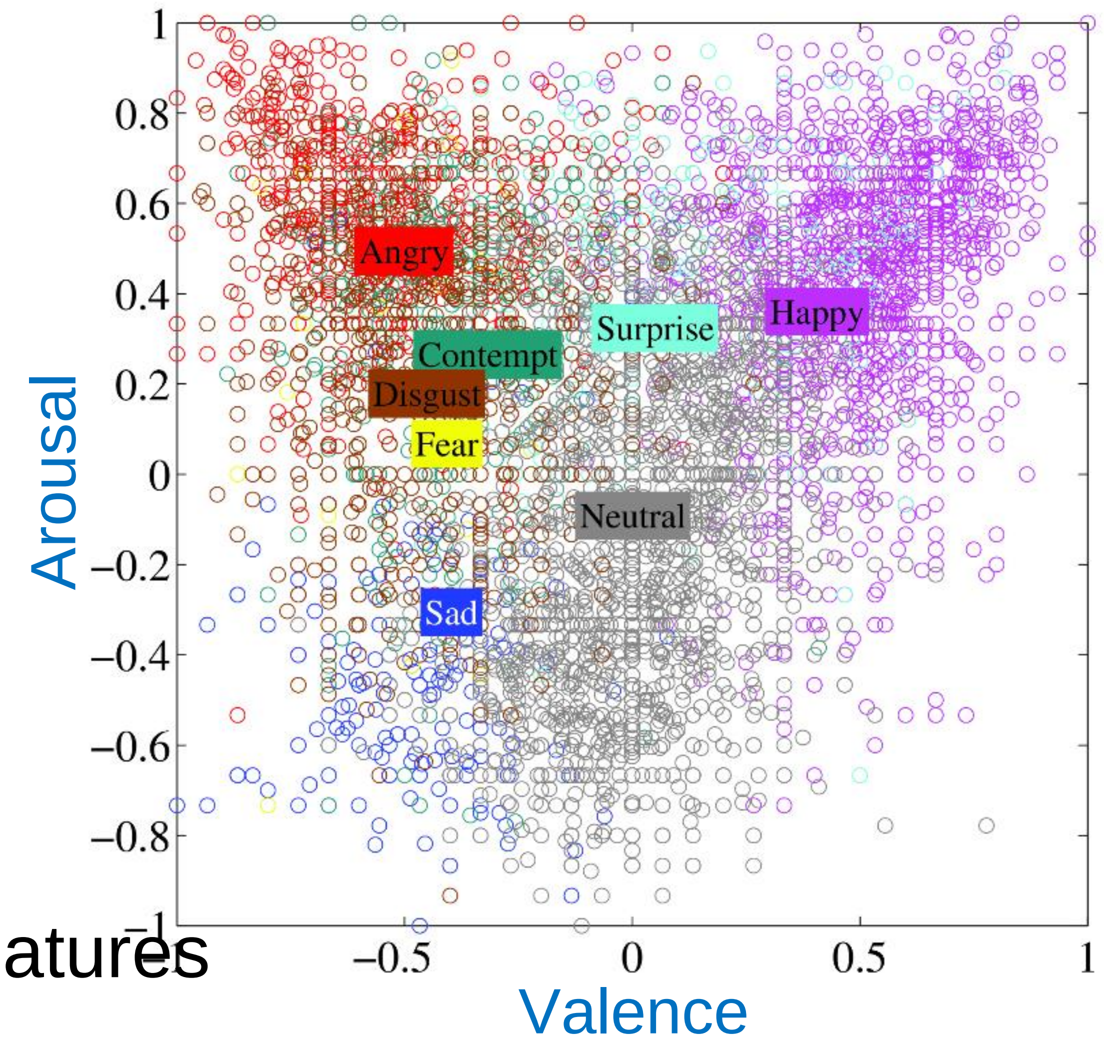
Reza Lotfian and Carlos Busso

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

October. 25th, 2017

Perception of Categorical Emotions

- Recognizing categorical emotions
  - Happiness, sadness or anger
- Typically one-hot classification problem
- Assumptions
  - Each sample -> one class
  - Same class samples share similar features



- Expressive behaviors tend to be ambiguous with blended emotions
  - Design the machine learning framework to captures the intrinsic ambiguity of emotional perception

Emotional Annotation Process

- Spontaneous corpora
  - Emotions are not predetermined during recording
  - Need to be emotionally annotated
- Emotional labels often come from perceptual evaluations from multiple evaluators
  - Compensate for outlier and individual variations
- Aggregating annotators' votes (consensus label)
  - Majority vote

- Hap
- Hap
- Sur
- Ang
- Neu

Noise

Hap

Emotion Annotation Process

- Evaluators disagree on the perceived emotion
  - Noise or information?
- Assigning a single emotion per sentence oversimplifies the subjectivity in emotion perception
- Goal: leverage information provided by multiples evaluators
  - Training emotion recognition with soft-labels
  - Soft-labels i.e., weighted label

Training with Soft Labels

- Straightforward approach
  - Use distribution of emotions assigned by evaluators [Fayek et al., 2016]

Sentence 1: happiness, happiness, neutral, happiness

$$\begin{bmatrix} neu \\ hap \end{bmatrix} = \begin{bmatrix} 0.25 \\ 0.75 \end{bmatrix}$$

Sentence 2: neutral, happiness, neutral, neutral

$$\begin{bmatrix} neu \\ hap \end{bmatrix} = \begin{bmatrix} 0.75 \\ 0.25 \end{bmatrix}$$

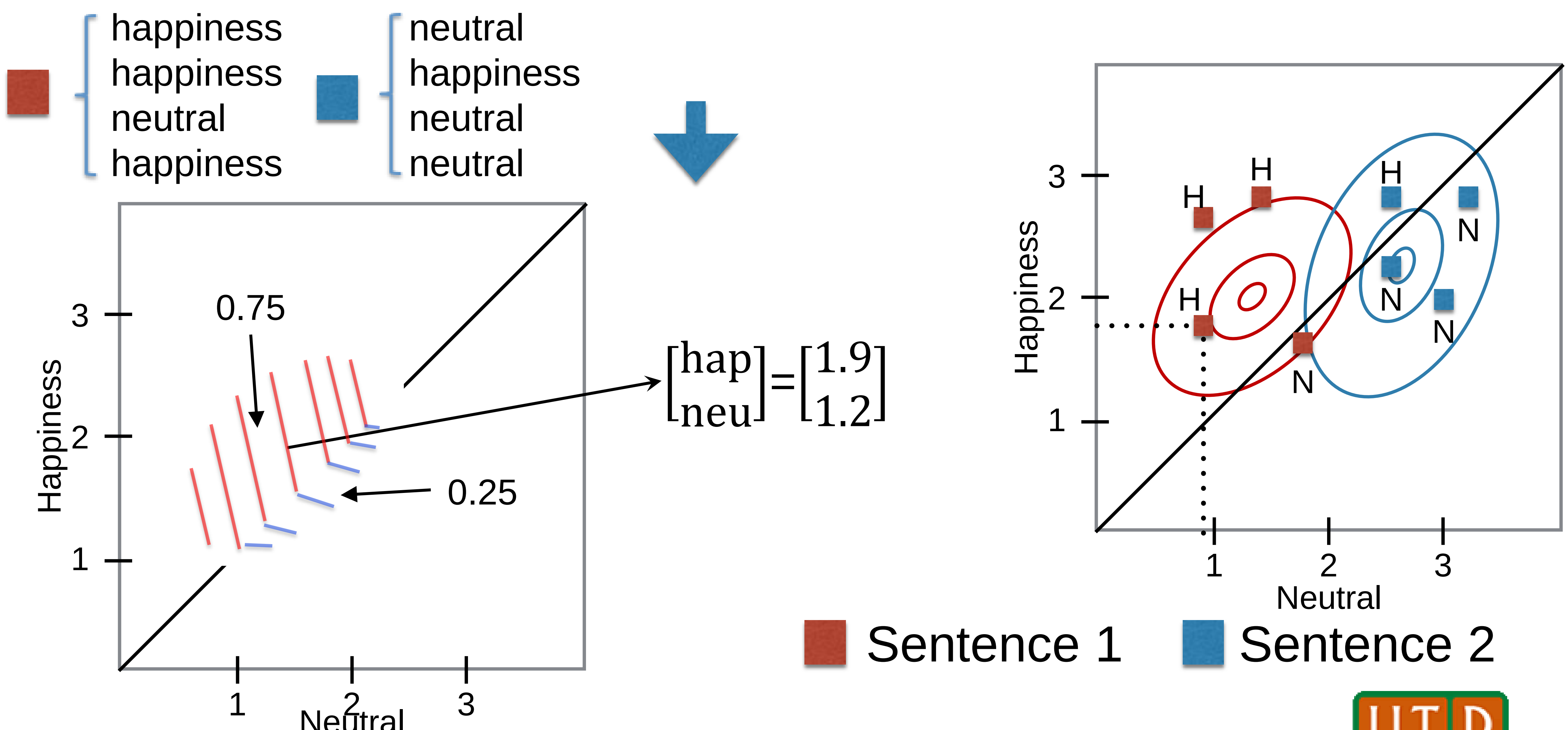
- This approach ignores relationship between emotional classes (orthogonal axes)
  - Anger → Disgust : low cost mistake
  - Anger → Happy: high cost mistake

Emotional Annotation Process

- Annotator perspective
  - Listen to a stimulus
  - Perceive the emotional content
  - Choose label that is the most relevant to the perceived emotion
- Implication to machine learning
  - Intrinsic relationship between emotional classes
  - Crucial when many choices present
  - Aware of the votes of all annotators
- Propose a method to fulfill these requirements

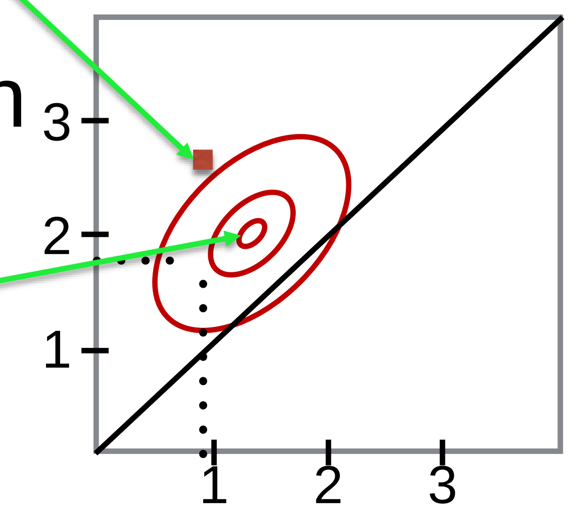
Emotion Perception as a Probabilistic Model

- 2-dimensional neutral-happiness space
- Each point: an individual evaluation (unobservable)



Theoretical Framework

- Stimuli vector (unobservable) $\boldsymbol{x} = [x_1, x_2, ..., x_D]^T$
- x realization of random vector X with distribution



$$\frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})^T]$$

- Probability of annotator selecting each class $\boldsymbol{p} = [p_1, p_2, ..., p_D]^T$

happiness
happiness
neutral
happiness

$\boldsymbol{p} = [0.75, 0.25]^T$

- An annotator selecting class j:

$$p_j = \boldsymbol{P}(X_1 \leq X_j, ..., X_{j-1} \leq X_j, X_{j+1} \leq X_j, ..., X_D \leq X_j)$$

$$x_j > x_i, \quad \forall i \neq j$$

Theoretical Framework

- Find the probability

$$p_j = \int_{-\infty}^{\infty}\int_{-\infty}^{x_j}\cdots\int_{-\infty}^{x_j} \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp \; [-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]dx_1 \underset{j}{..} dx_{j-1}dx_{j+1}...dx_D dx_j$$

$$\begin{aligned} Y_i &= X_i - X_j, \qquad \forall i \neq j \\ Y_j &= X_j \end{aligned}$$

$$\boldsymbol{Y} = \begin{array}{c} \\ \\ \\ j \\ \\ \\ \end{array} \underbrace{\begin{pmatrix} 1 & 0 & \dots & -1 & \dots & 0 \\ 0 & 1 & \dots & -1 & \dots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & -1 & \dots & 1 \end{pmatrix}}_{H} \boldsymbol{X}$$

- Find cumulative density function of $\mathcal{N}(\boldsymbol{y}; \boldsymbol{H}_j\boldsymbol{\mu}, \boldsymbol{H}_j\boldsymbol{\Sigma}\boldsymbol{H}_j^T)$

$$p_j = \boldsymbol{P}(X_1 \leq X_j, ..., X_{j-1} \leq X_j, X_{j+1} \leq X_j, ..., X_D \leq X_j)$$

$$\Downarrow$$

$$p_j = P(Y_1 \leq 0, \dots, Y_{j-1} \leq 0, Y_j < \infty, Y_{j+1} \leq 0, \dots, Y_D \leq 0)$$

Theoretical Framework

- Knowing the probabilities p_j, estimate μ and $\sum$

$$\forall j \quad p_j > 0; p_j = \boldsymbol{P}(X_1 \le X_j, ..., X_{j-1} \le X_j, X_{j+1} \le X_j, ..., X_D \le X_j)$$

- Adding a constant to all X does not change
- Extra constraint:
  - Intensity of *neutral* is reference: $\mu_{Neutral} = 1$
- $\sum$: covariance matrix is universal (*i.e.,* fixed for all speech samples) :
  - Capture dependencies between emotional categories

Estimating Covariance Matrix

- Use p instead of x:
  - Multiply by a constant to make $p_{Neutral} = 1$
  - Make zero mean $\hat{p}$

$$\widetilde{\boldsymbol{\Sigma}} = \left[\hat{\boldsymbol{p}}_{(1)}, \quad \hat{\boldsymbol{p}}_{(2)}, \quad \ldots, \quad \hat{\boldsymbol{p}}_{(n)}\right] \begin{bmatrix} \hat{\boldsymbol{p}}_{(1)}^T \\ \hat{\boldsymbol{p}}_{(2)}^T \\ \vdots \\ \hat{\boldsymbol{p}}_{(n)}^T \end{bmatrix}$$

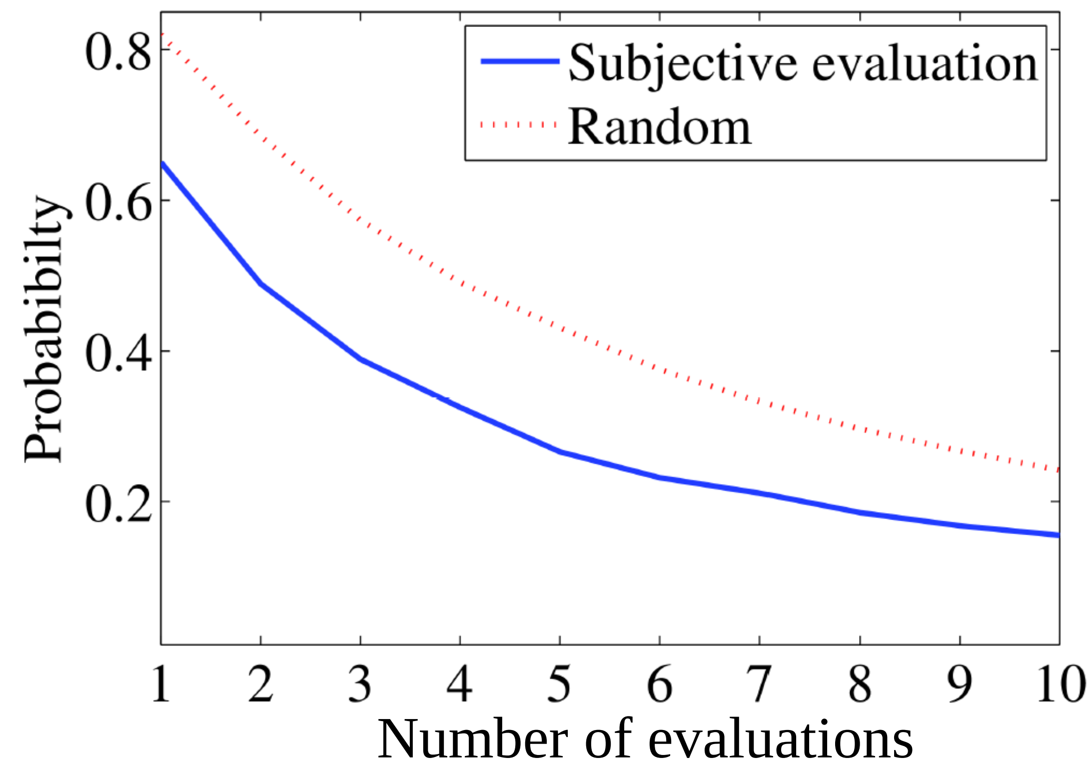
	ANG	SAD	HAP	SUR	DIS	CON	NEU
ANG	0.24	-0.02	-0.11	-0.02	0.04	0.03	-0.15
SAD	-0.02	0.13	-0.06	-0.01	-0.01	-0.01	-0.01
HAP	-0.11	-0.06	0.68	-0.02	-0.10	-0.11	-0.25
SUR	-0.02	-0.01	-0.02	0.16	-0.01	-0.02	-0.09
DIS	0.04	-0.01	-0.10	-0.01	0.18	0.03	-0.12
CON	0.03	-0.01	-0.11	-0.02	0.03	0.20	-0.10
NEU	-0.15	-0.00	-0.25	-0.09	-0.12	-0.10	0.79

Estimating Mean (Intensity of Emotions)

- Problem with $p_j = 0$:
  - No annotator select a category
  - Infeasible equality

$$\boldsymbol{P}(X_1 \leq X_j, ..., X_{j-1} \leq X_j, X_{j+1} \leq X_j, ..., X_D \leq X_j) = 0$$

- If one more label was available, what is the probability to capture a new label
  - Depends on number of evaluations
  - Leave one annotator out
- Scale probability of seen labels accordingly: $1 - \lambda(n)$



Estimating Mean (Intensity of Emotions)

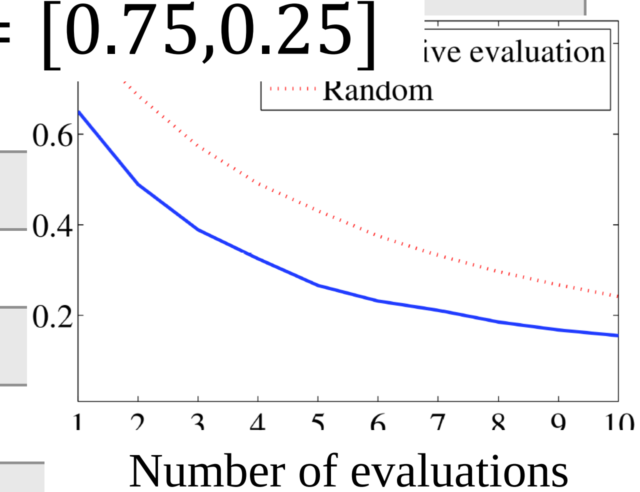
- For each sentence
- Initial expected values (from training set)

	ANG	SAD	HAP	SUR	DIS	CON	NEU
p	0.08	0.05	0.20	0.07	0.08	0.11	0.33
$\mu_{Initial}$	-0.07	0.30	-0.75	0.26	0.39	0.41	1.00

Input:
p: Probability of classes
$\tilde{\Sigma}$: Covariance matrix
k: Number of iterations
λ (n): Probability of unseen labels
$\mu_{Initial}$: initial intensity vector
Output: Estimated emotion intensity $\bar{\mu}$
$\hat{p} \leftarrow p/(1-\lambda(n))$
$\tilde{\mu} \leftarrow \mu_{Initial}$
for i <= k do
for j where $p_j > 0$ do
find $\bar{\mu}_j$ such that $F(0,0,\dots,\infty,\dots,0) = \tilde{p}_j$ for N (x; $H_j[\tilde{\mu}_0,\dots,\tilde{\mu}_{j-1},\bar{\mu}_j,\dots,\tilde{\mu}_D]^T, H_j \tilde{\Sigma} H_j^T$)
$\tilde{\mu}_j \leftarrow \bar{\mu}_j$
for j = 1 : D do
$\tilde{\mu}_j \leftarrow \tilde{\mu}_j + \tilde{\mu}_{Neutral} + 1$

happiness
happiness
neutral
happiness

$\boldsymbol{p} = [0.75, 0.25]$



Loss Function

- Measure the disagreement between the ground truth and predicted labels
  - Previously, categorical cross-entropy as the loss function for hard (one-hot) label and soft-labels
- Mahalanobis distance reflects a more meaningful measure for disagreement cost
  - Intensity value predicted by the network: θ

$$\mathcal{L}(\boldsymbol{\theta}, \tilde{\boldsymbol{\mu}}) = 1 - \exp[-\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\mu}})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \tilde{\boldsymbol{\mu}})^T]$$

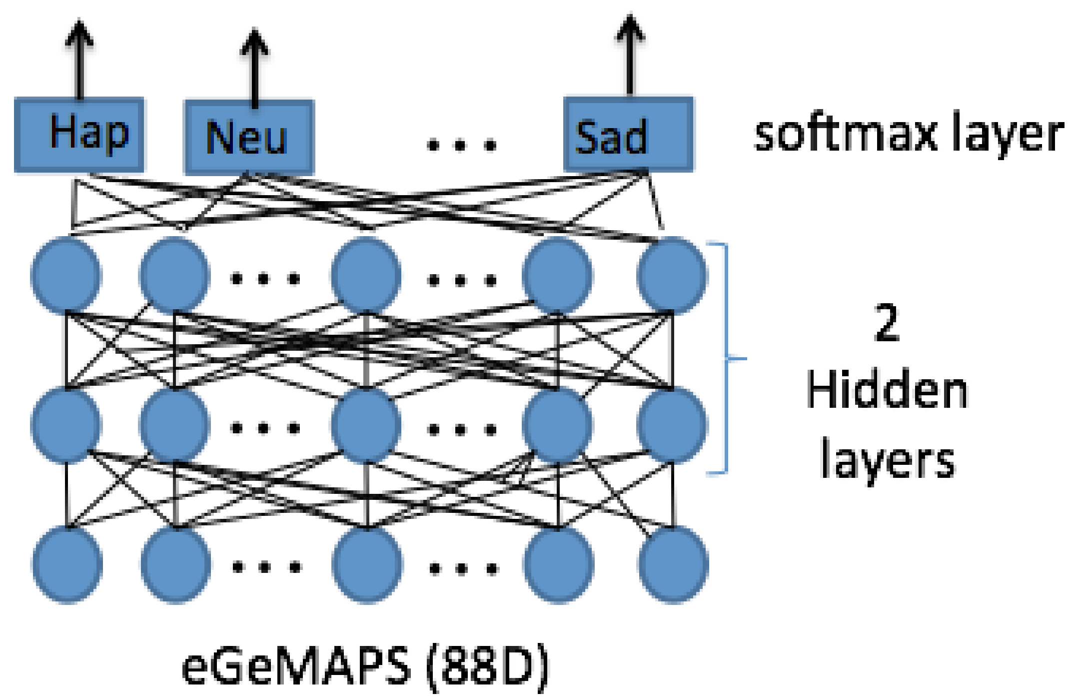
- Anger →Happy greater penalty Anger → Disgust

Experimental Setup

- Database: MSP-PODCAST (University of Texas at Dallas)
  - Speech segments from podcast recordings
  - One speaker, no background music, no telephone quality
  - Duration 2.75s< … <11s
  - Test: 4,283 Train: 7,289 Development: 1,860
  - Total 13,432 (21h 15min)
  - Evaluated through Amazon Mechanical Turk (at least 5 evaluations per sentence)
- 88 features: eGeMAPS [Eyben et al., 2016]

Classifier Configuration

- Seven-class problem: anger, sadness, happiness, surprised, disgust, contempt, and neutral (chances is 14%)
- Feed forward DNN with 2 hidden layers
- Each hidden layer 512 rectified linear unit (ReLU)
- Output softmax with one output per emotion
- Loss functions
  - Mahalanobis distance
  - Cross-entropy
- Trained 50 epoch



Experimental Evaluations

- Ground truth labels for test set from majority vote
- Predicted class: dimension with highest *expected intensity estimation* (SL-EIE)
- Classification Performance

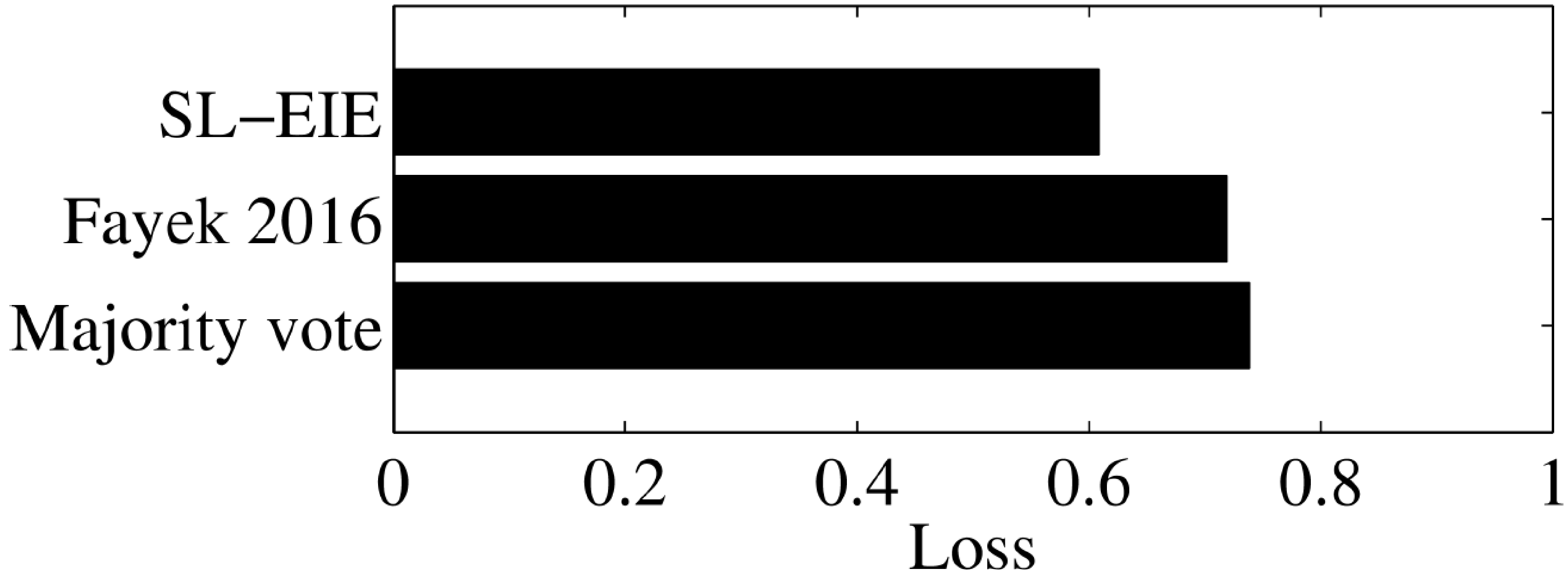
	Rec [%]	Pre [%]	F1-Score
Majority vote	25.7	24.2	**24.9**
Soft-label [Fayek, 2016]	27.2	23.7	**25.3**
SL-EIE [proposed]	28.1	24.5	**26.2**

- Human performance: Reference of difficulty
  - One annotator compared to the consensus label of the rest

	Rec [%]	Pre [%]	F1-Score
Human Performance	38.2	41.1	**39.6**

Experimental Evaluations

- The error between estimated labels and ground-truth based on the proposed loss function



- Better measure of performance
- Penalty considers relationship between emotions

Conclusions

- Framework to address the problem of classifying categorical emotions in spontaneous speech
- Soft-labels inspired by the emotion perception
- Non-observable multivariate Gaussian distribution
  - Dimensions correspond to the emotional categories
- Evaluations are points drawn from the distribution
- Selected category is the emotions with the highest intensity

Conclusions

- Benefit of using this representation for training emotional classifiers
- Future directions
  - Estimate covariance matrix for each sample
  - Probability of a new label depends on other parameter, not only number of evaluations
  - Better model considering the bias and reliability of individual evaluators

Thanks for your attention!

http://msp.utdallas.edu/