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

Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels

Reza Lotfian and Carlos Busso

ICASSP 2020
May 2020

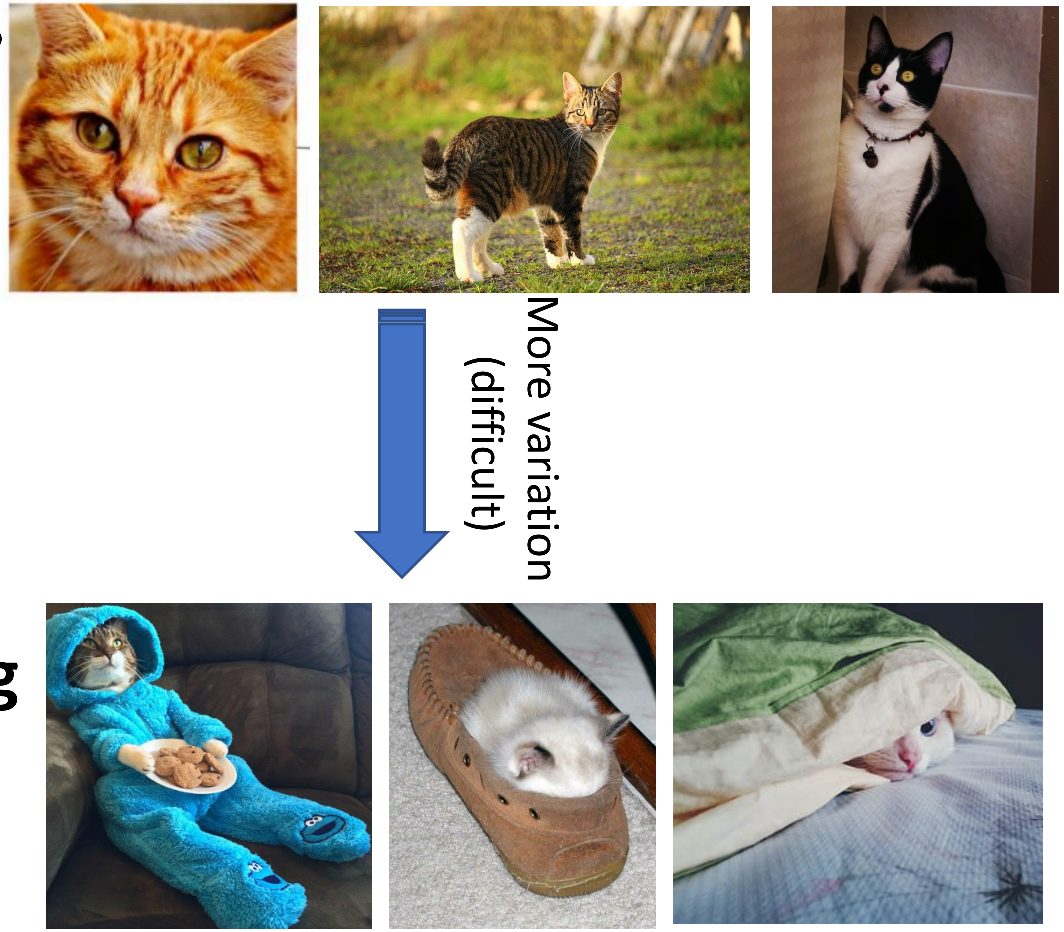
UTD CRSS

THE UNIVERSITY OF TEXAS AT DALLAS

UT Dallas MSP Multimodal Signal Processing Laboratory

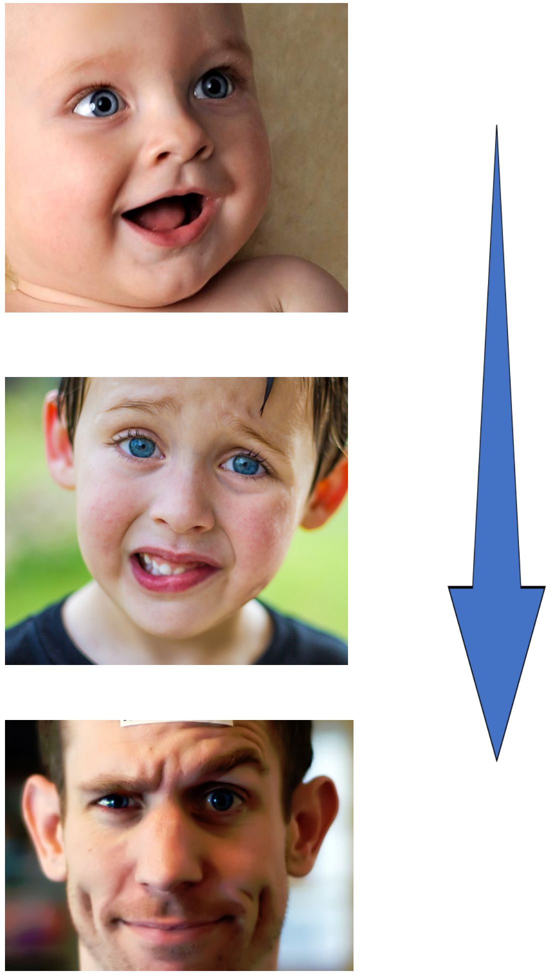
Curriculum Learning

- **One-pass learning: train using all samples together**
- **Curriculum learning: sequentially present training data from simple to complex**
- **Designing curriculum based on difficulty**
  - First, presenting simple to recognize samples
  - Later, increasing the difficulty during training
- **Lead to better local minima when training a classifier with a non-convex criterion**
  - Better generalization
  - Speed-up the convergence



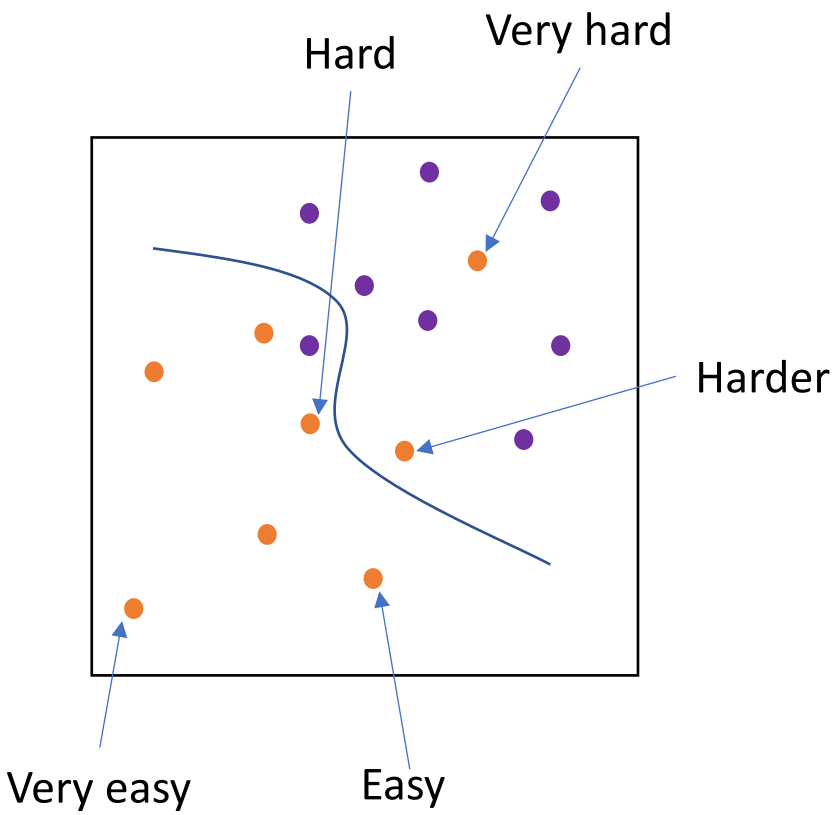
Curriculum Learning on Emotion Recognition

- **Why learning emotion with curriculum?**
- **Emotion recognition: complex problem, takes years to master its essential skills**
  - Infants start with limited capabilities
  - Over time, they develop more sophisticated emotional representations
- **Step-by-step process of acquiring the capability to perceive emotions**
  - Curriculum learning can benefit machines to learn emotions



Curriculum Policies

- **Natural policies**
  - e.g., Natural language processing: complex sentences with relative clauses, several phrases
- **When natural policy is not available:**
  - Error of predicted label:
    - Train a classifier using all training samples
    - Test it on the same set
    - Repeat the training starting with easy examples
- **Proposed method: Use human judgment to find curriculum policy**
  - Assumption: Hard sentences for human are also hard for computers [Busso et al. 2017]



Minmax Method for Crowdsourced Labels

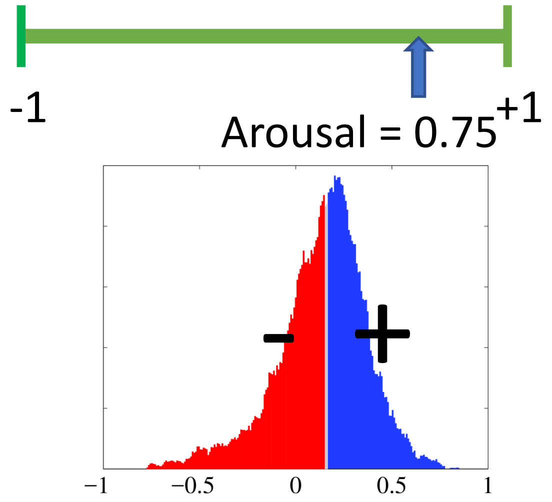
- **Crowdsourcing annotation**
  - Find the consensus label
- **Disagreements in the labels**
  - Annotators make more mistakes on difficult tasks
  - Low-skill or inattentive annotators make mistake too
- **Conditional minmax entropy method** [Zhou 2014, Zhou 2015]
  - Jointly learn label, worker ability, and **item difficulty**
  - Input: observed labels x_{ij}
  - Item difficulty output for item j: matrix $[\tau_j]$
  - $\tau_j(c,k)$; How likely class c is mistaken with class k for item j
  - Difficulty measure

$$d_j = \frac{\sum_k \tau_j(k,k)}{\sum_c \sum_k \tau_j(c,k)}$$

	item 1	item 2	…	item n
worker 1	x_{11}	x_{12}	…	x_{1n}
worker 2	x_{21}	x_{22}	…	x_{2n}
…	…	…	…	…
worker m	x_{m1}	x_{m2}	…	x_{mn}
	y_1	y_2	…	y_n

Experiments

- **Machine learning problems for emotion detection**
  - Regression of dimensional emotions
    - Predicting the attribute levels
  - Binary classification of dimensional emotions
    - Predicting high versus low class for attribute
  - Classification of categorical emotions
    - Predicting the most relevant category of emotion

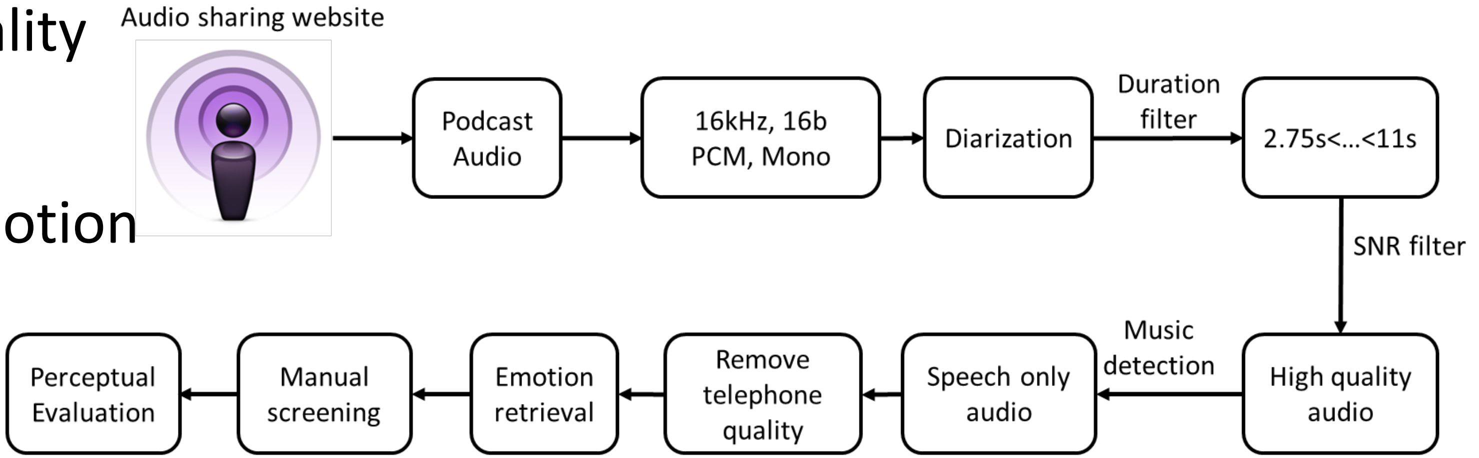


- [ ] Happy
- [x] Angry
- [ ] Sad
- [ ] Neutral

- **Features**
  - Utterance level features 6,373 (IS2013 ComParE set)
- **MSP-Podcast corpus**

MSP-Podcast Corpus

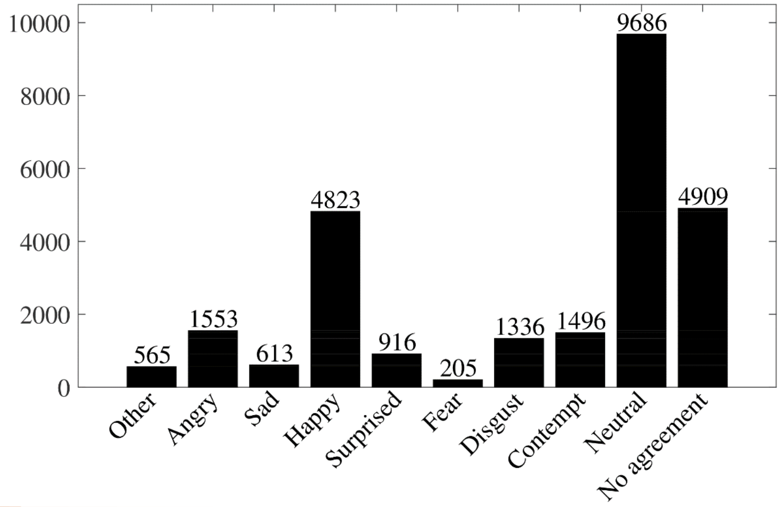
- Collecting audio recordings (Podcasts)
  - Natural, Creative Commons copyright license, diverse
- Automatic speaker diarization
  - Single speaker segments
- Low noise, remove telephone quality
- No background music
- Retrieve samples with desired emotion
- Manual screening
- Perceptual evaluation
  - Crowdsourced based method

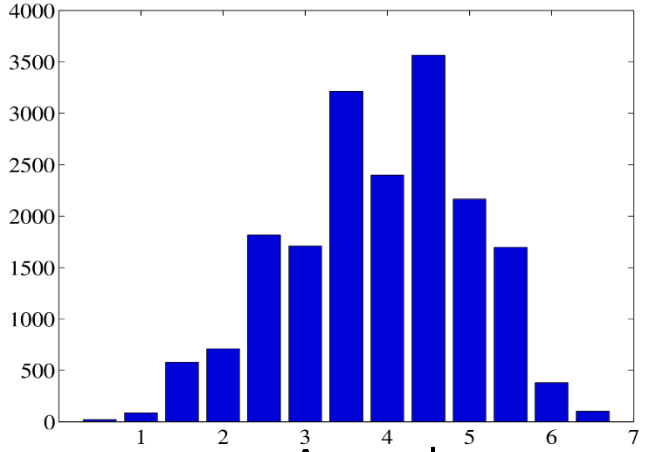


Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing, vol. 10, no. 4, pp. 471-483, October-December 2019.

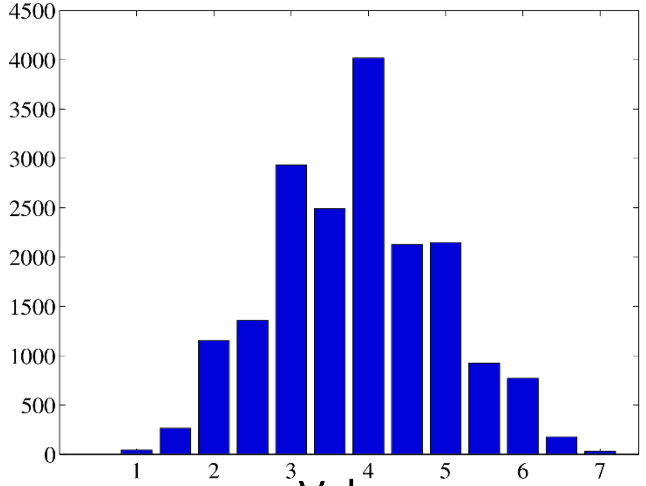
MSP-Podcast Corpus

- **Release 1.0**
  - Total number of samples: 20,045
    - Test set: 6,069 segments (50 speakers)
    - Development set: 2,226 segments (15 speakers)
    - Train set: 11,750 segments (rest of speakers)
  - Total time: 34 hours, 15 minutes
  - Total number of un-labelled samples: 541,975





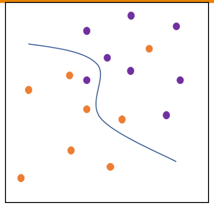
Arousal



Valence

Building Curriculum

- **Method 1**: Error of predicted label
  - Regression problem
  - Binary and multi-class classification



$$d_i = |y_i - y_i'|$$

$$d_i = \begin{cases} P(y_i = {y'}_i | x_i), \text{if } y_i = y_i' \\ -P(y_i = {y'}_i | x_i), \text{if } y_i \neq y_i' \end{cases}$$

- **Method 2**: Disagreement between annotators without considering level of expertise
  - Regression problem: variance of annotations
  - Binary and multi-class classification

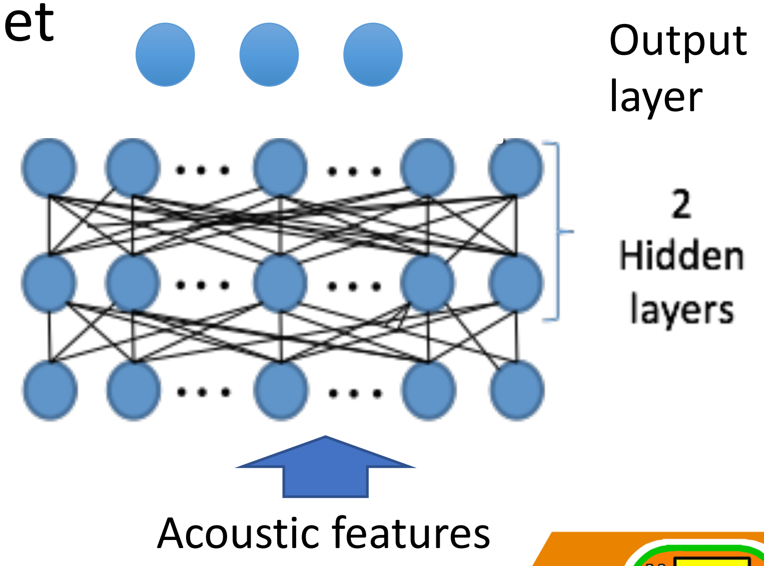
$$d_j = \frac{\sum_{k=mv_j} (k)}{\sum_k (k)}$$

- **Method 3**: Disagreement between annotators by considering level of expertise
  - Minmax conditional entropy inference

$$d_j = \frac{\sum_k \tau_j(k,k)}{\sum_c \sum_k \tau_j(c,k)}$$

Classifiers

- **Deep Neural Network**
  - Fully connected feed forward neural network with two hidden layers
  - Hidden layer 1,024 nodes with rectifies linear unit (ReLU)
  - Keras with TensorFlow as backend
  - Optimization Adaptive moment estimation (ADAM)
  - Learning rate for each step was found using validation set
  - 50 epochs each step
- **Cost function:**
  - Regression: Mean square error
  - Classification: Cross-entropy



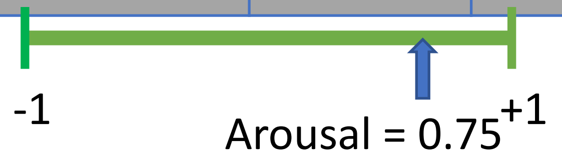
Results (Regression of Emotional Attributes)

- **Concordance correlation coefficient (CCC)**

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

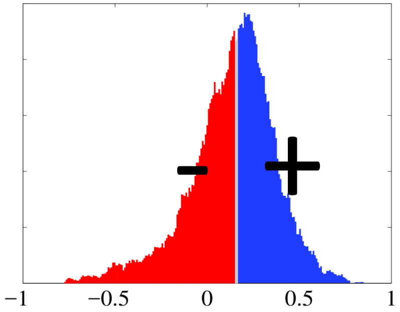
- **Observations**
  - Minmax difficulty curriculum learning leads to the highest improvement in CCC
  - Statistical significance test: one-tailed z-test on difference in population proportions (p-value = 0.05)

	Aro. [CCC]	**Val. [CCC]**	**Dom. [CCC]**
w/o curriculum	0.724	0.298	0.690
With random curriculum	0.729	0.293	0.686
Method 1: Error of predicted label	0.725	0.313	0.694
Method 2: Disagreement between annotators	0.730*	0.320*	0.696
Method 3: Minmax entropy	**0.745***	**0.325***	**0.705***



Results (Binary Classification of Emotional Attributes)

- **Binary problem (high versus low)**
  - Arousal, valence and dominance
  - F-score
- **Observations**
  - Using curriculum increases the performance
  - Best curriculum: Method 3 (minmax entropy curriculum)
  - Statistical significance test: one-tailed z-test on difference in population proportions (p-value = 0.05)



	Aro. F1-score	Val. F1-score	Dom. F1-score
w/o curriculum	0.778	0.592	0.685
With random curriculum	0 .771	0.591	0.685
Method 1: Error of predicted label	0.785	0.606*	0.684
Method 2: Disagreement between annotators	0.789*	0.616*	0.695*
Method 3: Minmax entropy	**0.791***	**0.616***	**0.696***

Results (Multi-class Categorical Emotion Classification)

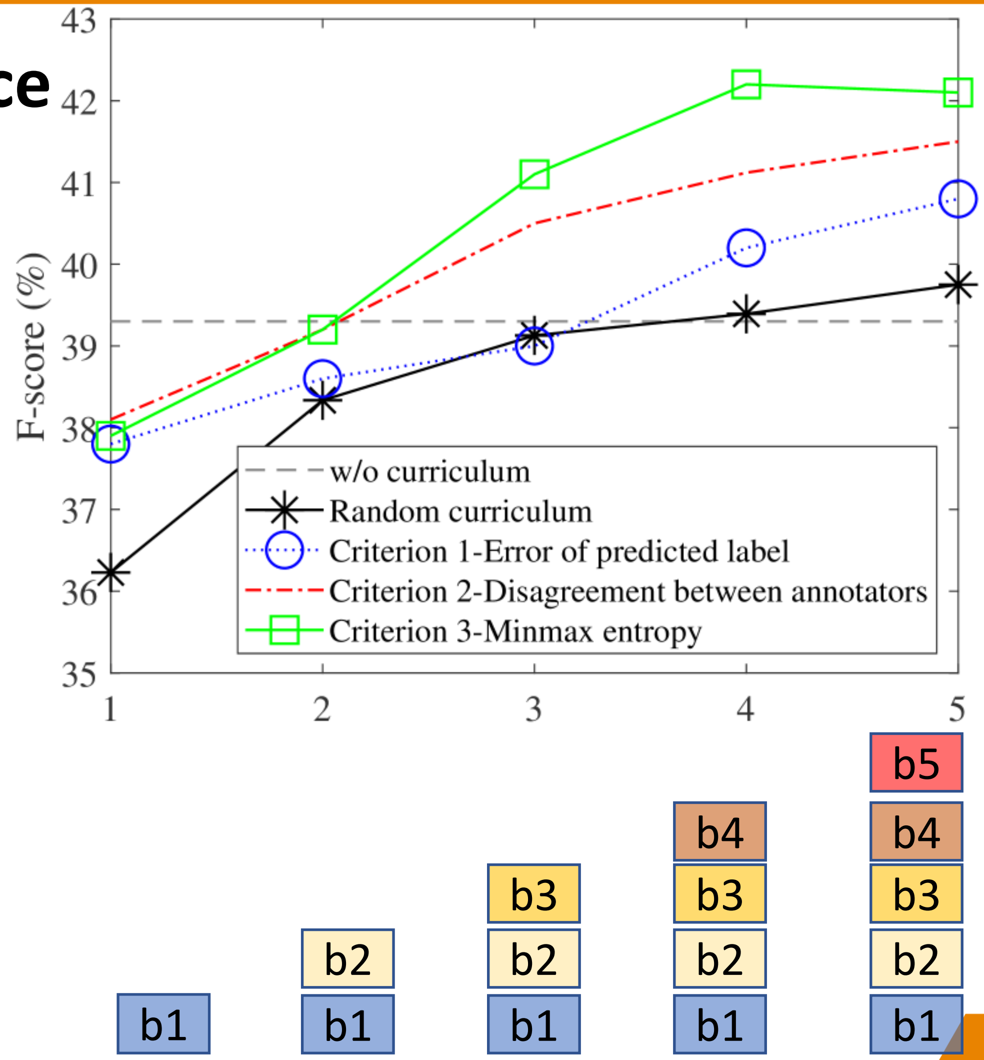
- **5-class classification**
- **Observations**
  - Using curriculum increases the performance
  - Best curriculum: Method 3 (minmax entropy curriculum)
  - 2.4% increase in F-score
  - Statistical significance test: one-tailed z-test on difference in population proportions (p-value = 0.05)

	F-score[%]
w/o curriculum	39.7
With random curriculum	39.8
Method 1: Error of predicted label	40.8
Method 2: Disagreement between annotators	41.5*
Method 3: Minmax entropy	**42.1***

- ❑ Happy
- ☑ Angry
- ❑ Sad
- ❑ Neutral

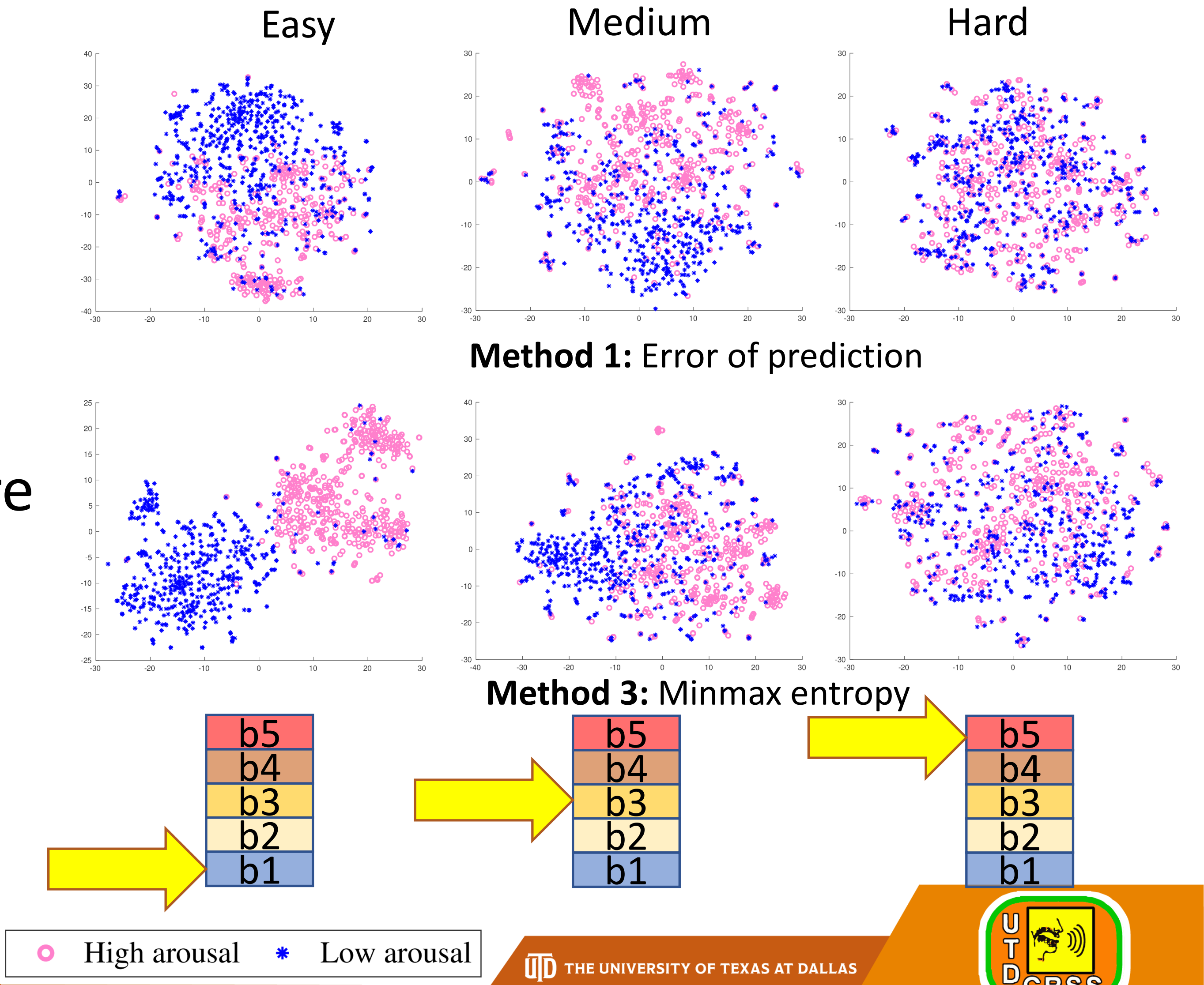
Results (Multi-class Categorical Emotion Classification)

- F-score improvement as we introduce more training samples
- 5-class classification problem
- Without curriculum: One-pass
- Randomly selected bins



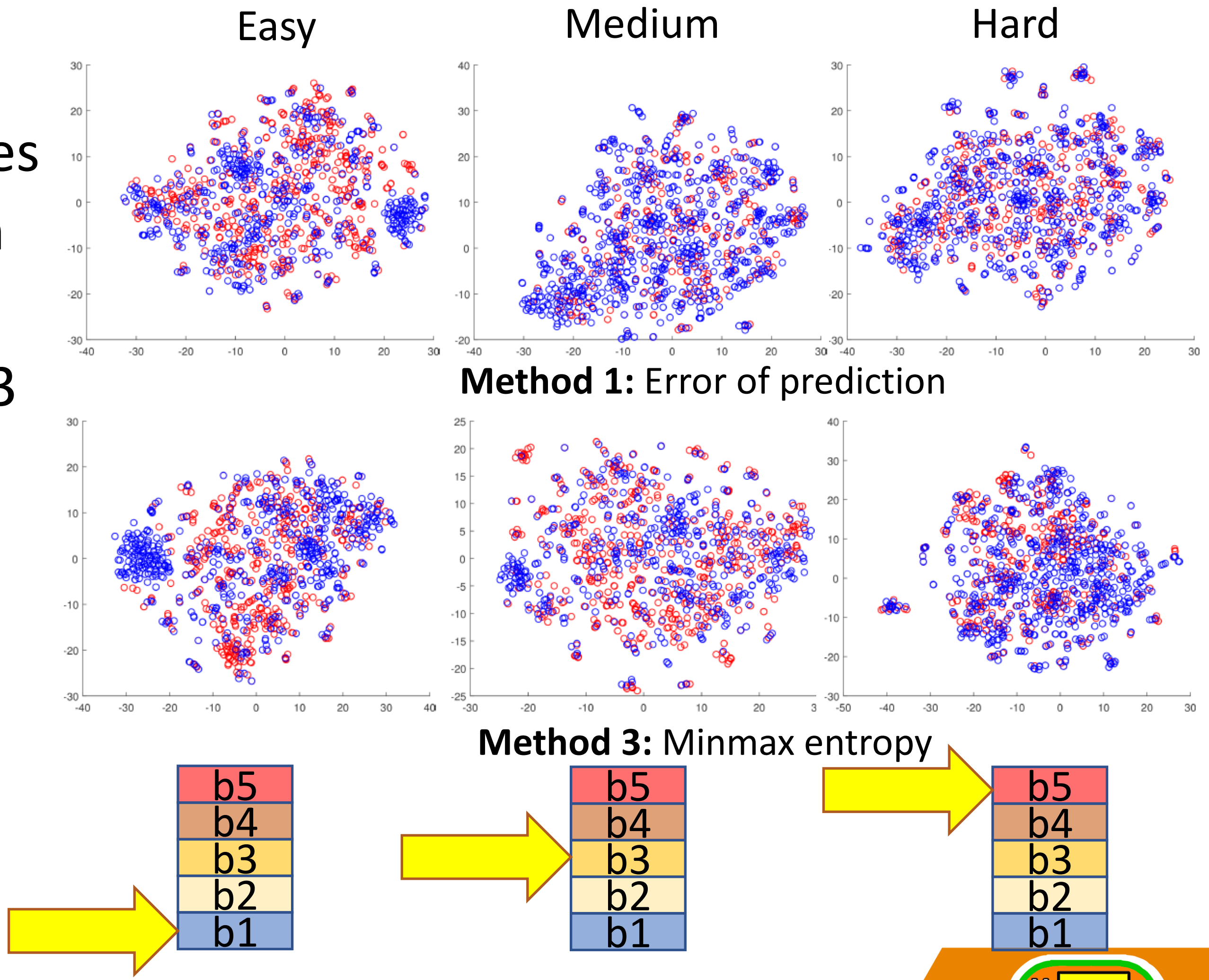
Analysis of Features (Arousal Binary Classification)

- **Is human perception of difficulty reflected on feature domain?**
- Applied to classification problems
- t-SNE: visualize high dimensional data
- Arousal: More separation in feature domain for easier samples
  - Bin1: easiest
  - Bin5: hardest



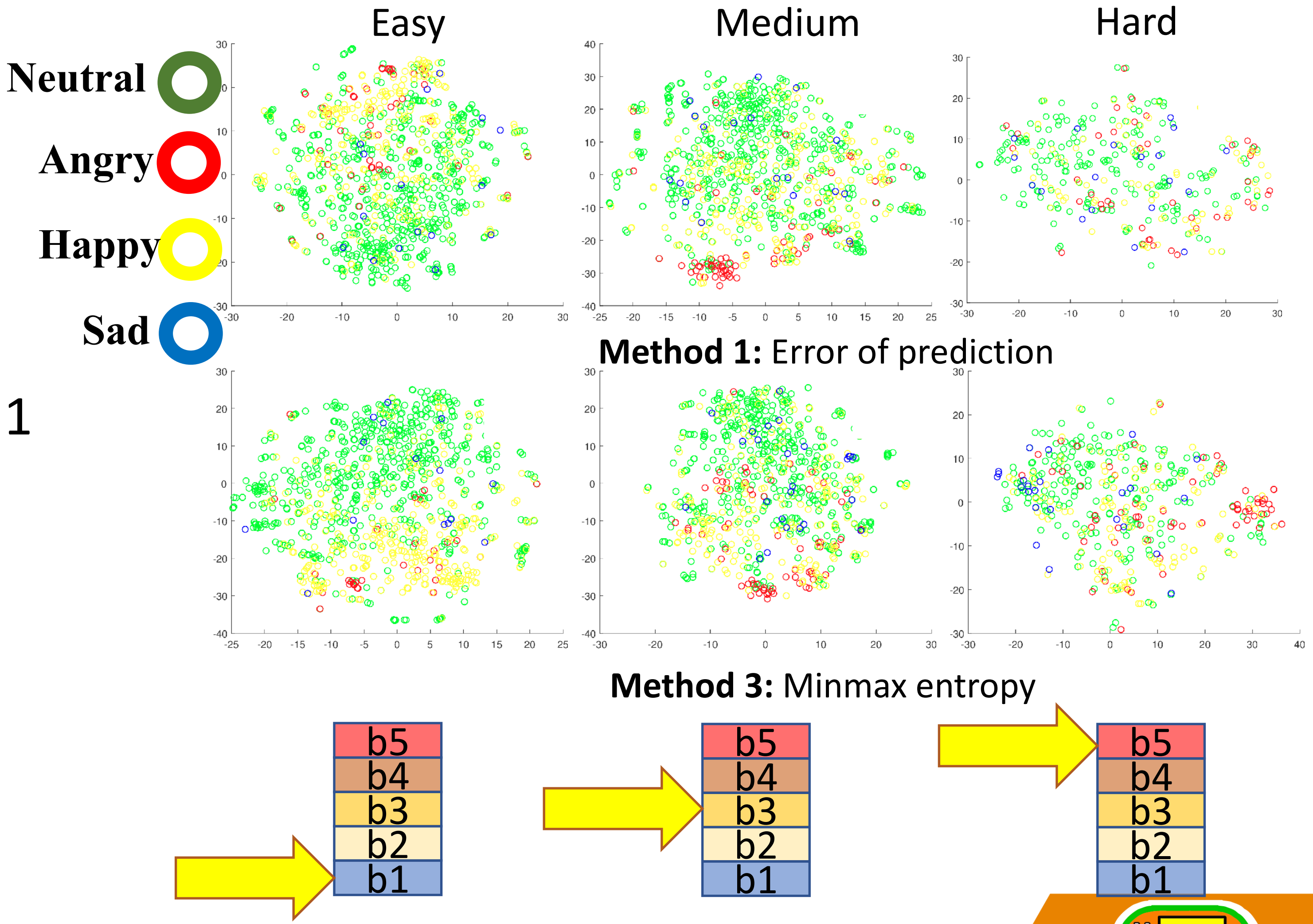
Analysis of Features (Valence Binary Classification)

- **Valence:**
  - Hardest problem from acoustic features
  - Human relies on semantic information
  - Bin 1 shows some separation
  - No separation between classes in bin 3 and bin 5



Analysis of Features (Categorical Emotions)

- **Categorical emotions:**
  - Only show 4 classes for better visualization
  - More visible in bin 1
  - Minmax method even better than Error of prediction on bin 1



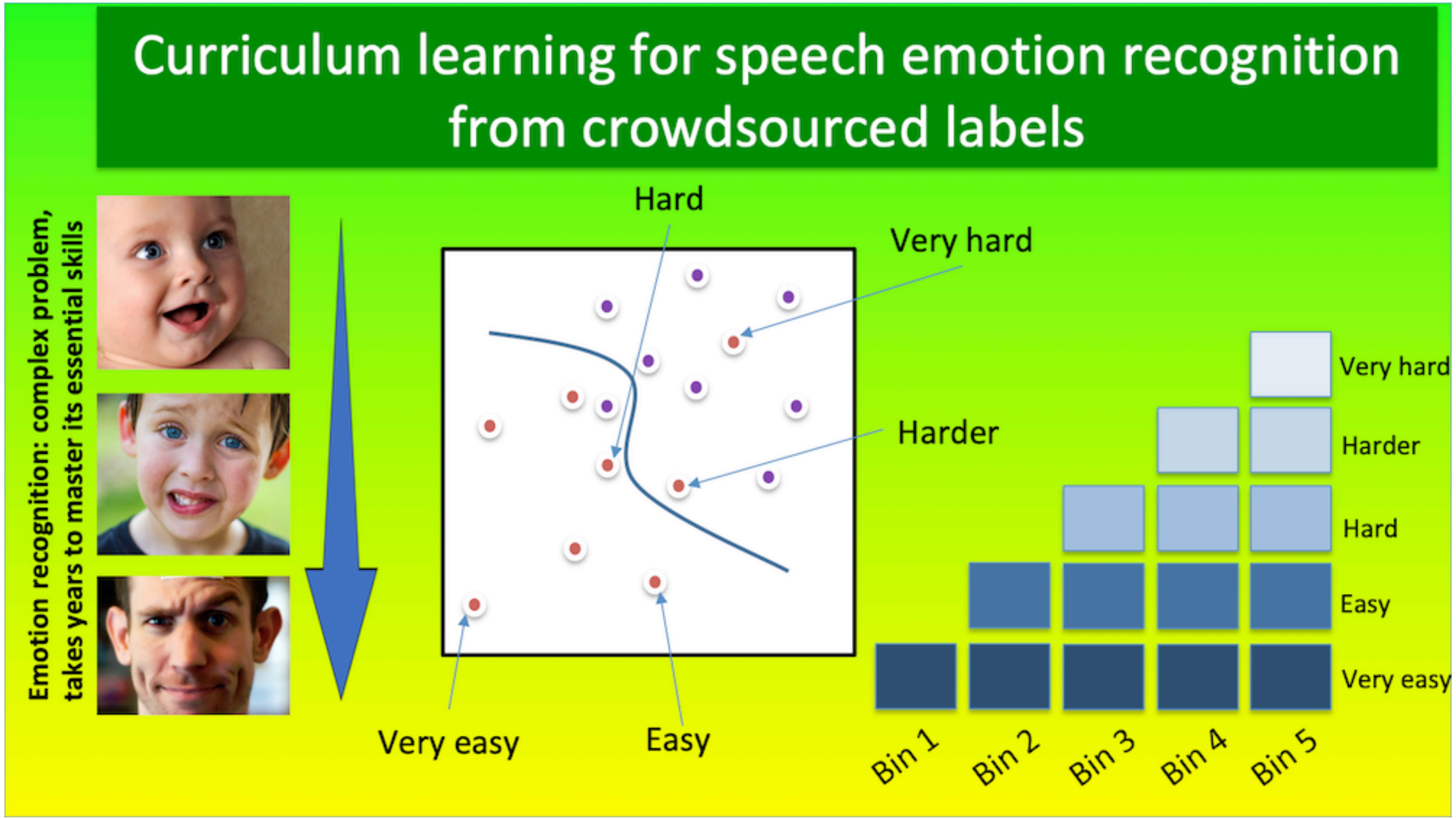
Conclusions

- **Curriculum learning for speech emotion recognition**
  - No implicit way to determine difficulty
- **Quantify the difficulty level by:**
  - Error of predicted label by a pre-trained model
  - Disagreement among annotators
  - Minmax entropy method
- **SER benefits from curriculum learning compared to no policy or random policy**
- **Best policy curriculum learning with Minmax entropy**
  - Find difficulty as a latent variable using labels from multiple raters

Future Directions

- **Find samples not worthy of learning (removing to increase performance)**
  - Too difficult to learn
  - No reliable labels generated by annotators
- **Use the difficulty measure to find training examples that negatively affect the performance of the models**
  - Select a subset of the data for supervised adaptation of speech emotional models
- **Exploring the effectiveness of the curriculum learning as the size of the training set increases**
- **Train with reject option**

Thank you for Your Attention



If you have questions, please send it to Reza Lotfian
rlotfian@cogitocorp.com

https://:msp.utdallas.edu