INTERSPEECH 2020
OCTOBER 25-29/ SHANGHAI, CHINA
SHANGHAI INTERNATIONAL CONVENTION CENTER

Ensemble of Students Taught by Probabilistic Teachers to Improve Speech Emotion Recognition

Kusha Sridhar & Carlos Busso

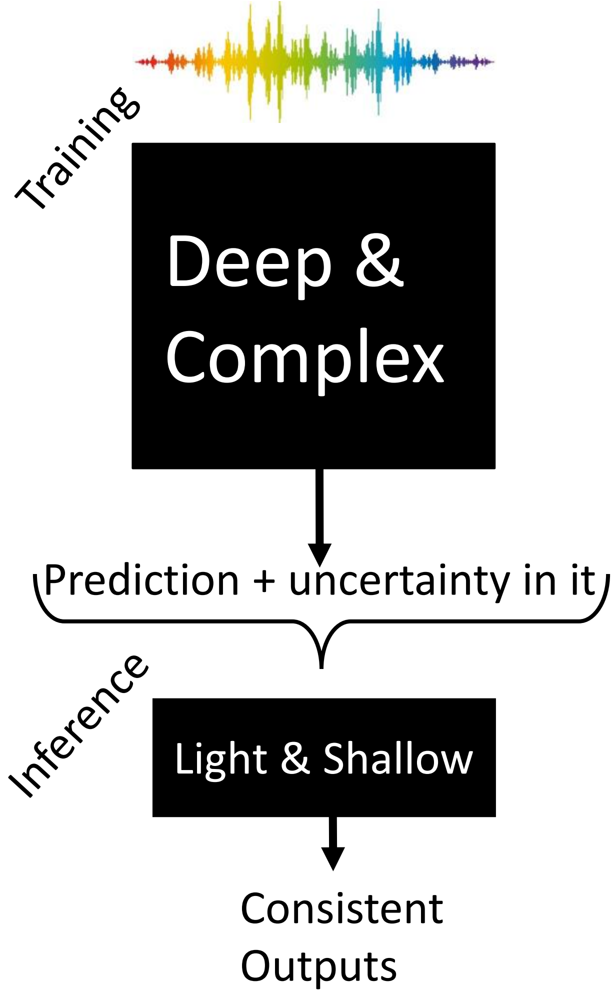
THE UNIVERSITY OF TEXAS AT DALLAS
UT Dallas MSP Multimodal Signal Processing Laboratory

Outline

1. Introduction (MC dropout → Uncertainty estimation)
2. Proposed T-S formulation for SER
3. Dataset: MSP-Podcast corpus
4. Experimental evaluations and analysis
5. Conclusion

Scalability and Consistency of SER Models

- **Application areas: Security and Defense, healthcare → mission critical**
  - SER should *generalize* well to new conditions
  - Be *scalable* and provide high test-retest *reliability*
- **Knowledge of uncertainty in model predictions**
  - It introduces diversity in model prediction
  - It creates robust models that are stable across diverse inputs
- **Knowledge transfer from deep to lighter models**
  - Flexible approach for generalization
    - Train deep, complex models on huge training data
    - Use light, shallow models at inference → **PREFERRED!**
  - Adapt to new conditions by learning from unlabeled data



Bayesian Inference with a Teacher-Student (T-S) Framework

Related Works

Speech, Language & Image tasks

Image Classification → Distilled Dropout Network (DDN) to transfer knowledge from T to S via MC samples of soft-targets generated by teacher

[Gaurau et. al. 2018]

ASR → Multi-task ensembles of T to reduce WER on telephone speech

[Wong et. al. 2017]

NLP → Multi-layer Knowledge distillation (KD) using embeddings from multiple intermediate layers of T (BERT) to train S

[Sun et. al. 2019]

Speech Emotion Recognition

Audio-visual SER with cross-modal distillation → Learn facial embeddings from T to train S on SER task. Reduction in labels noise with KD from faces to speech and robustness to ambiguous annotations

[Albaine et. al. 2018]

Preprocessing with emotion distillation to detect emotionally salient regions in audio-visual inputs

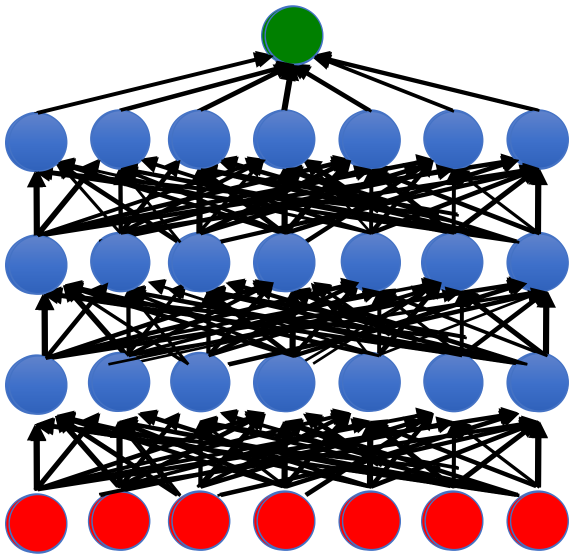
[Mower Provost et. al. 2012]

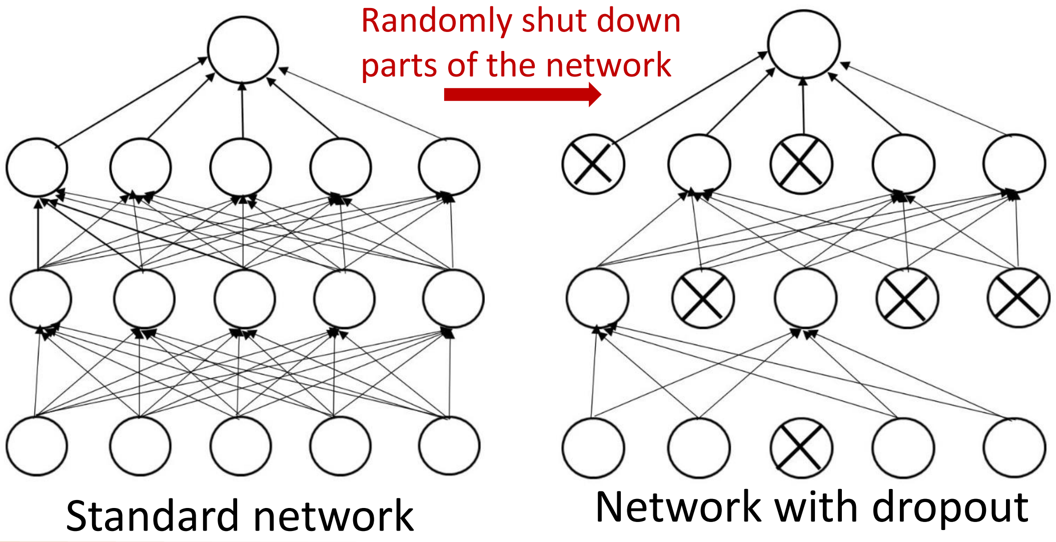
Motivation

- **Three main motivations:**
  - Transfer knowledge to a shallow, flexible model during inference
    - Leverage T-S framework in speech emotion recognition
    - Teacher is trained with deep, complex models trained on large amounts of training data
  - Create probabilistic distribution of embeddings to train student models
    - Use of an ensemble of teacher models
  - Capture model's uncertainty in its predictions
    - Use of MC dropout in T-S framework
    - Handle out-of-distribution inputs or inputs from sparse regions of the in-domain data
    - Obtain information about the reliability of the prediction

Monte Carlo Dropout

- **DNNs with dropout regularization can be used to quantify prediction uncertainty [Gal et al., 2016]**
  - Change the weights setup randomly by applying dropout
  - As such, different configurations of the network lead to slightly different prediction
  - Prediction uncertainty will be the variance of *N* step predictions
  - Multiple iterations through a network with dropout is analogous to obtaining predictions from an ensemble of thinner networks.
- **We can estimate the posterior distribution on the predictions during inferences by sampling weights in a Monte Carlo fashion**



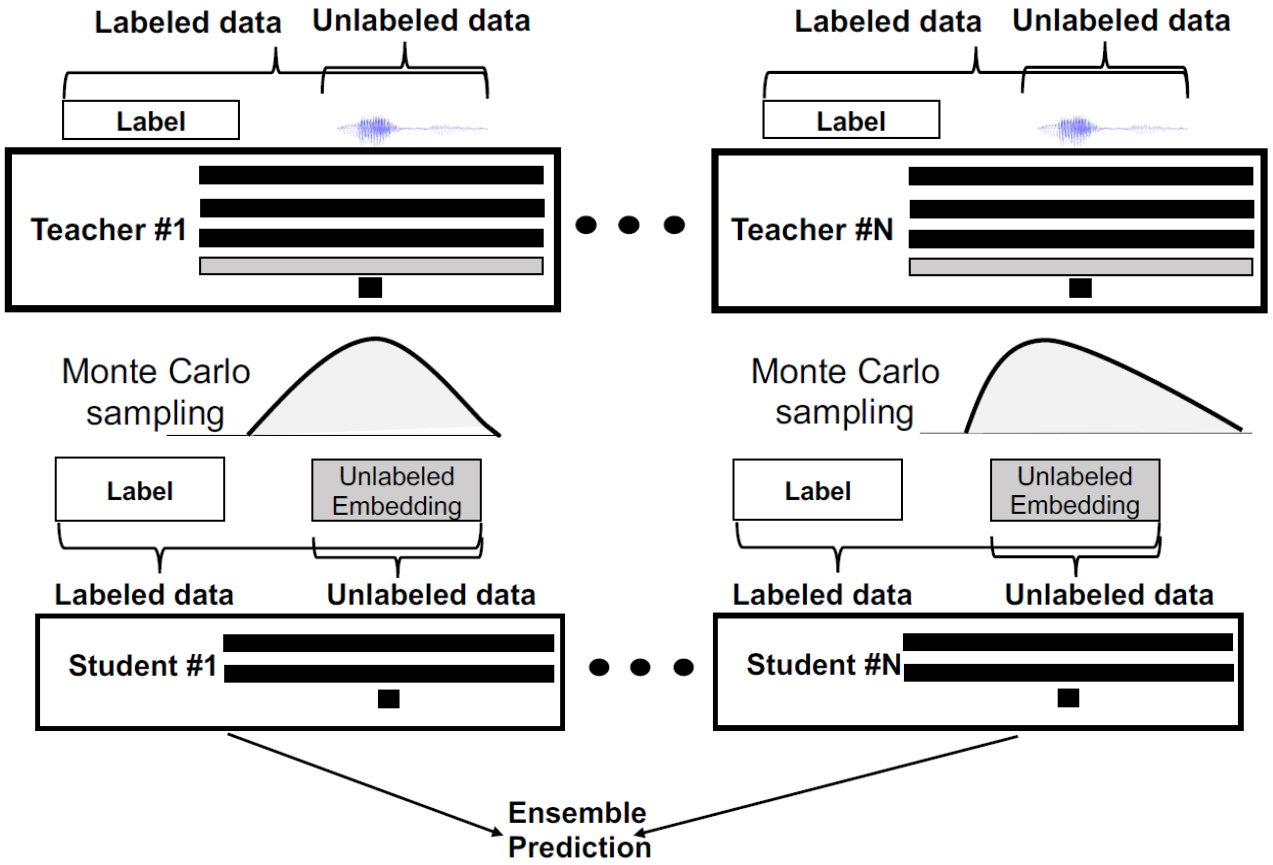


Posterior predictive distribution

$$p(x_{test}|X) \approx \int p(x_{test}|\omega)p(\omega|X)d\omega$$

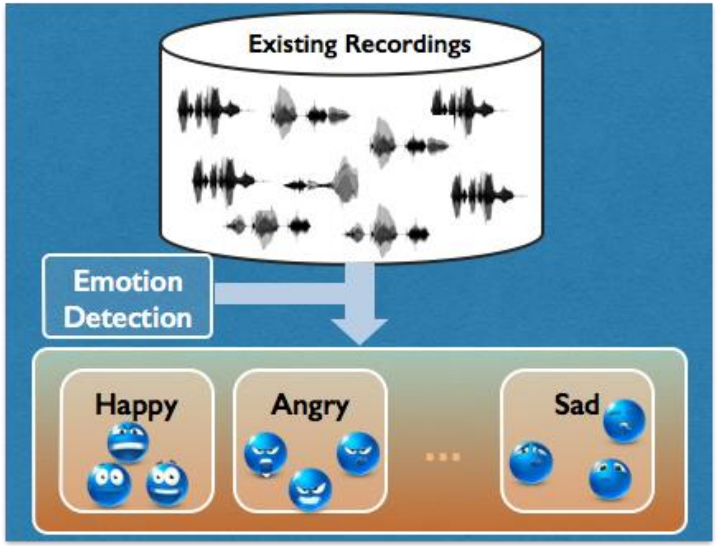
Teachers and Students

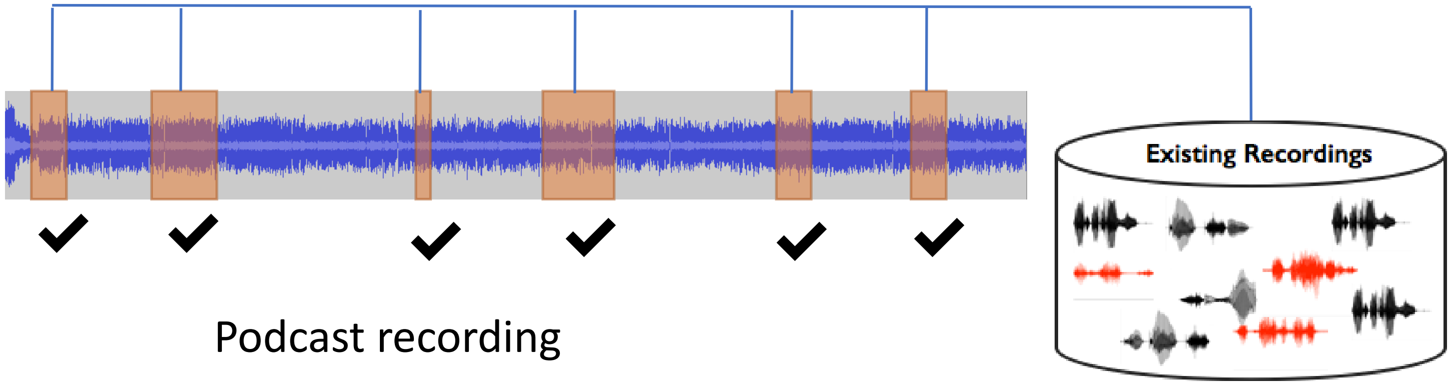
- **Teacher**
  - *N* (*N* = 5) teachers with different dropouts (MC dropout)
    - Model diversity giving complementary information
- **Average 100 MC teacher embeddings**
  - Preserves mean of the ensemble as well as captured uncertainty in predictions
- **Student**
  - *N* (*N* = 5) students learn from feature representations learned by teachers
  - Use unlabeled data + supervision from teachers
  - Final prediction is the average of the student ensemble predictions



The MSP-Podcast Database

- Use existing podcast recordings
- Divide into speaker turns
- Emotion retrieval to balance the emotional content
- Annotate using crowdsourcing framework

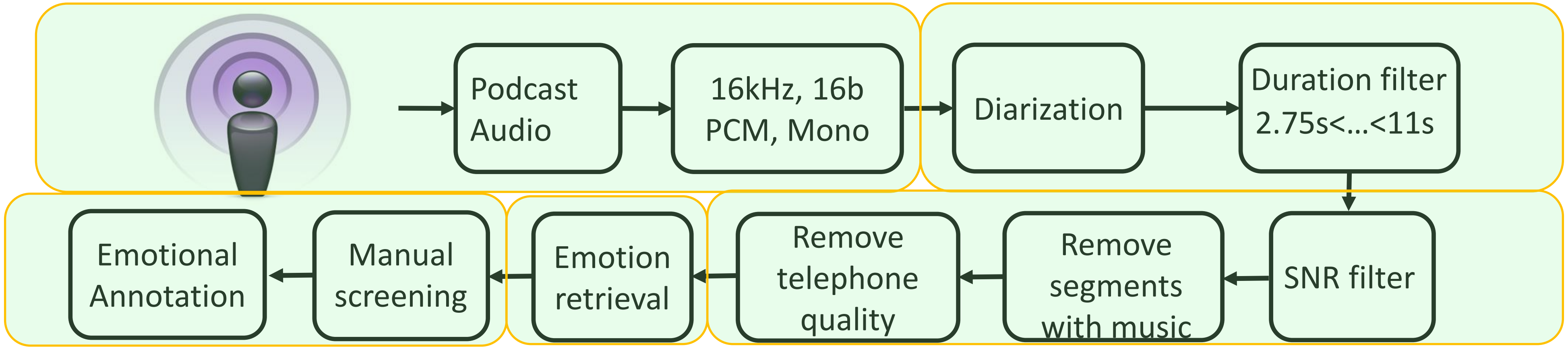




Reza Lotfian and Carlos Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," IEEE Transactions on Affective Computing, vol. To appear, 2018.

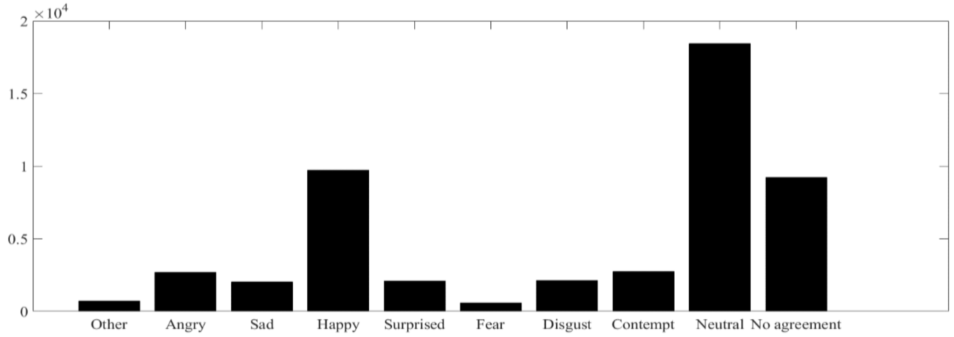
The MSP-Podcast Database

- **MSP-Podcast**
  - Collection of publicly available podcasts (naturalness and the diversity of emotions)
    - Interviews, talk shows, news, discussions, education, storytelling, comedy, science, technology, politics.
  - Creative Commons copyright licenses (**Available for sharing!**)
  - Single speaker segments, High SNR, no music, no phone quality
  - Developing and optimizing different machine learning framework using existing databases
    - Balance the emotional content
  - Emotional annotation using crowdsourcing platform

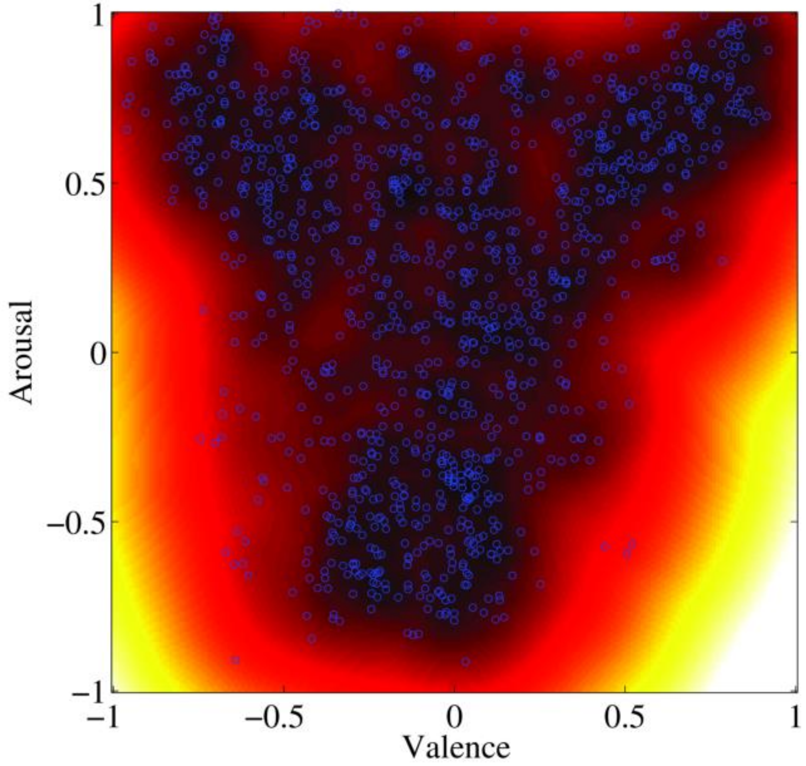


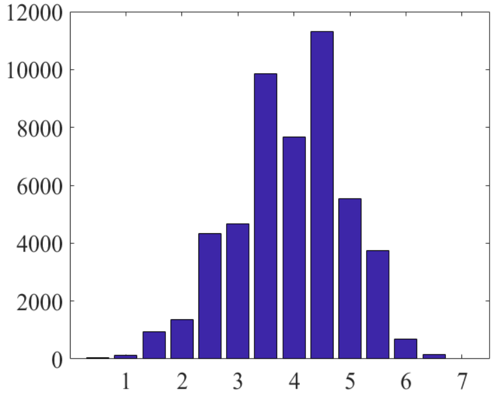
MSP-Podcast corpus version 1.6

With emotion labels:
50,362 sentences
(83h, 29m)

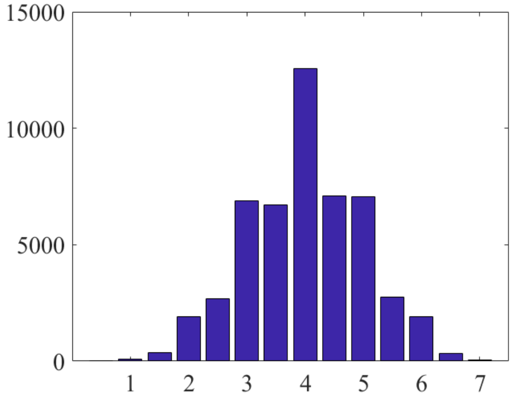


Primary emotional classes

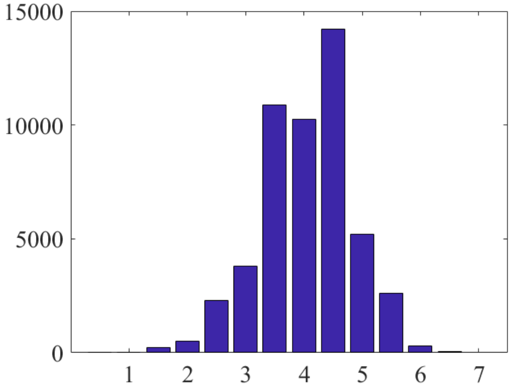




Arousal



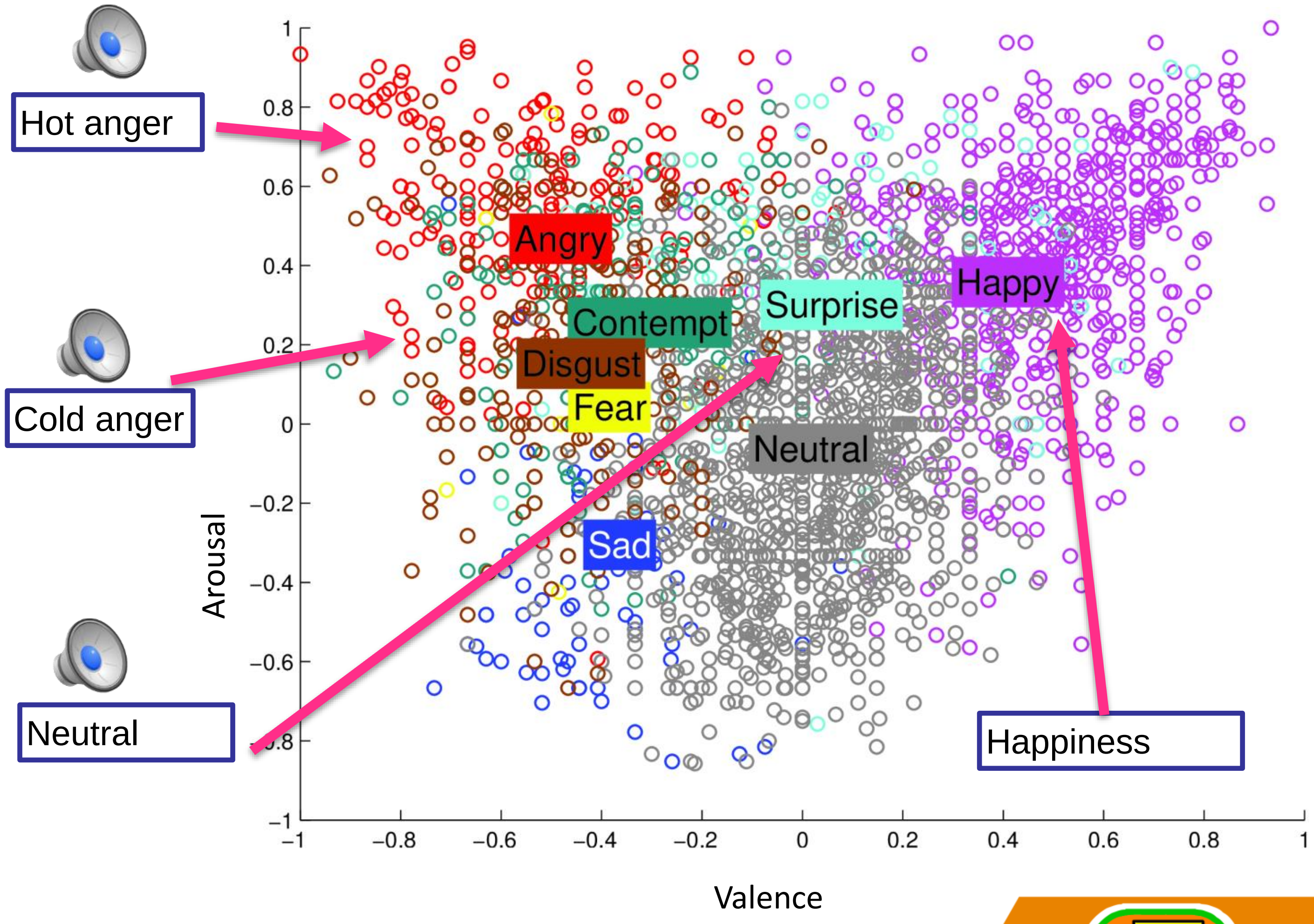
Valence



Dominance

MSP-Podcast Database

- Version 1.6 of the **MSP-Podcast** corpus
  - 50,362 (83h,29m)
- Corpus partition with aims to reduced speaker overlap in the sets:
  - Test data
    - 10,124 samples from 50 speakers (25 males, 25 females)
  - Validation data
    - 5,958 samples from 40 speakers (20 males, 20 females)
  - Train data
    - Remaining 34,280 samples



Acoustic Features

- **Interspeech 2013 Feature set**
  - 65 low level descriptors (LLD)
  - High Level Descriptors (HLDs) are calculated on LLDs resulting in total of 6,373 features
  - HLDs include:
    - Quartile ranges
    - Arithmetic mean
    - Root quadratic mean
    - Moments
    - Mean/std. of rising/ falling slopes

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	spectral
MFCC 1–14	cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F_0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice qual.
log. HNR, Jitter (local & δ), Shimmer (local)	voice qual.

Implementation Details

- **Train separate regression models each for arousal, valence and dominance**
- **Teacher:**
  - 5 teachers → DNN with 4 dense layers, 512 nodes per layer
  - MC dropout models with dropout rates: 0.45, 0.5, 0.55, 0.6, 0.65
  - SDG optimizer with learning rate equals to 0.001
  - Cost function: (1 - CCC)
  - Input: 6,373D feature vector
  - Output: 100 MC samples of the feature embeddings from the 4th dense layer
- **Student:**
  - 5 students → DNN with 2 dense layers, 512 nodes per layer
  - NADAM optimizer with learning rate equals to 0.0001
  - Loss = supervised loss + unsupervised loss → α . (1 - CCC) + β . (MSE)
  - Input: Feature embeddings from teacher (labeled) + Unlabeled data
  - Output: Predicted ensemble average score for arousal, valence and dominance

CCC

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

Performance or T-S models

- **Frameworks**
  - Baseline = 1 T without MC dropout
  - Teachers' MC ensemble = 5 T MC ensemble without S
  - T-S (test) = 5 T-S ensemble with test as unlabeled data
  - T-S (unlabeled) = 5 T-S ensemble with true unlabeled data
  - T-S (pseudo-label) = use S predictions on unlabeled data as labels and re-train S
  - T-S (top 75%) = use 75% of samples with lowest std.dev in the predictions from MC ensembles

Methods	Arousal	Valence	Dominance
Baseline	0.7045	0.3146	0.6336
Teachers' MC ensemble	0.7217	0.3184	0.6480
T-S framework (test)	**0.7345**	**0.3230**	**0.6652**
T-S framework (unlabeled)	**0.7322**	**0.3219**	**0.6625**
T-S framework (Pseudo-Label)	0.7290	0.3213	0.6558
T-S framework (Top 75%)	0.7279	0.3205	0.6508

Observations

- **Significant improvements ($p < 0.01$) over the baseline in terms of CCC with the use of unlabeled data at S training stage**
- **Relative increase in CCC:**
  - 4.25% for arousal, 2.67% for valence & 4.98% for dominance
- **Advantage of adding S (comparing row2 and row3)**
  - Relative increase in CCC upto 1.77% for arousal, 1.44% for valence & 2.65% for dominance

Ablation Studies

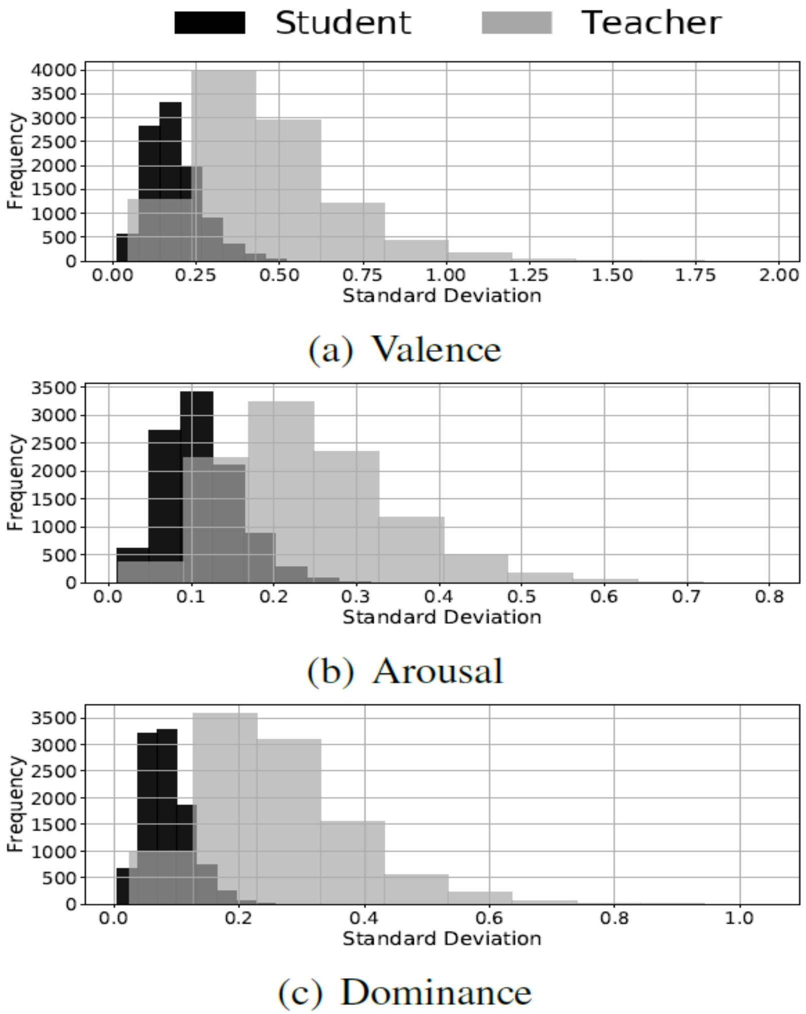
- **Systematic removal of contributing factors for our model**
  - Best with both labeled + unlabeled data, MC dropout and 5 T-S ensembles (row1)
  - Influence of unlabeled data on the generalization ability of our model (row2)
  - Importance of MC dropout ensembles → it contributes significantly to improvements over the baseline (row 3)
  - Usefulness of the ensemble approach (row 4)
  - Without MC dropout & ensemble → loss in CCC between 6.4% and 17.2% across A, V & D

A	B	C	Arousal	Valence	Dominance
✓	✓	5	0.7345	0.3230	0.6652
-	✓	5	0.7300	0.3211	0.6585
✓	-	5	0.7205	0.3154	0.6480
✓	✓	1	0.7240	0.3172	0.6512
-	✓	1	0.7219	0.3166	0.6556
✓	-	1	0.6873	0.2673	0.6198

A → Unlabeled data
B → MC dropout
C → No. of teachers and students in the ensemble

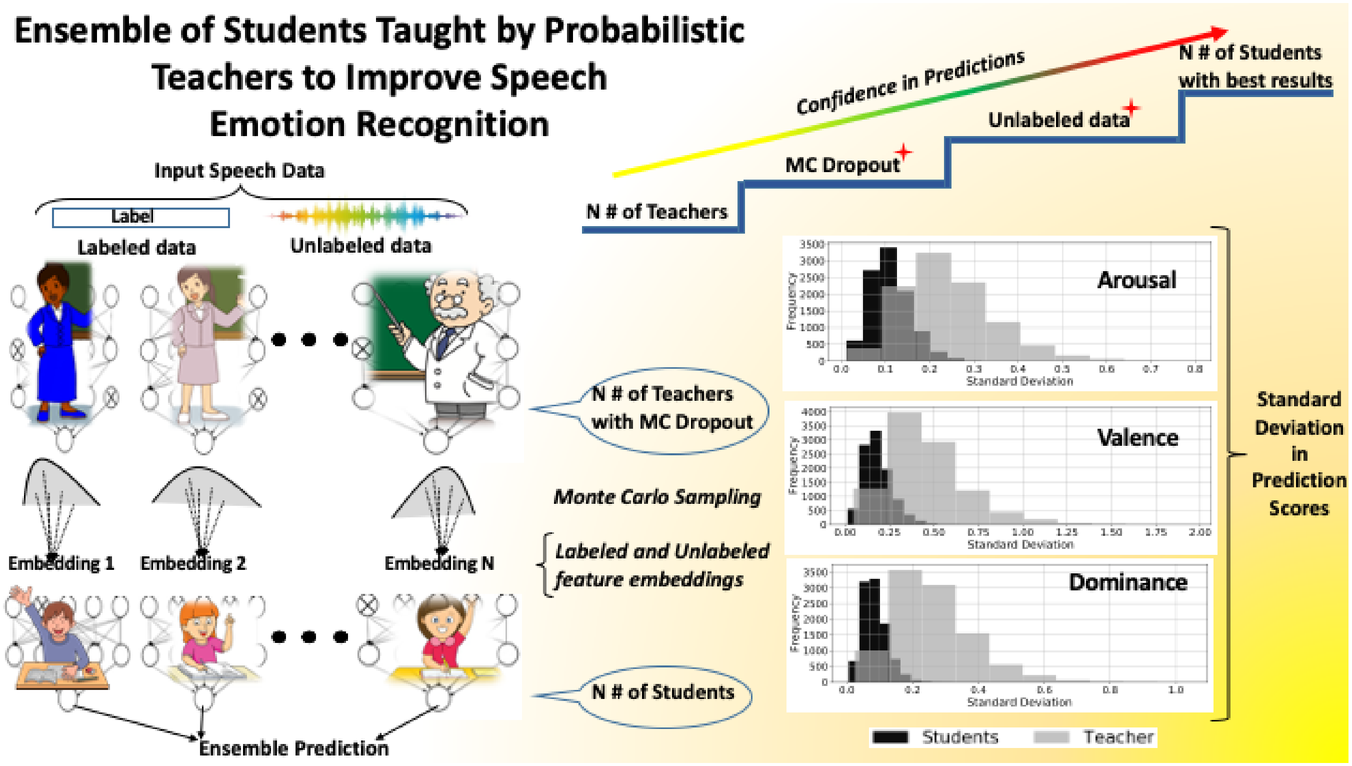
Analysis of Uncertainty in Predictions

- **Standard deviation (std.dev) in predictions to quantify uncertainty**
  - Teacher: select one MC sample per T and calculate std.dev across ensemble
  - Student: calculate std.dev across ensemble
- **Observations**
  - Std.dev for T are higher and dispersed
  - S predictions are more consistent
  - MC dropout is effective in guiding the student ensembles to give consistent predictions



Conclusions

- **Novel T-S framework for SER that:**
  - Improves prediction of emotional attributes
  - Gives consistent predictions
- **Knowledge distillation from T to S via MC ensemble of probabilistic features embeddings of T**
  - It leverages the learning of S on unlabeled data
- **Overall improvements in performance, generalizability and consistency in predictions**
- **Power of using MC ensembles + unlabeled data → up to 5% increase in CCC**



Release of the MSP-Podcast Corpus

- **Academic license**
  - Federal Demonstration Partnership (FDP) Data Transfer and Use Agreement
  - Free access to the corpus
- **Commercial license**
  - Commercial license through UT Dallas

https://msp.utdallas.edu

Thank you

- This work was funded by NSF CAREER Grant IIS-1453781

Questions or Contact: Kusha Sridhar
Kusha.Sridhar@utdallas.edu

Our Research: msp.utdallas.edu