Meta Learning for Robust Child/Adult Classification from Speech

Nithin Rao Koluguri, Manoj Kumar, So Hyun Kim, Catherine Lord, Shrikanth Narayanan

ICASSP2020 Barcelona

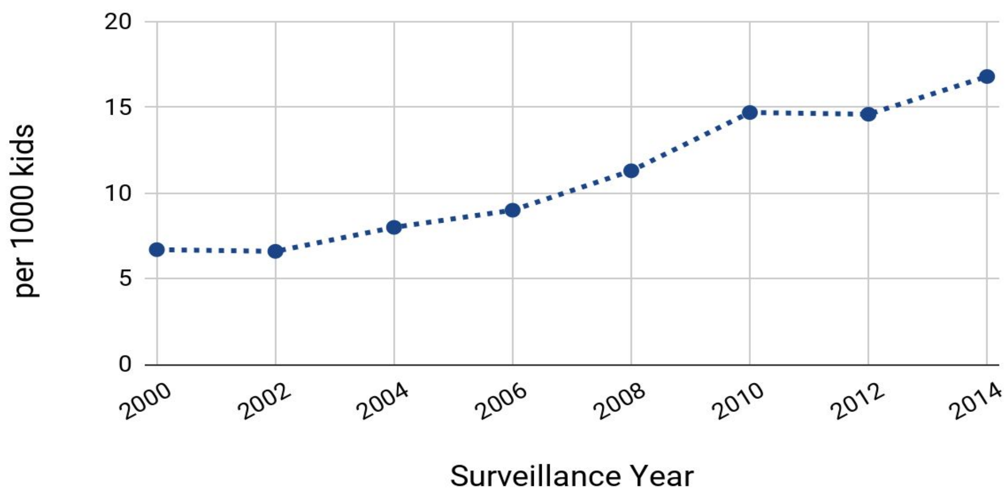
Presented by Manoj Kumar

Motivation: ASD

Autism Spectrum Disorder

- Heterogeneous group of complex neurodevelopmental disorders
- Rising reported prevalence among children in US

Reported prevalence of ASD among children (Baio et. al. 2018, CDC)

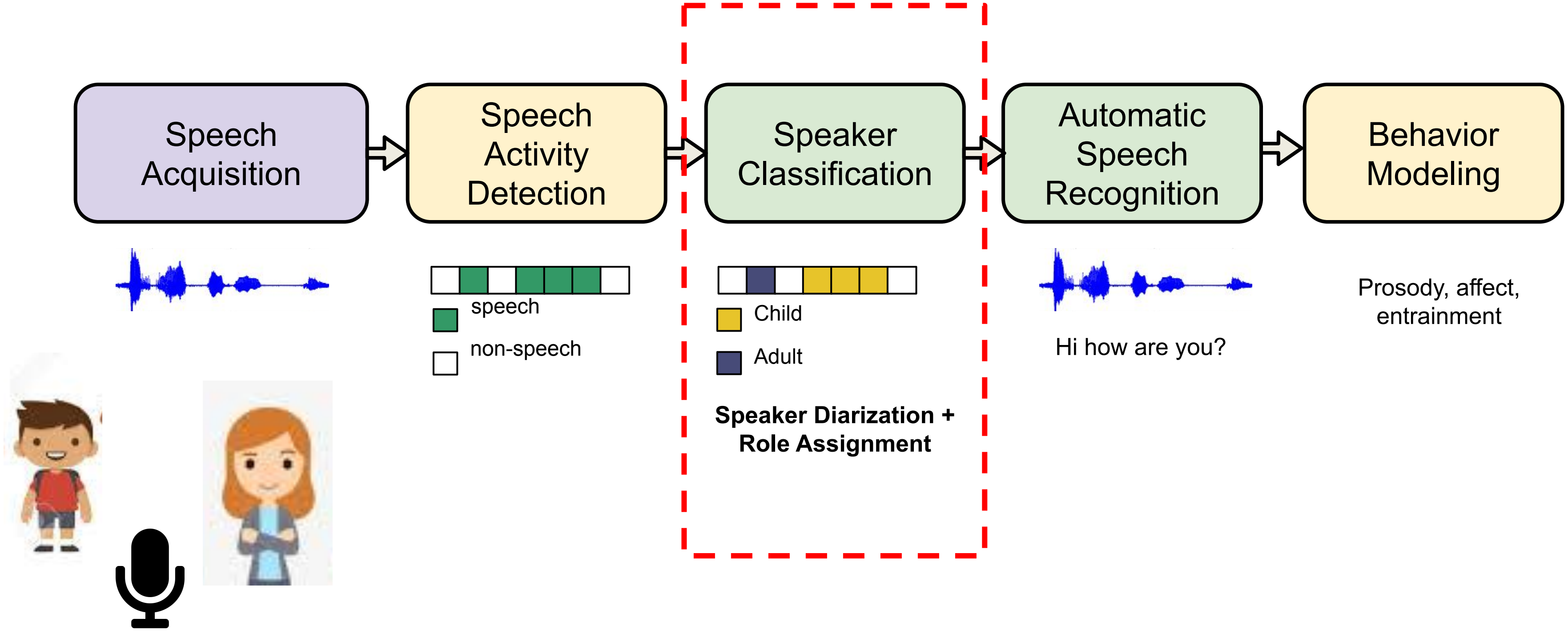


- Difficulties in communication and social interaction (Kenner 1943)

Child/Adult Classification from Speech

ASD Diagnosis and Assessment

- Primary tool: Semi-naturalistic conversations between child and clinician
- Automated analysis of diagnosis sessions can assist clinicians (Bone et al. 2016, Thabtah 2017, 2019)



Challenges

Factors that confound a conventional child/adult classification system:

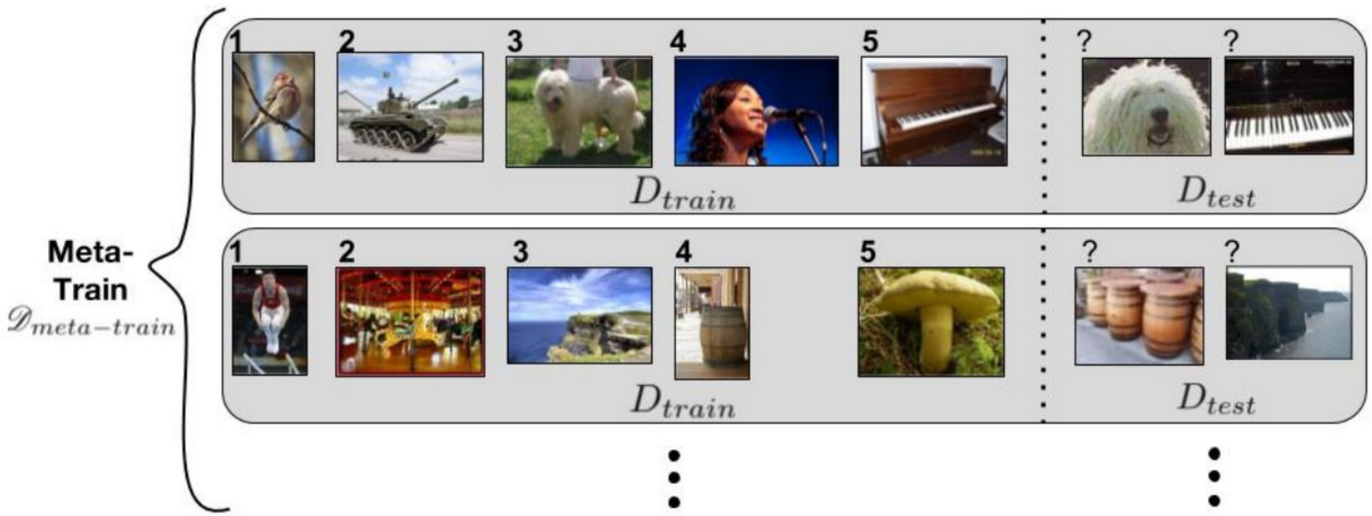
- Large within-class variability especially for child from age, gender, clinical symptom severity (Lee et. al., 1999, Gerosa et. al., 2009)
- Lack of sufficient amounts of balanced training data needed to tackle the above issue

Meta Learning: (Learning to learn) Paradigm of supervised learning developed for low-resource applications in computer vision (Finn et. al., 2017, Ravi et. al., 2016)

Meta Learning

Train Split *(Supports)*

Test Split *(Queries)*



Training Task 1

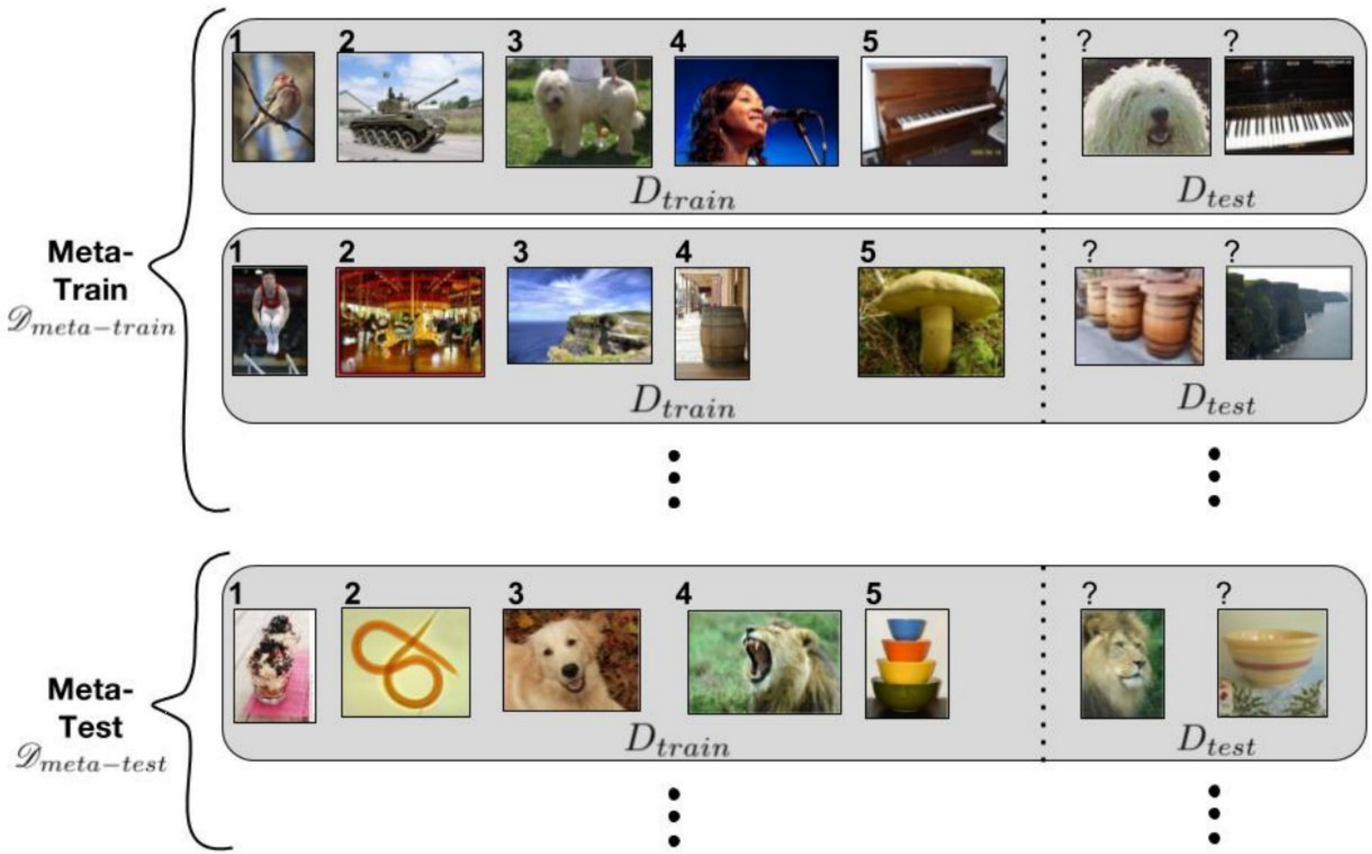
Training Task 2

Illustration modified from (Ravi et. al., 2016)

Meta Learning

Train Split *(Supports)*

Test Split *(Queries)*



Training Task 1

Training Task 2

Target Task

Illustration modified from (Ravi et. al., 2016)

Prototypical Networks: Illustration

Goal: Learn an embedding space to minimize distance-based task loss

Prototypical Networks: Represent each class using centroid (Snell et. al., 2016)

Prototypical Networks: Illustration

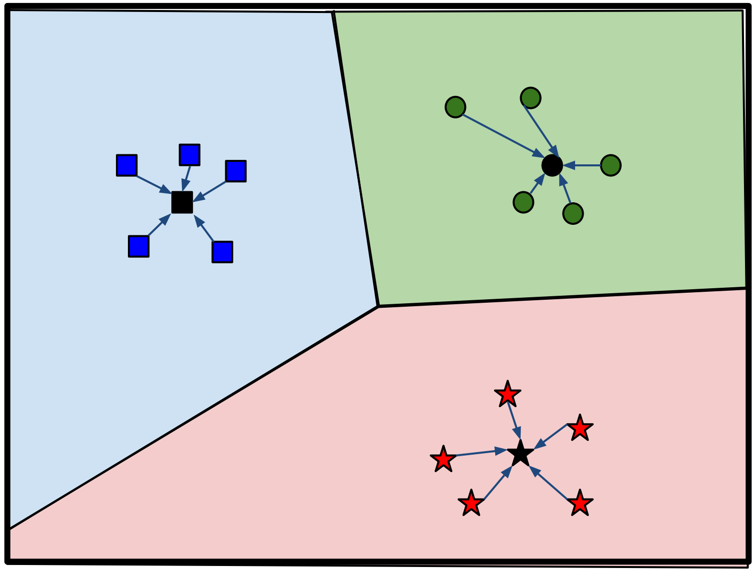
Goal: Learn an embedding space to minimize distance-based task loss

Prototypical Networks: Represent each class using centroid (Snell et. al., 2016)

Training Steps:

1. Compute prototypes

$$\mathbf{p}_c = \frac{1}{|S_c|} \sum_{(x_i, y_i) \epsilon S_c} f_\theta(\mathbf{x_i})$$



Prototypical Networks: Illustration

Goal: Learn an embedding space to minimize distance-based task loss

Prototypical Networks: Represent each class using centroid (Snell et. al., 2016)

Training Steps:

1. Compute prototypes

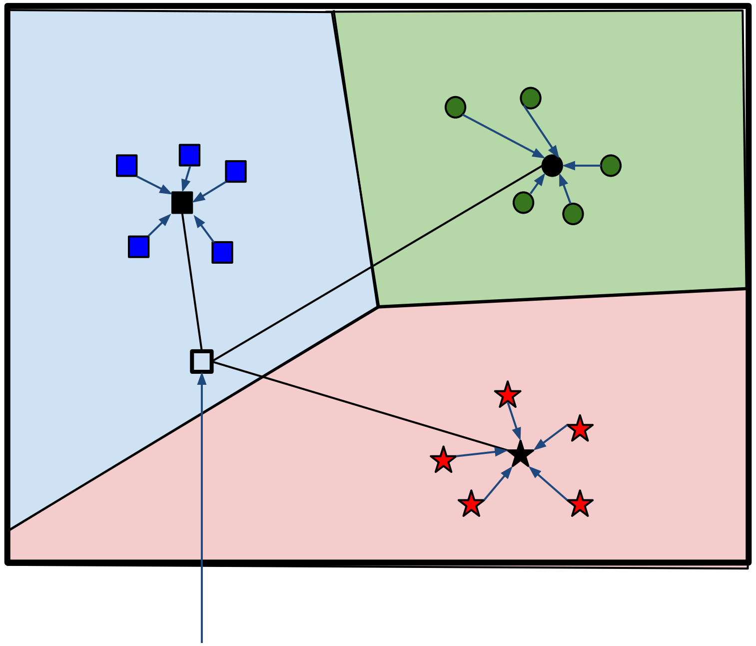
$$\mathbf{p}_c = \frac{1}{|S_c|} \sum_{(x_i, y_i) \epsilon S_c} f_\theta(\mathbf{x_i})$$

2. Estimate class posteriors

$$p_\theta(y = c|x) = \frac{\exp\left(-d_\varphi\left(f_\theta(\mathbf{x}), \mathbf{p}_c\right)\right)}{\sum_{c' \in C} \exp\left(-d_\varphi\left(f_\theta(\mathbf{x}), \mathbf{p}_{c'}\right)\right)}$$

3. Compute loss

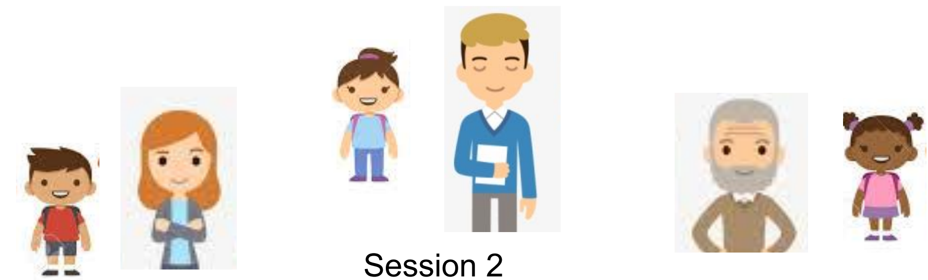
$$L(y, \mathbf{x}) = -\sum_{c=1}^{C} y_c \log(p_\theta(y = c \mid \mathbf{x}))$$



Test sample

Meta-Learning for Child/Adult Classification

Training Corpus
(Session ≡ Task)

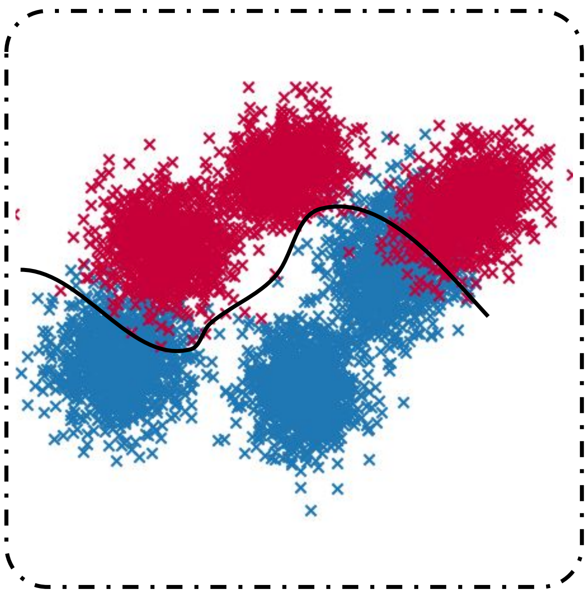


Session 1

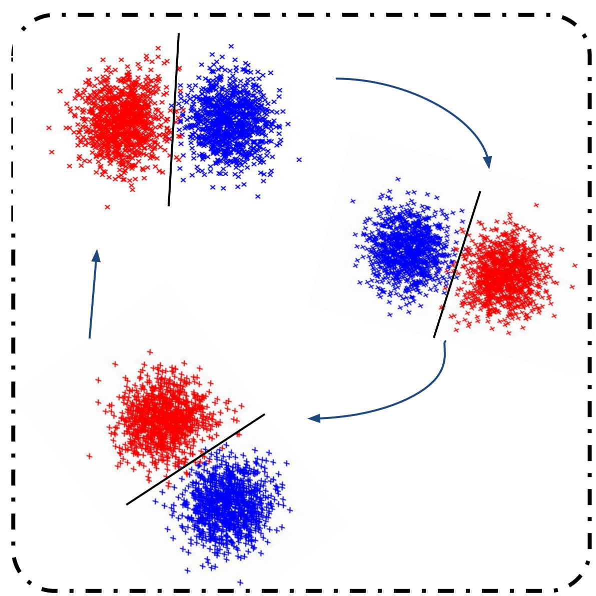
Session 2

Session 3

Conventional Learning



Meta Learning



Datasets

- Two categories of child-adult interactions used: ADOS & BOSCC
  - **Autism Diagnostic Observation Schedule** (Lord et. al. 2000): Gold-standard tool for autism diagnosis and assessment
  - **Brief Observation of Social Communication Change** (Grzadzinski et. al. 2016): Treatment outcome measure to assess social-communication (SC) and restricted & repetitive behaviors (RRB)

- Corpora division:
  - **ASD-Verbal:** Fully-verbal children (Train & Test)
  - **ASD-Infants:** Minimally-verbal toddlers & infants (Test only)

Table: Data statistics for ASD and ASD-Infants

Corpus	**Duration** *(min)* (mean ± std.)	**Child Age** *(yrs)* (mean ± std.)	**# Utts** **Child**	**Adult**
ASD-Verbal	17.76 ± 11.99	9.02 ± 3.10	11045	20313
ASD-Infants	10.35 ± 0.51	1.87 ± 0.78	1371	4120

Experiments

Features:

- X-vectors: State-of-the-art performance in speaker recognition (Snyder et. al., 2018) and speaker diarization (Sell et. al., 2018)
- DNN embeddings trained using speaker classification loss.
- In this work, pre-trained x-vectors used from the CALLHOME recipe[1]

Evaluation Settings:

- Classification: Standard low-resource evaluation (Ravi et. al., 2016)
  - *Weakly-supervised*: Randomly select 5 samples/class within each test session; Evaluation repeated 200 times to reduce bias.

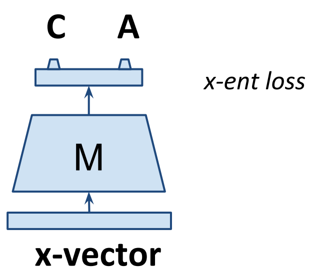
- Clustering: Standard speaker diarization evaluation (Sell et. al., 2018)
  - Cluster embeddings into #spkrs clusters within each test session.

1. https://kaldi-asr.org/models/m6

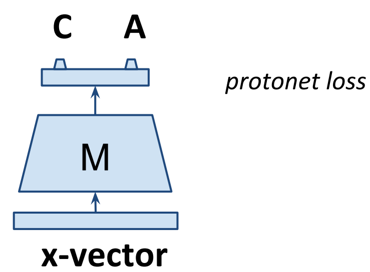
Experiments

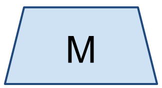
Classification Models:

Baseline



Protonet



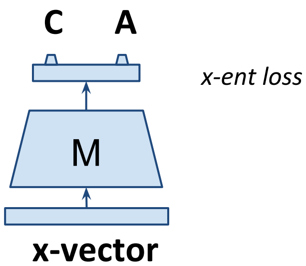


Architecture: 128x64x32
Non-linearity: ReLU
Optimizer: Adam
Regularization: BN, p_{drop}= 0.2

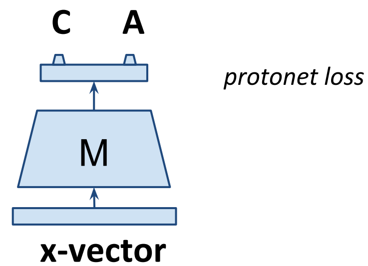
Experiments

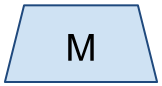
Classification Models:

Baseline



Protonet



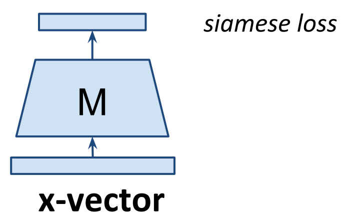


Architecture: 128x64x32
Non-linearity: ReLU
Optimizer: Adam
Regularization: BN, p_{drop}= 0.2

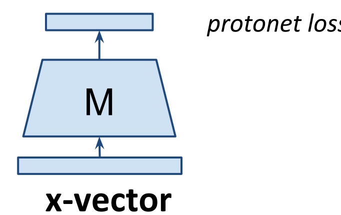
Clustering Models:

Baseline

Siamese Network (Snyder et. al., 2017)



Protonet



Results

Classification:

Table: Child/adult classification results (macro-F1, %)

Method	ASD-Verbal	ASD-Infants
Baseline (xent)	82.67	53.67
Baseline + test-backprop	78.64	56.20
Protonets	**86.66**	**61.30**

Results

Classification:

Table: Child/adult classification results (macro-F1, %)

Method	ASD-Verbal	ASD-Infants
Baseline (xent)	82.67	53.67
Baseline + test-backprop	78.64	56.20
Protonets	**86.66**	**61.30**

Clustering:

Table: Mean cluster purity (%) scores (SC: spectral clustering)

Method	ASD-Verbal		ASD-Infants	
	K-Means	SC	K-Means	SC
x-vectors	77.05	75.22	77.98	75.97
Siamese	78.22	79.18	78.30	76.86
Protonets	**81.39**	**80.70**	**85.51**	**85.55**

Qualitative Analysis

What do protonet embeddings learn?

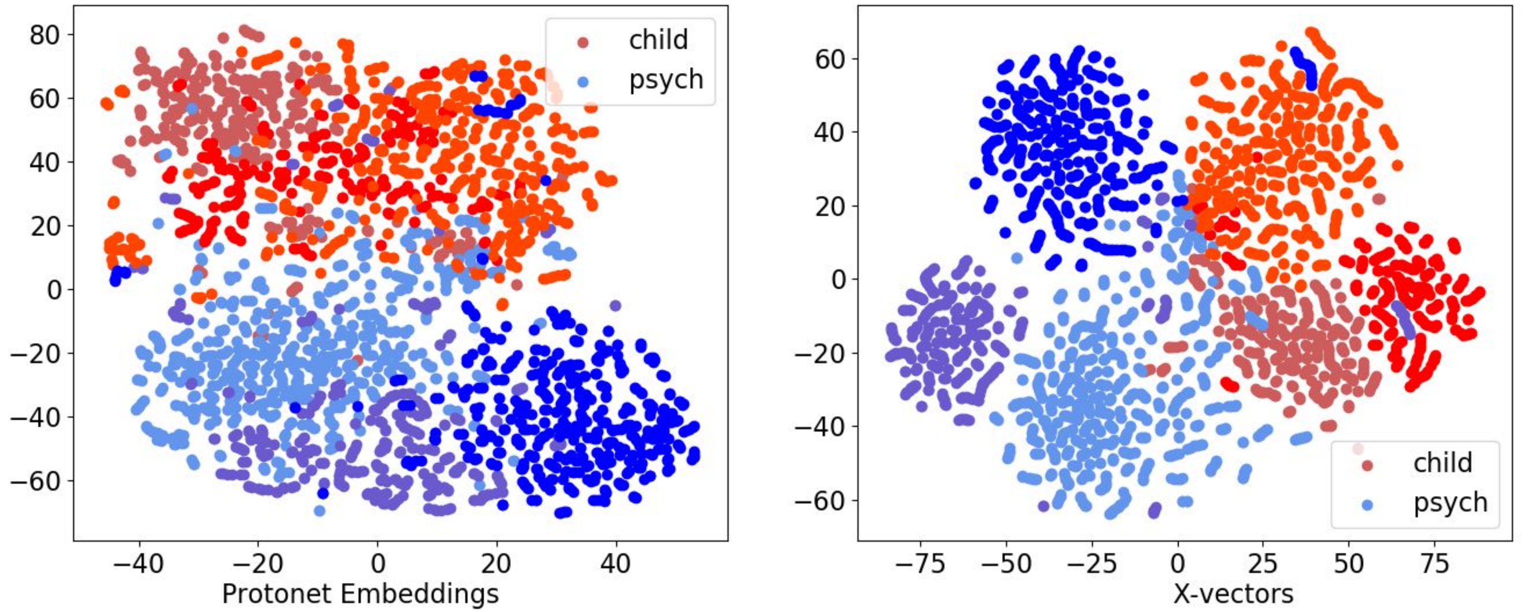


Figure: TSNE visualizations for protonet embeddings **(left)** and x-vectors **(right)** for 3 test sessions on the ASD corpora. Colors represents classes: **Child** and **Psych,** while shades within each color represent a session

Conclusions

- Modeling child/adult classification across sessions as multiple, related tasks ➔ Learn task-invariant representations using meta-learning
- How to extend this framework for a generic speaker embedding?
- Classification performance on ASD-Infants poor ➔ How to combine protonets within a domain adversarial framework?

Thank You

Manoj Kumar
prabakar@usc.edu