Voice Activity Detection with Teacher-Student Domain Emulation

Jarrod Luckenbaugh, Samuel Abplanalp, Rachel Gonzalez, Daniel Fulford, David Gard, Carlos Busso

UT Dallas MSP Multimodal Signal Processing Laboratory

THE UNIVERSITY OF TEXAS AT DALLAS

Multimodal Signal Processing Lab (MSP)

Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas, Richardson, Texas - 75080, USA

NIH Study supported by the NIH under grant 1R01MH122367-01

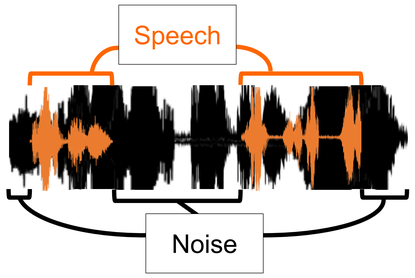
Background

Voice Activity Detection (VAD) :

- **Distinguishes between speech and noise**
  - Important first step for using speech data
  - Good performance on clean audio
- **NOT perfect: struggles with speaker distance and very loud, dynamic noise**

Our Work:

- Deep learning based VAD technique
  - Usage motivated by advances in automatic speech recognition (ASR)
- Domain adaption scales lab conditions to the real world for a medical application



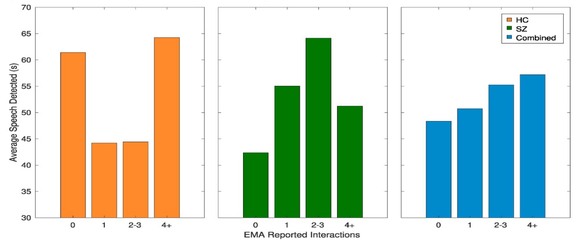
Target Domain (TD)

Social sensing with Digital Phenotyping:

- Uses datastreams from a patient's smartphone to make psychiatric assessments
- Seek to assess social isolation for those with schizophrenia spectrum disorders (SZ)

Data Collection:

- 2 groups: SZ and healthy controls (HC)
- 2 weeks: carried a phone with our program
- We gather ambient audio and spoken responses from prompted questions



- Previous work: voice activity detected tends to increase with number of social interactions

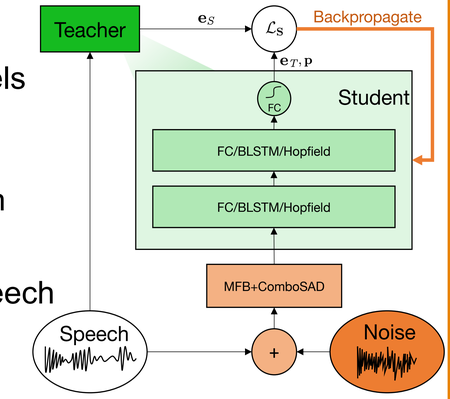
Teacher Student Domain Emulation

Teacher Student Learning

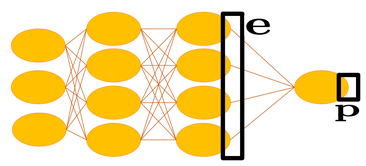
- Well performing teacher models can help a student improve performance
- Adapts student to new domain

Our Method:

- Teacher: Trained on clean speech
- Student: Trained on a noisy version of the same speech
- Teacher creates an "ideal" embedding before prediction
- Student is penalized for straying, in addition to the VAD task

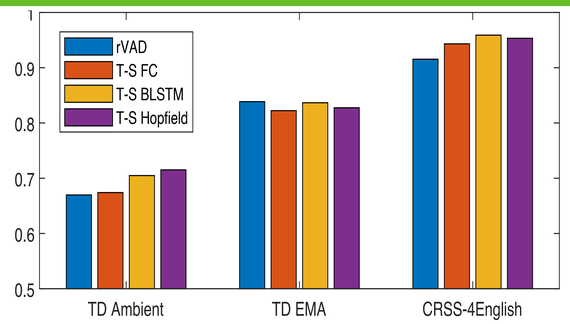


$$\mathcal{L}_{\mathrm{S}}(\mathbf{e}_S, \mathbf{e}_T, \mathbf{y}, \mathbf{p}) = \alpha \underbrace{\mathcal{L}_{\mathrm{emb}}(\mathbf{e}_S, \mathbf{e}_T)}_{\text{Embedding matching task}} + (1-\alpha) \underbrace{\mathcal{L}_{\mathrm{vad}}(\mathbf{y}, \mathbf{p})}_{\text{Binary Cross Entropy: VAD task}}$$



$$\frac{1}{J}\|\mathbf{e}_S - \mathbf{e}_T\|_2^2 \qquad -\sum_{i \in I} y_i \log(p_i)$$

VAD Performance / Transfer Analysis



- **Our method outperforms baseline for ambient audio and laboratory speech**
  - Left: Ambient (unconstrained, long) audio
  - Middle: Short, prompted user recordings
  - Right: Laboratory environment
- **Proposed implementations:**
  - LSTM best for shorter, prompted audio
  - CS-Hopfield best for long, ambient audio

Train \ Test	White 0dB T	White 0dB S	Babble 0dB T	Babble 0dB S	CHiME5 0dB T	CHiME5 0dB S
CRSS-4English14	0.992	0.970	0.992	0.988	0.992	0.985
+ White 0dB	**0.870**	**0.960**	0.859	0.799	0.870	0.695
+ White 10dB	**0.951**	**0.975**	0.951	0.945	0.951	0.915
+ Babble 0dB	0.434	0.248	**0.390**	**0.465**	0.434	0.353
+ Babble 10dB	0.796	0.587	**0.769**	**0.810**	0.796	0.709
+ CHiME5 0dB	0.897	0.845	**0.889**	**0.957**	**0.897**	**0.958**
+ CHiME5 10dB	0.957	0.919	**0.956**	**0.984**	**0.957**	**0.981**
+ TD Noise 0dB	0.889	0.777	**0.884**	**0.955**	**0.889**	**0.919**
+ TD Noise 10dB	0.962	0.868	**0.964**	**0.980**	**0.962**	**0.962**

- **Method improves performance when added training noise matches that of test condition**
  - Decreases otherwise. Also depends on noise power
- **Babble and CHiME5 noises are similar to TD noises**
  - Usage likely to improve performance in the target domain

Model	Test	Without T-S	With T-S
T-S BLSTM	CRSS-4English14	0.989	**0.990**
T-S BLSTM	TD-EMA	0.868	**0.875**
T-S BLSTM	TD-Ambient	0.750	**0.766**

- **Ablation study: Student trained without teacher vs with teacher**
  - Method Improves performance on all test sets

Conclusions

- Model agnostic, feature agnostic, teacher-student domain adaptation framework for training a VAD model
  - Improves performance most with unconstrained recording conditions
- Teacher-student domain embedding minimization is a complementary task to VAD