30 August – 3 September 2021
INTERSPEECH 2021
BRNO | CZECHIA
Speech everywhere!

Study supported by NIH
NIH
Grant 1R01MH122367-01

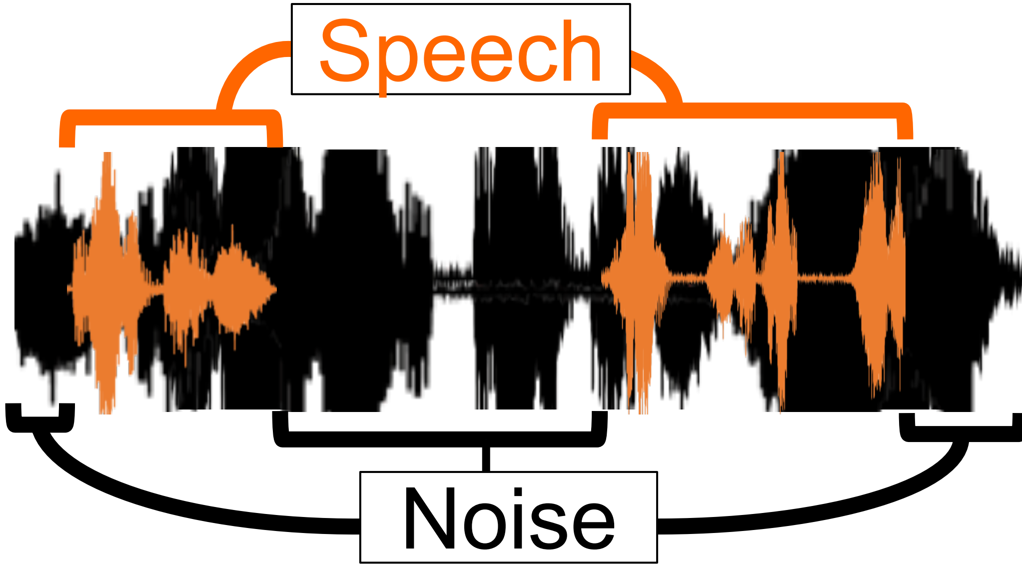
Voice Activity Detection with Teacher-Student Domain Emulation

J. Luckenbaugh, S. Abplanalp, R. Gonzalez,
D. Fulford, D. Gard, C. Busso

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

Voice Activity Detection

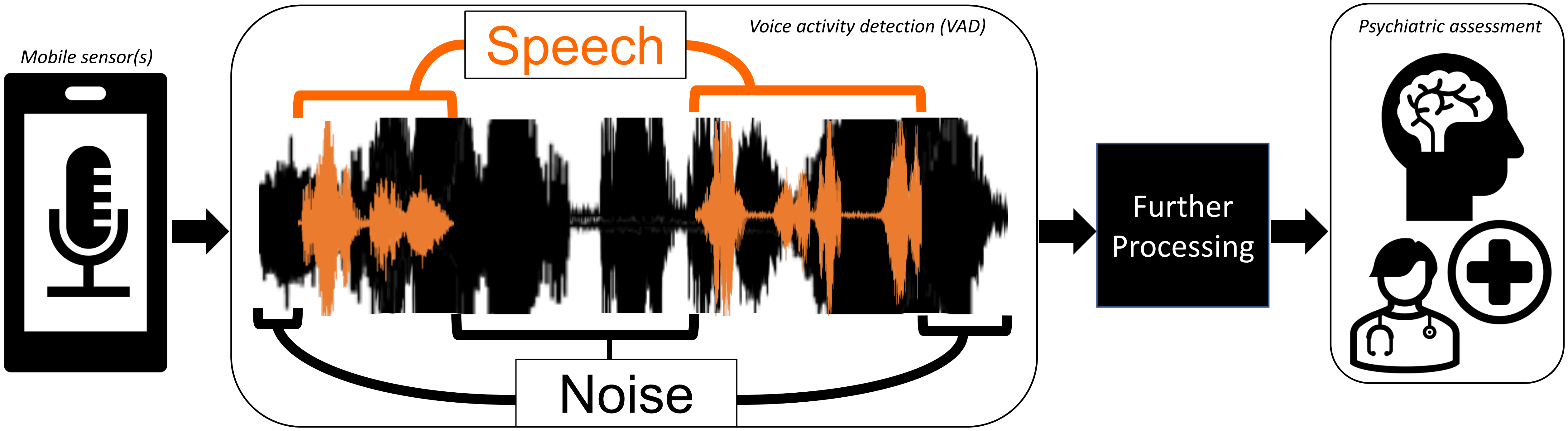
- **VAD – Classifies between speech and noise**
  - Essential pre-processing step for other speech tasks like ASR, SER, etc.
  - Open Problem: robust performance in realistic conditions
- **Brief history**
  - Statistical Methods – LP [1], PCA, etc.
    - Limited model capacities – often linear
    - Shorter evaluation windows
  - Deep learning
    - Learns nonlinear relations within sequences
    - Tends to be more robust
    - Requires training



[1] A. Benyassine, E. Shlomot, H. . Su, D. Massaloux, C. Lamblin, and J. Petit, "ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997

Voice Activity Detection "in the wild"

Practical applications require VAD that is robust to real world recording conditions



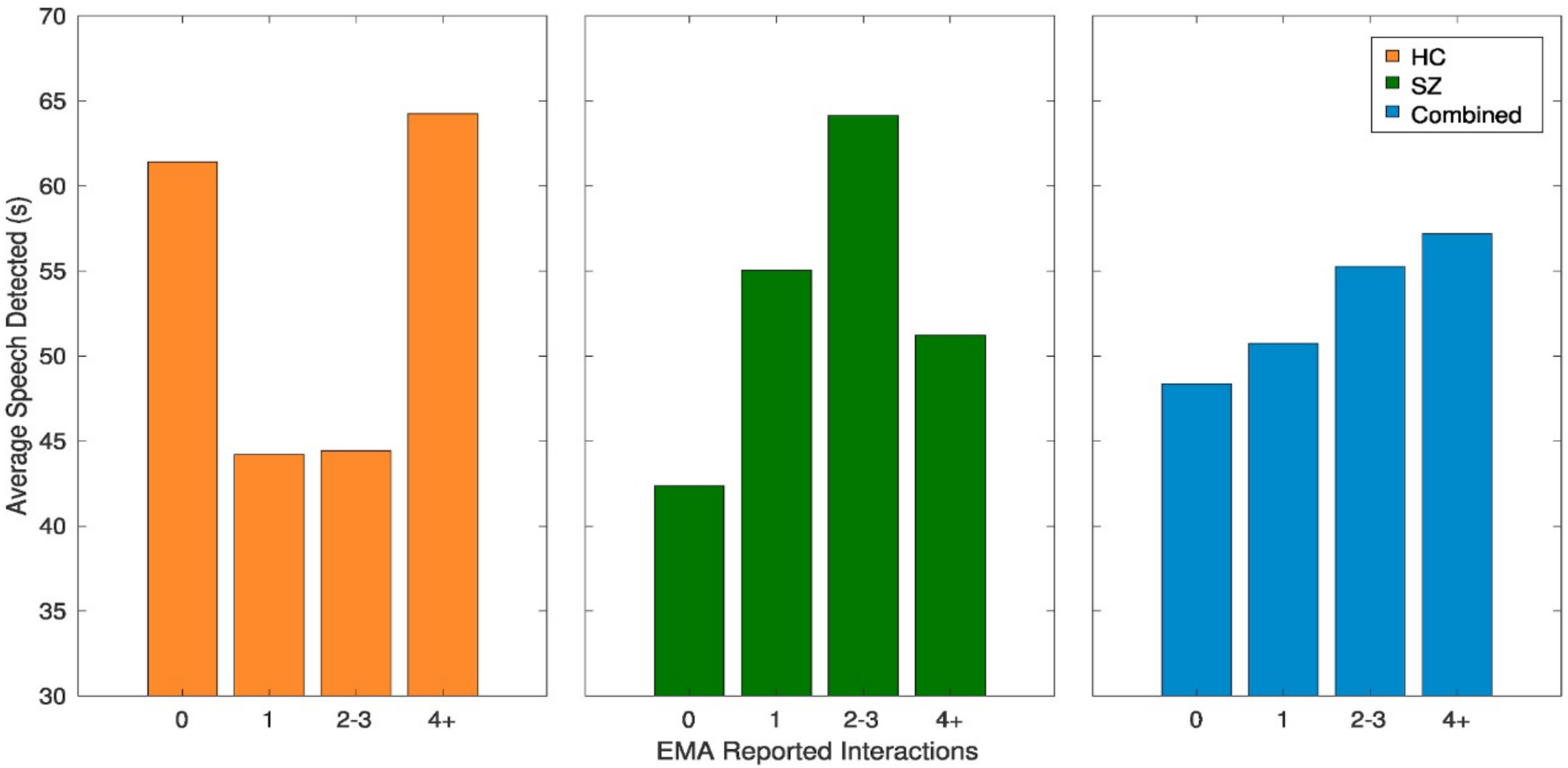
Transfer learning proves useful via the use of paired data in a teacher student framework

Target Domain (TD)

Two groups [HC & SZ] carried a phone with our program for two weeks

Schizophrenia Grp, n = 20 [SZ]

Control Grp, n = 15 [HC]

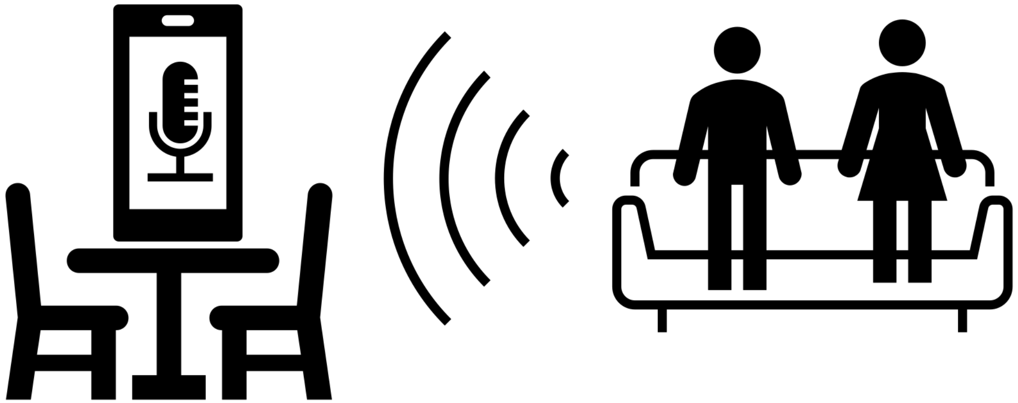


★ **Voice activity detected increases with self-reported number of conversations [2]**

[2] D. Fulford, J. Mote, R. Gonzalez, S. Abplanalp, Y. Zhang, J. Luckenbaugh, J.P. Onnela, C. Busso, D.E. Gard, "Smartphone sensing of social interactions in people with and without schizophrenia," *Journal of Psychiatric Research*, Volume 137, 2021, Pages 613-620

Target Domain (TD)

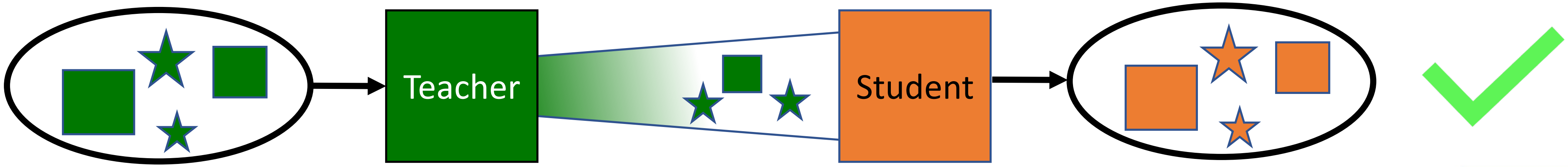
- **Ambient recordings [TD-ambient]**
  - Longer (5min), unprompted
  - Unknown microphone placements
  - Unknown number of speakers
  - Sparsely voiced



- **Ecological Momentary Assessments [TD-EMA]**
  - Shorter (~30sec), prompted
  - Microphone close to speakers
  - At least one speaker typically

Teacher Student Learning

- **The parameters of a well-trained model encodes task info**
- **Sequential feature representations become more task specific with depth**
- **These representations can be used in new models to transfer knowledge**
- **Teacher model can supervise the training of a student**
  - Can support generalization if tasks are similar
  - Can facilitate Supervised/unsupervised approaches
  - Can adapt a student to a slightly different domain



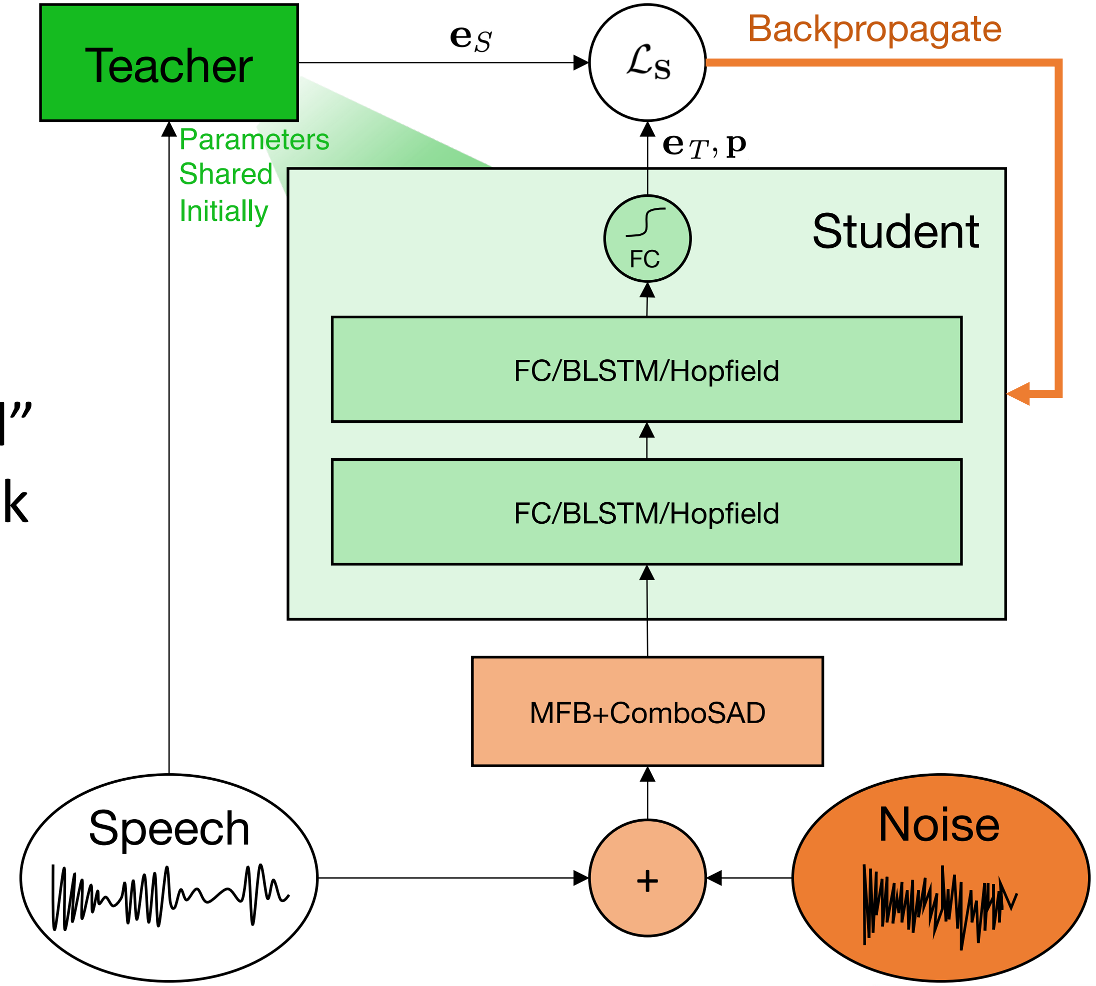
Method - Teacher Student Domain Emulation

- **Well performing teacher models can adapt a student to a new domain [3]**
- **Our Method:**
  - Teacher Training: Clean speech
  - Student Training: Noisy, paired speech
  - Student is penalized for straying from "ideal" teacher embedding, along with the VAD task

$$\mathcal{L}_{\mathrm{S}}(\mathbf{e}_S, \mathbf{e}_T, \mathbf{y}, \mathbf{p}) = \alpha \underbrace{\mathcal{L}_{\mathrm{emb}}(\mathbf{e}_S, \mathbf{e}_T)}_{\frac{1}{J}\|\mathbf{e}_S - \mathbf{e}_T\|_2^2} + (1-\alpha) \underbrace{\mathcal{L}_{\mathrm{vad}}(\mathbf{y}, \mathbf{p})}_{-\sum_{i \in I} y_i \log(p_i)}$$

Embedding matching task

Binary Cross Entropy: VAD task



[3] J. Li, M.L. Seltzer, X. Wang, R. Zhao, Y. Gong, "Large-Scale Domain Adaptation via Teacher-Student Learning," *Proc. Interspeech*, 2017, Pages 2386-2390

Method - Teacher Student Domain Emulation

- **Method relies on generating paired data to match target domain**
  - Clean speech easily collected in sound booth
  - Corrupting noise similar to the target domain

Speech

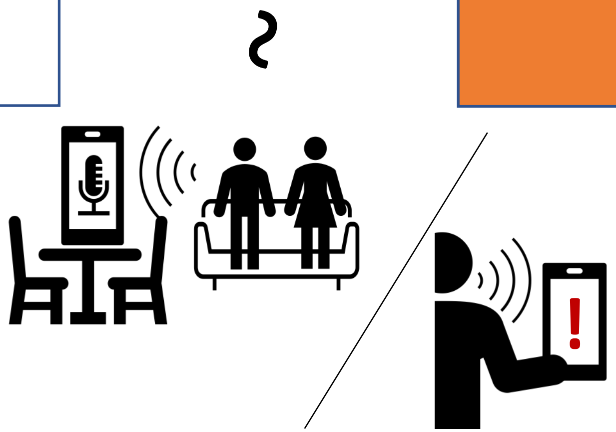
- CRSS4English14 [4]

- Clean, laboratory speech
- 130.3hr total
- 90%/5%/5%, Train/test/val

+

Noise

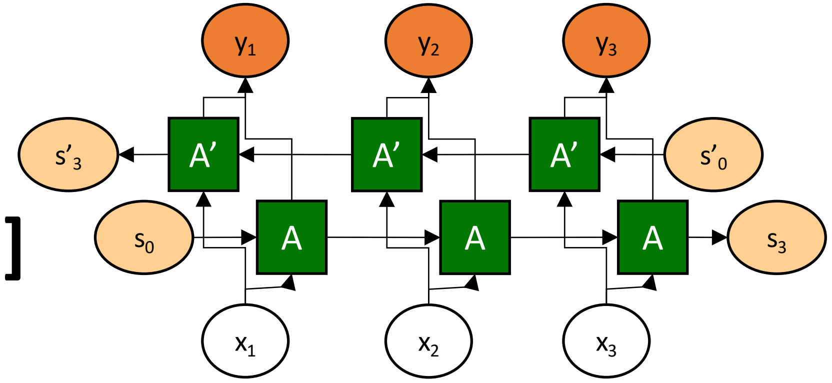
- TD-Noise (23.5hr)
- CHiME5 [5](77.4hr)
- Babble Noise (TIMIT)
- White Noise



[4] F.Tao, C.Busso,“Gating neural network for large vocabulary audiovisual speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1286– 1298, July 2018.
[5] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, “The Fifth ‘CHiME’ Speech Separation and Recognition Challenge: Dataset, Task and Baselines,” Interspeech 2018, Sep 2018.

Implementation - Temporal Models

- **Proposed method may be implemented for any model architecture**
- **Temporal models are best to handle sequential relations in feature windows**
- **Bidirectional Long Short-Term Memory (BLSTM)**
  - Extends LSTM to include a forward and backward pass
- **Continuous State Hopfield Network (CS-Hopfield) [6]**
  - Modern Hopfield network [7] with continuous states
  - More efficient than LSTM with performance similar to Transformers



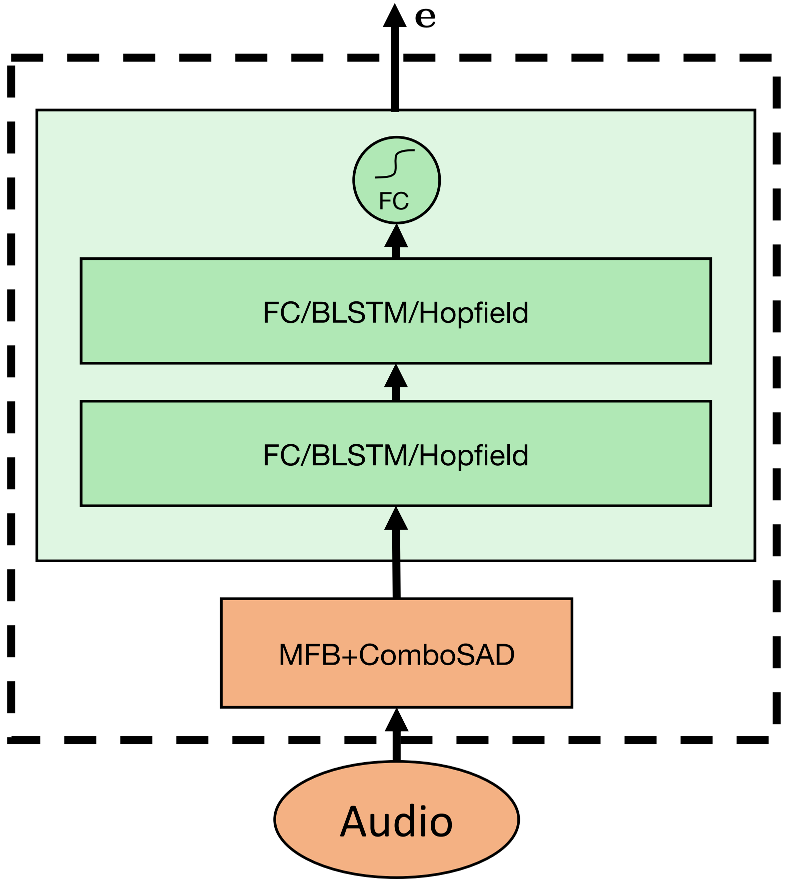
Hopfield Energy		New Energy		Update Rule		Transformer
$-\exp\left(\mathrm{lse}\left(1, \boldsymbol{\xi}^T \boldsymbol{X}\right)\right)$	$=$	$-\mathrm{lse}(\beta, \boldsymbol{\xi}^T \boldsymbol{X}) + \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} + c$	$\Rightarrow$	$\mathrm{softmax}\left(\beta\ \boldsymbol{\xi}^T \boldsymbol{X}\right) \boldsymbol{X}^T$	$=$	$\mathrm{softmax}\left(\frac{1}{\sqrt{d_k}}\ \boldsymbol{Q}\boldsymbol{K}^T\right)\boldsymbol{V}$

[6] Ramsauer, Hubert, et al. "Hopfield Networks is All You Need," *International Conference on Learning Representations*, 2021

[7] Krotov, Hopfield. "Dense associative memory for pattern recognition." Advances in neural information processing systems 29, 2016

Implementation – Experiments

- **Proposed method may be implemented for any model architecture, or feature set**
- **DNN Architectures - Two layers before sigmoid**
  - FC / Sequential layers (BLSTM/CS-Hopfield)
  - Fixed 0.6M parameters
  - ReLU activation; LayerNorm Regularization
- **Features – Window of 11 frames of 26 MFBs**
  - Explored addition of 5 ComboSAD [8] features
  - Frame size 20ms, Stride 10ms
- **Loss – Teacher: BCE, Student: Proposed**
  - Hyperparameter $\alpha = 0.2$
- **Training – ADAM(lr =1e-5), 4 epochs**

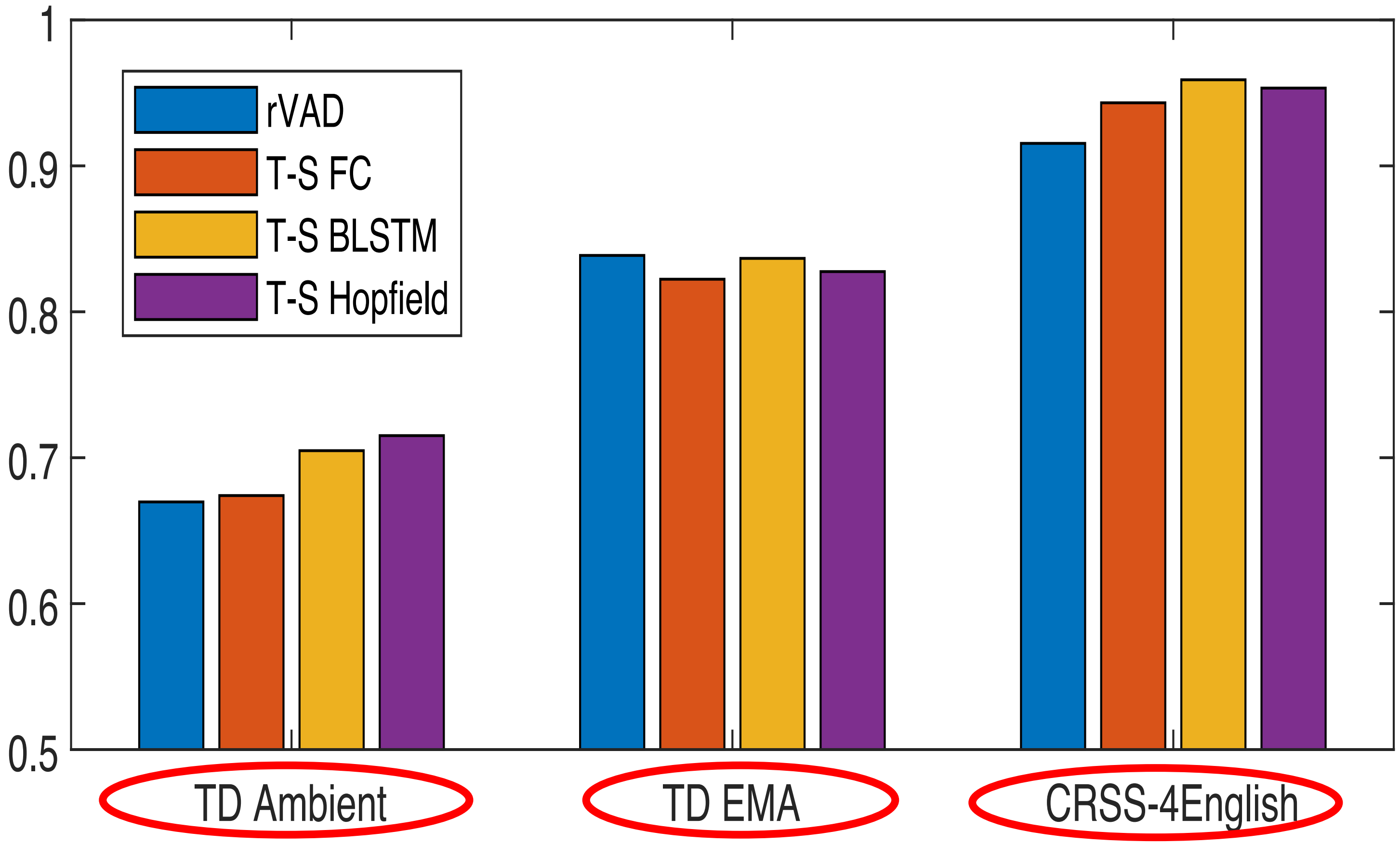


$$\alpha\mathcal{L}_{\text{emb}}(\mathbf{e}_S, \mathbf{e}_T) + (1-\alpha)\mathcal{L}_{\text{vad}}(\mathbf{y}, \mathbf{p})$$

[8] S. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, March 2013.

Results – Better real-world performance

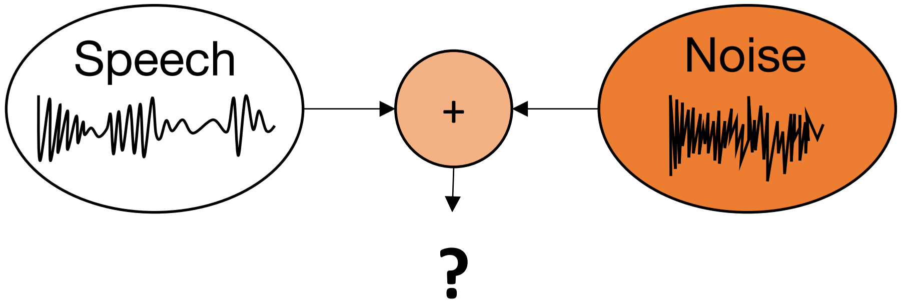
- **We achieve up to 7% higher F1 score than baseline for ambient audio and laboratory speech**
- **Best performing implementations:**
  - BLSTM for shorter, prompted audio
  - CS-Hopfield [9] for ambient audio



[9] Ramsauer, Hubert, et al. "Hopfield Networks is All You Need," *International Conference on Learning Representations,* 2021

Results – Domain Emulation

- **Method improves performance when added training noise matches that of test condition**
  - Generalization measured with AUPRG Scores [10]
  - Positive transfer highlighted



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**

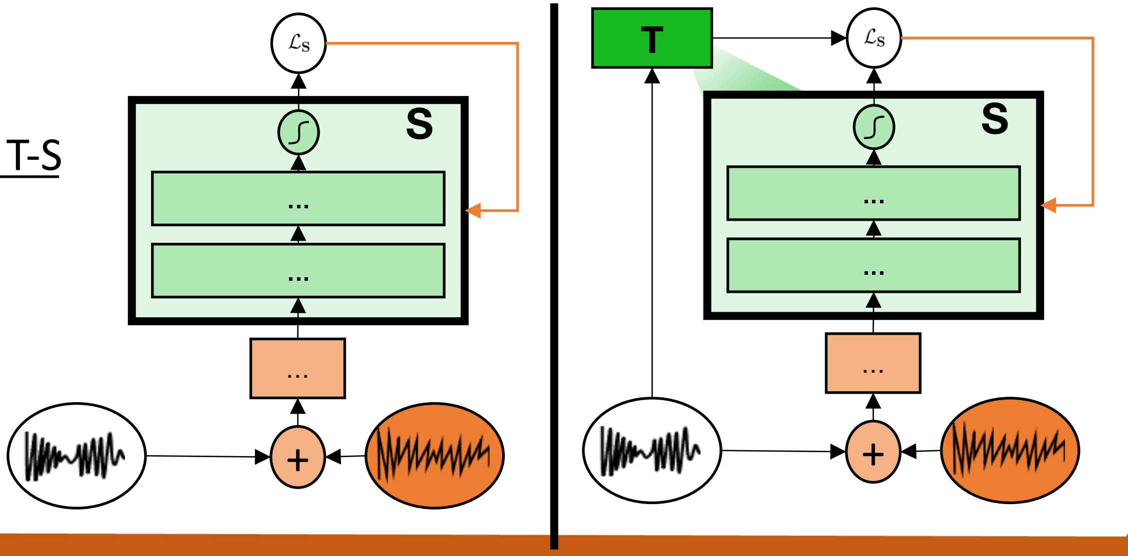
[10] P. Flach, M. Kull, Precision-Recall-Gain Curves: PR Analysis Done Right, *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2015, Vol. 28

Results – Ablation

- **Ablation study: Student trained without teacher vs with teacher**
- **Method improves generalization (AUPRG) - Higher values highlighted**

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**

Without T-S



With T-S

Results – Implementation Details

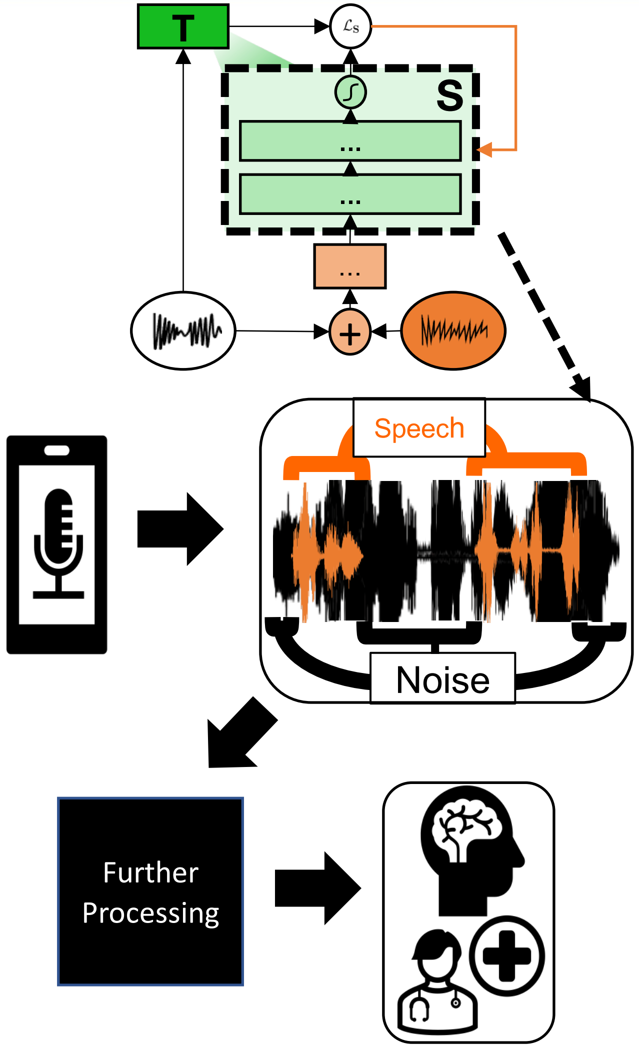
- **Ablation study: Models trained with vs without ComboSAD features**

Test	MFB		MFB+ComboSAD	
	T	S	T	S
CRSS 4English-14	0.994	0.989	0.994	0.990
TD-EMA	0.902	0.864	0.905	0.875
TD-Ambient	**0.747**	**0.759**	**0.734**	**0.766**

- **Analysis window size varied and tested on ambient set**
  - i.e. number of consecutive feature frames

Window	Test	T-S HF		T-S BLSTM	
		T	S	T	S
5	TD-Ambient	0.714	0.737	0.701	0.717
11	TD-Ambient	0.734	0.766	0.737	0.766
61	TD-Ambient	0.819	0.790	0.743	0.806

More details in our paper!



Thank you for attending!

Contact me at:

jvl170030@utdallas.edu
linkedin.com/in/jluckenbaugh2
github.com/jluckenbaugh2

Check out our lab!

msp.utdallas.edu

Study supported by NIH
Grant 1R01MH122367-01