INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Voice Activity Detection with Teacher-Student Domain Emulation

Jarrod Luckenbaugh**, Samuel Abplanalp*†*, Rachel Gonzalez*$^{\wedge}$*, Daniel Fulford*†‡*, David Gard*$^{\wedge}$*,*
Carlos Busso*

*Department of Electrical and Computer Engineering, The University of Texas at Dallas, USA
†Department of Occupational Therapy, Boston University, USA
‡Department of Psychological & Brain Sciences, Boston University, USA
$^{\wedge}$Psychology Department, San Francisco State University, USA

{jvl170030,busso}@utdallas.edu, {samabp,dfulford}@bu.edu, rachelgonzalez742@gmail.com, dgard@sfsu.edu

Abstract

Transfer learning is a promising approach to increase performance for many speech-based systems, including *voice activity detection* (VAD). Domain adaptation, a subfield of transfer learning, often improves model conditioning in the presence of a mismatch between train-test conditions. This study proposes a formulation for VAD based on the teacher-student training, where the teacher model, trained with clean data, transfers knowledge to the student model trained with a noisy, paired version of the corpus resembling the test conditions. The models leverage temporal information using *recurrent neural networks* (RNN), implemented with either *bidirectional long short term memory* (BLSTM) or the modern, continuous-state Hopfield network. We provide evidence that in-domain noise emulation for domain adaptation is viable under unconstrained audio channel conditions for VAD "in the wild." Our application domain is in healthcare, where multimodal sensors, including microphones, from portable devices are used to automatically predict social isolation in patients affected by schizophrenia. We empirically show positive results for domain emulation when the training conditions are similar to the target domain. We also show that the Hopfield network outperforms our best BLSTM for VAD on real-world benchmarks.

Index Terms: Voice Activity Detection, Speech Activity, Transfer learning, Domain Adaptation, Hopfield Network

1. Introduction

Voice activity detection (VAD), the binary classification task of distinguishing voiced segments in an audio stream, is an increasingly important building block for other speech processing tasks such as *automatic speech recognition* (ASR), speaker separation, and *speech emotion recognition* (SER). While current methods for VAD have shown to perform well within constrained recording environments, VAD under non-ideal conditions is still an open problem.

In unconstrained recording environments, VAD must remain robust to challenging, non-stationary noise at a low *signal-to-noise-ratio* (SNR). These conditions arise naturally in real applications, including our target domain. We are exploring the use of speech technologies as tools for healthcare, where mobile devices [1] are used to gather acoustic or environmental information to inform the well-being of a patient over a period of time [2]. Traditionally, VAD has relied on statistical relations between audio features by employing mathematical methods such as *principal component analysis* (PCA) [3], and *linear prediction* (LP) analysis [4], often employing dynamic approaches [5]. However, current methods for this task are now largely based on deep learning [6–10], where *deep neural networks* (DNNs) have demonstrated unprecedented levels of success on speech processing problems due to the ability to extract valuable, non-linear relationships between data points. This feature has led to increased robustness to noise and overall performance of VAD systems.

This paper proposes a supervised approach based on the teacher-student formulation to transfer knowledge by reducing the mismatch between source and target domains. Our approach emulates noise conditions in the target domain by creating a parallel corpus using additive noise. A teacher model, trained for the ideal source domain, is used to transfer knowledge to a student model designed to work on the target domain. During training, the student model leverages the teacher's representation, reducing the mismatch between train and test conditions. This formulation can be implemented with a variety of modern DNN architectures. We exploit temporal information by using *bidirectional long short term memory* (BLSTM) and the novel, *continuous-state Hopfield* (CS-Hopfield) network [11]. The proposed approach can increase the generalization of the student model, improving robustness.

The experimental evaluation shows that the proposed approach of incorporating paired data with similar noise conditions to the target domain during training can be useful to improve the performance of VAD systems. When the target domain and the emulating data are sufficiently similar, positive transfer occurs, improving the student performance. The approach is not effective when noise conditions are mismatched in the target domain. We show recurrent architectures such as BLSTM and CS-Hopfield networks lead to improved performance when a longer temporal context is available to the models. The proposed approach achieves a F1-score of 72% on our challenging, uncontrolled target domain, which is 7% better than our baseline system. The main contributions of our study are:

- We propose a model agnostic, feature agnostic, teacher-student domain adaptation framework for training a VAD model. This approach reduces the mismatch between source and target domain, boosting the performance of existing deep learning based VAD systems in the presence of noise.
- For the first time, the teacher-student domain embedding minimization is introduced in a VAD task, showing its complimentary benefits in a real healthcare application.

2. Related Works

2.1. Target Domain

This study is part of our multidisciplinary collaboration to investigate social isolation in patients affected by *schizophrenia spectrum disorders* (SZ) using portable devices [2]. People with

http://dx.doi.org/10.21437/Interspeech.2021-1234

SZ and healthy controls wear a smartphone with our program, which collects information during their daily activities. Relevant to this study is the audio captured by the cellphone, which includes prompted and unprompted recordings (Sec. 3). Unprompted speech is sampled randomly over a prolonged period in unconstrained environments, where we do not have control over the placement of the cellphone with respect to the subject. Therefore, robust speech models are required to handle the noisy audio conditions observed in naturalistic environments. Detecting speech activity is the first step before implementing other complex tasks, such as ASR or SER.

2.2. Voice Activity Detection in Noisy Environments

The primary challenge to accurately estimate voice activity in our target domain is the presence of arbitrary, non-stationary noise in naturalistic environments. This common issue has inspired similar efforts for effective VAD [1, 12, 13]. With a great amount of available speech data, DNNs have been effective in discriminating between complex speech signal and noise patterns [14–18]. Many solutions have attempted to mitigate the effects of noise using approaches such as speech enhancement and signal denoising [14–16, 19]. Other VAD approaches have incorporated related modalities [17, 18, 20] to increase prediction accuracy by leveraging discriminative features that are invariant under audible noise (e.g., facial features [18, 21]).

2.3. Transfer Learning using Teacher-Student Training

Teacher-student training is an effective domain adaptation method where a successful "teacher" model trained in ideal conditions is used to improve the generalization of a "student" model designed to work in a specific target domain. This method has led to improved performance on related speech tasks such as ASR [22–25] and SER [26]. In particular, domain adaptation has been shown to have positive effects on ASR [25] using parallel data to adapt a model between similar speech domains. The inductive bias passed when the student is initialized with the well performing teacher parameters improves the generalization ability of the overall system, given that the source and target domains are sufficiently similar. This paper applies these techniques to the task of VAD with novel modifications to incorporate speech denoising. To the best of our knowledge, this is the first study using this formulation for VAD tasks.

3. Resources

Our key assumption is that emulating the recording conditions in the target domain has a positive effect on the generalization of the proposed transfer learning method. We achieve this goal with additive noise recordings sourced from various naturalistic corpora expected to be similar to those in the target domain.

3.1. Target Domain Recordings

The target domain data corresponds to several datasets collected using a similar protocol for a total of 35 subjects. The protocol includes unprompted recordings, where the smartphone records the ambient audio for 5 minutes within 30-minute time windows with start times randomly sampled throughout each day (participants were asked to wear a round, visible pin, indicating they may be recording audio). As a result, the recordings are often sparsely voiced, with little to no speech activity in many segments. From this set we collect two subsets. The first is our test set for evaluating VAD performance. Multiple recordings are randomly selected for a total duration of 2.51 hours. Speech and silence labels are created manually for these segments. We refer to these recordings as *target domain - ambient* (TD-Ambient). The second subset contains noise recordings that we use for corrupting data. This in-domain noise is gathered by taking a random sample of the remaining ambient recordings and selecting those without speech data for a total duration of 23.5 hours. The set was partitioned into train (21.2 hours), test (1.12 hours), and validation (1.17 hours) sets. We refer to these recordings as *target domain - noise* (TD-noise).

The protocol also includes prompted recordings with better recording conditions. It uses *ecological momentary assessment* (EMA) [27], which are periodic surveys where the participant answers questions about their emotional state. This prompted nature often yields better recording conditions and higher speech intelligibility as users speak more directly into the smartphone's microphone for 2 to 30 seconds. We create a test set by randomly selecting from these recordings (1.08 hours). We refer to these recordings as *target domain - EMA* (TD-EMA).

3.2. Other Recordings

We use a subset of the CRSS-4English-14 corpus [28] to train the model with clean speech. Speech and silence labels for these segments were created using forced-alignment [29]. These labels serve as the ground truth for speech predictions. We use 117.1 hrs (train set), 6.60 hrs (test set), and 6.58 hrs (validation set) of clean speech from the data collected from the speakers with American accents. To collect naturalistic noise, we also use a subset of the CHiME5 dataset [30], a speech recognition challenge for conversational speech across multiple interlocutors in naturalistic, home environments. The corpus considers a dinner party scenario, where the audio is recorded, fully transcribed and time-aligned. We consider the segments with duration of at least 500ms without speech. This approach creates a noise corpus of 77.4 hours of audio across 178,224 segments split into train (69.3 hrs), test (4.08 hrs), and validation (4.02 hrs) sets. We also use babble noise generated by overlaying sentences taken from the TIMIT dataset. Speaker tracks are created by randomly concatenating sentences, mixing seven speakers. Krishnamurthy and Hansen [31] have shown that when audio from seven or more simultaneous speakers are mixed, individual words are indistinguishable in the opinion of 100% of participants surveyed, resulting in babble-like noise. To corrupt a speech corpus, we concatenate noise recordings and adjust the energy by fixing a target SNR when noise is added.

4. Proposed Approach

Our premise is that generating paired data points that emulate the domain of interest can improve performance on related speech problems [25]. Figure 1 presents our proposed method to achieve this goal, which is based on the *teacher-student* (T-S) formulation. The teacher model is trained with clean speech from the source domain (CRSS-4English-14 corpus). Knowledge gained by a well-trained teacher within a well-known domain is then leveraged to train a student in a more difficult domain. The student model is trained with a parallel corpus corrupted with additive, naturalistic noise. We expect that using a loss that optimizes performance via the influence of inductive priors taken from the teacher model aids the generalization ability of a student model. The priors established by a well-trained speech model in ideal (uncorrupted) cases encode a useful optimization criterion that can be used to minimize the divergence of the teacher and student predictive distributions.

4.1. Teacher Model

We train the teacher model, implemented with two sequential layers before classification, to discriminate between speech and

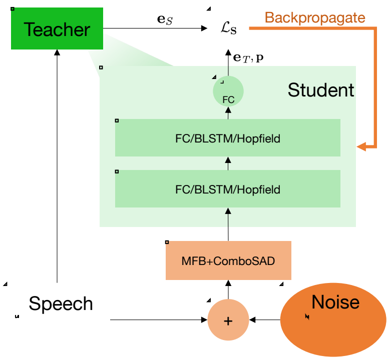

Figure 1: *Teacher-student approach for VAD using noise emulation.*

silence frames using a supervised deep learning framework. The loss function of the teacher model is the binary cross entropy criterion:

$$\mathcal{L}_{\text{vad}}(\mathbf{y}, \mathbf{p}) = -\sum_{i \in I} y_i \log(p_i) \tag{1}$$

where p_i is the i-th binary prediction generated by the model for the i-th feature frame in a segment of audio with y_i being the corresponding label generated through forced-alignment.

4.2. Student Model

To train the student model, the teacher parameters are copied before further training. We use a composite loss that combines the cross entropy loss to predict speech activity with a *mean square error* (MSE) loss that quantifies the distance between the teacher and student embeddings. The signal representation embedded at the input of the final classification layer of a DNN encodes useful, global information about the output distribution given the input. However, compressing this vector into a scalar during a many-to-one classification destroys a portion of this information. Since the prior produced by our teacher for speech predictions also involves the calculation of this final layer embedding, this information can be leveraged to regularize our VAD representation in the presence of noise using the paired, uncorrupted representation. This can be interpreted as a denoising effect as representations generated by noisy data are penalized for straying from the clean representation. This technique minimizes the MSE between the embedding produced by the teacher ($\mathbf{e}_T$) and student ($\mathbf{e}_S$) during training, where J in Equation 2 is the dimension of the embeddings. We use the output of the last recurrent network as our embedding.

$$\mathcal{L}_{\text{emb}}(\mathbf{e}_S, \mathbf{e}_T) = \frac{1}{J}\|\mathbf{e}_S - \mathbf{e}_T\|_2^2 \tag{2}$$

We train a student model by combining these loss functions with the goal of producing an optimal VAD performance on a non-ideal domain. We use the hyperparameter value $\alpha = 0.2$.

$$\mathcal{L}_S(\mathbf{e}_S, \mathbf{e}_T, \mathbf{y}, \mathbf{p}) = \alpha\mathcal{L}_{\text{emb}}(\mathbf{e}_S, \mathbf{e}_T) + (1-\alpha)\mathcal{L}_{\text{vad}}(\mathbf{y}, \mathbf{p}) \tag{3}$$

4.3. Temporal Model

VAD is a time series problem where considering the predictions from previous frames can benefit the voice activity prediction of future frames. Providing a temporal context to our model can help avoiding noisy transitions in the prediction between speech and silence. We implement our approach with two alternative *recurrent neural networks* (RNN). The first approach consists of BLSTM networks. While this model is not causal, as it considers a backward pass, the model can be easily implemented with LSTM if needed. The second approach is the CS-Hopfield network [11]. Ramsauer *et al.* [11] proposed a modified Hopfield energy function and update rule of the classical binary Hopfield network, finding an equivalence between this method and the attention mechanism of the Transformer [32]. The benefits of the modern CS-Hopfield network has been demonstrated in several applications [33, 34], so we hypothesize that it may be useful on VAD tasks. Furthermore, this approach is also more computationally efficient than BLSTM, which is important for VAD systems.

4.4. Implementation

The acoustic feature is a 26 *Mel-filterbank* (MFB) vector. We also consider adding the 5 hand-crafted features used in the ComboSAD proposed by Sadjadi and Hansen [3]: harmonicity, clarity, prediction gains, periodicity, and perceptual spectral flux. These features are extracted with an analysis window of 20ms with a stride of 10ms.

As shown in Figure 1, our proposed approach is implemented with two sequential layers (BLSTM or Hopfield networks). We formulate the problem as a many-to-one task, where an analysis window is used to predict the speech activity on the central frame. The default value for the windows size is 11 frames, but this parameter is studied in Table 4. We fix the number of parameters used for each model by adjusting the hidden state dimensionality for the BLSTM or the CS-Hopfield networks. The BLSTM is implemented with 128 nodes, and the CS-Hopfield network has a hidden dimension of 2,238. These configurations results in approximately 0.6M parameters. The recurrent layers are implemented with ReLU. We use layer normalization [35] to regularize model parameters and speed up convergence. Then, we add a fully connected layer with a single node, using a sigmoid activation. All other parameters for the Hopfield network such as the number of heads and head scaling are set according to the defaults set by Ramsauer *et al.* [11]. All models are optimized with the ADAM optimizer with a static learning rate of 0.00001. The models are trained on the train set, maximizing the performance on the validation set. The best models are then evaluated on the test set.

5. Experiments

Due to the natural class imbalance present due to the random sampling of a smartphone audio stream, we quantify the performance of our system in the precision-recall space. Flach and Kull [36] showed that naively plotting the precision and recall of the system for unbalanced data according to thresholds presents several issues. Instead, they measured performance with the *precision-recall-gain* (PRG) curve. The area under the *PRG-curve* (AUPRG) is analogous to a scaled version of the expected F1-score. Therefore, we assess model performance using either the AUPRG or F1-score.

First, we provide evidence of performance gains via domain emulation during model adaptation. This is accomplished by showing how model performance varies as training noise is varied across a set of speech domains. Table 1 shows the performance of three teacher-student model pairs when the student is trained on various noise conditions. All test cases are generated by applying the listed noise type to the same test set of clean speech drawn from the CRSS-4English-14 corpus. For simplicity, we only consider for this analysis the BLSTM-based T-S model using MFB features. When training the student, the corrupted training corpus is mixed at a fixed SNR of 0dB and trained for 4 epochs. This is true for all other experiments. Table 1 shows that the performance increases in all test conditions that match the noise type used to train the student model. Interestingly, we see generalization of some models to other noise

Table 1: *Performance of the BLSTM-based T-S model when the student models is trained and tested with different noise conditions. Performance is measured with AUPRG values. The table reports the results for the teacher (T) and student (S) models. Cases with positive transfer are highlighted with* ***bold*** *text.*

Test	White 0dB		Babble 0dB		CHiME5 0dB	
Train	T	S	T	S	T	S
CRSS-4English-14	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**

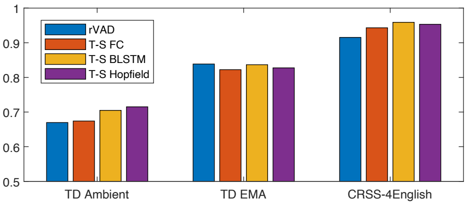

Figure 2: *F1 score for the baseline and proposed models when the student model is trained with the CHIME5 and TD Noise conditions.*

conditions, especially in the case of additive simulated babble noise and noise from the CHiME5 corpus. This positive transfer on cases with mismatched noise conditions may be attributed to some similarity between the train noise condition and the test set. We also see that model performance is also dependent on the similarities between the noise power of the train and test sets, with increased performance at a matched SNR. Note that the performance on clean conditions is expected to go down in all cases as the model parameters are shifted away from an ideal solution for clean speech. For the rest of the evaluation, we train the student models using noise from the TD-noise and the CHiME5 noise sets, which are the closest noise conditions to our target application.

Next, we evaluate the proposed models with real world performance in terms of F1 score by considering model predictions at a fixed threshold of 0.5 to obtain a binary VAD decision. We evaluate within the target domain, considering the TD-Ambient, and TD-EMA sets. We also consider the clean conditions of the CRSS-4English-14 corpus. All our models are trained with MFB and ComboSAD features. As a baseline, we compare our approaches using the *robust VAD* (rVAD) method proposed by Tan *et al.* [6], which is a speech enhancement-based model. We also implement our approach without modeling the temporal information, replacing the recurrent networks with fully connected layers (T-S FC). Figure 2 shows that our models outperform the baseline method on both clean and noisy target domains. The results shows that our models implemented with recurrent networks are better than the model implemented with fully connected networks, showing that modeling temporal information leads to improved VAD performance. The differences are particularly clear in the TD-Ambient conditions, where the recurrent architectures can more readily capture the temporal dynamics of longer audio segments.

Table 2: *Ablation study on the T-S model to assess transfer knowledge benefits. Performance is measured with AUPRG values. The models are trained with and without the T-S approach.*

Model	Test	Without T-S	With T-S
T-S BLSTM	CRSS-4English-14	0.989	**0.990**
T-S BLSTM	TD-EMA	0.868	**0.875**
T-S BLSTM	TD-Ambient	0.750	**0.766**

Table 3: *Ablation study on the BLSTM-based T-S model to assess the benefit of adding ComboSAD features. Performance is measured with AUPRG values for the teacher (T) and student (S) models. Positive transfer cases are marked in* ***bold***.

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**

Table 4: *Analysis of the proposed Hopfield and BLSTM based S-T models using different temporal window sizes. Performance is measured with AUPRG values. The table reports the results for the teacher (T) and student (S) models.*

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

We conduct ablation studies that show the benefits of (1) using the teacher-student model, and (2) adding the ComboSAD features. First, we compare our BLSTM-based T-S model with a model trained with the cross entropy criterion on noisy speech using the same DNN architecture without the teacher model parameters. Table 2 indicates that transferring knowledge to the students leads to consistent improvements, especially for the most challenging test domain (TD-Ambient). Second, we evaluate the benefits of augmenting our feature vector with ComboSAD features. We implement our BLSTM-based T-S model with and without the five extra features used in the ComboSAD framework (Sec. 4.4). Table 3 shows that combining MFB and ComboSAD features results in consistent performance gains.

Finally, we conduct an analysis of the performance when the evaluation feature window is varied. We train the systems with MFB and ComboSAD features, testing the performance on the TD-Ambient set. Table 4 shows the results of the BLSTM and CS-Hopfield networks trained with a feature window size of 5, 11, or 61 consecutive feature frames. Increasing the window size generally increases performance. We see positive transfer for both models. The BLSTM T-S model exhibits a clear trend where longer analysis windows leads to better performance. This trend is not as clear with the Hopfield model, where using 61 feature frames results in negative transfer.

6. Conclusions

This paper proposed a novel VAD training method for domain adaptation toward a real world healthcare application. The approach consists of the teacher-student formulation with domain emulation via paired data. The approach leverages temporal information using recurrent networks implemented with either BLSTM or CS-Hopfield networks. When the noise conditions are emulated, the approach leads to important performance gains, particularly for unconstrained audio. In future work, we will compare these two architectures and their flexibility to be adapted to unconstrained conditions using transfer learning.

7. Acknowledgements

Study supported by NIH under grant 1R01MH122367-01.

8. References

[1] D. Liaqat, R. Wu, A. Gershon, H. Alshaer, F. Rudzicz, and E. de Lara, "Challenges with real-world smartwatch based audio monitoring," in *ACM Workshop on Wearable Systems and Applications (WearSys 2018)*, Munich, Germany, June 2018, pp. 54–59.

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

[3] 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.

[4] A. Benyassine, H. S. E. Shlomot, D. Massaloux, C. Lamblin, and J.-P. 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, September 1997.

[5] B. Sharma, R. Das, and H. Li, "Multi-level adaptive speech activity detector for speech in naturalistic environments," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2015–2019.

[6] Z.-H. Tan, A. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, vol. 59, pp. 1–21, January 2020.

[7] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, Canada, May 2013, pp. 7378–7382.

[8] N. Kurpukdee, S. Boonkla, V. Chunwijitras, P. Sertsi, and S. Kasuriya, "Improving voice activity detection by using denoising-based techniques with convolutional LSTM," in *International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2019)*, Chiang Mai, Thailand, October-November 2019, pp. 1–6.

[9] Y. Zhuang, S. Tong, M. Yin, Y. Qian, and K. Yu, "Multi-task joint-learning for robust voice activity detection," in *International Symposium on Chinese Spoken Language Processing (ISCSLP 2016)*, Tianjin, China, October 2016, pp. 1–5.

[10] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1181–1185, August 2018.

[11] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, L. Gruber, M. Holzleitner, M. Pavlović, G. Kjetil Sandve, V. Greiff, D. Kreil, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "Hopfield networks is all you need," *ArXiv e-prints (arXiv:2008.02217)*, pp. 1–10, July 2020.

[12] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Odyssey 2012*, Singapore, June 2012, pp. 291–297.

[13] J. Hansen, A. Sangwan, A. Joglekar, A. Bulut, L. Kaushik, and C. Yu, "Fearless Steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 2758–2762.

[14] S. Thomas, G. Saon, M. V. Segbroeck, and S. S. Narayanan, "Improvements to the IBM speech activity detection system for the DARPA RATS program," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, South Brisbane, QLD, Australia, April 2015, pp. 4500–4504.

[15] A. Vafeiadis, E. Fanioudakis, I. Potamitis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Two-dimensional convolutional recurrent neural networks for speech activity detection," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2045–2049.

[16] Y. Zhang, Z.-M. Tang, Y.-P. Li, , and Y. Luo, "A hierarchical framework approach for voice activity detection and speech enhancement," *The Scientific World Journal*, vol. 723643, pp. 1–8, May 2014.

[17] L. Ferrer, M. Graciarena, and V. Mitra, "A phonetically aware system for speech activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5710–5714.

[18] F. Tao and C. Busso, "End-to-end audiovisual speech activity detection with bimodal recurrent neural models," *Speech Communication*, vol. 113, pp. 25–35, October 2019.

[19] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselý, and P. Matějka, "Developing a speech activity detection system for the DARPA RATS program," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1969–1972.

[20] Y. Chen, H. Dinkel, M. Wu, and K. Yu, "Voice activity detection in the wild via weakly supervised sound event detection," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 3665–3669.

[21] F. Tao and C. Busso, "End-to-end audiovisual speech recognition system with multi-task learning," *IEEE Transactions on Multimedia*, vol. 23, pp. 1–11, January 2021.

[22] Z. Meng, J. Li, Y. Gong, and B. Juang, "Adversarial teacher-student learning for unsupervised domain adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5949–5953.

[23] J. Wong and M. Gales, "Sequence student-teacher training of deep neural networks," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2761–2765.

[24] ——, "Multi-task ensembles with teacher-student training," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2017)*, Okinawa, Japan, December 2017, pp. 84–90.

[25] J. Li, M. Seltzer, X. Wang, R. Zhao, and Y. Gong, "Large-scale domain adaptation via teacher-student learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 2386–2390.

[26] K. Sridhar and C. Busso, "Ensemble of students taught by probabilistic teachers to improve speech emotion recognition," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 516–520.

[27] J. Mote and D. Fulford, "Ecological momentary assessment of everyday social experiences of people with schizophrenia: A systematic review," *Schizophrenia Research*, vol. 216, pp. 56–68, February 2020.

[28] F. Tao and 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.

[29] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 498–502.

[30] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 1561–1565.

[31] N. Krishnamurthy and J. H. L. Hansen, "Babble noise: Modeling, analysis, and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1394–1407, September 2009.

[32] A. Vaswani ***et al.***, "Attention is all you need," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 5998–6008.

[33] M. Widrich ***et al.***, "Modern Hopfield networks and attention for immune repertoire classification," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, Virtual, December 2020, pp. 18 832–18 845.

[34] Z. Shi, L. Liu, R. Liu, X. Mi, and K. Murase, "LoRRaL: Facial action unit detection based on local region relation learning," *ArXiv e-prints (arXiv:2009.10892)*, pp. 1–11, September 2020.

[35] L. Ba, J. Kiros, and G. Hinton, "Layer normalization," *ArXiv e-prints (arXiv:1607.06450)*, pp. 1–12, July 2016.

[36] P. Flach and M. Kull, "Precision-recall-gain curves: Pr analysis done right," in *Advances in neural information processing systems (NIPS 2015)*, vol. 28, Montreal, Canada, December 2015, pp. 1–9.