

# A Multimodal Approach for Dementia Detection from Spontaneous Speech with Tensor Fusion Layer

Loukas Ilias, Dimitris Askounis, and John Psarras

*Decision Support Systems Laboratory*

*School of Electrical and Computer Engineering*

*National Technical University of Athens*

15780 Athens, Greece

{lilias,askous,john}@epu.ntua.gr

**Abstract**—Alzheimer’s disease (AD) is a progressive neurological disorder, meaning that the symptoms develop gradually throughout the years. It is also the main cause of dementia, which affects memory, thinking skills, and mental abilities. Nowadays, researchers have moved their interest towards AD detection from spontaneous speech, since it constitutes a time-effective procedure. However, existing state-of-the-art works proposing multimodal approaches do not take into consideration the inter- and intra-modal interactions and propose early and late fusion approaches. To tackle these limitations, we propose deep neural networks, which can be trained in an end-to-end trainable way and capture the inter- and intra-modal interactions. Firstly, each audio file is converted to an image consisting of three channels, i.e., log-Mel spectrogram, delta, and delta-delta. Next, each transcript is passed through a BERT model followed by a gated self-attention layer. Similarly, each image is passed through a Swin Transformer followed by an independent gated self-attention layer. Acoustic features are extracted also from each audio file. Finally, the representation vectors from the different modalities are fed to a tensor fusion layer for capturing the inter-modal interactions. Extensive experiments conducted on the ADReSS Challenge dataset indicate that our introduced approaches obtain valuable advantages over existing research initiatives reaching Accuracy and F1-score up to 86.25% and 85.48% respectively.

**Index Terms**—Alzheimer’s Disease, Dementia, log-Mel spectrogram, BERT, Swin Transformer, Gated Self-Attention, Tensor Fusion Layer

## I. INTRODUCTION

Alzheimer’s disease (AD) constitutes a progressive brain disorder, which is ranked as the seventh leading cause of death in the United States and is the main cause of dementia among older adults [1]. Dementia comes with a group of symptoms and affects memory, behaviour, thinking and social abilities [2]. Nowadays, researchers use spontaneous speech for detecting AD patients, since the detection of AD patients through Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), etc. requires access to medical centers, and demands time and money. Recently, there have been proposed shared tasks [3], [4], where the researchers can introduce models for detecting dementia from spontaneous speech.

Although there have been proposed several studies, which adopt multimodal approaches by exploiting both speech and transcripts, there are still substantial limitations. More specifically, existing research initiatives add or concatenate the

representation vectors obtained by different modalities [5] during training, treating equally each modality and consequently obtaining suboptimal performance. In addition, research works employ early [6], [7] and late fusion approaches [8]–[12]. Employing early fusion approaches means that the representation vectors of the different modalities are concatenated at the input level, while exploiting late fusion approaches means that multiple models are trained for each modality separately and the final result is obtained either by a straightforward majority-vote approach or by combining the results of each classifier through a weighted manner. Except for increasing the training time, none of these approaches captures the inter- and intra-modal interactions.

Motivated by these limitations, we introduce deep neural networks, which are trained in an end-to-end trainable manner and take into account both the inter- and intra-modal interactions. First, we convert each audio file into a log-Mel spectrogram, its delta, and delta-delta. Thus, we create an image consisting of three channels. We pass each transcript through a BERT model followed by a self-attention mechanism incorporating a gating model for capturing the intra-modal interactions. Similarly, we pass each image through a Swin Transformer followed by an independent self-attention mechanism with a gate model. We extract also acoustic features from the audio files. Next, the representation vectors from the three different modalities are passed to a tensor fusion layer, which captures effectively the inter-modal interactions. The output of the tensor fusion layer is passed through a series of dense layers for classifying the subject into an AD patient or a non-AD one.

Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first study proposing a Swin Transformer in the task of dementia detection from spontaneous speech.
- We introduce the tensor fusion layer, which can capture the inter-modal interactions. There is no prior work utilizing a tensor fusion layer for the task of dementia detection from spontaneous speech.
- Our proposed architectures achieve comparable results to existing state-of-the-art approaches.

## II. TASK AND DATA

Given a labelled dataset consisting of AD and non-AD patients, the task is to identify if an audio file along with its transcript belongs to an AD patient or to a non-AD one.

We use the ADReSS Challenge dataset for conducting our experiments [3]. This dataset comprises speech recordings and transcripts of spoken picture descriptions elicited from participants through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [13]. We opted for this dataset, since it minimizes biases often overlooked in evaluations of AD detection methods, including repeated occurrences of speech from the same participant, variations in audio quality, and imbalances of gender and age distribution. It consists of a train and a test set. The train set consists of 54 AD and 54 non-AD patients, while the test set includes 24 AD and 24 non-AD patients. One further advantage of using the ADReSS Challenge dataset is pertinent to its speaker independent nature.

## III. PREDICTIVE MODELS

**BERT+Gated Self-Attention:** We use the Python library *PyLangAcq* [14] for having access to the transcripts, since the manual transcripts have been annotated using the CHAT coding system [15]. We pass each transcript through a BERT model [16]. Let  $Z \in \mathcal{R}^{N \times d}$  be the output of the BERT model. After this, we add positional encodings to the output of BERT [17]. Next, we pass  $Z$  through a gated self-attention mechanism [18], as described by the equations below:

$$Q = Z, K = Z, V = Z \quad (1)$$

Next, we adopt the gating model introduced by [18] as follows:

$$M = \sigma \left( FC_q^g \left( FC_q^g(Q) \odot FC_k^g(K) \right) \right) \quad (2)$$

where  $FC_q^g, FC_k^g \in \mathbb{R}^{d \times d_g}$ ,  $FC^g \in \mathbb{R}^{d_g \times 2}$  are three fully-connected layers, and  $d_g$  denotes the dimensionality of the projected space corresponding to 64 units.  $\odot$  denotes the element-wise product function and  $\sigma$  the sigmoid function. In addition,  $M \in \mathbb{R}^{N \times 2}$  corresponds to the two masks  $M_q \in \mathbb{R}^N$  and  $M_k \in \mathbb{R}^N$  for the features  $Q$  and  $K$  respectively.

Next, the two masks  $M$  and  $K$  are tiled to  $\tilde{M}_q, \tilde{M}_k \in \mathbb{R}^{N \times d}$  and then used for computing the attention map as following:

$$A^g = \text{softmax} \left( \frac{\left( (Q \odot \tilde{M}_q) \left( K \odot \tilde{M}_k \right)^T \right)}{\sqrt{d}} \right) \quad (3)$$

$$H = A^g V \quad (4)$$

Next, we pass  $H$  through an average pooling layer followed by a dense layer with 128 units. Let  $z^t \in \mathcal{R}^{128}$  be the representation vector corresponding to the transcripts, i.e., textual modality.

**Swin Transformer+Gated Self-Attention:** We use the Python library *librosa* [19] and convert each audio file into an image consisting of three channels, namely log-Mel spectrogram, its delta, and its delta-delta. For all the experiments conducted, we

use 224 Mel bands, hop length equal to 1024, and a Hanning window. Each image is resized to  $(224 \times 224)$  pixels. Next, each image is passed through a Swin Transformer [20]. Let  $Z \in \mathcal{R}^{T \times d}$  be the output of the Swin Transformer. After this, we add positional encodings to the outputs of the Swin Transformer. Next, we pass  $Z$  through an independent gated self-attention, as described via the equations 1-4. The output of the gated self-attention mechanism is passed through an average pooling layer followed by a dense layer consisting of 32 units. Let  $z^v \in \mathcal{R}^{32}$  be the representation vector corresponding to the visual modality.

**eGeMAPS:** We use the openSMILE Toolkit [21] and extract the acoustic feature set, namely eGeMAPSv02 (functionals), per audio file. In this way, each feature set per audio file has a dimensionality of 88d. Then, we use a dense layer and project the dimensionality to 32. Let  $z^\alpha \in \mathcal{R}^{32}$  be the representation vector for the acoustic modality.

**Tensor Fusion Layer:** We pass  $z^t, z^v, z^\alpha$  through a tensor fusion layer [22] for capturing the inter-modal interactions, as described via the equations below.

$$\left\{ (z^t, z^v, z^\alpha) \mid z^t \in \begin{bmatrix} z^t \\ 1 \end{bmatrix}, z^v \in \begin{bmatrix} z^v \\ 1 \end{bmatrix}, z^\alpha \in \begin{bmatrix} z^\alpha \\ 1 \end{bmatrix} \right\} \quad (5)$$

$$z^m = \begin{bmatrix} z^t \\ 1 \end{bmatrix} \otimes \begin{bmatrix} z^v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} z^\alpha \\ 1 \end{bmatrix} \quad (6)$$

, where  $\otimes$  indicates the outer product between the vectors. Let  $z^m \in \mathcal{R}^{129 \times 33 \times 33}$  be the output of the tensor fusion layer.

**Output Layer:** We pass  $z^m$  to a dropout layer with a rate of 0.6 followed by a dense layer of 128 units with a ReLU activation function. We use a dropout layer with a rate of 0.2. Finally, we use a dense layer consisting of two units.

The proposed architecture (**Transcript+Image+Acoustic**) is illustrated in Fig. 1.<sup>1</sup>

## IV. EXPERIMENTS

### A. Comparison with state-of-the-art approaches

We compare our introduced architectures with the following research works, which have proposed multimodal approaches and have reported the results on the ADReSS test set: **(1)** Fusion Maj. (3-best) [9], **(2)** System 3: Phonemes and Audio [23], **(3)** Fusion of system [6], **(4)** Bimodal Network (Ensembled Output) [24], **(5)** GFI, NUW, Duration, Character 4-grams, Suffixes, POS tag, UD [7], **(6)** Acoustic & Transcript [10], **(7)** Dual BERT (Concat/Joint, BERT large) [5], **(8)** Model C [25], **(9)** Majority vote (NLP + Acoustic) [26], **(10)** LSTM with Gating (Acoustic + Lexical + Dis) [27], **(11)** Ensemble [11], and **(12)** Attempt 4 [12].

<sup>1</sup>We experiment with multiple inputs, namely **Transcript+Acoustic** and **Transcript+Image**.

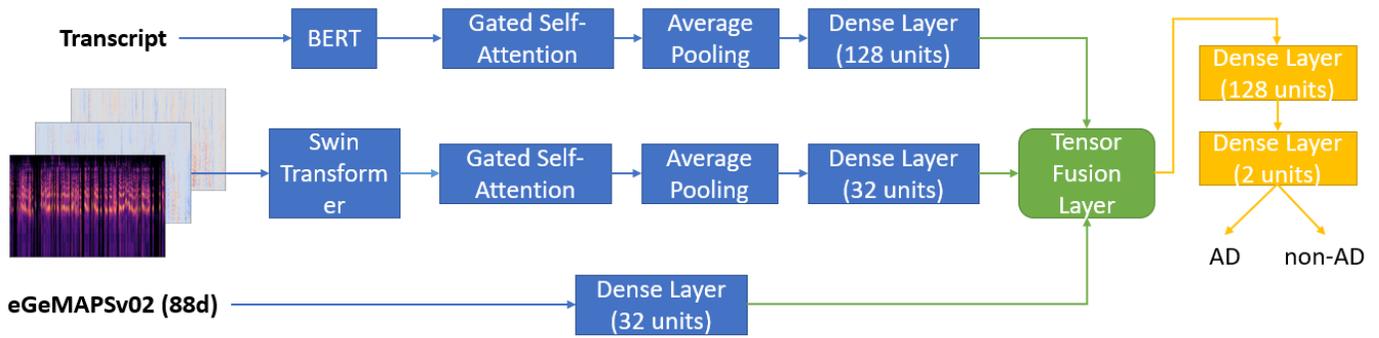


Fig. 1: Our introduced architecture

### B. Experimental Setup

We follow a similar training strategy to the one adopted by [5]. Firstly, we divide the train set provided by the Challenge into a train and a validation set (65%-35%). Next, we train the proposed architectures five times with an Adam optimizer and a learning rate of  $1e-5$ . We minimize the cross-entropy loss function. We apply *ReduceLROnPlateau*, where we reduce the learning rate by a factor of 0.1, if the validation loss has stopped decreasing for three consecutive epochs. Also, we apply *EarlyStopping* and stop training, if the validation loss has stopped decreasing for six consecutive epochs. We test the proposed models using the test set provided by the Challenge. All models are created using the PyTorch library [28]. We use the Swin Transformer<sup>2</sup> and the BERT base uncased version from the Transformers library [29]. All experiments are conducted on a single Tesla P100-PCIE-16GB GPU.

### C. Evaluation Metrics

Accuracy, Precision, Recall, F1-Score, and Specificity have been used for evaluating the results of the introduced architectures. These metrics have been computed by regarding the dementia class as the positive one. We report the average and standard deviation of these metrics over five runs.

## V. RESULTS

Table I reports the results of our introduced architectures. In this table, we also compare the results of our approaches with the multimodal state-of-the-art ones.

Regarding our proposed models, one can observe that *Transcript+Image+Acoustic* constitutes our best performing model in terms of Precision, F1-score, Accuracy, and Specificity. Although there are models surpassing *Transcript+Image+Acoustic* in Recall, *Transcript+Image+Acoustic* surpasses them in F1-score, which is the weighted average of Precision and Recall. In addition, F1-score is a more important metric than Specificity in health-related problems, since high Specificity and low F1-score means that AD patients are misdiagnosed as non-AD ones. *Transcript+Image+Acoustic* obtains an Accuracy and F1-score of 86.25% and 85.48% respectively. It outperforms the other introduced models in Accuracy by 1.67-3.75%, in

F1-score by 0.40-2.14%, in Precision by 7.37-11.29%, and in Specificity by 10.00-14.16%.

In comparison with the existing research initiatives, one can observe that our best performing model, namely *Transcript+Image+Acoustic*, outperforms the other approaches in terms of Accuracy and F1-score. Specifically, *Transcript+Image+Acoustic* surpasses the research initiatives in Accuracy by 1.05-13.33%. Also, it surpasses the state-of-the-art approaches in F1-score by 0.08-15.72%. With regards to *Transcript+Image*, it surpasses the research initiatives, except [9], in Accuracy by 1.58-11.66% and in F1-score by 2.08-15.32%. In terms of *Transcript+Acoustic*, it outperforms the multimodal state-of-the-art approaches, except [5], [9], [11], [26], in Accuracy by 1.25-9.58%, while it outperforms the research initiatives, except [9], in F1-score by 0.34-13.58%.

Therefore, one can observe that our introduced advanced transformer-based networks along with techniques capturing the inter-modal interactions obtain comparable or better performance than existing research initiatives, while they require less time for training.

## VI. LIMITATIONS

**GPU resources - Hyperparameter Tuning:** Due to limited access to GPU resources, we did not perform hyperparameter tuning. To be more precise, the number of units in the dense layers, the number of the dense layers, the learning rate, etc. were not chosen via tuning. Performing hyperparameter tuning often leads to an increase in the classification performance.

**Explainability - Misclassifications:** In this study, we did not apply explainability techniques for providing to the reader an error analysis, i.e., find the samples that our introduced model fails to classify correctly. In the future, we aim to apply explainability techniques, such as integrated gradients, for rendering the introduced model explainable. In this way, we will be able to understand why our model fails to classify some subjects correctly.

**Generalizability:** In this paper, we used one dataset for conducting our experiments. For proving the generalizability of our introduced model, we aim to exploit more datasets in the future.

<sup>2</sup>microsoft/swin-tiny-patch4-window7-224

TABLE I: Performance comparison among proposed models and state-of-the-art approaches on the ADReSS Challenge test set. Reported values are mean  $\pm$  standard deviation. Results are averaged across five runs.

Architecture	Evaluation Metrics				
	Prec	Rec	F1-score	Acc	Spec
<b>Multimodal state-of-the-art approaches</b>					
[9]	-	-	85.40	85.20	-
[23]	81.82	75.00	78.26	79.17	83.33
[6]	94.12	66.67	78.05	81.25	95.83
[24]	89.47	70.83	79.07	81.25	91.67
[7]	-	-	-	77.08	-
[10]	70.00	88.00	78.00	75.00	63.00
[5]	83.04 $\pm 3.97$	83.33 $\pm 5.89$	82.92 $\pm 1.86$	82.92 $\pm 1.56$	82.50 $\pm 5.53$
[25]	78.94	62.50	69.76	72.92	83.33
[26]	-	-	-	83.00	-
[27]	81.82	75.00	78.26	79.17	83.33
[11]	83.00	83.00	83.00	83.00	-
[12]	-	-	-	79.17	-
<b>Our introduced models</b>					
Transcript+	79.59	87.50	83.34	82.50	77.50
Acoustic	$\pm 2.66$	$\pm 2.64$	$\pm 2.35$	$\pm 2.49$	$\pm 3.33$
Transcript+	83.51	87.50	85.08	84.58	81.66
Image	$\pm 4.25$	$\pm 3.73$	$\pm 1.84$	$\pm 2.49$	$\pm 7.73$
Transcript+	90.88	80.83	85.48	86.25	91.66
Image+Acoustic	$\pm 3.60$	$\pm 2.04$	$\pm 0.76$	$\pm 1.02$	$\pm 3.73$

## VII. CONCLUSION AND FUTURE WORK

In this paper, we present the first study for the task of dementia detection from spontaneous speech, which employs a Swin Transformer and a tensor fusion layer for extracting visual information and capturing the inter-modal interactions respectively. Specifically, the representation vectors corresponding to the three modalities, i.e., textual, visual, and acoustic, are passed to a tensor fusion layer, which models the inter-modal interactions. Our model achieves comparable performance to existing research initiatives reaching Accuracy and F1-score up to 86.25% and 85.48% respectively.

In the future, we aim to create an application, which will incorporate the introduced model and detect AD patients. In addition, we plan to propose advanced transfer learning techniques, in order to employ our introduced models in other tasks, including the detection of Parkinson’s disease. Also, explainability techniques, including integrated gradients, are one of our plans for understanding the reasons of misclassifications.

## REFERENCES

[1] National Institute on Aging, “Alzheimer’s Disease Fact Sheet,” Available online at: <https://www.nia.nih.gov/health/alzheimers-disease-fact-sheet>, 2021.

[2] S. Gauthier, P. Rosa-Neto, J. A. Morais, and C. Webster, “World alzheimer report 2021: Journey through the diagnosis of dementia,” 2021.

[3] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Alzheimer’s Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge,” in *Proc. Interspeech 2020*, 2020, pp. 2172–2176.

[4] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, “Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge,” in *Proc. Interspeech 2021*, 2021, pp. 3780–3784.

[5] Y. Zhu, X. Liang, J. A. Batsis, and R. M. Roth, “Exploring deep transfer learning techniques for alzheimer’s dementia detection,” *Frontiers in Computer Science*, vol. 3, p. 22, 2021. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fcomp.2021.624683>

[6] A. Pompili, T. Rolland, and A. Abad, “The INESC-ID Multi-Modal System for the ADReSS 2020 Challenge,” in *Proc. Interspeech 2020*, 2020, pp. 2202–2206.

[7] M. Martinc and S. Pollak, “Tackling the ADReSS Challenge: A Multimodal Approach to the Automated Recognition of Alzheimer’s Dementia,” in *Proc. Interspeech 2020*, 2020, pp. 2157–2161.

[8] A. Mittal, S. Sahoo, A. Datar, J. Kadiwala, H. Shalu, and J. Mathew, “Multi-modal detection of alzheimer’s disease from speech and text,” 2021.

[9] N. Cummins, Y. Pan, Z. Ren, J. Fritsch, V. S. Nallanthighal, H. Christensen, D. Blackburn, B. W. Schuller, M. Magimai-Doss, H. Strik *et al.*, “A comparison of acoustic and linguistics methodologies for alzheimer’s dementia recognition,” in *Interspeech 2020*. ISCA-International Speech Communication Association, 2020, pp. 2182–2186.

[10] R. Pappagari, J. Cho, L. Moro-Velázquez, and N. Dehak, “Using State of the Art Speaker Recognition and Natural Language Processing Technologies to Detect Alzheimer’s Disease and Assess its Severity,” in *Proc. Interspeech 2020*, 2020, pp. 2177–2181.

[11] U. Sarawgi, W. Zulfikar, N. Soliman, and P. Maes, “Multimodal Inductive Transfer Learning for Detection of Alzheimer’s Dementia and its Severity,” in *Proc. Interspeech 2020*, 2020, pp. 2212–2216.

[12] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, “Automated Screening for Alzheimer’s Dementia Through Spontaneous Speech,” in *Proc. Interspeech 2020*, 2020, pp. 2222–2226.

[13] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, “The Natural History of Alzheimer’s Disease: Description of Study Cohort and Accuracy of Diagnosis,” *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 06 1994. [Online]. Available: <https://doi.org/10.1001/archneur.1994.00540180063015>

[14] J. L. Lee, R. Burkholder, G. B. Flinn, and E. R. Coppess, “Working with chat transcripts in python,” Department of Computer Science, University of Chicago, Tech. Rep. TR-2016-02, 2016.

[15] B. MacWhinney, “The CHILDES Project: Tools for Analyzing Talk (third edition): Volume I: Transcription format and programs, Volume II: The database,” *Computational Linguistics*, vol. 26, no. 4, pp. 657–657, 12 2000. [Online]. Available: <https://doi.org/10.1162/coli.2000.26.4.657>

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[18] Z. Yu, Y. Cui, J. Yu, D. Tao, and Q. Tian, “Multimodal unified attention networks for vision-and-language interactions,” *arXiv preprint arXiv:1908.04107*, 2019.

[19] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, A. Weiss, D. Hereñú, F.-R. Stöter, P. Friesch, M. Vollrath, T. Kim, and Thassilo, “librosa/librosa: 0.9.1,” Feb. 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6097378>

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10012–10022.

[21] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM ’10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: <https://doi.org/10.1145/1873951.1874246>

[22] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in*

*Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114. [Online]. Available: <https://aclanthology.org/D17-1115>

- [23] E. Edwards, C. Dognin, B. Bollepalli, and M. Singh, “Multiscale System for Alzheimer’s Dementia Recognition Through Spontaneous Speech,” in *Proc. Interspeech 2020*, 2020, pp. 2197–2201.
- [24] J. Koo, J. H. Lee, J. Pyo, Y. Jo, and K. Lee, “Exploiting Multi-Modal Features from Pre-Trained Networks for Alzheimer’s Dementia Recognition,” in *Proc. Interspeech 2020*, 2020, pp. 2217–2221.
- [25] P. Mahajan and V. Baths, “Acoustic and language based deep learning approaches for alzheimer’s dementia detection from spontaneous speech,” *Frontiers in Aging Neuroscience*, vol. 13, p. 20, 2021. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnagi.2021.623607>
- [26] Z. Shah, J. Sawalha, M. Tasnim, S.-a. Qi, E. Stroulia, and R. Greiner, “Learning language and acoustic models for identifying alzheimer’s dementia from speech,” *Frontiers in Computer Science*, vol. 3, p. 4, 2021. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fcomp.2021.624659>
- [27] M. Rohanian, J. Hough, and M. Purver, “Multi-Modal Fusion with Gating Using Audio, Lexical and Disfluency Features for Alzheimer’s Dementia Recognition from Spontaneous Speech,” in *Proc. Interspeech 2020*, 2020, pp. 2187–2191.
- [28] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [29] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>