IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Alzheimer's Dementia Detection from Audio and Language Modalities in Spontaneous Speech

Edward L. Campbell[1], *Laura Docío-Fernández*[1], *Javier Jiménez-Raboso*[2], *Carmen García-Mateo*[1]

[1]GTM research group, AtlanTTic Research Center, University of Vigo, Spain
[2]acceXible

ecampbell@gts.uvigo.es, ldocio@gts.uvigo.es, javij@accexible.com, carmen.garcia@uvigo.es

Abstract

Automatic detection of Alzheimer's dementia (AD) by speech processing is enhanced when features of both the acoustic waveform and the content are extracted. Audio and text transcription have been widely used in health-related tasks, as spectral and prosodic speech features, as well as semantic and linguistic content, convey information about various diseases. Hence, this paper describes and compares the performance of different Alzheimer's disease detection approaches based on both the patient's voice and message transcription. To this effect, five different individual systems are analysed: three of them are speech-based and the other two systems are text-based. Specifically, as speech-based systems the x-vector and i-vector paradigm to characterise speech, and a set of rhythmic-based hand-crafted features are proposed. And, for transcription analysis, two systems are proposed, one which uses pre-trained BERT models and the other which uses knowledge-based linguistic and language modelling features. Also, to examine if acoustic and content features are complementary intra-modality and inter-modality score fusion strategies are studied. Experiments in the framework of Interspeech 2020 ADReSS challenge show that the BERT-based system outperforms other individual systems for the AD detection task. Furthermore, the fusion of acoustic- and transcription-based systems provides the best result, suggesting that the two modalities are complementary to some extent.

Index Terms: Alzheimer's disease detection, i-vector, x-vector, speech fluency, BERT, score level fusion, ADReSS challenge

1. Introduction

The Alzheimer's disease (AD) is a neurodegenerative illness that represents the most common cause of dementia in the world. It is provoked by the damage of neurons involved in thinking, learning and memory. This disease has three main stages. In the first one, called preclinical, patients do not present clear symptoms because the brain initially compensate for them, enabling individuals to continue to function normally. The second one is defined as Mild Cognitive Impairment (MCI). In this stage, patients show greater cognitive decline then expected for their age, having problems to express and connect ideas. However, these changes may be only noticeable to family members and friends. The critical phase is the last one, which is called as dementia. It is characterized by noticeable memory, thinking and behavioural symptoms that impair a person's ability to function in daily life [1][2].

Common signs of AD are related to problems with uttering words; consequently, people with AD may have trouble following or joining a conversation, they may stop in the middle of a sentence and have no idea how to continue. As a result, analysis of speech and its transcription may represent a suitable mechanism for detecting the AD during the second or third stage of the disease [1] [3]. According to the literature[4][5][6], using information from the patient's voice, as well as from its transcription would ease the early AD detection task.

Different multimodal approaches for AD recognition have been proposed in the ADReSS challenge [7]. In them, speech is commonly represented using x-vectors, a large set of functionals obtained from low level descriptors, and other deep learning based speech representations. Using the speech transcriptions, best systems [8] are those based on deep language embeddings such as Bidirectional Encoder Representations from Transformers (BERT) [9].

In this work, three speech-based systems and two text-based systems are proposed for automatic distinction between individuals with and without AD. Those based on the speech signal are compound by four approaches to represent their spectral and prosodic content, namely: i-vector and x-vector embeddings, and rhythmic features. The first two use support vector machine (SVM) as classifiers and the last one uses a linear discriminant analysis (LDA) classifier. As for systems that use speech transcriptions, one is based on fine-tuning BERT model [9] for text classification, and the other one is based on features extracted by language modelling, using a SVM as classifier. Finally, an intra-modality and inter-modality score fusion strategy was done to improve the final results. All the individual systems, and their fusion, are evaluated on the AD recognition task within the framework of the ADReSS Challenge [7]. This challenge targets the AD detection using spontaneous speech. The data used in the Challenge consists of speech recordings, and their transcripts, corresponding to the description of the Cookie Theft picture. Specifically, it is a selection of Alzheimer and control patients from of the DementiaBank's Pitt corpus [1].

The rest of the paper is organized as follows. Sections 2 and 3 describe the speech-based and text-based systems, respectively. Section 4 outlines the experimental framework. The experimental results are exposed and discussed in Section 5. Finally, Section 6 draws some conclusions and future work.

2. Speech-based systems

In this section, a description of speech feature extraction strategies and classification methods is done, with special attention to different statistical features extracted from the patient's voice.

2.1. Speech embedding features

Speech embedding features are considered as state-of-art speech representation for speaker recognition application. These speech representations can be applied for AD detection,

[1]http://dementia.talkbank.org/

10.21437/IberSPEECH.2021-57

as long as they preserve those spectral patterns in the speaker's voice that allow the distinction between individuals with and without AD. In this paper, two strategies were analyzed, the first one uses the i-vector paradigm [10], and the second one uses an x-vector [11] based representation. The main characteristics of these approaches are briefly described below.

A. *i-vectors*

To extract the i-vectors a universal background model (UBM) and a total variability matrix $\mathbf{T}$ model must be trained. As speech parameterization these models use 13 perceptual linear prediction (PLP) cepstral coefficients, combined with two pitch-related features (F0 and voicing probability) [12]. This features are augmented with their delta and acceleration coefficients, leading to vectors of dimension 45. These features were chosen since that combination achieves a representation of speech that includes spectral information and prosodic features such as rhythm or intonation that are embedded in the fundamental frequency. The UBM has a diagonal covariance, 512-component gaussian mixture model (GMM), and was trained with data from outside this task; $\mathbf{T}$ was trained using the task training data. The dimension of the i-vectors was set to 125 and they were length-normalized.

B. *x-vectors*

A pretrained time delay deep neural network (TDNN) [2] with 5 time delay layers and two dense layers was used. The network was trained to discriminate between speakers, using the nnet3 neural network library of the Kaldi Speech Recognition Toolkit [13] on augmented VoxCeleb 1 and VoxCeleb 2 datasets [14]. The input to the TDNN are 30 mel-frequency cepstral coefficients (MFCC), and the embeddings are extracted from the first dense layer with a dimensionality of 512. The output of this layer (x-vector) is first projected using latent discriminant analysis (LDA) into a 200 dimensional space and then length-normalized.

Instead of extracting an i-vector or x-vector (embeddings) to represent the entire audio signal, a set of these vectors is obtained applying a sliding window. In this way, each audio file is represented by a certain number of embeddings, which are then used for classification. The optimal window length and overlap were tuned experimentally.

Both systems use as classifier an SVM with a linear kernel. Since a number of embeddings are extracted from each audio, there will also be a set of classification results (one for each embedding), which must be combined to obtain a patient's classification in AD or non-AD. In this work, the mean of the classifications was used as score for the final decision.

2.2. Speech fluency

The lack of speaking fluency is a common pattern in patient with AD, being the rhythm a viable clue to detect that behavior. However, prosodic information (e.g., mean energy) was also used because the correct pronunciation of words does not only depends on the rhythm but also on intonation, tone and stress. For this system, the selected parameters were based on [4] [15], and they are as follows:

- Number of syllables
- Rate of speech (syllables / original duration)
- Speaking duration
- Average fundamental frequency
- Median of fundamental frequency
- Minimum fundamental frequency
- Pronunciation posterior probability
- Average voice interval duration
- Average duration of pairs [3]
- Mean energy
- The ratio between the energy mean and its standard-deviation

The extraction process was done using the Python library My-Voice Analysis [4], the voice activity detection of the SIDEKIT software [16], and our own algorithms.

The classification process was done by the LDA algorithm, projecting the rhythmic feature vector into an one-dimensional space where every projected point represents a classification score.

3. Text-based systems

Two different text-based approaches were analysed. The first approach uses a pre-trained Bidirectional Encoder Representation from Transformers (BERT) sequence classification model [9]. The second approach manually extract linguistic information for creating input features for a classifier.

3.1. Text preprocessing

The transcripts contained in the dataset are in CHAT format [17], which facilitates speech annotation and analysis. They include the transcript of both the subject and the investigator in charge of the test, as well as additional non-speech annotations such as times, pauses, errors, morphological or syntactic analysis. This information is represented with special characters (such as "[//]" for pauses), and since they are very specific for this format they probably are not present in the BERT original tokenizer. We included these tokens in the BERT tokenizer, so a representation of them can be learned during fine-tuning.

Transcripts are divided into several interventions, i.e. sentences or parts of complete sentences with meaning, and this granularity has been maintained in the preprocessing. Metadata, researcher interventions and linguistic analysis included in the files have been removed. We keep both subject's words and some annotations from the transcription (for pauses, errors and subject's actions such as laughing) as input for the classification.

3.2. BERT model

The first approach consists on fine-tuning a pre-trained version of BERT at intervention level, by classifying if a given sentence belongs to an AD subject. By using the interventions as independent samples the training set is increased to a size of 1492 records from the 108 subjects. Then, the probability that the subject had AD given all his/her interventions is given by

$$p(\text{AD}) = \frac{\sum_i l_i s_i}{\sum_i l_i}, \tag{1}$$

where l_i is the length in tokens of the i-th intervention of the subject and s_i is its score (between 0 and 1) estimated by the

[2] http://kaldi-asr.org/models.html

[3] Pairs: consecutive voiced and unvoiced segments

[4] https://github.com/Shahabks/my-voice-analysis

model. With this weighted mean, more importance is given to longer interventions.

Input sequences are tokenized and padded to a maximum length of 40. We used the uncased version of BERT [18][5] for automated feature extraction. The resulting vector of dimension 768 is mapped to a final linear layer to perform the binary classification. A dropout rate of 0.3 is added to the last layer to prevent overfitting. The whole model is trained for 3 epochs, using a batch size of 16 and AdamW optimizer [19] with an initial learning rate of 5e-5 and linear scheduling.

3.3. Model based on linguistic features

In this approach, several linguistic features and indicators are built from subject's interventions and, using this feature vector as input, an SVM is trained. Unlike the previous method, the classification here is performed at subject level, taking the full transcript of each participant, and only subject's words are considered.

Previous works [20] have shown that certain linguistic features are useful for detecting AD using Cookie Theft test. Here 13 features have been built, grouped into 4 categories:

- Extension information such as the number of interventions, number of words per intervention and mean word length.
- Vocabulary richness, by measuring the number of unique words used by the subject.
- Presence of key informational concepts: kitchen, mother, stool, boy and girl.
- Frequency of verbs, nouns, adjectives and pronouns from POS-tagging.

Each feature is then rescaled with min-max normalization in range [0, 1] and an SVM with radial basis function (RBF) kernel and C = 1.0 is trained, whose output is the probability that the subject had AD given the 13-dimensional feature vector.

4. Experimental framework

The training dataset [7] consists of the recordings and manual transcripts of 108 subjects performing the test known as Cookie Theft, whose objective is to describe an image. Out of the 108 participants, 54 are patients diagnosed with Alzheimer's. Table 1 shows the training data distribution in detail.

Table 1: *ADReSS training dataset*

	AD		**non-AD**	
Age interval	Male	Female	Male	Female
$[50, 55)$	1	0	1	0
$[55, 60)$	5	4	5	4
$[60, 65)$	3	6	3	6
$[65, 70)$	6	10	6	10
$[70, 75)$	6	8	6	8
$[75, 80)$	3	2	3	2
Total	24	30	24	30

The evaluation metrics for the AD classification task are: $Accuracy = \frac{TN+TP}{N}$, Precision $\pi = \frac{TP}{TP+FP}$, Recall $\rho = \frac{TP}{TP+FN}$ and F-1 score $F_1 = 2\frac{\pi*\rho}{\pi+\rho}$.

where N is the number of patients, TP, FP and FN are the number of true positives, false positives and false negatives, respectively.

[5]https://github.com/huggingface/transformers

5. Results

This section presents the results in the AD classification for both the leave-one-subject-out (LOSO) and test settings of the ADReSS challenge.

5.1. Results for LOSO setting

All the systems have been trained using leave-one-subject-out (LOSO) cross-validation strategy for measuring the generalization error. Therefore, models use 107 subjects as training data and are tested on the held-out subject.

Table 2 illustrates the individual results achieved by the three speech-based systems. These results show that the x-vectors and fluency based systems have similar performance regarding the accuracy, as well as Area Under Curve (AUC). The i-vector was the weakest model, being the only one with an accuracy under 70%, although it is still above the challenge baseline results [7].

Table 2: *AD classification results of the proposed speech-based systems (LOSO cross-validation).*

	class	Precision	Recall	F1 Score	Accuracy	AUC
i-vector	non-AD AD	0.7000 0.6724	0.6481 0.7222	0.6730 0.6964	0.6851	0.6798
x-vector	non-AD AD	0.6923 0.7906	0.8333 0.6296	0.7563 0.7010	0.7314	0.7568
fluency	non-AD AD	0.7272 0.7358	0.7407 0.7222	0.7339 0.7289	0.7314	0.7613

The results for both text-based systems are shown in Table 3. Concerning BERT-based model, the AUC in held-out set is 0.9078. For a selected threshold probability, we also obtain an accuracy of 0.8518, recall of 0.8333 and F1-score of 0.8490. For the linguistic model, the AUC in held-out set is 0.7510. For a selected threshold probability the accuracy is 0.7129, precision of 0.8965, recall of 0.4814 and F1-score of 0.6265.

Table 3: *AD classification results of the proposed text-based systems (LOSO cross-validation).*

	class	Precision	Recall	F1 Score	Accuracy	AUC
BERT model	non-AD AD	0.8392 0.8653	0.8703 0.8333	0.8545 0.8490	0.8518	0.9077
Linguistic model	non-AD AD	0.6455 0.8965	0.9444 0.4818	0.7669 0.6265	0.7129	0.7510

Moreover, three different score fusion strategies were carried out. In the first one (referred as Fusion I), the scores of every system were normalized by z-norm, and merged by a weighted sum fusion rule. Weights are chosen from the accuracy of each individual AD detection system. In the second one (referred as Fusion II), the same normalization strategy was used to first, using also a weighted sum, merge the speech-based and text-based system scores separately, and then these new scores were again merged by a new weighted sum. Finally, in the last one (referred as Fusion III) instead of use text-based fused scores, only the BERT-based scores were fused with the speech-based fused scores. Figure 2 shows the ROC curves of the described fusion systems. The Fusion I model had an accuracy of 0.8611, a F1-score of 0.8543, and an AUC of 0.9355. The Fusion II model presented an accuracy of 0.8611, a F1-score of 0.8543 and an AUC of 0.9372. Lastly, the Fusion III model presented an accuracy of 0.8796, a F1-score of 0.8807

and an AUC of 0.9405. The results show that the fusion of both text-based and speech-based modalities improves the detection of AD.

For further insight, Figures 1 and 2 compares the ROC curves of all individual systems and their fusion, respectively.

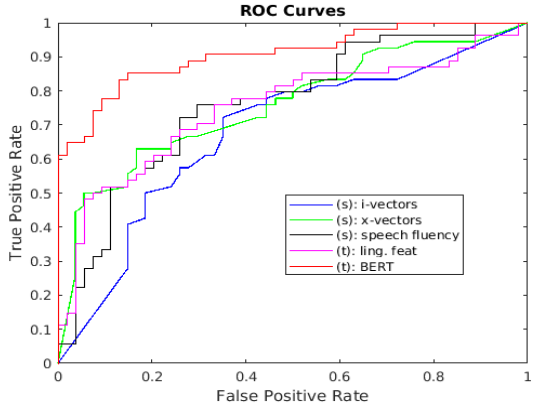

Figure 1: *ROC curves of all individual systems.*

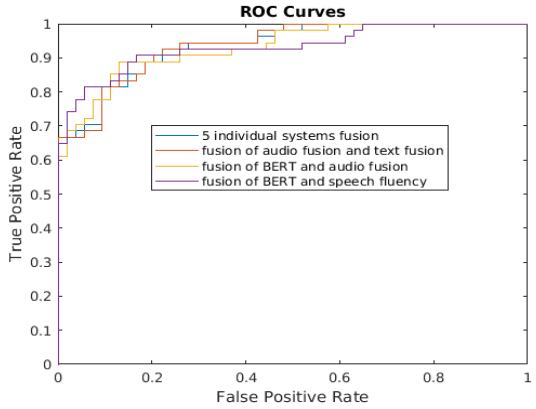

Figure 2: *ROC curves for the different fusion strategies.*

5.2. AD classification results for test setting.

Based on the LOSO cross-validation results the three best individual systems and all the fusion strategies were evaluated on the testing dataset. Table 4 shows the performance achieved by them on the testing dataset, which consist of recordings of 48 subjects.

The results achieved by the BERT model were very similar in validation and testing, which is a good sign of the lack of overfitting. However, the submitted speech-based systems had a significant decrease in performance. This fact shows that acoustic systems, require more data in order to improve the predictive ability of speech embeddings, and thus improve their generalization capacity avoiding overfitting issues. As a result, the performance of the fusion I and fusion II strategies also declines, but fusion III improves the accuracy of the BERT model.

6. Conclusions and Further work

Two lines of work have been developed on the ADReSS Challenge dataset: one based on speech processing and the other based on text processing.

Table 4: *Results on the testing dataset*

	class	Precision	Recall	F1 Score	Accuracy
Fusion I	non-AD AD	0.8182 0.7692	0.7500 0.8333	0.7826 0.8000	0.7917
Fusion II	non-AD AD	0.8000 0.8260	0.8333 0.7917	0.8163 0.8085	0.8125
Fusion III	non-AD AD	0.8636 0.8077	0.7917 0.8750	0.8261 0.8400	**0.8333**
BERT model	non-AD AD	0.7778 0.8571	0.8750 0.7500	0.8235 0.8000	0.8125
fluency	non-AD AD	0.6250 0.6250	0.6250 0.6250	0.6250 0.6250	0.6250
x-vector	non-AD AD	0.5417 0.5417	0.5417 0.5417	0.5417 0.5417	0.5417

It is important to highlight the following points when assessing both solutions:

- **Performance:** x-vector and fluency speech features have shown a competitive performance in the LOSO cross-validation. However, their performance decreases on the test setting. This would be a result of a low generalization level achieved at the training stage. On the other hand, the text-based BERT model has obtained superior results both in the LOSO and testing settings. The Fusion III strategy improves the accuracy of the BERT model and also obtains more balanced performance measures. Then, multimodal systems demonstrate to be a better strategy for the detection of AD than individual systems.
- **Complexity:** The x-vector and BERT systems, as deep-learning-based models, need a large amount of training data to be able to generalize well.
- **Explainability:** deep learning models are black-box models, being very difficult to interpret from a human perspective. On the contrary, the linguistic and fluency features are explainable and one could determine the weight of them in the classification.

Several future research lines have been identified for further work. Firstly, investigation on improved classification based on deep neural networks and novel acoustical feature extraction algorithms. Secondly, addition of new linguistic features and non-verbal information (breaks, silence duration, word mistakes, etc.) in text-based systems. Thirdly, analysis of strategies for increasing the generalization capacity of the proposed speech-based systems; for example: to find new rhythmic parameters more discriminatory between patients with and without AD or to adapt the deep learning models to the characteristics of elderly speech. It is also important to conduct experiments in other experimental settings, for example using other questions of the mini-mental state examination test, to validate the results obtained and, above all, to increase the size of the data settings.

7. Acknowledgements

This work has received financial support from the Spanish "Ministerio de Economia y Competitividad" through the project Speech&Sign RTI2018-101372-B-100, and also from Xunta de Galicia (AtlanTTic and ED431B 2018/60 grants) and European Regional Development FundERDF.

8. References

[1] A. Association, "2019 alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 15, no. 3, pp. 321–387, 2019.

[2] P. J. Nestor, P. Scheltens, and J. R. Hodges, "Advances in the early detection of alzheimer's disease," *Nature medicine*, vol. 10, no. 7, pp. S34–S41, 2004.

[3] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease," *Brain*, vol. 136, no. 12, pp. 3727–3737, 2013.

[4] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana, "Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia?" *Dementia and Geriatric Cognitive Disorders*, vol. 37, no. 5-6, pp. 327–334, 2014.

[5] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognitive Computation*, vol. 7, no. 1, pp. 44–55, 2015.

[6] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS Challenge," in *Proceedings of INTERSPEECH 2020*, Shanghai, China, 2020. [Online]. Available: https://arxiv.org/abs/2004.06833

[8] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 2162–2166. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2516

[9] 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://www.aclweb.org/anthology/N19-1423

[10] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[12] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2014, pp. 2494–2498.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," 2011.

[14] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech and Language*, vol. 60, 2020.

[15] M. Ajili, S. Rossato, D. Zhang, and J.-F. Bonastre, "Impact of rhythm on forensic voice comparison reliability," in *Odyssey 2018: The Speaker and Language Recognition Workshop*. ISCA, 2018, pp. 1–8.

[16] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5095–5099.

[17] B. Macwhinney, "The childes project: tools for analyzing talk," *Child Language Teaching and Therapy*, vol. 8, 01 2000.

[18] 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

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations (ICLR 2019)*. OpenReview.net, 2019.

[20] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify alzheimer's disease in narrative speech," *Journal of Alzheimer's disease : JAD*, vol. 49, no. 2, pp. 407–422, 2016.