INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multiscale System for Alzheimer's Dementia Recognition through Spontaneous Speech

Erik Edwards, Charles Dognin, Bajibabu Bollepalli, Maneesh Singh

Verisk Analytics

erik.edwards4@gmail.com, charles.dognin@verisk.com, bajibabu.bollepalli@aalto.fi, msingh@verisk.com

Abstract

This paper describes the Verisk submission to The ADReSS Challenge [1]. We analyze the text data at both the word level and phoneme level, which leads to our best-performing system in combination with audio features. Thus, the system is both multi-modal (audio and text) and multi-scale (word and phoneme levels). Experiments with larger neural language models did not result in improvement, given the small amount of text data available. By contrast, the phoneme representation has a vocabulary size of only 66 tokens and could be trained from scratch on the present data. Therefore, we believe this method to be useful in cases of limited text data, as in many medical settings.

Index Terms: Dementia detection, voice classification, computational paralinguistics

1. Introduction and Related Work

Alzheimer's disease (AD) is the most common cause of dementia, a group of symptoms affecting memory, thinking and social abilities. Detecting and treating the disease early is important to avoid irreversible brain damage. Several machine-learning (ML) approaches to identify probable AD and MCI (Mild Cognitive Impairment) have been developed in an effort to automate and scale diagnosis. A comprehensive review of medical-imaging-based approaches was provided by [2], but methods that are less invasive and expensive still require exploration.

Acoustic Approaches: Detection of AD using only audio data could provide a lightweight and non-invasive screening tool that does not require expensive infrastructure, and can be used in peoples' homes. Speech production with AD differs qualitatively from normal aging or other pathologies, and such differences can be used for early diagnosis of AD [3]. Several studies have been proposed to detect AD using speech signals. [4] showed that spectrographic analysis of temporal and acoustic features from speech can characterise AD with high accuracy. [5] used only acoustic features extracted from the recordings of DementiaBank for AD detection, and reported accuracy results of up to 97%.

Linguistic Approaches: There has also been recent work in text-based diagnostic classification approaches; these techniques use either engineered features or deep features.

Engineered Features: [6] showed that classifiers trained on automatic semantic and syntactic features from speech transcripts were able to discriminate between semantic dementia, progressive nonfluent aphasia, and healthy controls. This work was later extended to AD vs healthy control classification [7] using lexical and n-gram linguistic biomarkers.

Deep Features: Deep learning models to automatically detect AD have also recently been proposed. Orimaye et al. [8] introduced a combination of deep language models and deep neural networks to predict MCI and AD. One limitation of a deep-learning-based approach is the paucity of training data typical in medical settings. [9] attempted to interpret what the neural models learned about the linguistic characteristics of AD patients. Text embeddings of transcribed text have also been recently explored for this task. For instance, Word2Vec and GloVe have been successfully used to discriminate between healthy and probable AD subjects [10], while, more recently, multi-lingual FastText embedding combined with a linear SVM classifier has been applied to classification of MCI versus healthy controls [11].

Multimodal Approaches using representations from images have been recently used to detect AD [12, 13]. [14] used lexicosyntactic, acoustic and semantic features extracted from spontaneous speech samples to predict clinical MMSE scores (indicator of the severity of cognitive decline associated with dementia). The work of [15] extended this approach to classification, and obtained state-of-the-art results on DemantiaBank-fused linguistic and acoustic features extracted into a logistic regression classifier.

Multimodal and Multiscale Deep Learning Approaches to AD detection have been applied using medical imaging data [16]. Inspired by this, we propose an Acoustic-Linguistic approach with late fusion to classify AD vs healthy controls. Our contributions are as follows:

1. We introduce a multiscale approach for linguistic features by learning phoneme-level representation from scratch using FastText [17] and Sent2Vec [18]. We show that this phoneme-level embedding can be learned with a very small amount of data, which is a considerable advantage over existing work and ideally suited for clinical settings.
2. We combine speech and text domains to obtain a novel multiscale and multimodal approach to AD recognition. We find that subword (phoneme) information helps the classifier discriminate between healthy and ill participants.

2. Dataset

The dataset was provided by the ADReSS Challenge [1]. The participants were asked to describe the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [19]. Both the speech and corresponding text transcripts were provided. It was released in two parts: train and test sets. The train data had 108 subjects (48 male, 60 female) and the test data had 48 subjects (22 male, 26 female). For the train data, 54 subjects were labeled with AD and 54 with non-AD. The speech transcriptions were provided in CHAT format [20], with 2169 utterances in the train data (1115 AD, 1054 non-AD), and 934 in the test data.

http://dx.doi.org/10.21437/Interspeech.2020-2781

Table 1: *Acoustic features and their dimensions. CFS denotes correlation feature selection and RFECV denotes recursive feature elimination using cross-validation.*

Feature	Dim. (All)	Dim. (CFS)	Dim. (RFECV)
GEMAPS	64	53	3
eGEMAPS	90	76	4
emobase	979	626	6
emobase2010	1583	995	19
emolarge	6511	1810	21
ComParE2016	6375	3592	54
MRCG	6914	367	5

3. Acoustic Systems

All audio started as 16-bit WAV files at 44.1 kHz sample rate. These were provided as 'chunks', which were sub-segments of the above speech dialog segments that had been cropped to 10 seconds or shorter duration (2834 chunks: 1476 AD, 1358 non-AD). In general, the audio data was found to be very noisy and some of the chunks were unintelligible to the human ear. For example, a basic audio classification into 'speech' vs. 'other' using pyAudioAnalysis [21] found only 49.8% of audio chunks were clearly 'speech'.

We applied a basic speech-enhancement technique using VOICEBOX [22], which slightly improved the audio results, but is not essential to our method. We also tried rejecting noisy chunks, or using a 3-category classification scheme to separately identify the noisiest chunks. These attempts did not significantly improve the results, however, and so were not pursued further. We also attempted using voice activity detection, using OpenSMILE [23] or rVAD [24], and weighting audio results accordingly. This led to small improvements for some analyses, but was also not included in the final results, as it was apparent that more radical changes in methodology would be required to deal with these noise levels (e.g., a windowing into fixed-length frames). We decided therefore to use the noisy audio 'chunks' as given, with only the basic speech enhancement applied, and to defer additional improvements to future work.

3.1. Acoustic Features

Acoustic features were extracted on the enhanced speech segments downsampled to 16-kHz sample rate. We used the same feature sets as in the baseline Challenge paper [1], along with a few additional sets, but also added a stage of feature selection. Features are computed every 10-ms to give "low-level descriptors" (LLDs) and then statistical functionals of the LLDs (such as mean, standard deviation, kurtosis, etc.) are computed over each audio chunk of 0.5-10 sec duration (chunks shorter than 0.5 s were rejected). Using OpenSMILE [23], we extracted the following sets of functionals: emobase [25], emobase2010, GeMAPS [26], extended GeMAPS (eGeMAPS), and ComParE2016 (a minor update of numerical fixes to the ComParE2013 set [27]). Using code from the Cacophony Project (https://github.com/TheCacophonyProject), we extracted multi-resolution cochleagram (MRCG) LLDs [28], and then several statistical functionals of these. The dimensions of each functionals set are given in Table 1, and details can be found in the cited references.

3.2. Acoustic Feature Selection

As the dimensionality of each functionals set was large (Table 1), we explored feature selection techniques to improve subsequent classification. First, we used correlation feature selection (CFS), which discards highly-correlated features. Second, we used recursive feature elimination with cross validation (RFECV), where a classifier is employed to evaluate the importance of the each feature dimension. In each recursion, the feature that least improves or most degrades classifier importance is discarded, leading to a supervised ranking of features.

Table 2: *Accuracy scores of feature selection. These numbers calculated by taking majority vote on segments.*

Feature	All	CFS	RFECV
GEMAPS	0.490	0.472	0.629
eGEMAPS	0.453	0.462	0.620
emobase	0.555	0.555	0.657
emobase2010	0.555	0.574	0.601
emolarge	0.595	0.629	0.666
ComParE2016	**0.601**	**0.629**	**0.694**
MRCG	0.546	0.509	0.611

Table 3: *Accuracy scores of the ComParE2016 acoustic feature set with different classifiers. LR: Logistic regression, SVM: support vector machine, and LDA: linear discriminant analysis.*

Feature	LR	SVM	LDA
ComPareE2016	0.694	**0.740**	**0.740**

Table 1 shows the raw ("All") feature dimensions and after each feature selection method. We further appended age and gender to each acoustic feature set. With CFS, we discarded features with correlation coefficient ≥ 0.85. For RFECV, we used logistic regression (LR) as the base classifier with leave-one-subject-out (LOSO) cross validation. CFS reduced the dimensionality by 15-95%, and the RFECV method further brought the dimensionality down to 3-54 for all sets.

Table 2 shows the performance of feature selection methods employed in this study, assessed with LOSO cross-validation on the train set. There is considerable improvement in accuracy after the CFS and RFECV methods. Since the performance of the ComParE2016 features is best among the acoustic feature sets, we used only the ComParE2016 features for further experiments. However, it is noted that equivalent performance could be obtained with emobase2010 using other feature selection methodology (not included here).

Table 3 presents the accuracy scores achieved by the ComParE2016 features using different ML classification models. SVM (support vector machine) and LDA (linear discriminant analysis) models gave better performance than LR. The best accuracy obtained using acoustic features alone is 0.74. For our ensemble models, we used the posterior probabilities from the LDA model averaged over all chunks for each subject.

4. Linguistic Systems

The linguistic system contains two parts: the natural language representation and the phoneme representation.

4.1. Natural Language Representation

We applied a basic text normalisation to the transcriptions by removing punctuation and CHAT symbols and lower casing. Table 4 shows the accuracy and F_1 score results on a 6-fold cross validation of the training data-set (segment level). For each model used, hyper-parameter optimisation was performed to allow for fair comparisons.

4.1.1. Engineered Features

Following [7] and [9], we extract seven features from text segments: richness of vocabulary (measured by unique word count), word count, number of stop words, number of coordinating conjunction, number of subordinated conjunction, average word length, and number of interjections. Using CHAT symbols, we extract four more features: number of repetitions (using [/]), number of repetitions with reformulations (using [//]), number of errors (using [*]), and number of filler words (using [&]).

4.1.2. Deep Learning Features

We experimented with three different settings: Random Forest with deep pre-trained Features (DRF), fine-tuning of pre-trained models (FT) and training from scratch (FS).

Deep Random Forest Setting: We extract features using three pre-trained embeddings: Word2Vec (CBOW) with subword information [29] (pre-trained on Common Crawl), GloVe [30] pre-trained on Common Crawl and Sent2Vec [31] (with uni-grams) pre-trained on English Wikipedia. The procedure is the same for each model: each text segment is represented by the average of the normalised word embeddings. The segment embeddings are then fed to a Random Forest Classifier. In this setting the best performing model is Sent2Vec with unigram representation. Sent2Vec is built on top of Word2Vec, but allows the embedding to incorporate more contextual information (entire sentences) during pre-training.

Training from Scratch Setting: In this setting, models are trained from scratch on the given dataset. The only model fast enough to allow us to find the best hyper-parameters while being a good baseline is FastText. With an embedding dimension as low as 5 and with as low as 16 words in its vocabulary, FastText performs competitively compared to most of the Deep Random Forest Settings. Subword information determined by character n-grams are keys to this result as explained below.

Fine-Tuning Setting: For this final setting, pre-trained embeddings (Word2Vec, GloVe, Sent2Vec) or models (Electra [32], Roberta [33]) are fine-tuned on the data. Electra uses a Generator/Discriminator pre-training technique more efficient than the Masked Language Modeling approach used by Roberta. Though the results of the two models are approximately the same at the segment level, Electra strongly outperforms Roberta at the participant level. The best models still remain the ones using subword information: GloVe (FT) and Word2Vec (FT). Both of those pre-trained embeddings are fine-tuned with the FastText classifier. The later turn sentences into character-n-gram augmented sentences (we found that a maximum character n-grams of 6 was optimal). Though FastText from scratch also have the sub-word information, it does not have the pre-trained representation of those sub-words learnt using GloVe or CBOW (Word2Vec).

4.1.3. Interpretation and Discussion

Subword Information appears to be a key discriminative feature for effective classification. As Figure 1 shows, not using subword information is detrimental to the discriminative power of the model. As a result, we can make the hypothesis that in low resource settings like in this case of medical data, taking into account subword information might be the key to good performance. We explore even further this hypothesis by transforming sentences into phoneme level sentences.

Table 4: *Best performance after hyper-parameters optimisation for each model, metrics are the average of accuracy and f1 scores across 6-fold cross-validation, participant level (softmax average).*

Model	Dim.	Accuracy	F1-score
Random (DRF)	11	0.463	0.482
Engineered Feat (DRF)	11	0.704	0.68
Sent2Vec (FT)	600	0.787	0.758
GloVe (FT)	300	0.861	0.865
Word2Vec (FT)	300	**0.926**	**0.923**
Word2Vec (DRF)	300	0.787	0.785
GloVe + EF (DRF)	311	0.796	0.792
Sent2Vec (DRF)	600	0.833	0.83
GloVe (DRF)	300	0.824	0.822
FastText (FS)	5	0.796	0.776
Roberta-Base (FT)	768	0.787	0.753
Electra-Base (FT)	768	0.861	0.845

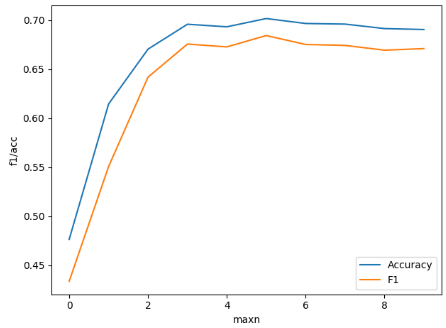

Figure 1: *F1 and Accuracy on 6-fold cross validation as a function of the maximum size of character n-grams (maxn) using FastText supervised classifier*

Word Order: When word order is important, FastText tends to not perform well as it averages the word embeddings of the input sentences without accounting for their original position. We confirmed this hypothesis by measuring the impact of adding word n-grams as additional features to the classifiers. Figure 2 shows that adding word n-grams, thus introducing temporality, does not impact the performance or even degrade it.

Performance of Transformers Though Transformers have subword information through the use of Byte Pair Encoding tokenizer for Roberta and WordPiece tokenizer for Electra, there are too few data points for their large number of parameters.

Experiment Details For the Random Forest (RF), we found that the best results on the 6-fold cross validation were obtained using 200 estimators, entropy criterion, square root for the maximum number of features. A StandardScaler (subtracting the mean and scaling to unit variance) was also applied to the features. FastText From Scratch (FS) hyper-parameters are: wordNgrams=1, 100 epochs, max number of character n-grams=6, minimum number of word occurences=100, learning rate of 0.05 and embedding dimension of 5. We kept the same hyper-parameters for FastText fine-tuned except for the dimension that we set to 300 for Word2Vec and GloVe and 600 for Sent2Vec. Roberta-Base and Electra-Base performance was measured on the best hyper-parameters found. The hyper-parameters that were found to work best are: a batch size of 16, 5 epochs, a maximum token length of 128 and a learning rate of 2e-05.

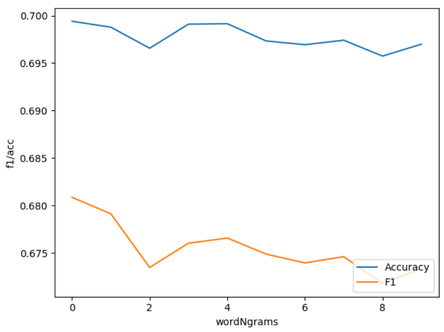

Figure 2: *F1 and Accuracy on 6-fold cross validation as a function of the word n-grams (wordNgrams) features using FastText supervised classifier*

Table 5: *Results of 9-fold CV on the Train set for several combined systems, using simple LR on posterior probability outputs. Audio represents the LDA posterior probabilities of ComParE2016. Word2Vec and GloVe were text (word-based) systems (Section 4.1) and Phonemes are as in Section 4.2. Age and speaking rate were added to each system.*

Model	Accuracy
GloVe + Phonemes	0.8981
GloVe + Phonemes + Audio	0.9074
Word2Vec + Phonemes	**0.9352**
Word2Vec + Phonemes + Audio	**0.9352**

4.2. Phonetic Representation

The discriminative importance of the subword information was confirmed by our phoneme transcription experiments. We transcribed the segment text into phoneme written pronunciation using CMUDict [34]. The most likely pronunciation is used for words with multiple pronunciations. Thus, "*also taking cookies*" becomes "*ao1 l s ow0 t ey1 k ih0 ng k uh1 k iy0 z*". In several experiments, it always helped to include vowel stress in the pronunciation (0 is no stress, 1 is full stress, 2 is part stress). With stress, there were 66 phones total.

Several text classifiers were trained on the phoneme representation (FastText, Sent2Vec, StarSpace), and FastText was again found to perform best (and fastest). Our best performance on the Test set (Table 6) used only the phoneme representation and FastText classification, along with the audio. However, in 9-fold CV tests with the Train set, the best result was a combination of natural language and phonetic representation (Table 5).

The numbers appended to vowel phonemes are stress indicators according to the convention of CMUdict. Our experiments showed that removing stress always caused a decrease in performance. The discriminative importance of phonetic and articulatory representation in AD patient is in accord with previous medical research (e.g., [35]), and deserves future experimentation for ML purposes.

Experiment Details For the phonetic experiments, we used FastText supervised classifier with the following hyperparameters: 4 wordNgrams, an embedding dimension of 20, a learning rate of 0.05, 300 epochs, and a bucket size of 50000. The other hyperparameters were at default. We did not use character n-grams (many phones are already characters).

Table 6: *Challenge Test Set Results*

Model	Class	Precision	Recall	F1 Score	Accuracy
System 1	non-AD	0.6316	0.5	0.5581	0.6042
	AD	0.5862	0.7083	0.6415	
System 2	non-AD	0.7407	0.8333	0.7843	0.7708
	AD	0.8095	0.7083	0.7556	
System 3	non-AD	0.7692	0.8333	0.8	**0.7917**
	AD	0.8182	0.7500	0.7826	
System 4	non-AD	0.7308	0.7917	0.76	0.75
	AD	0.7727	0.7083	0.7391	
System 5	non-AD	0.75	0.75	0.75	0.75
	AD	0.75	0.75	0.75	

5. Discussion

- **System 1**: Audio (LDA posterior probabilities of ComParE2016 features)
- **System 2**: Phonemes (as in Section 4.2)
- **System 3**: Phonemes and Audio
- **System 4**: Phonemes and Word2vec (as in Section 4.1)
- **System 5**: Phonemes and Audio and Word2Vec

For each combined system (Tables 5 and 6), we appended the age and speaking rate as auxiliary features. Those two variables are well studied for identifying AD (see [36] for the positive correlation with age and [37] for the negative correlation with speech rate).

Acoustic Features alone are not as discriminative as text features alone. There is indeed a 15 points difference in accuracy between System 1 which mainly use acoustic features and System 4 which mainly uses text features. However, the audio was very noisy in this set; new feature sets and robustness measures should be explored.

Deep learning text systems easily overfit for small data. RoBERTa and Electra models performed worse than Word2Vec on this small dataset (Table 4), and systems 4 and 5 perform worse on the final Test set than just Phonemes alone (Table 6). However, 9-fold CV on the Train set (Table 5) found that the best performing system was multiscale (Word2Vec and Phonemes) as well as multimodal (text and audio) (Table 5). We believe this would also give the best result for the Test set if the amount of data were larger.

Using Phoneme/Subword is key. The effectiveness of using subword features to discriminate between AD and non-AD people can be understood as analogous to data augmentation. Splitting tokens into subwords or mapping them to phonemes reduces the size of the vocabulary and at the same time expands the number of tokens in the training set. Also, several studies like [35] have found that AD patients show articulatory difficulties and patterns which would show on the phonetic transcription. Phoneme representations also capture many simple aspects of word-based text models, noting that phoneme 4-grams as used here already include many basic words.

6. Conclusions

We propose a multiscale approach to the problem of automatic Alzheimer's Disease (AD) detection. We find that subword information, and in particular phoneme representation, helps the classifier discriminate between healthy and ill participants. This finding could prove useful in many medical or other settings where lack of data is the norm.

7. References

[1] 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. [Online]. Available: https://arxiv.org/abs/2004.06833

[2] J. Wen, E. Thibeau-Sutre, M. Diaz-Melo, J. Samper-Gonzalez, A. Routier, S. Bottani, D. Dormont, S. Durrleman, N. Burgos, O. Colliot *et al.*, “Convolutional neural networks for classification of Alzheimer’s disease: overview and reproducible evaluation,” *Medical image analysis*, p. 101694, 2020.

[3] M. L. B. Pulido, J. B. A. Hernández, M. Á. F. Ballester, C. M. T. González, J. Mekyska, and Z. Smékal, “Alzheimer’s disease and automatic speech analysis: a review,” *Expert systems with applications*, p. 113213, 2020.

[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] S. Al-Hameed, M. Benaissa, and H. Christensen, “Simple and robust audio-based detection of biomarkers for Alzheimer’s disease,” in *Proc SLPAT Workshop*, 2016, pp. 32–36.

[6] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, “Automated classification of primary progressive aphasia subtypes from narrative speech transcripts,” *Cortex*, vol. 55, pp. 43–60, 2014.

[7] S. O. Orimaye, J. S. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, “Predicting probable Alzheimer’s disease using linguistic deficits and biomarkers,” *BMC bioinformatics*, vol. 18, no. 1, p. 34, 2017.

[8] S. O. Orimaye, J. S.-M. Wong, and C. P. Wong, “Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia,” *PloS one*, vol. 13, no. 11, 2018.

[9] S. Karlekar, T. Niu, and M. Bansal, “Detecting linguistic characteristics of Alzheimer’s dementia by interpreting neural models,” *arXiv:1804.06440*, 2018.

[10] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, “Detecting signs of dementia using word vector representations.” in *Proc INTERSPEECH*. ISCA, 2018, pp. 1893–1897.

[11] K. C. Fraser, K. L. Fors, and D. Kokkinakis, “Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment,” *Computer speech & language*, vol. 53, pp. 121–139, 2019.

[12] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen *et al.*, “Multimodal classification of Alzheimer’s disease and mild cognitive impairment,” *Neuroimage*, vol. 55, no. 3, pp. 856–867, 2011.

[13] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, P. Skudlarski, E. Cavedo, G. B. Frisoni, W. Hoffmann *et al.*, “Multimodal imaging in Alzheimer’s disease: validity and usefulness for early detection,” *The Lancet neurology*, vol. 14, no. 10, pp. 1037–1053, 2015.

[14] M. Yancheva, K. C. Fraser, and F. Rudzicz, “Using linguistic features longitudinally to predict clinical scores for Alzheimer’s disease and related dementias,” in *Proc SLPAT Workshop*, 2015, pp. 134–139.

[15] 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.

[16] D. Lu, K. Popuri, G. W. Ding, R. Balachandar, and M. F. Beg, “Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer’s disease using structural MR and FDG-PET images,” *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” *arXiv:1607.01759*, 2016.

[18] M. Pagliardini, P. Gupta, and M. Jaggi, “Unsupervised learning of sentence embeddings using compositional n-gram features,” *arXiv:1703.02507*, 2017.

[19] 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, 1994.

[20] B. MacWhinney, *The CHILDES project: tools for analyzing talk*. Psychology Press, 2014, vol. I.

[21] T. Giannakopoulos, “pyAudioAnalysis: an open-source python library for audio signal analysis,” *PloS one*, vol. 10, no. 12, 2015.

[22] M. Brookes, “VOICEBOX: speech processing toolbox for matlab,” Imperial College London, UK, 2010. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[23] F. Eyben, M. Wöllmer, and B. Schuller, “OpenSMILE: the Munich versatile and fast open-source audio feature extractor,” in *Proc ACM Conf Multimedia*, 2010, pp. 1459–1462.

[24] Z.-H. Tan, N. Dehak *et al.*, “rVAD: an unsupervised segment-based robust voice activity detection method,” *Computer speech & language*, vol. 59, pp. 1–21, 2020.

[25] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” in *Proc ICASSP*, vol. 1. IEEE, 2004, pp. 577–581.

[26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[27] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, “On the acoustics of emotion in audio: what speech, music, and sound have in common,” *Frontiers in psychology*, vol. 4, p. 292, 2013.

[28] J. Chen, Y. Wang, and D. Wang, “A feature study for classification-based speech separation at low signal-to-noise ratios,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1993–2002, 2014.

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[30] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global Vectors for word representation,” in *Proc Conf EMNLP*, 2014, pp. 1532–1543.

[31] P. Gupta, M. Pagliardini, and M. Jaggi, “Better word embeddings by disentangling contextual n-gram information,” in *Proc NAACL-HLT Conf*, vol. I. ACL, 2019, pp. 933–939.

[32] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “Electra: pre-training text encoders as discriminators rather than generators,” *arXiv:2003.10555*, 2020.

[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: a Robustly optimized BERT pretraining approach,” *arXiv:1907.11692*, 2019.

[34] R. Weide, “The Carnegie Mellon pronouncing dictionary [CMU-dict. 0.6],” *Pittsburgh, PA: Carnegie Mellon Univ.*, 2005.

[35] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson, “Phonological and articulatory impairment in Alzheimer’s disease: a case series,” *Brain and language*, vol. 75, no. 2, pp. 277–309, 2000.

[36] R. Guerreiro and J. Bras, “The age factor in Azheimer’s disease,” *Genome medicine*, vol. 7, no. 1, p. 106, 2015.

[37] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski, “Speaking in Alzheimer’s disease, is that an early sign? importance of changes in language abilities in Alzheimer’s disease,” *Frontiers in aging neuroscience*, vol. 7, p. 195, 2015.