Impact of ASR on Alzheimer's Disease Detection: All Errors are Equal, but Deletions are More Equal than Others

Aparna Balagopalan
Winterlight Labs
Toronto, Canada
aparna@winterlightlabs.com

Ksenia Shkaruta
Georgia Tech
Atlanta, USA
k.shkaruta@gatech.edu

Jekaterina Novikova
Winterlight Labs
Toronto, Canada
jekaterina@winterlightlabs.com

Abstract

Automatic Speech Recognition (ASR) is a critical component of any fully-automated speech-based dementia detection model. However, despite years of speech recognition research, little is known about the impact of ASR accuracy on dementia detection. In this paper, we experiment with controlled amounts of artificially generated ASR errors and investigate their influence on dementia detection. We find that deletion errors affect detection performance the most, due to their impact on the features of syntactic complexity and discourse representation in speech. We show the trend to be generalisable across two different datasets for cognitive impairment detection. As a conclusion, we propose optimising the ASR to reflect a higher penalty for deletion errors in order to improve dementia detection performance.

1 Introduction

There is a rapid growth in the number of people living with Alzheimer's disease (AD) (Alzheimer's Association, 2018). Clinical research has shown that quantifiable signs of cognitive decline associated with AD and mild cognitive impairment (MCI) are detectable in spontaneous speech (Bucks et al., 2000; Sajjadi et al., 2012). Machine learning (ML) models have proved to be successful in detecting AD using speech and language variables, such as syntactic and lexical complexity of language extracted from the transcripts of the speech (Fraser et al., 2016; Meilán et al., 2012; Rentoumi et al., 2014). Since transcripts should be accurate enough to properly represent syntactic and linguistic characteristics, current approaches (Fraser et al., 2013; Zhu et al., 2019) frequently rely on 100% accurate human-created transcripts produced by trained transcriptionists. However in real-life speech-based applications of AD detection, ASR is used and it produces noisy, error-prone transcripts (Yousaf et al., 2019). To our best knowledge, while the importance of well-performing ASR in speech classification has been studied in depth (Zhou et al., 2016), no prior research was done to understand what patterns of speech are influenced the most by ASR errors such as word deletions and substitutions, and how this impacts performance of AD detection using ML models.

In this paper, we focus on this issue and study the *effect of deletion, insertion and substitution errors on lexico-syntactic language features* and their resulting *effect on classification performance*. The effect of these errors on binary AD-healthy classification performance is studied and suggestions are provided on how to improve ASR in order to maintain reasonable AD classification performance.

We identify that deletion errors affect the classification more than substitution and insertion errors on two datasets of spontaneous impaired speech. The effect of these deletion errors are most profound on features related to syntactic complexity and discourse representations in speech, such as production rules, word-level structure and repetitions. These features are also identified as being the most important for the classification task using a feature gradient-based importance metric.

2 Data and Setup

2.1 Datasets

DementiaBank (DB) The DementiaBank[1] dataset is a large dataset of pathological speech. It consists of narrative picture descriptions from participants aged between 45 to 90 (Becker et al., 1994). Out of the 210 participants in the study, 117 were diagnosed with AD (180 samples of speech) and 93 were healthy (HC, 229 samples). Voice recordings and manual transcriptions (following CHAT protocol (MacWhinney, 2000)) are available for all

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

Proceedings of the 2020 EMNLP Workshop W-NUT: The Sixth Workshop on Noisy User-generated Text, pages 159–164
Online, Nov 19, 2020.

Dataset		Del (%)	Ins (%)	Sub (%)
DB	HC	54.14	4.27	41.59
	AD	56.98	3.89	39.13
HA	HC	24.37	13.11	62.52
	MCI	21.78	14.81	63.40

Table 1: Rates of ASR errors on DB and HA datasets.

samples. This dataset is used for the experiments in Section 4, 5, and 6.

Healthy Aging (HA) The Healthy Aging dataset (Balagopalan et al., 2018) consists of speech samples of 97 participants with no cognitive impairment diagnosis, all older than 50 years. Every participant describes a picture, analogous to the DB dataset. The dataset constitutes 8.5 hours of audio with manual transcriptions. Each speech sample is associated with a score on the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005). Based on published cut-off scores (Nasreddine et al., 2005) for presence of MCI (minimum score for healthy participants is 26), we obtain class-labels for this dataset.

2.2 ASR Setup

The Automatic Speech Recognition (ASR) system we use for this work is based on the open-source Kaldi toolkit (Povey et al., 2011). ASR uses ASPiRE chain model trained on multi-condition Fisher English corpus as a 3-gram language model.

Rates of ASR errors for healthy and impaired speakers for DB and HA datasets are in Table 1. Majority of errors arise from deletions and substitutions for both datasets and groups.

3 Methodology

3.1 Feature Extraction and Aggregation

Following previous studies (Fraser et al., 2016; Balagopalan et al., 2018), we automatically extract 507 lexico-syntactic and acoustic features. To simplify the presentation, the extracted features are aggregated into the following major groups:

Syntactic Complexity: features to analyze the syntactic complexity of speech, such as number of occurrence of various production rules, mean length of clause (in words) etc.

Lexical Complexity and Richness : measures of lexical density and variation, such as average familiarity scores of all nouns, age of word acquisition, frequency of POS tags etc.

Discourse mapping: features that help identify cohesion in speech using a *speech graph*-based representation of message organization in speech (Mota et al., 2012). Examples of features include the number of edges in the graph, number of self-loops, cosine-distance across unique utterances etc.

Additionally, we extract features quantifying difficulty in finding the right words (e.g. filled pauses), measures related to description of content in the picture (e.g. number of content units), coherence in speaking at local and global level, and acoustic measures. such as MFCC and Zero Crossing Rate related voice representations (full list in App.A.1).

3.2 Error and Noise Addition

3.2.1 Artificial ASR Errors

We introduce artificial ASR errors to understand if any specific error type influences the classification performance more than others. In previous research it was shown that lexical and syntactic groups of features extracted from transcripts of speech have different predictive power in dementia classification (Novikova et al., 2019). As such, we hypothesize that different ASR error types may influence the features differently and would cause different effects on classification performance. The non-artificial output of ASR combines the errors of deletion, insertion and substitution in some proportion, thus not allowing analysis of the individual effects of each error type separately. This is why we generate each type of errors artificially.

3.3 Error Addition Method

We follow a method similar to the one used by Fraser et al. (2013) to artificially add errors to manual transcripts at predefined 20%, 40% and 60% WER rates. All altered words w, where w refers to a word in gold-standard manual transcripts, are selected at random. The following modifications are done: a) *deletion* - word instance w is deleted, b) *insertion* - new word w_1 is added after the word w, c) *substitution* - word w is replaced with a new word w_1.

For ***deletion*** we simply delete random words from manual transcript at a specified rate.

To ***substitute*** word w, we select a unigram from 2,000 most used unigrams from Fisher language model that has the smallest Levenshtein distance with word w based on the phonemic model from The Carnegie Mellon Pronouncing on Pronouncing Dictionary (Weide, 1998). If word w is not found in the Fisher language model a random unigram from the top 2,000 is used for substitution.

For ***insertion***, we select a word from the bigram

list from the language model that has the highest probability to follow after word w and insert it if it does not match the following word in transcript. In case of a match, the next most probable word is inserted. If word w is not found in bigram list a random unigram is used for insertion.

To verify if simulated errors are a fair approximation of what is seen on a true ASR output, we have calculated the BLEU score (Papineni et al., 2002) between the manual and ASR-generated transcripts and compared them to the BLEU score between the manual transcripts and the transcripts with artificially simulated errors. The correlation between these two BLEU scores is strong and significant for both datasets (Spearman $\rho = 0.72, p < 0.001$ for DB; $\rho = 0.66, p < 0.001$ for HA), i.e. transcripts with simulated errors are corrupted with respect to the manual transcripts in a similar manner as the ASR-generated transcripts are.

3.3.1 Noise Addition

We perturb all lexico-syntactic features or equivalently features that could be affected by ASR errors such as deletions, insertions, and/or substitutions, to mimic random sources of errors using Gaussian noise. We do this to compare and differentiate from the consequences of ASR errors. This modification is implemented by adding a randomized number to the extracted feature values where the mean of the number added to a given feature is zero and the standard deviation varies depending on the amount of noise we add (see App.A.2 for details).

3.4 Classification Setup

Model: All our experiments are based on predictions obtained from a 2-hidden layer neural network (see App.A.3 for details). We chose this model type and parameter-setting since it attained performance on-par with previously published results (Fraser et al., 2016) with 10-fold cross-validation on gold-standard manual DB transcripts.

4 Changes in Classification Performance Due to Simulated Errors

We evaluate performance of classifying samples of speech to two classes - AD or healthy - using the DB dataset.

Figure 1 shows that deletion errors affect classification performance significantly more than insertion and substitution errors do. 40% of deletions reduce F1 score by more than 10%, while 40% of insertions only result in 2.8%, and 40% of substitutions - in 6.3% of F1 score reduction. These differences become even more pronounced with adding a bigger amount of errors. Trajectory of F1 score with varying levels of noise is substantially different from that with varying deletion errors but not that with insertions or substitutions, showing that insertion and substitution errors influence classification performance in a way that is similar to a random noise. Deletion errors, however, have a significantly stronger effect on classification. It is also interesting to note that the model utilizing automatic transcripts from ASR retains a level of performance at 74.96% (Table 2), which is comparable to the potential decrease in performance due to the rate of ASR deletion errors.

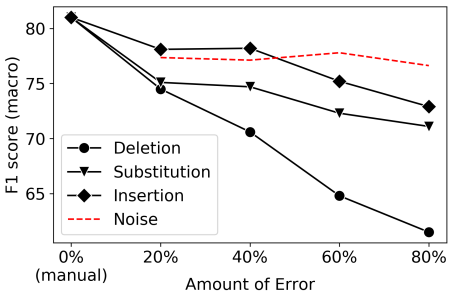

Figure 1: Effect of a controlled amount of ASR errors and random noise on classification performance.

Transcript	Accuracy	F1 (macro)	Sensitivity	Specificity
Manual	80.20	79.76	79.29	80.83
ASR-based	74.96	73.97	76.01	74.36

Table 2: Effect of original ASR on classification performance with the DB dataset.

Different effects of errors on classification performance suggest that some features, extracted from the speech samples and used as an input for the classification algorithm, are affected far more substantially by deletions rather than any other type of errors. This leads us to inspect the correlation of feature values and the amount of deletions.

5 Distinctive Effects of Deletion Errors

In order to understand why deletions errors influence the classification performance significantly more than other error types, we identify features maintaining higher correlation with the amount of deletions than that with the amount of insertions and substitutions. We observe 18 features in total that distinctively correlate with deletions. Out of these, the absolute majority of 15 features (83.33%

of all selected) are associated with syntactic complexity (production rules of a constituency parser) and discourse phenomena (graph self-loop with 3 edges) and 3 (16.7%) - with lexical richness in speech. Other feature groups, such as acoustic features or those associated with word finding difficulty, do not meet the required conditions. Such results show that syntactic structure of language is much more vulnerable to deletions than to other ASR errors. This can be explained by the fact that insertions and substitutions use words from the language model (i.e. most probable words) for the modifications, which to some extent helps maintain basic syntactic rules and structure.

Correlation between the number of deletions and features of syntactic structure shows the vulnerability of the feature group representing syntactic complexity and discourse phenomena to ASR deletion errors. However, it does not explain a decrease in classification performance when adding deletion errors. In Section 6 we inspect if features of syntactic complexity are more influential in AD detection than other characteristics of speech.

6 Model-based Analysis of Feature Importance

In order to quantify the importance of input features for classification, we obtain the gradient of the output prediction loss with respect to input features on a manually-transcribed version of the DB dataset.

We define gradient-based importance for feature k for an input, $X_{i,j}$, in the training set for a classification model as:

$$imp^{i,j,k} = \frac{\partial L(y_{i,j}, p_{i,j})}{\partial X_{i,j,k}} \quad (1)$$

where L denotes the loss criterion (binary cross-entropy loss), $y_{i,j}$ is the ground-truth label, $p_{i,j} \subset [0, 1]$ is the prediction probability; $p_{i,j} > 0.5$ denotes an AD prediction, k is a given feature (1 to D), and i is a number of samples (1 to N_j) in the training set in fold j of the DB dataset classification setup. Hence, to obtain the average importance for feature k in a single fold, we compute:

$$imp^{j,k} = 1/N_j \sum_{i=1}^{N_j} \frac{\partial L(y_{i,j}, p_{i,j})}{\partial X_{i,j,k}} \quad (2)$$

This importance is then averaged across the 10-folds to obtain the final importance, i.e.:

$$imp^{k} = 1/10 \sum_{j=1}^{10} imp^{j,k} \quad (3)$$

In order to interpret high-level patterns of input importance, we aggregate the feature importances into the groups defined in Section 3.1, where aggregation of importances involves averaging the absolute gradient-importance, $|imp^k|$, of features belonging to that group.

Feature group	Importance of top-10 features HC	AD	#features	Group rank
Syntactic complexity and Discourse phenomena	0.94	0.95	37	1
Lexical richness	0.91	0.92	18	2

Table 3: Importance of the two feature groups, summarised as the mean value of the top-10 most important features selected for HC and AD components, number of features having significant Spearman correlation with deletion errors, and the rank of each group.

Results provided in Table 3 show that the average normalised importance of the features associated with syntactic complexity and discourse is higher than the average importance of lexical richness features, when top-10 most important features across all the groups are selected for comparison.

To conclude, the feature group of syntactic complexity and discourse phenomena is affected significantly and distinctively the most by deletion errors as seen in Section 5. This group is also important for classification as seen in Table 3, indicating why classification is affected significantly by deletion errors. Hence, we track the effects from the initial step of adding artificial errors of different amounts to obtaining the final predictions in this manner.

7 Generalisability Evaluation

In order to test how well our conclusions generalise to a different dataset of impaired speech, we repeat the same experiments performed on DB on the HA dataset (Section 2.1).

We follow the same method, as described in Section 3 to extract the features and classify samples. Similarly to the results obtained on DB data, with HA deletion errors affect classification performance the most. Furthermore, deletion errors differentiate the same feature group of syntactic complexity and discourse phenomena: with HA dataset, 39 features correlate with deletions stronger than with insertions or substitutions, with 79.49% of

features belonging to the aggregate group of syntactic complexity and discourse, and 20.51% - to the group of lexical richness. The rank of feature groups, based on the average absolute Spearman correlation of all the features included in the groups, correspond to the rank observed with DB dataset, with a stronger significant correlation corresponding to the group of syntactic complexity, rather than lexical richness.

8 Conclusions

We observe that simulated deletion errors have a strong effect on classification performance when detecting cognitive impairment from speech and language, which can be traced back to their effect on syntactic complexity and discourse representations. With this observation in mind, the practical suggestion would be to optimise the ASR to reflect a higher penalty for deletion errors to improve dementia detection performance. For example, the decoder can be parametrised to find a balance between insertions and deletions, so that the number of deletion errors is minimised.

However, dealing with deletions in training time is not trivial, so in future work, we will focus on the optimisation of ASR performance and its effect on AD detection. Careful ASR error management, following previous work by Simonnet et al. (2017), could help enable strong fully-automated speech-based predictive models for dementia detection.

References

Alzheimer's Association. 2018. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429.

Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, and Marzyeh Ghassemi. 2018. The Effect of Heterogeneous Data for Alzheimer's Disease Detection from Speech. In *NIPS Workshop on Machine Learning for Health ML4H*.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594.

Romola S Bucks, Sameer Singh, Joanne M Cuerden, and Gordon K Wilcock. 2000. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91.

Kathleen Fraser, Frank Rudzicz, Naida Graham, and Elizabeth Rochon. 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia. In *Proceedings of the fourth workshop on speech and language processing for assistive technologies*, pages 47–54.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Brian MacWhinney. 2000. *The Childes Project: Tools for Analyzing Talk: Vol. II: The Database*. Mahwah.

Juan JG Meilán, Francisco Martínez-Sánchez, Juan Carro, José A Sánchez, and Enrique Pérez. 2012. Acoustic markers associated with impairment in language processing in Alzheimer's disease. *The Spanish journal of psychology*, 15(2):487–494.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Natalia B Mota, Nivaldo AP Vasconcelos, Nathalia Lemos, Ana C Pieretti, Osame Kinouchi, Guillermo A Cecchi, Mauro Copelli, and Sidarta Ribeiro. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4):e34928.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2005. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

Jekaterina Novikova, Aparna Balagopalan, Ksenia Shkaruta, and Frank Rudzicz. 2019. Lexical Features Are More Vulnerable, Syntactic Features Have More Predictive Power. In *EMNLP Workshop on on Noisy User-generated Text*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron

Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The Kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society.

Vassiliki Rentoumi, Ladan Raoufian, Samrah Ahmed, Celeste A de Jager, and Peter Garrard. 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42(s3):S3–S17.

Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J Nestor. 2012. Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, 26(6):847–866.

Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, Yannick Estève, and Renato de Mori. 2017. Asr error management for improving spoken language understanding. In *Interspeech 2017*.

Robert L Weide. 1998. The CMU pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict*.

Kanwal Yousaf, Zahid Mehmood, Israr Ahmad Awan, Tanzila Saba, Riad Alharbey, Talal Qadah, and Mayda Abdullateef Alrige. 2019. A comprehensive study of mobile-health based assistive technology for the healthcare of dementia and alzheimer's disease (ad). *Health Care Management Science*, pages 1–23.

Luke Zhou, Kathleen C Fraser, and Frank Rudzicz. 2016. Speech Recognition in Alzheimer's Disease and in its Assessment. In *INTERSPEECH*, pages 1948–1952.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2019. Detecting cognitive impairments by agreeing on interpretations of linguistic features. *NAACL*.