Original Paper

Evaluating Web-Based Automatic Transcription for Alzheimer Speech Data: Transcript Comparison and Machine Learning Analysis

Thomas Soroski[1], BSc; Thiago da Cunha Vasco[2]; Sally Newton-Mason[1], BSc; Saffrin Granby[2], BSc; Caitlin Lewis[1], BA; Anuj Harisinghani[2], BSc; Matteo Rizzo[2], BSc, MSc; Cristina Conati[2], MSc, PhD; Gabriel Murray[3], MSc, PhD; Giuseppe Carenini[2], MSc, PhD; Thalia S Field[1], MHSc, MD; Hyeju Jang[2], MS, PhD

[1]Vancouver Stroke Program and Division of Neurology, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada

[2]Department of Computer Science, Faculty of Science, University of British Columbia, Vancouver, BC, Canada

[3]School of Computing, University of the Fraser Valley, Abbotsford, BC, Canada

Corresponding Author:
Hyeju Jang, MS, PhD
Department of Computer Science
Faculty of Science
University of British Columbia
201-2366 Main Mall
Vancouver, BC, V6T 1Z4
Canada
Phone: 1 604 822 3061
Email: hyejuj@cs.ubc.ca

Abstract

Background: Speech data for medical research can be collected noninvasively and in large volumes. Speech analysis has shown promise in diagnosing neurodegenerative disease. To effectively leverage speech data, transcription is important, as there is valuable information contained in lexical content. Manual transcription, while highly accurate, limits the potential scalability and cost savings associated with language-based screening.

Objective: To better understand the use of automatic transcription for classification of neurodegenerative disease, namely, Alzheimer disease (AD), mild cognitive impairment (MCI), or subjective memory complaints (SMC) versus healthy controls, we compared automatically generated transcripts against transcripts that went through manual correction.

Methods: We recruited individuals from a memory clinic ("patients") with a diagnosis of mild-to-moderate AD, (n=44, 30%), MCI (n=20, 13%), SMC (n=8, 5%), as well as healthy controls (n=77, 52%) living in the community. Participants were asked to describe a standardized picture, read a paragraph, and recall a pleasant life experience. We compared transcripts generated using Google speech-to-text software to manually verified transcripts by examining transcription confidence scores, transcription error rates, and machine learning classification accuracy. For the classification tasks, logistic regression, Gaussian naive Bayes, and random forests were used.

Results: The transcription software showed higher confidence scores (P<.001) and lower error rates (P>.05) for speech from healthy controls compared with patients. Classification models using human-verified transcripts significantly (P<.001) outperformed automatically generated transcript models for both spontaneous speech tasks. This comparison showed no difference in the reading task. Manually adding pauses to transcripts had no impact on classification performance. However, manually correcting both spontaneous speech tasks led to significantly higher performances in the machine learning models.

Conclusions: We found that automatically transcribed speech data could be used to distinguish patients with a diagnosis of AD, MCI, or SMC from controls. We recommend a human verification step to improve the performance of automatic transcripts, especially for spontaneous tasks. Moreover, human verification can focus on correcting errors and adding punctuation to transcripts. However, manual addition of pauses is not needed, which can simplify the human verification step to more efficiently process large volumes of speech data.

(JMIR Aging 2022;5(3):e33460) doi: 10.2196/33460

KEYWORDS

Alzheimer disease; mild cognitive impairment; speech; natural language processing; speech recognition software; machine learning; neurodegenerative disease; transcription software; memory

Introduction

Identifying individuals with Alzheimer disease (AD) and mild cognitive impairment (MCI) early is beneficial for patient care, family support, and resource planning for the health care system [1]. Identification of individuals who are in the earliest stages of neurodegenerative disease, before irreversible brain changes have occurred, may also allow for the use of disease-modifying therapies when they would be most effective [2].

Analysis of speech to aid in the identification of individuals with early neurodegenerative disease can be a promising strategy, as speech recording is noninvasive, scalable, and easily repeated over time. This contrasts with the current methods for screening for AD or MCI, such as nuclear medicine scans or spinal fluid analysis, which can be both expensive and invasive [3]. Short samples of spontaneous or prompted speech can be collected remotely by telephone or videoconference. To date, speech and language have shown promising results in a significant number of studies aiming to classify AD or MCI [4].

For AD classification using speech, transcription is a key step to leverage the wealth of information contained in lexical data [5,6]. DementiaBank [7], the largest cohort of MCI and AD speech data for research, is entirely manually transcribed. Manual transcription, while highly accurate, is very low throughput (eg, requiring 4 minutes of transcriber time for each minute of audio [8]), limiting the potential scalability and cost savings associated with language-based screening for MCI and AD. As a result, there is a move toward automatically preprocessing medical speech as opposed to manual transcription.

To date, some groups have investigated AD/MCI classification using only automatically generated transcripts produced by transcription software [9,10]. While automatic transcription allows high-throughput speech transcription for a very low cost per sample, these systems can vary in their accuracy (ranging from 68% to 87% in past work [11]), which may affect the performance of downstream linguistic analysis [12]. Furthermore, the impact of automatic preprocessing on classification is not fully understood and should be investigated before continuing downstream investigations.

To better understand the use of automatic transcription for AD/MCI classification, we compared the automatically generated transcripts from Google speech-to-text [13] (“automatic transcripts”) against automatic transcripts that went through a second stage of manual correction (“manually corrected transcripts”). These manually corrected transcripts were used as ground truth.

Specifically, we first examined a confidence metric in the transcription software for transcribing speech recordings from memory clinic patients versus healthy controls. Second, we measured the word-level accuracy of the automatic transcripts against ground truth. Third, we compared classification performances of machine learning models using data from automatic versus manually-corrected transcripts. Based on these results, we discuss accuracy trade-offs associated with manual transcript verification in the context of dementia classification, and we suggest more efficient manual verification methods to improve the performance of automatically generated transcripts.

This investigation aims to highlight differences in human versus automatically processed transcripts to drive future automatic transcription–based research. Therefore, we focus here on comparing transcription methods using existing machine learning algorithms rather than building a novel model that outperforms state-of-the-art models.

This work has 4 main contributions addressing knowledge gaps in the existing literature. First, we evaluate automatic transcription and manual transcription on a data set of older adults for AD/MCI classification using 3 measures: transcription confidence, error rates, and machine learning classification accuracy. To our knowledge, this approach for evaluating transcriptions has not been used previously.

Second, our investigation is novel in that we are exploring the robustness of automatic transcription in a cohort of older adults, including those with cognitive impairment and dementia. The aging process includes changes to voice and speech (eg, presbyphonia, word-finding difficulties), which may affect automatic transcription. However, previous investigations on transcription methods have focused solely on younger or heterogeneous cohorts [12,14]. To our knowledge, this is the first investigation on the impact of transcription methods in a cohort of older adults.

Third, based on the evaluation results, we make practical suggestions about how to use automatic transcription. These suggestions will help researchers to better leverage automatic transcription for building natural language processing (NLP)–based screening methods using large data sets for AD/MCI or subjective memory complaints (SMC), which can be a prodromal state for MCI and AD [15].

Finally, while our results are generated with an AD/MCI data set, our findings could also be extrapolated to other neurological and psychiatric conditions where speech analysis is being investigated as a classification tool. This includes stroke [16], Parkinson disease [17], concussion [18], anxiety [19], bipolar disorder [20], depression, and suicidal ideation [21,22].

Methods

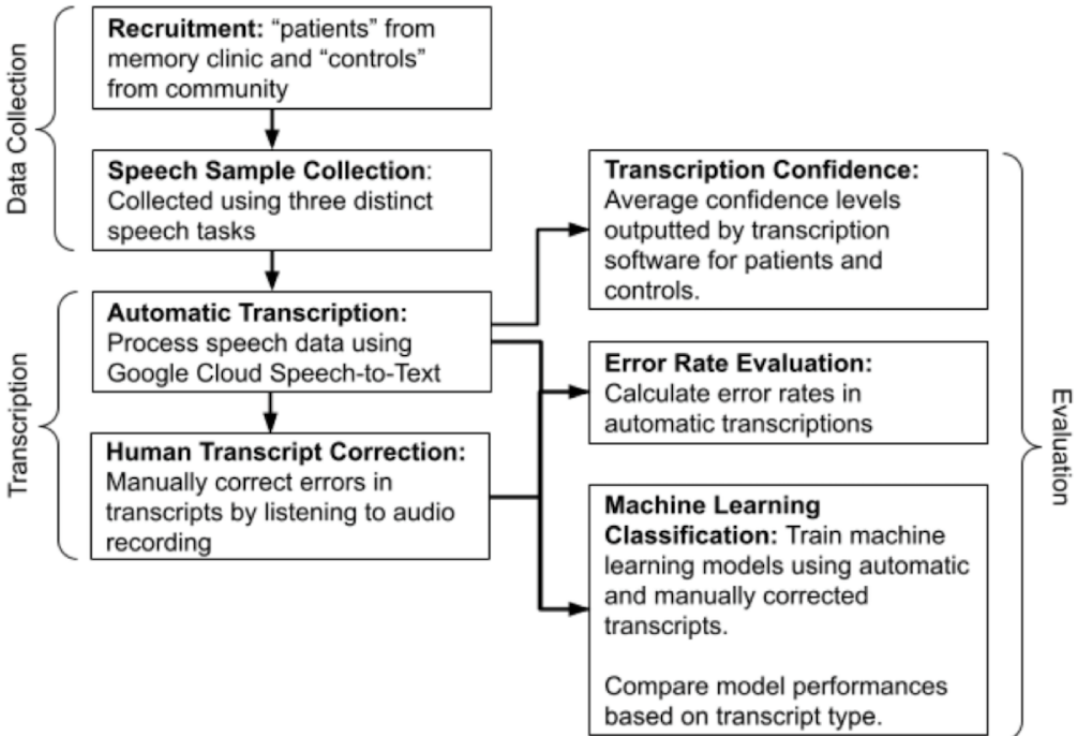
Overview

This study involved 3 main phases: (1) data collection, (2) transcription, and (3) evaluation. Our workflow is summarized in Figure 1. As part of a larger study examining machine learning algorithms for classification of memory clinic patients versus healthy controls, we recruited participants with a clinical diagnosis of mild-to-moderate AD, MCI, or SMC (“patients”)

from a subspecialty memory clinic and healthy volunteer controls from the community. Participants underwent a test battery that included describing the "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination, a reading task incorporating a sixth-grade level paragraph from the International Reading Speed Texts (IReST), and recounting a pleasant past experience. Their speech was recorded, and we used Google Cloud speech-to-text (STT) to automatically transcribe speech data. We then manually corrected errors in the automatic transcripts.

For evaluation, we first aggregated transcription confidence levels provided by the software to determine whether transcription software confidence levels vary between patients and controls. Using manually corrected transcripts as the gold standard, we calculated the error rate of automatic transcripts. Then, we compared the performance of machine learning models trained with either automatic or manually corrected transcripts in classifying transcripts as belonging to "patients" versus "controls."

Figure 1. Diagram of our methods and process.



Data Collection

Recruitment

Patients were recruited from a memory clinic in British Columbia, Canada, and diagnosed with AD, MCI, or SMC. Control participants were recruited from the community, with efforts made to age- and sex-match patient participants. All participants were conversationally fluent in English, could engage in a spontaneous conversation, and were aged 50 or older (mean 68.8, SD 9.5 years). Clinic patients were excluded if they had psychiatric medication changes under 18 months ago or neurological conditions other than SMC, MCI, or AD. We report data from 72 memory clinic patients, of which 44 (30%) were diagnosed with mild-to-moderate AD, 20 (13%) with MCI, and 8 (5%) with SMC (mean age 71.9, SD 8.9 years), along with 77 (52%) healthy volunteers (mean age 65.7, SD 9.1 years).

Diagnoses were made by specialist clinicians using standard-of-care guidelines. The diagnostic process involves a combination of cognitive testing, neuroimaging, laboratory data, medical history, physical exam, and collateral information collected from individuals close to the patient.

Speech Sample Collection

Participants underwent a 10-minute computer-based battery. They were asked to complete a total of 3 speech tasks while their voice was recorded. Participants described the Cookie Theft photo [23], read a standardized paragraph from the IReST, and recalled a pleasant past experience. All tasks were carried out in English. During these spontaneous speech tasks, the audio was recorded using the Logitech C922x ProStream webcam. The Cookie Theft picture description task is a validated spontaneous speech task used extensively in prior work for AD/MCI classification [6,24-26]. This task has also been used for predicting the future risk of developing AD in cognitively normal individuals [27].

For the reading task, a single paragraph was selected from IReST, a collection of short paragraphs (<200 words) designed to be readable at a sixth-grade level [28]. To recreate a natural reading environment such as a book or newspaper, the entire paragraph was presented on the screen at the same time rather than displaying each sentence individually, as in some other investigations [29]. For the final task, participants were asked to describe a pleasant past experience ("experience description task"). Several examples were given to participants prior to

starting the task, such as their first pet, how they met their best friend, or a place they had traveled.

Automatic Transcription

Following the speech tasks, participant audio data was labeled with a unique anonymized identifier and converted to the Waveform audio file format. Next, participant audio was uploaded to the Google Cloud STT platform using US English and 16000 Hz settings, with word-level time stamps enabled, to output the automatic transcripts.

Each transcribed word was labeled as being within a specific task or as being extraneous from all tasks. Words spoken outside of tasks were removed in downstream experiments.

Human Transcript Correction

After automatic transcript files were generated, human transcribers listened to the recorded audio files and made manual corrections to the transcripts based on the recorded audio. This manual transcription involved 3 steps: fixing transcription errors, adding punctuation, and adding filled pauses and silent pause annotations.

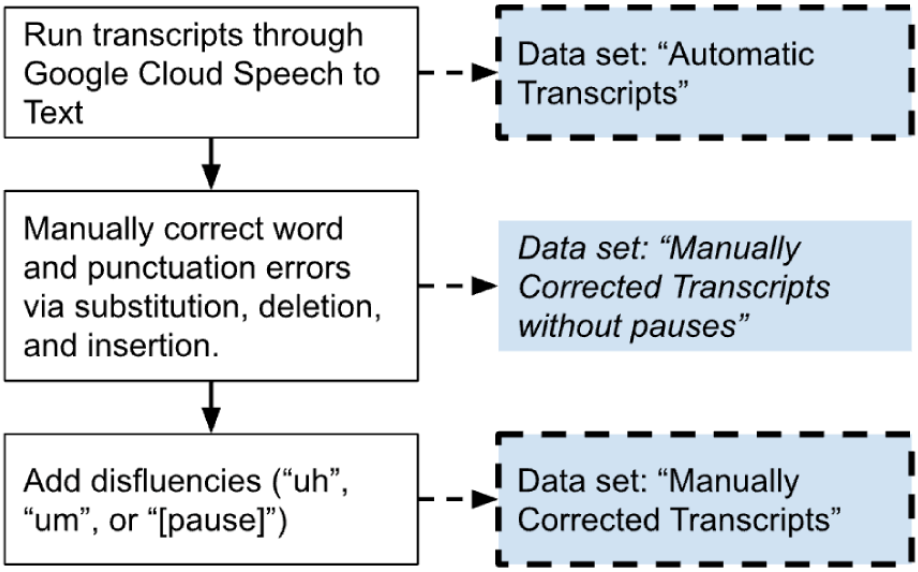
For the first step, which involved fixing transcription errors**,** human transcribers manually substituted incorrectly transcribed words (eg, change "cookie far" to "cookie jar"), inserted missed words (eg, change "cookie" to "cookie jar"), and deleted extra words (eg, change "cookie key jar" to "cookie jar").

The second step entailed adding punctuation. While Google STT adds punctuation, it is very rare, with some transcripts having as few as 0 automatically added punctuation marks. As NLP preprocessing (eg, parsing) benefits from fully formed sentences, human transcribers manually added punctuation (ie, ".", "!", and "?") to the transcripts.

For the third step, which consisted of adding filled pauses and silent pause annotations, human transcribers manually added both filled and silent pauses. A filled pause was considered to be any utterance of "uh" or "um." Filled pauses were consistently transcribed as "uh" or "um" regardless of the length of the pause. Silent pauses were specially labeled as "[pause]" to distinguish this from the word "pause." Silent pauses were considered to be any break or silence in speech for 0.25 seconds or longer, following Goldman-Eisler [30] and Park [31]. Instances where the participant was not speaking but was not silent were not labeled as a pause (eg, coughing or laughing). The duration of pauses was not differentiated.

Figure 2 summarizes the transcription process. Acoustic data were transcribed with Google Cloud STT to generate "automatic transcripts." Then, human transcribers fixed spoken words and added punctuation based on the audio recording to generate "manually corrected transcripts without pauses." Finally, human transcribers manually added both filled and silent pauses to generate the "manually corrected transcripts" data set.

Figure 2. Diagram illustrating how the 3 different transcript data sets were generated.



Ethics Approval

This study was approved by the University of British Columbia Clinical Research Ethics Board (H17-02803). All participants provided their written informed consent prior to participating in this study. Baseline demographic characteristics of the patients and controls are summarized in Multimedia Appendix 1.

Evaluation

Transcription Confidence

For a given audio clip, Google STT outputs transcribed words and a confidence level between 0 and 1. This is calculated by aggregating the likelihood values assigned to each word in the audio. A higher number indicates that the words were more likely to be transcribed accurately. We used these confidence levels to determine whether transcription software confidence levels vary between patients and controls and to determine if patient speech was more difficult to transcribe than control speech.

Error Rate Evaluation

To examine the error rate of automatic transcripts, we compared these to manually corrected transcripts without pauses. We chose not to include pauses because automatic transcripts do not transcribe pauses at all; thus, not denoting a pause would not be considered an error.

We calculated standard measures of transcription accuracy, including word error rate (WER) and match error rate (MER) [32], using a Python package, JiWER (v2.1.0, Vassen [33]). These metrics take into account the number of substitutions (eg, “cookie far” to “cookie jar”), deletions (eg, “cookie key jar” to “cookie jar”), and insertions (eg, “cookie” to “cookie jar”) in the manually corrected transcript.

WER represents the rate of errors to the number of input words. This is calculated as follows:

$$WER = \frac{Substitutions + Deletions + Insertions}{Automatic\ Transcript\ Words + Substitutions + Deletions}$$

WER does not weigh insertions and deletions equally. For example, a 6-word transcript with 30 insertion errors has a WER value of 5, while a 36-word transcript with 30 deletion errors has a WER of 0.83.

MER represents the probability of a given word match being incorrect and is calculated as follows:

$$MER = \frac{Substitutions + Deletions + Insertions}{Automatic\ Transcript\ Words + Substitutions + Deletions + Insertions}$$

For example, a MER of 0.25 means that 1 out of 4 word matches between the manually corrected transcript and automatic transcript will be an error. MER is calculated similarly to WER. However, MER takes into account the maximum number of words between both the automatic and manually edited transcripts, as opposed to only the number of words in only the automatic transcript. MER also weighs insertions and deletions equally.

WER and MER were calculated for each individual transcript. Then, the average and standard deviation of these values were calculated for patients and controls and for each task (eg, picture description, reading, and experience description tasks).

Machine Learning Classification

To determine whether manual correction impacts machine learning classification of patients versus controls, we performed experiments using both the automatic and manually corrected transcript data sets.

Table 1 outlines the entire feature set by task. For the picture description task and the experience description task, we extracted features from transcripts following the text-based features in previous work [6,34]. These features are based on grammar rules, vocabulary, or psycholinguistics. For the experience description task, we did not include information units used for the picture description task, each of which correspond to visual features in the Cookie Theft picture, such as cookie, jar, boy, or girl.

Table 1. Features for machine learning classification models.

Task	Feature groups and number of features (n) in each group
Picture description	• Cookie Theft image information units (13) • Part-of-speech (15), context-free-grammar rules (44), syntactic complexity (24), vocabulary richness (4), psycholinguistic (5), repetitiveness (5)
Reading	• Syllable count (1), pause count (1)[a], total duration (1), total time spent speaking (1), proportion of time spent speaking (1), speech rate (1), average syllable duration (1), pauses per syllable (1)[a], pause rate (1)[a], pause duration (3)[a]
Experience description	• Part-of-speech (15), context-free-grammar rules (44), syntactic complexity (24), vocabulary richness (4), psycholinguistic (5), repetitiveness (5)

[a]These features were computed using acoustic data and transcript data and are also affected by method of pause detection (ie, acoustic vs text data).

For the reading task, we used 12 reading-task–specific features based on the work of Fraser et al [35]. Extracting text features from reading task data may be counterintuitive because each participant reads an identical prompt. However, transcripts may contain repeated words, incorrectly read words, or filled pauses, making transcribed text features potentially informative. Since automatic transcripts do not contain pause information, we first compared automatic transcripts and manually corrected transcripts by using acoustic data to detect unfilled pauses. As an additional comparison for the reading task, we compared using unfilled pauses detected from audio and using unfilled pauses annotated in manually corrected transcripts to determine whether manually adding pauses to transcripts is useful for the reading task or not.

To parse text data and tag parts of speech, we used Stanford CoreNLP [36]. Psycholinguistic features were generated using the MRC database [37], which provides concreteness, familiarity, and imageability scores of English words. Pauses in the reading task were detected using pydub (v0.25.1 [38]), a Python audio processing package. Syllables were detected using Syllables (v1.0.3 [39]), a Python package.

Based on these features, we performed binary classification to distinguish patients from controls. We chose to perform binary classification due to the data size. The number of data samples for finer classes (MCI and SMC) was too small for multiclass classification. To investigate the usefulness of manual correction, we first compared the performance of automatic to manually corrected transcripts. To determine the importance of

pause annotation we compared the performance of manually corrected transcripts with and without pauses.

We tested with 3 classification algorithms that have shown best performances in previous work on dementia classification [40]: logistic regression (LR), random forest (RF), and Gaussian naive Bayes (GNB). In addition, we tested with an end-to-end fine-tuned pretrained model using Bidirectional Encoder Representations from Transformers (BERT) [41] for the picture description and experience description tasks. Note that we did not try BERT models for the reading task because participants read the same text. We used the Python package scikit-learn (v0.19.1 [42]) to perform classification. We used a stratified 10-fold cross-validation approach and repeated this process 10 times in total with differently stratified splits, each generated with a unique random seed. We report the classification performance in terms of area under the receiver operating characteristic curve (AUROC). AUROC is an evaluation metric for classification at various threshold settings and is commonly used for evaluating diagnostic accuracy [43]. The performance metric was averaged over the 10 folds and 10 runs. To remove highly pairwise correlated features and features poorly correlated with the label, we performed correlation feature selection [44]. Highly correlated features were defined as having a Pearson correlation coefficient greater than 0.85, while poorly correlated features had a Pearson correlation coefficient less than 0.20.

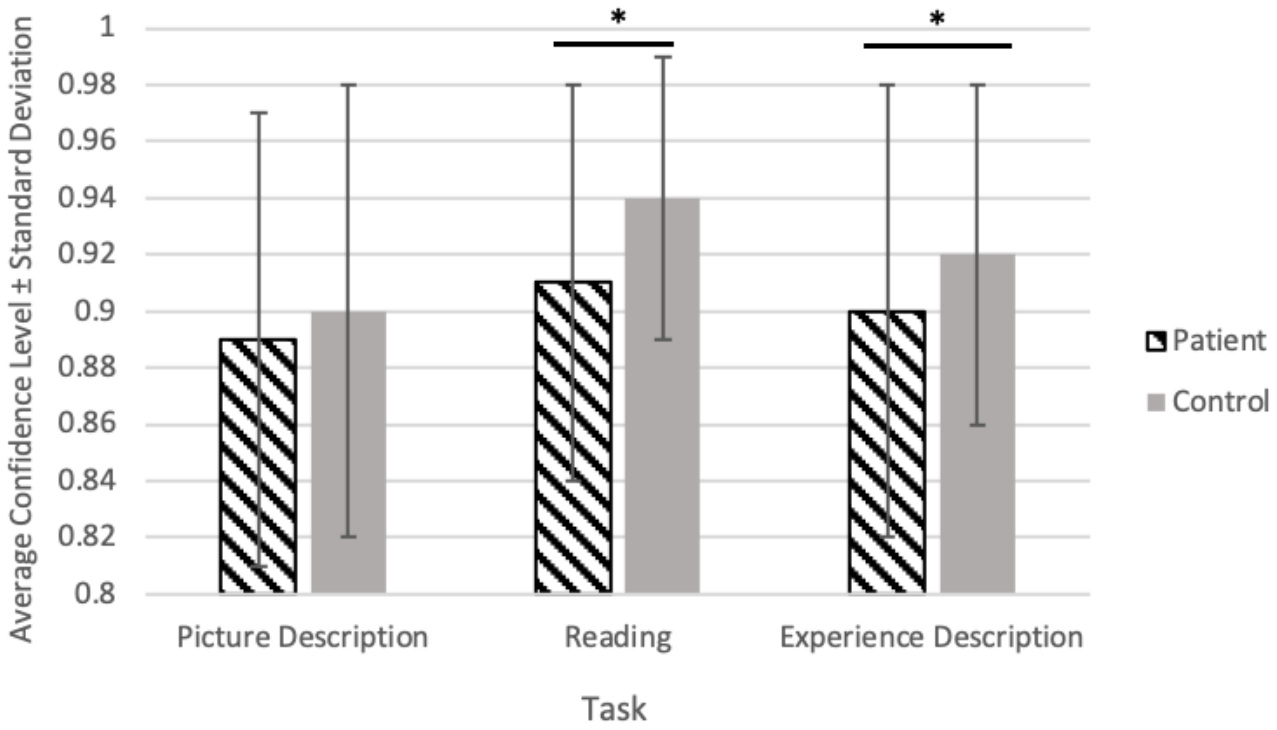
We performed a statistical analysis on the model results to determine if the different transcript data sets led to significant changes in model performance. For each classification algorithm for a given task, we ran a double-sided *t*-test using the null hypothesis that the mean AUROC was no different for automatic and manually corrected transcripts.

Results

Transcription Confidence Results

Google confidence level results are shown in Figure 3. Generally, Google STT produced a higher confidence level when transcribing audio from controls. In the reading task, for example, the average confidence level was 0.94 (SD 0.05) for controls, compared to 0.91 (SD 0.07) for patients. Both the reading and experience description tasks showed a significantly higher confidence level for controls than patients.

Figure 3. Google speech-to-text confidence results. Error bars represent the standard deviation. * represents *P*<.001, calculated by *t*-test.



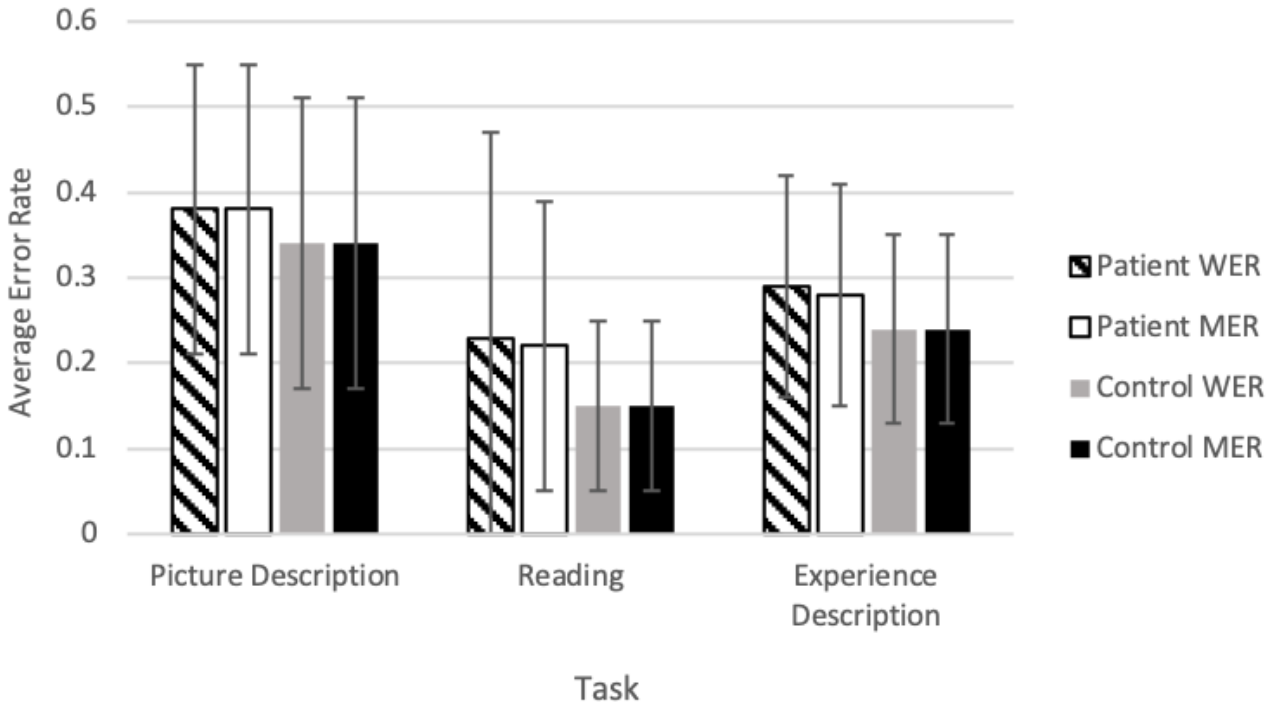
Error Rate Evaluation Results

Figure 4 shows the error rate results. In general, automatic transcription had a lower error rate when transcribing control speech compared to patient speech, as shown by the lower average WER and MER.

The reading task was the most accurate overall, showing an average MER of 0.15 (SD 0.10) for controls and 0.22 (SD 0.19) for patients. This could be because people tend to enunciate when they are asked to read a text aloud. WER and MER results were largely similar overall, suggesting that there were not disproportionately high rates of insertion errors. In other words, manual correction did not involve more word addition as opposed to word deletion or word substitution.

The picture description task was found to have the highest error rate overall when compared to the reading and experience description tasks. This indicates more manual corrections or poorer accuracy of automatic transcription, but it is not clear why this is the case.

Figure 4. Average error rates by task and participant type. Error bars represent the standard deviation. There were no significant differences in error rates between or within tasks. MER: match error rate; WER: word error rate.



Machine Learning Model Results

Models trained on manually corrected transcripts from the picture description and experience description tasks significantly outperformed models trained on automatic transcripts (Table 2). However, there was no significant difference in model performance trained using either transcription method from the reading task. This finding was true regardless of whether pause-related features were included or not (Table 3).

Table 4 shows results of the models using manually corrected transcripts, with and without pauses for the picture description and experience description tasks. There was no clear trend or significant change in any AUROC result when comparing transcripts with and without pauses.

Table 2. Machine learning classification results of models trained on automatic transcripts compared to results of models trained on manually corrected transcripts.

Task and model type	Automatic transcripts AUROC[a]	Manually corrected transcripts AUROC	Change in AUROC[b]
Picture description			
RF[c]	0.617	0.687	0.070[d]
GNB[e]	0.662	0.725	0.063[d]
LR[f]	0.671	0.743	0.072[d]
BERT[g]	0.618	0.686	0.068[d]
Experience description			
RF	0.503	0.636	0.133[d]
GNB	0.549	0.677	0.128[d]
LR	0.543	0.674	0.131[d]
BERT	0.630	0.650	0.020[d]

[a]AUROC: area under the receiver operating characteristic curve.

[b]Positive change in AUROC indicates that the manually corrected transcript model outperformed the automatic transcript model.

[c]RF: random forest.

[d]Indicates $P<.001$.

[e]GNB: Gaussian naive Bayes.

[f]LR: logistic regression.

[g]BERT: Bidirectional Encoder Representations from Transformers.

Table 3. Machine learning classification results of models trained on reading task data with pause features computed using acoustic data or computed using text data.

Reading task	(1) Automatic transcripts AUROC[a,b]	(2) Manually corrected transcripts AUROC[b]	(3) Manually corrected transcripts AUROC[c]	Change in AUROC (3)–(1)
RF[d]	0.638	0.655	0.662	0.024
GNB[e]	0.677	0.677	0.693	0.016
LR[f]	0.589	0.587	0.568	−0.021

[a]AUROC: area under the receiver operating characteristic curve.

[b]Pauses detected from acoustic data.

[c]Pauses detected from text data.

[d]RF: random forest.

[e]GNB: Gaussian naive Bayes.

[f]LR: logistic regression.

Table 4. Machine learning classification results of models trained on manually corrected transcripts without pauses compared to results of models trained on manually corrected transcripts (with pauses).

Task and model type	Transcripts without pauses AUROC[a]	Transcripts with pauses AUROC	Change in AUROC[b]
Picture description			
RF[c]	0.666	0.687	0.021
GNB[d]	0.730	0.725	−0.005
LR[e]	0.755	0.743	−0.012
BERT[f]	0.686	0.691	0.005
Experience description			
RF	0.631	0.636	0.005
GNB	0.676	0.677	0.001
LR	0.692	0.674	−0.018
BERT	0.622	0.649	0.027

[a]AUROC: area under the receiver operating characteristic curve.

[b]Positive change in AUROC indicates that the pause model outperformed the no-pause model.

[c]RF: random forest.

[d]GNB: Gaussian naive Bayes.

[e]LR: logistic regression.

[f]BERT: Bidirectional Encoder Representations from Transformers.

Discussion

Transcription Confidence

The transcription confidence results showed that the automatic transcription software was consistently more confident in transcribing the speech of controls compared to patients. This may indicate that patient speech differs from the speech used to train the automatic transcription software (which was likely trained using speech from a more general population, including younger or cognitively unimpaired individuals). This may be attributed to the fact that people with Alzheimer disease often have impaired speech production [45], such as distortions (eg, “ook” instead of “cookie”) and phonological paraphasias (eg, “tid” instead of “kid”) [46]. It is especially interesting that the confidence difference between the 2 groups was highest and most significant for the reading task. This confirms that reading task speech is effective for distinguishing AD/MCI patients from controls, as also shown in prior work [35,47,48].

Error Rate Evaluation

Automatic transcriptions were more accurate for healthy controls compared to patients with AD or MCI, as shown by higher error rate and information loss in patient transcripts. This result is logical in the context of the confidence result, as patient transcripts had significantly lower confidence, meaning that the transcription software was more unsure about its output.

Our results are markedly different from Google’s reports on the error rates of their own software (Google Cloud STT has not disclosed the composition of their training data set). According to Google, their transcription program achieved a WER of 6.7% using 12,500 hours of voice search data and a WER of 4.1% in a dictation task [49]. In contrast, for spontaneous speech tasks, we found a WER range of 24% to 34% for controls and 29% to 38% for patients. The reading task showed a lower WER of 15% for controls and 23% for patients.

While our error rate results differ from Google’s reported results, they are comparable to the results of other investigations using Google STT derived from simulated medical encounters. Kim et al [50] used audio data from 12 simulated patient and medical student interactions. In this investigation, Google STT showed an average WER of 34%, similar to our WER result of 34% for controls and 38% for patients completing the picture description task. Miner et al [14] recorded audio from 100 patients aged 18-52 (mean age 23) during therapy sessions and found that Google STT had an average WER of 25%. This result is comparable to the WER for our experience description task, which was 29% for patients and 24% for controls. Both therapy-related discussion and the experience description tasks typically involve spontaneous speech with minimal prompting.

Surprisingly, the experience description task showed lower error rates than the picture description task. This might be because Google STT repeatedly transcribed certain phrases or words related to the picture incorrectly across participants, leading to a higher average error rate in this task. It is also possible that the experience description is easier for automatic transcription because it is more conversational, like the material Google STT may have been trained on. Further investigation into to the discrepancy in performance between different spontaneous speech tasks is warranted.

The reading task WER for our cohort was notably higher than previous research. Kepuska et al [51] used Google STT to transcribe audio from 630 speakers reading 10 sentences each and found an average WER of 9%. This is markedly lower than the results of our investigation, in which we found that the

9-sentence reading task produced a WER of 23% for patients and 15% for controls. One possible reason for this large disparity is that Google STT is not specific to a particular population (eg, older adults may experience normal age-related changes to the larynx and vocal cords over time, known as presbyphonia) and may produce more accurate transcriptions in a more generalized sample.

Machine Learning Models

Despite our previous results showing that automatic transcription for our data set is more inaccurate than values reported by Google, our machine learning model results still show that automatic transcripts are discriminative for AD/MCI. Other studies using automatic transcription for classification experiments have noted that inaccuracies or errors in the audio-to-text transcription do not necessarily affect classification results [52].

However, manually correcting the picture description and experience description task transcriptions led to significantly higher performances in the machine learning models. By comparison, both automatic and manually corrected reading task transcripts showed similar performance, likely due to the majority of reading features being computed from audio data. To address this concern, we examined text versus audio-based silent pause detection and again found no significant changes in performance. This indicates that using either audio or text to detect pauses will produce similar results and that manually correcting transcripts does not significantly change model performance.

Surprisingly, the addition of filled and silent pauses did not significantly change performance for any of the tasks and algorithms. Moreover, using the pauses from the transcripts showed similar classification results to using pauses detected from audio data for the reading task. Previous studies have shown that people with Alzheimer disease demonstrate a multitude of disfluencies in their speech, including pauses [53-55]. However, manually adding pauses as either words (“um” or “uh”) or tokens (“[pause]”) to transcripts did not seem to have any effect on classification models. This could be because older adults also experience age-related changes in their speech, such as an increase in silent pauses [56], potentially weakening the association of pauses to either the patient or control category. Alternatively, this result may be due to the fact that there are no features that “directly” model pauses for the description tasks, weakening the association of the tasks with pauses.

Limitations

Some limitations with our cohort include varying language ability and variations between transcribers. In our cohort, English was not the first language of 13% of the patients and 21% of the control group, which could potentially contribute to transcription errors. Additionally, our use of 3 different transcriptionists may have introduced interrater variability, especially for more subjective correction steps such as adding punctuation, although variation in manual transcription was controlled via inter-transcriptionist review and protocol development for standardized transcription. Another limitation of our investigation is the size of the data set (N=149), which is quite small for machine learning experiments. However, this is an issue facing most work on using machine learning for dementia classification, especially with newly built data sets (N=55-82) [5,29,35]. While the DementiaBank and ADReSS data sets are larger (N=287 with 687 samples and N=156, respectively), they were originally created in the mid-1980s and are limited by the diagnostic practices of that time. The work described herein aims to mitigate this challenge. Our best practice suggestions for automatic transcription will facilitate data collection at a much faster rate in the future.

It is also important to note that this investigation was completed using Google speech-to-text software in an English-speaking cohort. Competitor speech-to-text software may produce different results, so readers should be wary when applying our conclusions to other software. Applying a similar method to a non-English data set may also produce different results, especially because automatic transcription in other languages might not be as advanced as English. Finally, speech-to-text software is continually being refined and improved. In the future, automatically generated transcripts may be indistinguishable from human-generated transcripts. In the meantime, it is still valuable to understand the impacts of automatic transcription, especially for medical speech data sets.

Conclusion

Our results showed that automatically transcribed speech data from a web-based speech recognition platform can be effectively used to distinguish patients from controls. According to our results, to improve the classification performance of automatically generated transcripts, especially those generated from spontaneous speech tasks, a human verification step is recommended. Our analyses indicate that human verification should focus on correcting errors and adding punctuation to transcripts and that manual addition of pauses is not needed, which can simplify the human verification step to more efficiently process large volumes of speech data.

Acknowledgments

Vancouver Coastal Health Research Institute, Centre for Aging + Brain Health Innovation, Alzheimer’s Society, and the Canadian Consortium on Neurodegeneration in Aging funded research personnel and equipment necessary to recruit participants, collect data, perform analyses, and synthesize results.

Authors' Contributions

TS, SNM, CC, GM, GC, TSF, and HJ contributed to the conception and design of the study. TS, SNM, and CL recruited study participants and administered study assessments. TS organized the database. TdCV, SG, AH, and HJ designed machine features

and performed all machine learning analyses. TdCV, AH, and HJ performed the statistical analyses. TS performed other analyses. TS and HJ wrote the first draft of the manuscript. TdCV and AH contributed to sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Summary baseline characteristics of patients and controls.
[DOCX File , 13 KB-Multimedia Appendix 1]

References

1. Rasmussen J, Langerman H. Alzheimer's disease - why we need early diagnosis. Degener Neurol Neuromuscul Dis 2019;9:123-130 [FREE Full text] [doi: 10.2147/DNND.S228939] [Medline: 31920420]
2. Mahase E. FDA approves controversial Alzheimer's drug despite uncertainty over effectiveness. BMJ 2021 Jun 08;373:n1462. [doi: 10.1136/bmj.n1462] [Medline: 34103308]
3. Watson JL, Ryan L, Silverberg N, Cahan V, Bernard MA. Obstacles and opportunities in Alzheimer's clinical trial recruitment. Health Aff (Millwood) 2014 Apr;33(4):574-579 [FREE Full text] [doi: 10.1377/hlthaff.2013.1314] [Medline: 24711317]
4. Petti U, Baker S, Korhonen A. A systematic literature review of automatic Alzheimer's disease detection from speech and language. J Am Med Inform Assoc 2020 Nov 01;27(11):1784-1797 [FREE Full text] [doi: 10.1093/jamia/ocaa174] [Medline: 32929494]
5. Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I. Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Comput Speech Lang 2019 Jan;53:181-197. [doi: 10.1016/j.csl.2018.07.007]
6. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. J Alzheimers Dis 2016;49(2):407-422. [doi: 10.3233/JAD-150520] [Medline: 26484921]
7. Becker JT, Boller F, Lopez OL, Saxton J, McGonigle KL. The natural history of Alzheimer's disease. Description of study cohort and accuracy of diagnosis. Arch Neurol 1994 Jun;51(6):585-594. [doi: 10.1001/archneur.1994.00540180063015] [Medline: 8198470]
8. Bazillon T, Esteve Y, Luzzati D. Manual vs assisted transcription of prepared and spontaneous speech. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation.: European Language Resources Association (ELRA); 2008 May Presented at: LREC'08; May 2008; Marrakech, Morocco URL: https://aclanthology.org/L08-1522/
9. Mirheidari B, Blackburn D, Walker T, Venneri A, Reuber M, Christensen H. Detecting signs of dementia using word vector representations. 2018 Sep Presented at: Interspeech 2018; September 2018; Hyderabad, India p. 1893-1897 URL: https://www.isca-speech.org/archive/interspeech_2018/mirheidari18_interspeech.html [doi: 10.21437/Interspeech.2018-1764]
10. Luz S, Haider F, de la Fuente S, Fromm D, MacWhinney B. Detecting cognitive decline using speech only: the ADReSSo challenge. 2021 Aug Presented at: Interspeech 2021; August 2021; Brno, Czechia p. 3780-3784 URL: https://www.isca-speech.org/archive/interspeech_2021/luz21_interspeech.html [doi: 10.21437/Interspeech.2021-1220]
11. Chlebek P, Shriberg E, Lu Y, Rutowski T, Harati A, Oliveira R. Comparing speech recognition services for HCI applications in behavioral health. In: Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers. USA: Association for Computing Machinery; 2020 Presented at: 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 12; Virtual Event p. 483-487 URL: https://dl.acm.org/doi/10.1145/3410530.3414372 [doi: 10.1145/3410530.3414372]
12. Pentland S, Spitzley L, Fuller C, Twitchell D. Data quality relevance in linguistic analysis: the impact of transcription errors on multiple methods of linguistic analysis. 2019 Presented at: Americas' Conference on Information Systems 2019; August 15; Cancun, Mexico.
13. Google Cloud Speech-to-Text. URL: https://cloud.google.com/speech-to-text [accessed 2019-11-19]
14. Miner AS, Haque A, Fries JA, Fleming SL, Wilfley DE, Terence Wilson G, et al. Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ Digit Med 2020;3:82 [FREE Full text] [doi: 10.1038/s41746-020-0285-8] [Medline: 32550644]
15. Kaup AR, Nettiksimmons J, LeBlanc ES, Yaffe K. Memory complaints and risk of cognitive impairment after nearly 2 decades among older women. Neurology 2015 Nov 24;85(21):1852-1858 [FREE Full text] [doi: 10.1212/WNL.0000000000002153] [Medline: 26511452]
16. Rigas E, Papoutsoglou M, Tsakpounidou K, Serdari A, Mouza E, Proios H. Analysis of spontaneous speech using natural language processing techniques to identify stroke symptoms. Encephalos 2020 Jan;57(1):1-12 [FREE Full text]

17. Moro-Velazquez L, Gomez-Garcia J, Arias-Londoño J, Dehak N, Godino-Llorente J. Advances in Parkinson's disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. Biomed Signal Process Control 2021 Apr;66:102418. [doi: 10.1016/j.bspc.2021.102418]
18. Falcone M, Yadav N, Poellabauer C, Flynn P. Using isolated vowel sounds for classification of mild traumatic brain injury. : IEEE; 2013 Presented at: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing; May 26-31; Vancouver, BC p. 7577-7581 URL: https://ieeexplore.ieee.org/document/6639136 [doi: 10.1109/ICASSP.2013.6639136]
19. McGinnis EW, Anderau SP, Hruschak J, Gurchiek RD, Lopez-Duran NL, Fitzgerald K, et al. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. IEEE J Biomed Health Inform 2019 Nov;23(6):2294-2301 [FREE Full text] [doi: 10.1109/JBHI.2019.2913590] [Medline: 31034426]
20. Karam Z, Provost E, Singh S, Montgomery J, Archer C, Harrington G, et al. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. 2014 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014; May 4; Florence, Italy p. 4858-4862 URL: https://ieeexplore.ieee.org/document/6854525 [doi: 10.1109/icassp.2014.6854525]
21. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Communication 2015 Jul;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
22. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cogn Emot 2004 Dec;18(8):1121-1133. [doi: 10.1080/02699930441000030]
23. Goodglass H, Kaplan E. BDAE: the Boston Diagnostic Aphasia Examination. proedinc.com. Philadelphia, PA: Lippincott Williams & Wilkins; 1972. URL: https://www.proedinc.com/Products/11850/bdae3-boston-diagnostic-aphasia-examinationthird-edition.aspx?bCategory=TBI!APXDY [accessed 2022-09-06]
24. Karlekar S, Niu T, Bansal M. Detecting linguistic characteristics of Alzheimer’s dementia by interpreting neural models. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1; New Orleans, LA p. 701-707. [doi: 10.18653/v1/N18-2110]
25. Kong W, Jang H, Carenini G, Field T. A neural model for predicting dementia from language. In: Proceedings of Machine Learning Research. 2019 Presented at: 4th Machine Learning for Healthcare Conference; August 8; Ann Arbor, MI p. 270-286.
26. Cummings L. Describing the Cookie Theft picture. Pragmat Soc 2019 Jul 5;10(2):153-176. [doi: 10.1075/ps.17011.cum]
27. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. EClinicalMedicine 2020 Nov;28:100583 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100583] [Medline: 33294808]
28. Trauzettel-Klosinski S, Dietz K, IReST Study Group. Standardized assessment of reading performance: the New International Reading Speed Texts IReST. Invest Ophthalmol Vis Sci 2012 Aug 13;53(9):5452-5461. [doi: 10.1167/iovs.11-8284] [Medline: 22661485]
29. Biondi J, Fernandez G, Castro S, Agamennoni O. Eye movement behavior identification for Alzheimer's disease diagnosis. J. Integr. Neurosci 2018 Nov 15;17(4):1702-00837. [doi: 10.31083/j.jin.2018.04.0416]
30. Goldman-Eisler F. Speech production and the predictability of words in context. Q J Exp Psychol (Hove) 2018 Jan 01;10(2):96-106. [doi: 10.1080/17470215808416261]
31. Park S. Measuring fluency: Temporal variables and pausing patterns in L2 English speech. Dissertation, Perdue University. 2016 Apr. URL: https://docs.lib.purdue.edu/open_access_dissertations/692/ [accessed 2021-01-01]
32. Errattahi R, El Hannani A, Ouahmane H. Automatic speech recognition errors detection and correction: a review. Procedia Comput Sci 2018;128:32-37. [doi: 10.1016/j.procs.2018.03.005]
33. Vaessen N. JiWER: Similarity measures for automatic speech recognition evaluation. Python Package Index. 2018 Jun 19. URL: https://pypi.org/project/jiwer/ [accessed 2021-01-01]
34. Barral O, Jang H, Newton-Mason S, Shajan S, Soroski T, Carenini G, et al. Non-invasive classification of Alzheimer’s disease using eye tracking and language. In: Proceedings of the 5th Machine Learning for Healthcare Conference. 2020 Presented at: 5th Machine Learning for Healthcare Conference; August 7; Virtual p. 813-841.
35. Fraser KC, Lundholm Fors K, Eckerström M, Öhman F, Kokkinakis D. Predicting MCI status from multimodal language data using cascaded classifiers. Front Aging Neurosci 2019;11:205. [doi: 10.3389/fnagi.2019.00205] [Medline: 31427959]
36. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics; June; Baltimore, MD p. 55-60 URL: https://stanfordnlp.github.io/CoreNLP/ [doi: 10.3115/v1/P14-5010]
37. MRC Psycholinguistic Database: machine usable dictionary. University of Western Australia. 1987 Apr 01. URL: https://websites.psychology.uwa.edu.au/school/MRCDatabase/mrc2.html [accessed 2020-01-01]
38. Robert J. Pydub. Python Package Index. 2011 May 03. URL: https://github.com/jiaaro/pydub [accessed 2019-12-31]
39. Day D. Syllables: A fast syllable estimator for Python. Python Package Index. 2018 Nov 25. URL: https://github.com/prosegrinder/python-syllables [accessed 2020-01-01]

40. Masrani V. Detecting dementia from written and spoken language. Doctoral Dissertations, The University of British Columbia. Vancouver, BC; 2018 Jan 08. URL: https://doi.org/10.14288/1.0362923 [accessed 2020-12-31]
41. Devlin J, Chang M, Lee K, Toutanova K. 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. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics; June; Minneapolis, MN p. 4171-4186 URL: https://aclanthology.org/N19-1423/ [doi: 10.18653/v1/N19-1423]
42. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
43. Šimundić AM. Measures of diagnostic accuracy: basic definitions. EJIFCC 2009 Jan;19(4):203-211 [FREE Full text] [Medline: 27683318]
44. Hall MA. Correlation-based feature selection for machine learning. Doctoral Dissertation, The University of Waikato. Waikato, New Zealand; 1999. URL: https://researchcommons.waikato.ac.nz/handle/10289/15043 [accessed 2021-01-01]
45. Ahmed S, Haigh AF, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain 2013 Dec;136(Pt 12):3727-3737 [FREE Full text] [doi: 10.1093/brain/awt269] [Medline: 24142144]
46. Kohn S. Phonological production deficits in aphasia. In: Phonological Processes and Brain Mechanisms. New York, NY: Springer; 1988:93-117.
47. Mirzaei S, El Yacoubi M, Garcia-Salicetti S, Boudy J, Kahindo C, Cristancho-Lacroix V, et al. Two-stage feature selection of voice parameters for early Alzheimer's disease prediction. IRBM 2018 Dec;39(6):430-435. [doi: 10.1016/j.irbm.2018.10.016]
48. Martínez-Sánchez F, Meilán JJG, Vera-Ferrandiz JA, Carro J, Pujante-Valverde IM, Ivanova O, et al. Speech rhythm alterations in Spanish-speaking individuals with Alzheimer's disease. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 2017 Jul;24(4):418-434. [doi: 10.1080/13825585.2016.1220487] [Medline: 27684109]
49. Chiu C, Sainath T, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, et al. State-of-the-art speech recognition with sequence-to-sequence models. 2018 Presented at: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing; April 15; Calgary, AB p. 4774-4778. [doi: 10.1109/ICASSP.2018.8462105]
50. Kim J, Liu C, Calvo R, McCabe K, Taylor S, Schuller B, et al. A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech. In: 12th International Workshop on Spoken Dialog System Technology Proceedings. 2021 Nov 15 Presented at: 12th International Workshop on Spoken Dialog System Technology; 15 Novmeber 2021; Singapore. [doi: 10.1007/978-3-030-14596-5_8]
51. Këpuska V. Comparing speech recognition systems (Microsoft API, Google API And CMU Sphinx). Int J Eng Res Appl 2017 Mar;07(03):20-24. [doi: 10.9790/9622-0703022024]
52. Murray G, Carenini G. Summarizing spoken and written conversations. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.: Association for Computational Linguistics; 2008 Presented at: 2008 Conference on Empirical Methods in Natural Language Processing; Honolulu, HI p. 773-782. [doi: 10.3115/1613715.1613813]
53. Pistono A, Jucla M, Barbeau EJ, Saint-Aubert L, Lemesle B, Calvet B, et al. Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. J Alzheimers Dis 2016;50(3):687-698 [FREE Full text] [doi: 10.3233/JAD-150408] [Medline: 26757034]
54. Pistono A, Pariente J, Bézy C, Lemesle B, Le Men J, Jucla M. What happens when nothing happens? An investigation of pauses as a compensatory mechanism in early Alzheimer's disease. Neuropsychologia 2019 Feb 18;124:133-143 [FREE Full text] [doi: 10.1016/j.neuropsychologia.2018.12.018] [Medline: 30593773]
55. Sluis RA, Angus D, Wiles J, Back A, Gibson T, Liddle J, et al. An automated approach to examining pausing in the speech of people with dementia. Am J Alzheimers Dis Other Demen 2020 Jul 10;35:1533317520939773 [FREE Full text] [doi: 10.1177/1533317520939773] [Medline: 32648470]
56. Bóna J. Temporal characteristics of speech: the effect of age and speech style. J Acoust Soc Am 2014 Aug;136(2):EL116-EL121. [doi: 10.1121/1.4885482] [Medline: 25096134]

Abbreviations

AD: Alzheimer disease
AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
GNB: Gaussian naive Bayes
IReST: International Reading Speed Texts
LR: logistic regression
MCI: mild cognitive impairment
MER: match error rate
NLP: natural language processing

RF: random forest
SMC: subjective memory complaints
STT: speech-to-text
WER: word error rate

Edited by T Leung, J Wang; submitted 08.09.21; peer-reviewed by S Kim, M Burns, X Zhou, Y Liu; comments to author 14.03.22; revised version received 11.07.22; accepted 23.07.22; published 21.09.22

Please cite as:
Soroski T, da Cunha Vasco T, Newton-Mason S, Granby S, Lewis C, Harisinghani A, Rizzo M, Conati C, Murray G, Carenini G, Field TS, Jang H
Evaluating Web-Based Automatic Transcription for Alzheimer Speech Data: Transcript Comparison and Machine Learning Analysis
JMIR Aging 2022;5(3):e33460
URL: https://aging.jmir.org/2022/3/e33460
doi: 10.2196/33460
PMID: