INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech recognition in Alzheimer's disease and in its assessment

Luke Zhou[1], *Kathleen C. Fraser*[1], *Frank Rudzicz*[2,1]

[1]Department of Computer Science, University of Toronto;
[2]Toronto Rehabilitation Institute, University Health Network;
luke.zhou@mail.utoronto.ca, {kfraser,frank}@cs.toronto.edu

Abstract

Narrative, spontaneous speech can provide a valuable source of information about an individual's cognitive state. Unfortunately, clinical transcription of this type of data is typically done by hand, which is prohibitively time-consuming. In order to automate the entire process, we optimize automatic speech recognition (ASR) for participants with Alzheimer's disease (AD) in a relatively large clinical database. We extract text features from the resulting transcripts and use these features to identify AD with an SVM classifier. While the accuracy of automatic assessment decreases with increased WER, this is weakly correlated (-0.31). This relative robustness to ASR error is aided by selecting features that are resilient to ASR error.

Index Terms: speech recognition; older voices; Alzheimer's disease; assessment

1. Introduction

Diagnosing and screening for Alzheimer's disease (AD) is an expensive and laborious process. Current approaches to the detection of AD and other dementias are unsustainable, given that the incidence of dementia is expected to rise significantly as the global population ages [1]. To ease this burden, modern approaches are being proposed to automate relevant aspects of assessment. Since AD can diminish vocabulary, syntactic complexity, and speech fluency, even in the earliest stages, various systems have been proposed to automatically detect signs of cognitive impairment from speech [2, 3, 4, 5, 6, 7, 8].

Currently, to extract clinically useful lexicosyntactic measures, spontaneous speech must be transcribed by professionals, which is impossible on a very large scale. Unfortunately, no available automatic speech recognition (ASR) system exists for people with AD or cognitive disorders. Moreover, even an optimized system would undoubtedly produce errorful transcripts, which may deteriorate automated assessment performance. For example, estimating the number of nouns in a narrative sample depends on those nouns being correctly recognized.

In this work, we first produce alternative ASR systems to optimize performance in older adults with and without dementia. We then use the resulting transcripts in our existing automatic assessment [8] to classify speakers by diagnostic group (healthy vs. AD), and discuss the features which are both relevant to binary classification and robust to the noisy recognition.

2. Previous work

In this study, we use speech recognition as the input to a system that can analyze a spoken narrative and predict whether the speaker is cognitively healthy or has AD. Other studies in this area have used manually transcribed transcripts [2, 4, 7, 8]. One strategy which combines ASR technology with manual transcripts is to use forced-alignment to measure features such as rate of speech [3, 9]. However, for a speech analysis system to be available online or as part of an in-home continuous monitoring system, there must be no reliance on manual transcriptions at the word-level, which forced-alignment requires.

In general, the accuracy of ASR systems on elderly voices tends to decrease with the age of the speaker [10]. Elderly voices typically have increased breathiness, jitter, shimmer, and a decreased rate of speech [10]. Older speakers may also exhibit articulation difficulties, changes in fundamental frequency, and decreased voice intensity [11]. These factors can result in speech that is less intelligible to both human listeners and ASR systems. For example, Hakkani-Tur *et al.* [12] found that in automatic scoring of a speech-based cognitive test, their ASR system had a higher WER for healthy speakers over the age of 70 than for those under the age of 70, with WERs between 26.3% and 34.1% for the elderly speakers, depending on the task and the gender of the speaker, while the error rates ranged between 21.1% and 28.2% for the younger speakers.

Effective speech recognition can be further challenged by the presence of linguistic impairments such as those occurring in dementia; however, there have been relatively few results reported in this area. Peintner et al. [13] analyzed speech from patients with frontotemporal lobar degeneration, two variants of which affect language: progressive nonfluent aphasia (PNFA) and semantic dementia (SD). They achieved a WER of 61% for speakers with PNFA and 37% for those with SD. They also tested a control group, who had an average WER of 20%.

In previous work, we used commercial ASR to generate transcripts of narrative speech from patients with primary progressive aphasia, which is a neurodegenerative language impairment [14]. We used Nuance Dragon NaturallySpeaking 12.5 Premium with the 'older voices' model, and tested it both with the default vocabulary and with a reduced vocabulary of words relevant to our task. Counter-intuitively, Dragon gave higher WER with the reduced vocabulary (97.5%) than with the default vocabulary (67.5%). Because that system is proprietary, we were unable to investigate the changes to the underlying models that led to this result. To overcome this issue, we now use the open-source Kaldi ASR toolkit and a relatively large corpus of speech in AD.

3. Data

Our data are derived from the DementiaBank (**DB**) corpus[1] [15]. These data were collected between 1983 and 1988 at the University of Pittsburgh. Information about the study cohort is

[1] Downloaded from https://talkbank.org/DementiaBank/ on November 22, 2013

http://dx.doi.org/10.21437/Interspeech.2016-1228

available from Becker *et al.* [16]. From the 'AD' group, we include participants with a diagnosis of 'possible' or 'probable' AD, resulting in 240 samples from 167 participants over 298.34 minutes. We also include 'CTRL' (control) participants, resulting in 233 additional files from 97 speakers over 220.09 minutes. Narrative speech was elicited using the standard 'Cookie Theft' picture description task from the Boston Diagnostic Aphasia Examination [17]. Each speech sample was recorded and manually transcribed at the word level following the CHAT protocol [18]. The AD participants produced an average of 104.3 ($\sigma = 59.0$) words per narrative, while the control participants produced an average of 114.4 ($\sigma = 59.5$) words per narrative. We convert the audio associated with each transcript from MP3 to 16-bit, 16 kHz mono WAV.

To train a baseline system, we also use the Wall Street Journal (phase 2) corpus (WSJ) [19], which consists of 78,000 training utterances ($\sim$73 hours of speech), 4,000 of which are the result of spontaneous dictation by journalists.

4. Experiments

Two experiments are conducted. In the first, we optimize ASR for individuals with AD performing the Cookie Theft task. In the second, we perform binary classification of AD using both manual transcripts, as in our previous work, and transcripts derived from the optimized ASR.

4.1. Experiment I: ASR for Alzheimer's disease

Four experimental systems are produced with Kaldi, using various subgroups of the WSJ and DB data. The *mono* model uses monophones, *tri1* is a triphone model with δ and $\delta\delta$ features, *tri2* adds linear discriminant analysis and maximum likelihood linear regression (MLLR) transforms, and *tri3* also adds speaker-adaptive training using full feature space MLLR [20].

All experiments use 10-fold cross-validation. In all cases, speakers are partitioned across folds so that individuals never appear in both the training and test sets, and so that each fold is balanced by the diagnosis of the speaker where relevant. In all cases, lexicons and finite-state transducer grammars are obtained on training sets. Insertion penalties (i.e., the cost of inserting a new word into the transcript) are empirically swept from 0.0 to 1.0. The language model (LM) weights, which regulate the importance of the language model relative to the acoustic model, vary along the default range from 9 to 20.

We first verify our methodology by training and testing on WSJ data, using the cross-validation strategy outlined above. Our best WER is 3.72%, using the *tri3* model, which is comparable to state-of-the-art for WSJ [21], indicating that our models work well on traditional speech data.

4.1.1. *Train on WSJ, test on DB*

We test how well the models trained on the WSJ data generalize to older voices and the Cookie Theft task. Here, the models are trained on WSJ data but tested on DB data. We again use the 10-fold cross validation strategy, to allow for appropriate comparisons across experiments. We test with both CTRL and AD data separately, to observe any differences in WER between groups, as shown in Table 1. In both cases, the best WER is achieved with the *tri3* model and an insertion penalty of 1.0, but even the best WER is extremely high: 91.74% among CTRLs and 93.54% among those with AD.

Tr.	Te.	Model	Pen 0.0	Pen 0.5	Pen 1.0
WSJ	DB CTRL	mono	97.65 (0.34)	97.33 (0.23)	97.20 (0.16)
		tri1	94.94 (0.65)	94.49 (0.47)	94.23 (0.33)
		tri2	93.88 (0.56)	93.61 (0.43)	93.46 (0.29)
		tri3	92.15 (0.49)	91.88 (0.41)	**91.74** (0.32)
	DB AD	mono	98.47 (0.52)	98.08 (0.37)	97.89 (0.26)
		tri1	96.37 (0.93)	95.84 (0.70)	95.53 (0.52)
		tri2	95.51 (0.78)	95.14 (0.62)	94.93 (0.49)
		tri3	94.07 (0.60)	93.75 (0.59)	**93.54** (0.39)
DB CTRL	DB CTRL	mono	42.81 (0.73)	45.07 (0.66)	47.66 (0.87)
		tri1	39.14 (1.64)	40.57 (1.06)	42.12 (0.71)
		tri2	42.84 (1.45)	44.24 (0.93)	45.61 (0.65)
		tri3	**36.28** (1.63)	37.06 (1.23)	37.92 (0.93)
DB AD	DB AD	mono	54.92 (0.99)	57.66 (0.81)	60.15 (0.73)
		tri1	49.30 (1.99)	50.51 (1.32)	51.90 (0.85)
		tri2	51.99 (2.25)	53.24 (1.55)	54.67 (1.10)
		tri3	**44.97** (2.26)	45.59 (1.91)	46.33 (1.47)
DB BOTH	DB BOTH	mono	51.08 (0.92)	53.41 (1.10)	55.71 (1.08)
		tri1	41.46 (1.81)	42.43 (1.23)	43.51 (0.82)
		tri2	43.38 (2.08)	44.35 (1.59)	45.38 (1.18)
		tri3	**38.24** (1.80)	38.69 (1.55)	39.26 (1.24)

Table 1: WER for 4 models and 3 insertion penalties, averaged over 10 folds and 13 LM weights. Results in **bold** are optimal, within each configuration of training (tr.) and testing (te.) data.

4.1.2. *Train on DB, test on DB*

We now train models on DB data alone. While we do not expect these models to generalize beyond the Cookie Theft task, we do expect that constraining the vocabulary to the specific task, and the acoustic models to the specific recording conditions, will improve accuracy. This is confirmed by the greatly improved WER (Table 1). The lowest WERs are again achieved using the *tri3* models, but with an insertion penalty of 0. The best WER for controls (36.28%) is significantly lower ($t(22) = 10.79, p < 0.001$) than for people with AD (44.97%), in keeping with previous work [22].

Although we separated CTRL and AD data to observe the difference in WER between groups, in real-life screening or diagnosis, such as those described in Section 1, which ASR model to use will not be known *a priori*. Therefore, we also train a model which combines data from both the CTRL and AD data, and test on a combination of these data. In this case, we obtain a best average WER of 38.24%. This result is closer to the CTRL result than the AD result when trained separately, possibly a benefit of increased training set size.

Ultimately, our results suggest that for this particular task and population, training on a relatively small amount of in-domain data achieves better results than training on a large amount of out-of-domain data.

4.2. Experiment II: Diagnosis with noisy transcripts

With the ASR transcripts from the previous section, we can now classify according to diagnosis. We use transcripts generated by models trained on a combination of AD and control data, to mimic expected real-life scenarios. We use the same 10-fold cross-validation framework for assessing binary classification accuracy, using the same folds as in the ASR experiments. This ensures that no information from any test set is present in the training data. Each datum represents one session.

4.2.1. *Diagnostic accuracy*

To test the effect of WER on classification, we consider each combination of LM weight and insertion penalty separately, and we report the results for each fold. We use language-based fea-

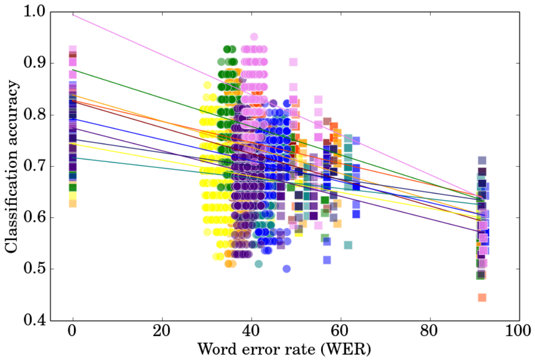


Figure 1: Classification accuracy by WER. Individual points are colored according to the folds to which they belong. Round points are generated from the final experiment in Section 4.1. Square points are generated from gold transcripts and a subset of the other experiments in Section 4.1, and included to show how accuracy changes over the extremes of the WER range.

tures, described in [8], derived from the transcripts as input to an SVM classifier with a second-degree polynomial kernel [23]. We do not consider acoustic and prosodic features, as we have in the past, in order to focus on text-based features that can be affected by ASR errors. We select the N most relevant features (N ranging from 5 to 120) using a filter method based on the correlation between feature and class in the training set [8].

Binary classification results are plotted against WER in Figure 1. Clearly, there is a large range of classification accuracies associated with a fairly narrow range of WER. By coloring the points associated with each fold separately, we can see that some of this variance is due to the variability between folds. An n-way ANOVA shows no effect of insertion penalty ($F_{2,4679} = 1.00, p = 0.50$), acoustic scale ($F_{11,4679} = 1.56, p = 0.10$), WER ($F_{1,4679} = 3.31, p = 0.07$), or N ($F_{1,4679} = 2.23, p = 0.14$) on classification accuracy.

However, WERs generated by ASR trained on DB (29%<WER<49%) do not cover the full range of possible WERs. For this reason, Figure 1 also includes accuracies achieved using the gold transcripts (WER=0%), transcripts using the *mono* model from the experiments outlined in Section 4.1.2 (44%<WER<64%), and transcripts using the models trained on WSJ (WER≈90%). The gap occurring between roughly 64% to 90% is due to the fact that none of the experiments produced WERs in that range.

Using these additional data, lines-of-best-fit are applied to each fold, with an average $R^2 = 0.17$, and an overall Pearson correlation between WER and accuracy of $r = -0.31$, indicating a *weak* negative correlation between WER and diagnostic accuracy. When we include data from the additional models and the gold transcripts in an ANOVA, there are significant effects of WER ($F_{1,5199} = 738.52, p < 0.001$) and N ($F_{1,5199} = 5.86, p < 0.05$).

4.2.2. Selected features

It is important to know which language features are still relevant to diagnosis, given errorful ASR transcripts. Since it gives near-optimum accuracy, we restrict our analysis to the $N = 10$ case, and consider only the data trained and tested on DB using the *tri3* model, as well as the gold-standard data. Since different features can be selected in each fold, we report the proportion of folds in which each feature is selected. We subdivide the training folds by WER, to determine if the selected features vary with WER, as shown in Figure 2.

Three features are selected across all folds and WERs, including gold transcripts: *frequency* (i.e., the frequency with which a word occurs in the 'SUBTL' speech corpus [24]), *word length* in characters, and the frequency of the construction *VP* → *VBG PP* (verb phrases as gerund verb and prepositional phrase). The number of *verbs* and the informational feature *concept: window* are also selected more than 50% of the time across WERs. These features appear to be robust to ASR error. Other features are selected very frequently in the gold transcripts but rarely in ASR transcripts, including *NP* → *PRP* (noun phrases as prepositional phrases), *not-in-dictionary'* (NID, aka out-of-vocabulary) words, and *pronoun ratio* (i.e., the ratio of pronouns to pronouns + nouns). Other features are selected *only* in the ASR transcripts, e.g., *keyword: window* and *concept: girl*. That these features distinguish CTRL from AD only in those transcripts suggests an asymmetry in how related words are recognized in the two groups.

Interestingly, the *concept: window* feature (which includes mentions of words relating to windows, such as "frame" or "glass") is selected across the range of WERs, but the *keyword: window* feature (which includes only mentions of the word "window") is only selected in ASR transcripts. In fact, both of these features are significantly different between the groups in the gold transcripts, but the difference is greater for the *concept*, and the *keyword* is not generally selected until we allow $N = 20$. In the ASR transcripts, there are fewer significant differences, and so *keyword: window* is selected earlier.

4.2.3. Significance of features

To compare how features calculated from gold transcripts differ when calculated from ASR transcripts (trained on DB, LM weight=9, insertion penalty=1.0), we construct bubble plots comparing the p-values obtained, in each condition, from two-tailed, heteroscedastic t-tests on each feature between the AD and CTRL populations. Low p-values in each dimension of these plots indicate features that differentiate CTRL from AD in gold and ASR transcripts, respectively. Each bubble represents a feature, and its size and color indicate the correlation between feature values in the gold and ASR transcripts. Small bubbles indicate that the values are relatively uncorrelated, while large bubbles indicate that the values are highly correlated. Because values of p span a wide scale, we show bubble plots with both linear (Figure 3a) and log-scaled axes (Figure 3b). Due to the large number of features, only select features are labelled.An interactive version of these plots is online at `http://www.cs.toronto.edu/~kfraser/testplot.html`.

In Figure 3a, most visible bubbles are *not* significant at $\alpha = 0.05$; however, the *position* of the bubbles is useful. Specifically, bubbles in the upper-right quadrants are features which have no diagnostic utility in either the gold or ASR transcripts. Bubbles in the lower-right correspond to features that are more useful (lower p) in the ASR transcripts than the gold transcripts. Bubbles in the upper-left are more useful in the gold transcripts than the ASR transcripts, and bubbles in the lower-left (as $p \to 0$ on each axis) are features that are always useful.

This pattern of interpretation extends to Figure 3b, where we can more easily see features whose differentiating power is always significant, e.g., *word length* and *frequency*, as sug-

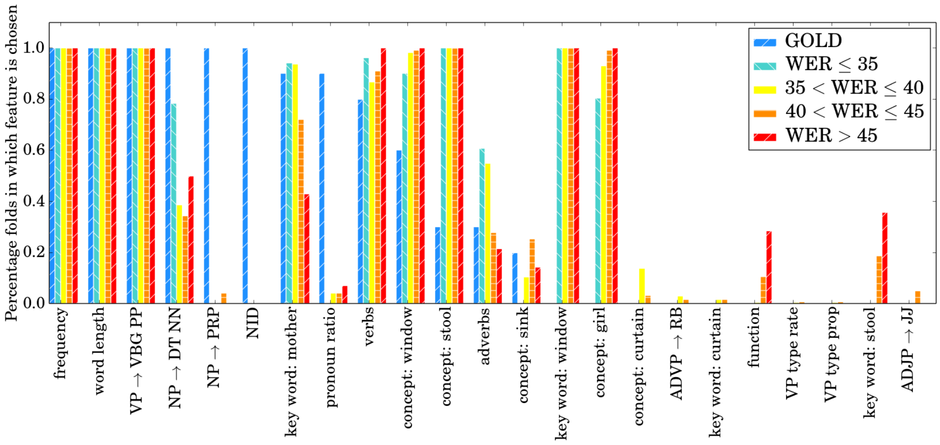


Figure 2: Percentage of folds ($N = 10$) in which each indicated feature was selected, across the given range of word error rates.

gested by Figure 2. Other features in Figure 2 also appear here. In general, bubbles are arranged linearly, with some exceptions such as *NID*, which is more significant given gold transcripts.

5. Discussion and future work

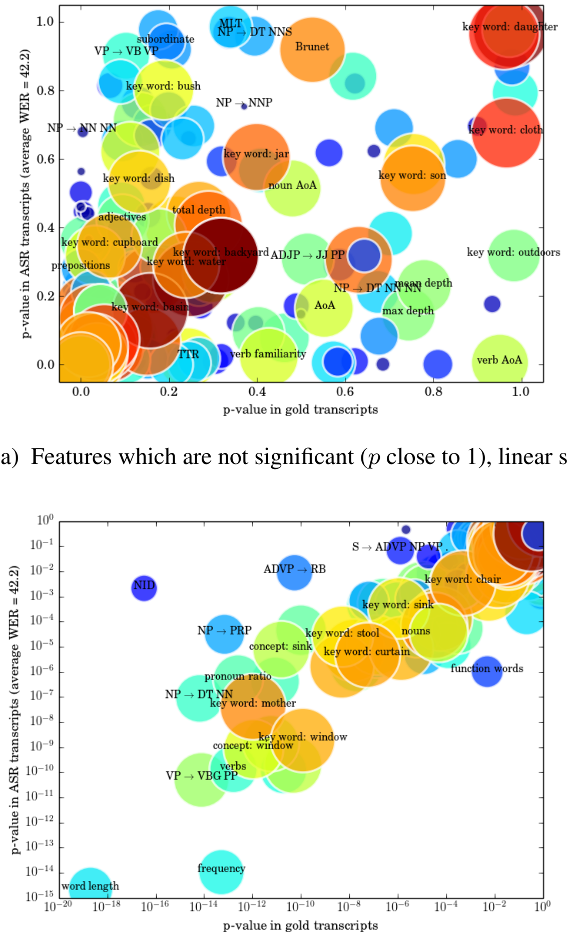
Our ASR results improve on previous work, and provide useful information for diagnosis. However, we are limited by the somewhat poor quality of the audio in DementiaBank. Some files are dominated by noise, and are difficult to decipher even by a human listener. Preliminary attempts to reduce noise using least squares amplitude estimation did not significantly improve the WER, and this remains the subject of ongoing work.

This paper is the first to detail the impact of ASR WER on the usefulness of text-based features in clinical diagnosis. While the broad trends in our results are expected (diagnostic accuracy decreases as WER increases), we observe a wide variance in accuracy associated with narrow bands of WER. Considering only the transcripts derived from our most successful ASR experiments, there is no significant effect on accuracy of any of the independent variables we examined. This underscores the fact that two ASR transcripts can have the same WER *without* containing the same diagnostically relevant information. Future work will seek to determine common properties of the transcripts which are associated with higher diagnostic accuracies, potentially including demographic information.

We have identified several features which appear to be robust to ASR error, but further work is required to better understand *why* these features remain significantly different in the ASR transcripts (and conversely, why some features are fragile to the recognition process). Further work will also explore the features which are significant only in the ASR transcripts. While these features cannot inform our knowledge of pathological speech patterns in dementia, if the differences are systematic, then they could still have utility in a diagnostic framework.

6. Acknowledgements

This research is funded by the Natural Sciences and Engineering Research Council of Canada (including grant RGPIN 435874) and Dr. Rudzicz's Young Investigator award from the Alzheimer Society of Canada.



(a) Features which are not significant (p close to 1), linear scale.

(b) Features which *are* significant ($p \ll 1$), log scale.

Figure 3: Diagnostic significance of each feature in gold vs. ASR transcripts. Smaller radii and cooler colors indicate smaller correlation between values in the gold and ASR transcripts.

7. References

[1] M. Prince, R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, and C. P. Ferri, "The global prevalence of dementia: a systematic review and metaanalysis," *Alzheimer's & Dementia*, vol. 9, no. 1, pp. 63–75, 2013.

[2] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech," in *Proceedings of the IEEE International Conference on Mechatronics and Automation*, 2005, pp. 1569–1574.

[3] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2081–2090, 2011.

[4] A. Habash and C. Guinn, "Language analysis of speakers with dementia of the Alzheimer's type," in *Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium*, 2012, pp. 8–13.

[5] J. J. Meilán, F. Martínez-Sánchez, J. Carro, J. A. Sánchez, and E. Pérez, "Acoustic markers associated with impairment in language processing in Alzheimer's disease," *The Spanish Journal of Psychology*, vol. 15, no. 02, pp. 487–494, 2012.

[6] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014, pp. 27–36.

[7] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard, "Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimers disease with and without additional vascular pathology," *Journal of Alzheimer's Disease*, vol. 42, pp. S3–S17, 2014.

[8] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features differentiate Alzheimer's from controls in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2015.

[9] S. V. Pakhomov, G. E. Smith, S. Marino, A. Birnbaum, N. Graff-Radford, R. Caselli, B. Boeve, and D. D. Knopman, "A computerized technique to asses language use patterns in patients with frontotemporal dementia," *Journal of Neurolinguistics*, vol. 23, pp. 127–144, 2010.

[10] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *Proceedings of INTERSPEECH*, 2008, pp. 2550–2553.

[11] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology*, vol. 22, no. 2, pp. 99–112, 2010.

[12] D. Hakkani-Tur, D. Vergyri, and G. Tur, "Speech-based automated cognitive status assessment," in *Proceedings of INTERSPEECH*, 2010, pp. 258–261.

[13] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. L. G. Tempini, and J. Ogar, "Learning diagnostic models using speech and language measures," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 4648–4651.

[14] K. C. Fraser, F. Rudzicz, N. Graham, and E. Rochon, "Automatic speech recognition in the diagnosis of primary progressive aphasia," in *Proceedings of SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, Grenoble, France*, 2013, pp. 47–54.

[15] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland, "Aphasiabank: Methods for studying discourse," *Aphasiology*, vol. 25, no. 11, pp. 1286–1307, 2011.

[16] J. Becker, F. Boiler, O. Lopez, J. Saxton, and K. 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.

[17] H. Goodglass and E. Kaplan, "The Boston Diagnostic Aphasia Examination," 1983.

[18] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk.*, 3rd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000.

[19] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 357–362. [Online]. Available: http://dx.doi.org/10.3115/1075527.1075614

[20] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[21] Y. Miao, "Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN," arXiv:1401.6984, 2014.

[22] F. Rudzicz, R. Wang, M. Begum, and A. Mihailidis, "Speech interaction with personal assistive robots supporting aging at home for individuals with alzheimer&rsquo;s disease," *ACM Trans. Access. Comput.*, vol. 7, no. 2, pp. 6:1–6:22, May 2015. [Online]. Available: http://doi.acm.org/10.1145/2744206

[23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[24] M. Brysbaert and B. New, "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," *Behavior Research Methods*, vol. 41, no. 4, pp. 977–990, 2009.