Learning diagnostic models using speech and language measures

Bart Peintner*
William Jarrold*
SRI International
Menlo Park, CA, USA
{peintner,jarrold}@ai.sri.com

Dimitra Vergyri†
Colleen Richey‡
SRI International
Menlo Park, CA, USA
{dverg,colleen}@speech.sri.com

Maria Luisa Gorno Tempini§
Jennifer Ogar˜
University of California San Francisco
San Francisco, CA, USA
{marilu,jogar}@memory.ucsf.edu

***Abstract*— We describe results that show the effectiveness of machine learning in the automatic diagnosis of certain neurodegenerative diseases, several of which alter speech and language production. We analyzed audio from 9 control subjects and 30 patients diagnosed with one of three subtypes of Frontotemporal Lobar Degeneration. From this data, we extracted features of the audio signal and the words the patient used, which were obtained using our automated transcription technologies. We then automatically learned models that predict the diagnosis of the patient using these features. Our results show that learned models over these features predict diagnosis with accuracy significantly better than random. Future studies using higher quality recordings will likely improve these results.**

I. INTRODUCTION

Disease-modifying treatments for dementia will soon be available. These drugs will only be effective at the early stages of the disease, making correct differential diagnosis imperative. Tools currently available to clinicians are based on one-time measurements taken in the clinic, including MRI data, neuropsychological tests, and interviews with the patient and relatives. Access to data obtained in the patients' home can potentially provide clinicians with more timely and representative data.

One important stream of data that can be unobtrusively and inexpensively collected is speech audio. These qualities, along with the growing evidence that common neurodegenerative diseases alter speech and language production, motivate our research into which automatically produced measures of speech and language give evidence for specific neurodegenerative conditions.

This study analyzed raw audio from 9 healthy control subjects and 30 patients diagnosed with one of three clinical variants of Frontotemporal Lobar Degeneration. From this data, we extracted features of the audio signal and the words the patient used, which were obtained using our automatic speech recognition (ASR) algorithms, our phoneme duration measurement tools, and existing linguistic content analysis (LCA) tools. Using off-the-shelf machine learning algorithms, we learned models that predict the diagnosis of the patient using subsets of the extracted features.

* Computer Scientist, Arificial Intelligence Center
† Research Engineer, Speech Technology and Research Laboratory
‡ Research Linguist, Speech Technology and Research Laboratory
§ Associate Professor, Memory and Aging Center
˜ Speech and Language Pathologist, Memory and Aging Center

Our results show that many of our extracted features vary significantly as a function of diagnosis, and that learned models over these features predict diagnosis with accuracy considerably better than random. Furthermore, we believe that with better data (i.e., longer audio samples with better sound quality) further improvements will occur. The results suggest that deployable, in-home speech analysis systems may provide significant value to clinicians and their patients in the near future.

II. GROUPS STUDIED AND DATA COLLECTED

In this work, we focused on three clinical presentations of frontotemporal lobar degeneration (FTLD). The prevalence of FTLD is similar to Alzheimer's Disease (AD) in patients under the age of 65 years [1], yet FTLD remains understudied relative to AD. Misdiagnosis of FTLD is common, even at research centers [2]. There are three clinical subtypes in the Neary criteria for FTLD [3], one with altered social conduct, known as the behavioral variant of frontotemporal dementia (bvFTD), and two with deficits in language: semantic dementia (SD) and progressive nonfluent aphasia (PNFA). With respect to speech and language, bvFTD is characterized by a progressive decline in social and emotional processing [4]. SD is characterized by naming difficulties and semantic memory difficulties but preserved fluency and grammar [4]. PNFA typically presents with motor speech and grammar difficulties [3], [4].

Tab. I shows demographic information about the study participants. For each patient in these groups, we analyzed 3-5 minutes of speech recorded via a wireless lapel microphone and digital video camcorder while the patient performed Part I of the Western Aphasia Battery [5]. This task elicits spontaneous speech as part of a standardized speech and language assessment protocol. This data was obtained from the UCSF Memory and Aging Center database.

TABLE I
DEMOGRAPHIC INFORMATION FOR PARTICIPANTS

	bvFTD	PNFA	SD	Controls
Male/Female	5/4	1/7	6/7	3/6
Age	63.0(8.3)	62.9(7.8)	65.2(6.6)	61.7(6.0)
Education	17.3(1.7)	16.1(2.3)	16.5(2.5)	17.3(2.1)

TABLE II

PHONEME DURATION FEATURES

Mean and Standard deviation of...	mean voiced fricative, vowel, nasal, sonorant, phone, approx, voiceless cons, voiceless fricative, sonorant+voiceless fricative, obstruent, consonant, sonorant consonant, fricative, voiced obstruent, voiced consonant, voiceless stop, voiced stop, stop
Other features	Phonemes per sec of speech, All pauses per spurt Filled pauses/speech, Filled pauses/spurt, Phonemes per sec of spurt

TABLE III

POS FEATURES AND A SUBSET OF THE LIWC FEATURES

POS Features	Frequencies of: Interjections, Verbs, Adverbs Adjectives, Pronouns, DetermPerNoun, VerbPerNoun PronounPerNoun, Function Words, Nouns, All other
LIWC Features (subset)	Six-letter words, Function Words, Personal Pronoun I, We, You, SheHe, They, Articles, Past tense, Present tense, Future tense, Swear, Social, Family, Friend, Affect, Positive emotion, Negative emotion, Anger, Sad, Insight, Sexual, Motion, Death, Money

III. FEATURE EXTRACTION TECHNIQUES

Three feature sets were extracted from each audio sample. One contained features based on phoneme duration measurements. The other two were based on word count frequencies as described below. The three corresponding extraction methods – all of which rely, in part, on ASR – are described below. The three feature sets were used to learn models that predict diagnosis. The "best" featues in these three compose a fourth set, called the *composite* set.

A. Automatic speech recognition

Given an audio speech signal, ASR produces a written sequence of words consisting of the transcription of that speech signal. We refer to that transcription as the Automatic Transcription (AT) and to its manual counterpart as the Human Transcription (HT). The accuracy of AT is typically measured by word error rate (WER), which computes the percentage of word errors in AT compared to HT.

The average WER of AT for these categories was 20% for controls, 61% for PNFA, 37% for SD, and 30% for bvFTD. Worst WER (100%) occurred with a PNFA patient exhibiting very impaired speech. We expected the high WER for this data because the models used were trained on data from healthy adults ages 18-65. In order to achieve high performance in ASR, we need training data similar to our test data. Previous studies [6], [7], [8], [9] have shown that ASR on elderly speech deteriorates significantly relative to that of younger adults. This complication is compounded by the patient's condition, which frequently affects that patient's speech quality. Our system is described in detail in [10].

B. Phoneme duration extraction

By default, our ASR system produces start and end times for each word. Moreover, we get time alignments of the phonemes in each word's pronunciation, which can be used to estimate the mean and variance of duration of certain phoneme categories, like consonants, pauses, and others, which we used as features in our machine learning algorithms. In total, each user's phoneme duration profile consisted of 41 features, listed in Tab. II.

C. Linguistic analysis

Computer-based linguistic content analysis (LCA) provides a quantitative characterization of word usage patterns. These patterns reflect thoughts, feelings, and neurobehavioral outcomes often in ways not anticipated by common sense [11], [12]. For example, the Nun Study [13] showed that LCA of writings early in life can predict cognitive test scores late in life. Gottschalk et al. [14] use LCA to predict cognitive impairment and other neurobehavioral variables.

The remaining two feature sets were obtained by analyzing each transcript via two linguistic content analysis software applications. The first determined frequencies of nouns, verbs, pronouns, and so on via an automatic part-of-speech (POS) tagger [15]. The second set was obtained via Linguistic Inquiry and Word Count (LIWC) [16]. LIWC determines word frequencies organized along more than 75 different categories such as psychological processes (e.g., emotional or cognitive), linguistic dimensions (e.g., articles, negations), or relativity (in time and space). Tab. III shows the 11 features in the POS feature set and a subset of the 81 features in the LIWC set.

IV. ANALYSIS

Our analysis had two steps. First, we performed feature selection by filtering the set of features to those that varied significantly as a function of diagnosis. This was done using a one-way analysis of variance (ANOVA) on each feature w.r.t. diagnosis, i.e., bvFTD, SD, PNFA, or healthy control.

Then, using the selected features, we "learned" computer models that diagnose a given patient when provided with that patient's profile of either (a) POS features, (b) LIWC-based features, (c) phoneme duration features, or (d) a composite set using the best features from all categories. We evaluated the learned models using cross-validation.

We tried three different learning algorithms from the Weka machine learning toolkit [17]: SimpleLogistic (logistic regression), J48 (decision tree), and MultiLayeredPerceptron (neural network). Default parameters were used for all learners. Using each algorithm, we learned separate model instances for each of five prediction problems:

- Four-way prediction: Healthy Controls vs. PNFA vs. FTD vs. SD
- Binary prediction: Healthy Controls vs. Impaired (PNFA, FTD, or SD)
- Binary prediction: Healthy Controls vs. PNFA
- Binary prediction: Healthy Controls vs. FTD
- Binary prediction: Health Controls vs. SD

The features in the composite model were chosen by anaylzing the models produced for the four-way prediction using the other three feature sets. The most prominent features (e.g., those with higher weights) were included.

A. ANOVA results

Of the 41 phoneme features (i.e., mean and standard deviation of different classes of phonemes), 25 features were

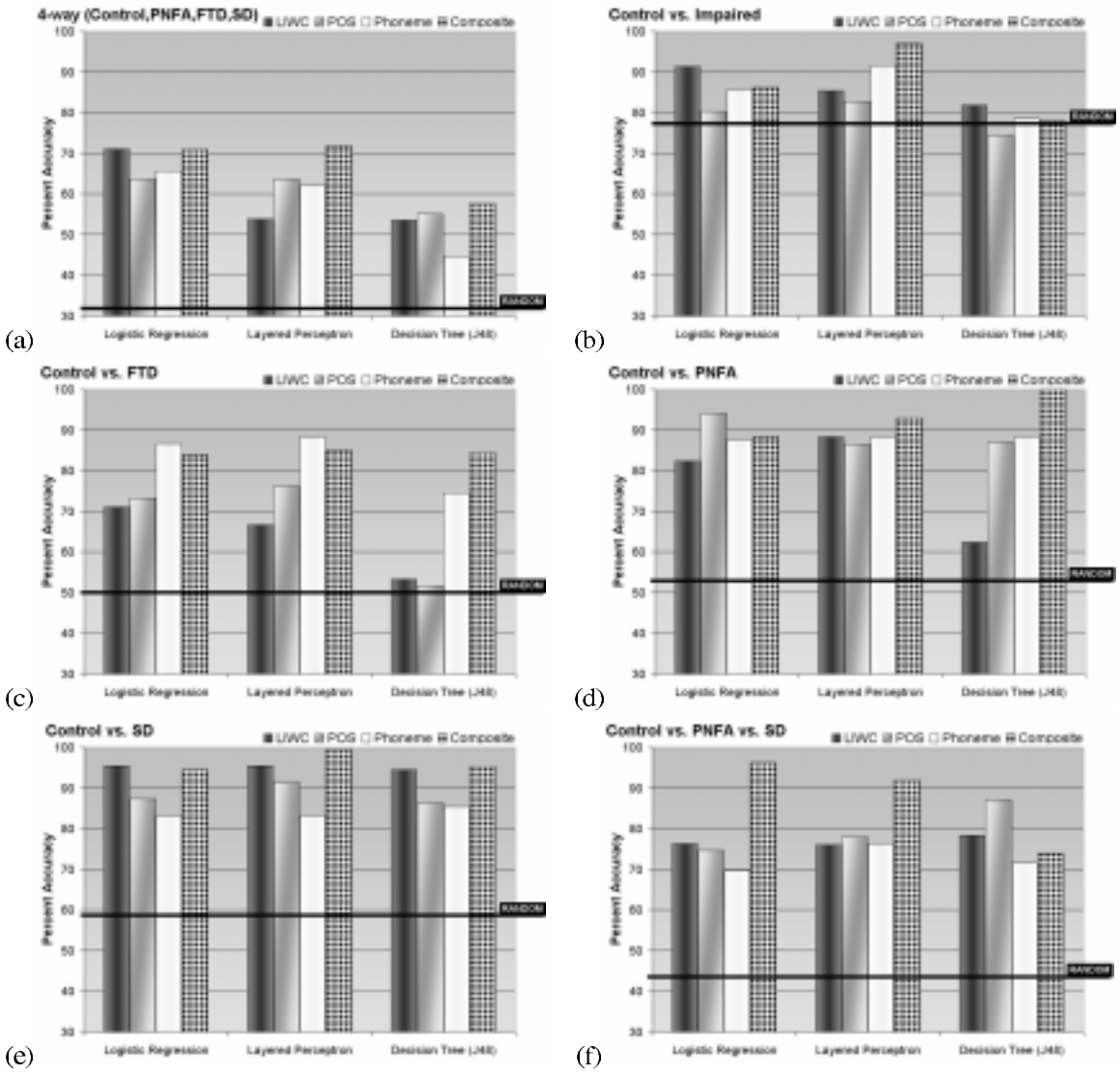

Fig. 1. Predictive accuracy of learned models using 10 fold cross-validation shows excellent performance. "Composite" refers to models learned on all features (i.e., phoneme, POS, and LIWC).

significant at the 0.05 level (Benjamini-Hochberg adjusted). Regarding LCA features, there are several more POS and LIWC-based differences than predicted by chance. For example, Pronouns, Adverbs, Interjections, and Connectives all showed significant ($p \leq .005$) differences as a function of diagnosis; among POS measures, only Adjectives and Verbs were not statistically significant. For LIWC features, 22 of 81 features were statistically significant at the 5% level, with $p \leq 0.005$ for 17 of them.

B. Machine learning results

We learned and evaluated models for each combination of learning algorithm, feature set, and prediction problem. Each model was evaluated via a 10-fold cross-validation, and we report the predictive accuracy of each model. Fig. 1 shows the result of this experiment. Each graph in Fig. 1 summarizes the results for a prediction problem. For each learning algorithm, four bars represent the accuracy for the LIWC, POS, Phoneme duration, and composite models, respectively. The theoretical accuracy of a random predictor is represented by a thick horizontal line labeled "RANDOM". Typically, a random predictor randomly chooses from N possible classifications, resulting in an accuracy of $1/N$. However, we report a higher value, achievable if the predicter that takes group sizes into account. Thus, we use the ratio of the size of the largest group to the total # of samples as the value. For Controls vs PNFA, 9 Controls vs 13 PNFA implies random accuracy of 13/(9+13).

Fig. 1(a) shows the 4-way prediction results. All models scored significantly better than random, with two composite models achieving over 70% accuracy. We find this result particularly exciting given that three of the four diagnostic groups are closely related (all are subtypes of FTLD).

Fig. 1(b) shows the ability of our methods to distinguish between healthy and impaired individuals. This ability is key to the home-monitoring applications we envision, whose

purpose will be to monitor healthy adults for changes that suggest the onset of any dementia. Although the accuracy is high for this prediction problem, only a few models are significantly better than random. The random bar is high in this case because of the relatively high number of impaired individuals (i.e. 77%).

Fig. 1(c-e) shows the ability of our methods to distinguish between healthy controls and each specific FTLD subtype. Note that we distinguish PNFA from healthy and SD from healthy with extremely high accuracy—well above 80% with most models and above 90% with many. However, ability to distinguish FTD from Healthy Controls is less impressive. Looking deeper into the data, we find that PNFA has signficantly longer phoneme durations and has significantly fewer function words. These differences, which are predicted by prior literature [18], account for much of our success.

Because we had less success with distinguishing Healthy Controls from FTD, we evaluated our methods on a final 3-way prediction problem between Healthy Controls, PNFA, and SD. The results, shown in Fig. 1(f) were promising, with the composite model achieving greater than 90% accuracy using two learning algorithms.

We performed the same experiments (all six prediction problems) using features extracted from the HTs. In general, the accuracies for models based on the HTs were slightly lower than the corresponding AT models. There are several possible reasons for this, but we will need more data and analysis before reaching concrete conclusions.

While these results are promising, we realize that our small sample size is a factor. We suspect that the lack of sufficient data explains some of the inconsistent facts in Fig. 1. For example, the composite/logistic regression model in Fig. 1(f) was more accurate than the corresponding model in Fig. 1(d), despite that fact that the prediction problem for the former is intuitively more difficult.

V. CONCLUSION

We have shown that simple measures of speech and language can be used to predict the diagnosis of a patient with significant accuracy. If we can continue to improve results, e.g., by adapting ASR to produce lower WER, or by testing our system on larger samples, we will be closer to a deployable, in-home speech analysis system, which can provide significant value to clinicians and their patients in the near future. The experiments in this paper were meant to provide a baseline of performance for other, more sophisticated measures, including measures of agrammaticism, n-grams, and prosodic extraction. We intend to explore these measures in future work and focus on collecting data from more patients and from patients with Alzheimer's disease and Parkinson's disease.

Despite promising results, our approach does suffer from some limitations. One limitation is the limited accuracy of ASR. This is partially addressable with more data because we can adapt the accoustic and linguistic models used by our ASR system to better recognize the speech of elderly and impaired speakers. Further, our linguistic content analysis features were taken (unadapted) from other applications. By further incoporating specific knowledge of the diseases in question, we believe that new linguistic measures will improve predictive accuracy.

VI. ACKNOWLEDGMENTS

The authors gratefully acknowledge Hal Javitz for help with statistical analysis, Jennifer Ogar for help with the clinical data, and Rafael B. de Leon and Naomi House for transcription and audio/video support.

REFERENCES

[1] Ratnavalli, E., Brayne, C., Dawson, K., Hodges, J.R., The prevalence of frontotemporal dementia. *Neurology*, 2002. 58(11): p. 1615-21.

[2] Mendez, M.F., Selwood, A., Mastri, A.R., Frey, W.H. 2nd, Pick's disease versus Alzheimer's disease: A comparison of clinical characteristics, *Neurology*, 1993. 43(2): p. 289-92.

[3] Neary, D., Snowden, J.S., Gustafson, L., Passant, U., Stuss, D., Black, S., Freedman, M., Kertesz, A., Robert, H., Albert, M., Boone, K., Miller, B.L., Cummings, J., Benson, D.F., Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria, *Neurology*, 1998. 51(6): p. 1546-54.

[4] Rosen, H.J., Gorno-Tempini, M.L., Goldman, W.P., Perry, R.J., Schuff, N., Weiner, M., Feiwell, R., Kramer, J.H., Miller, B.L., Patterns of brain atrophy in frontotemporal dementia and semantic dementia, *Neurology*, 2002. 58(2): p.198-208.

[5] Kertesz, A., *Western Aphasia Battery*. 1980, London, Ontario: University of Western Ontario Press.

[6] Wilpon, J.G., Jacobsen, C.N., A study of speech recognition for children and the elderly, *IEEE Trans. Acoust., Speech, Signal Processing*, 1996. p.349-52

[7] Yato, F., Inoue, N., Hashimoto, K., A study of speech recognition for the elderly, *Proceedings of European Conference on Speech Communication and Technology*, 1999. 1: p.101-4.

[8] Anderson, S., Liberman, N., Bernstein, E., Forster, S., Cate, E., Levin, B., Recognition of elderly speech and voice-driven document retrieval, in *IEEE Trans. Acoust., Speech, Signal Processing*, 1999. 1: p.145-8.

[9] Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shiakno, K., Elderly acoustic model for large vocabulary continuous speech recognition, in *Proceedings of European Conference on Speech Communication and Technology*, 2001. p.1657-60.

[10] Stolcke, A., Boakye, K., Cetin, O., Janin, A., Magimai-Doss, M., Wooters, C., Zheng, J. The SRI-ICSI Spring 2007 meeting and lecture recognition system, in *Proc. NIST 2007 Rich Transcription Workshop*, 2007.

[11] Chung, C.K., Pennebaker, J.W., The psychological function of function words, in K. Fiedler (Ed.), *Social communication: Frontiers of social psychology*, 2007. p. 343-59.

[12] Pennebaker, J.W., Mehl, M.R., Niederhoffer, K., Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 2003. 54: p. 547-77.

[13] Snowdon, D.A., Kemper, S.J., Mortimer, J.A., Greiner, L.H., Wekstein, D.R., Markesbery, W.R., Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Assoc.*, 1996. 275: p. 528-32.

[14] Gottschalk, L.A., Bechtel, R.J., Maguire, G.A., Katz, M.L., Levinson, D.M., Harrington, D.E., Nakamura, K., Franklin, D.L.. Computer detection of cognitive impairment and associated neuropsychiatric dimensions from the content analysis of verbal samples. *American Journal of Drug and Alcohol Abuse*, 2002. 28(4): p. 653-70.

[15] Toutanova, K., Klein, D., Manning, C., Singer, Y., Feature-rich part-of-speech tagging with a cyclic dependency network, in *Proceedings of HLT-NAACL*, 2003, p. 252-9.

[16] Pennebaker, J.W., Francis, M.E., Booth, R.J., *Linguistic Inquiry and Word Count: LIWC2001*, 2001. Mahwah, NJ: Erlbaum Publishers.

[17] Witten, I.H., Frank, E. *Data mining: Practical machine learning tools and techniques*, 2005. San Francisco: Morgan Kaufmann. 2nd edition.

[18] Saffran, E.M., Berndt, R.S., Schwartz, M.F., The quantitative analysis of agrammatic production: Procedure and data, *Brain and Language*, 1989. 37(3): p. 440-79.