Aided Diagnosis of Dementia Type through Computer-Based Analysis of Spontaneous Speech

William Jarrold
Nuance Communications[1]
william.jarrold@gmail.com

Bart Peintner
Soshoma[1]
bpeintner@gmail.com

David Wilkins
Language & Linguistic Consulting
wilkinsdavidp@gmail.com

Dimitra Vergryi and **Colleen Richey**
SRI International
dverg@speech.sri.com,
colleen.richey@sri.com

Maria Luisa Gorno-Tempini and **Jennifer Ogar**
University of California, San Francisco
{marilu|jogar}@memory.ucsf.edu

Abstract

This pilot study evaluates the ability of machined learned algorithms to assist with the differential diagnosis of dementia subtypes based on brief (< 10 min) spontaneous speech samples. We analyzed recordings of a brief spontaneous speech sample from 48 participants from 5 different groups: 4 types of dementia plus healthy controls. Recordings were analyzed using a speech recognition system optimized for speaker-independent spontaneous speech. Lexical and acoustic features were automatically extracted. The resulting feature profiles were used as input to a machine learning system that was trained to identify the diagnosis assigned to each research participant. Between groups lexical and acoustic differences features were detected in accordance with expectations from prior research literature suggesting that classifications were based on features consistent with human-observed symptomatology. Machine learning algorithms were able to identify participants' diagnostic group with accuracy comparable to existing diagnostic methods in use today. Results suggest this clinical speech analytic approach offers promise as an additional, objective and easily obtained source of diagnostic information for clinicians.

1 Introduction

Accurately differentiating certain neurodegenerative disorders such as *Alzheimer's Disease* (AD) and variants of *Fronto-temporal Lobar Degeneration* (FTLD) is extremely difficult (Varma et al., 1999). Differential diagnosis is often left to tertiary care settings (e.g. Research I Universities with medical schools). While the most definitive diagnosis is made post-mortem using brain tissue samples, the treatment and prognostic implications of living patients are often determined in large part on the basis of language assessment.

Although language is clearly not the exclusive diagnostic factor for AD, existing literature suggests it is an important one. Studies show significant differences in the written language abilities of AD patients and healthy older adults (Pestell et al., 2008 and Platel et al., 1993). The speech of patients with AD is partly characterized by word-finding difficulties, smaller vocabularies, and problems with semantic processing (Forbes at el., 2002). These symptoms appear early in the disease's progression, however language assessment of AD patients can fail to identify early symptoms that family members report to be present in their conversations (Crockford and Lesser, 1994).

FTLD has a prevalence similar to AD in patients under the age of 65 years (Mendez at el., 1993). Misdiagnosis of FTLD is common Mendez at el., 1993). Three variants are defined by the widely adopted Neary criteria (Neary at el., 1998); one with altered social conduct, the behavioral variant of *frontotemporal dementia* (bvFTD); the second characterized by a deterioration of conceptual-semantic knowledge, *semantic dementia* (SD); and the third marked by a disorder of expressive language fluency, *progressive non-fluent aphasia* (PNFA).

Clinicians diagnose using a wide array of evidence including patient history, imaging and neuropsychological assessment in which speech and language diagnostics feature prominently. In AD, cognitive disturbance is a required diagnostic feature and language impairment one several sufficient signs of such impairment. In the case

[1] Research conducted while at SRI International

Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 27–37,
Baltimore, Maryland USA, June 27, 2014. ©2014 Association for Computational Linguistics

of SD and PNFA, changes in speech and language are core diagnostic features, with changes in lexical content features being highly diagnostic of SD, and changes in the acoustic properties of speech being highly diagnostic of PNFA. Even in bvFTD, where changes in social behavior are the defining features, analysis of language-based differences is important, because language is an essential mediator of social behavior. To be sure, the clinician does not diagnose exclusively on language features -- patient history, imaging, memory functioning and more play a role. However, language does feature prominently in the differential diagnosis of AD, FTLD and its three subtypes. For this reason, computerized analysis of speech may offer an important aid to the clinical diagnosis of these syndromes.

Prior work in clinical speech analytics supports the possibility of computer-based diagnosis of dementia related syndromes. Singh (2001) describes a means of quantifying the degree of speech deficits derived from human transcriptions of the speech of patient with AD. Machine Learning has already been applied to distinguish AD from controls using human transcribed spontaneous speech (Thomas at el., 2005). Abel et al. (2009) applied a connectionist net that models patient speech errors (naming and repetition disorders) to the problem of diagnosis. Tur et al. (2010) have shown the ability to automatically score patient speech from a story recall and picture description task that is on par with human performance. Lehr et al. (2012) have developed a system that automatically transcribes and scores patient speech obtained during the story recall portion of the Wechsler Logical Memory test. The evaluation demonstrated it could distinguish mild cognition impairment from typical controls at performance level comparable to human scorers.

Our work builds upon these prior studies along a number of dimensions. First, we distinguish between a wider array of dementia subtypes, i.e. not only AD vs controls, but also the three subtypes of FTLD. Second, we use not just lexical features but also acoustic/prosodic related features. Third, in order to shed light on the opaque "black box" nature of many machine-learned classifiers, we identify relationships between model features and symptoms from the clinical literature. Fourth, our approach can claim to be more ecologically valid because it analyzes spontaneous speech as input rather than recall of a remembered passage. Fifth, we do not require human transcription - a labor-intensive step that hinders broader use in a clinical setting. Sixth we provide a comparison of our system performance against benchmarks obtained from practicing clinicians. Our paper is the first we know of to exhibit all of the above properties.

In sum we used computational techniques to analyze acoustic and lexical features of the speech of patients with AD and FTLD variants, and we investigated whether models derived from these features via machine learning could accurately identify a patient's diagnosis.

2 METHOD

2.1 Participant Recruitment and Diagnosis

We obtained spontaneous speech data from 9 controls, 9 AD patients and 30 FTLD patients—9 with frontotemporal dementia (bvFTD), 13 with semantic dementia (SD), and 8 with progressive nonfluent aphasia (PNFA). Table 1 shows demographic information.

Data were collected in an ongoing series of NIH-funded studies being performed at the UCSF Memory and Aging Center. Patients were diagnosed by expert clinicians at the center by applying current clinical criteria. Patients underwent detailed standard speech and language, cognitive, emotional, genetic, pathological, and neuroimaging evaluations. Age-matched healthy controls were community volunteers obtained by SRI In-

	bvFTD	PNFA	SD	AD	Controls
Male/Female	5/4	1/7	6/7	5/4	3/6
Age	63.00(8.25)	62.88(7.75)	65.23(6.61)	59.11(7.47)	61.7(6.0)
Education *	17.33(1.73)	16.13(2.30)	16.45(2.54)	15.44(2.30)	17.27(2.1)
MMSE	24.4(5.85)	22.0(9.34)	17.09(8.15)	18.67(7.53)	Not Administered

Table 1. Demographic information for participants

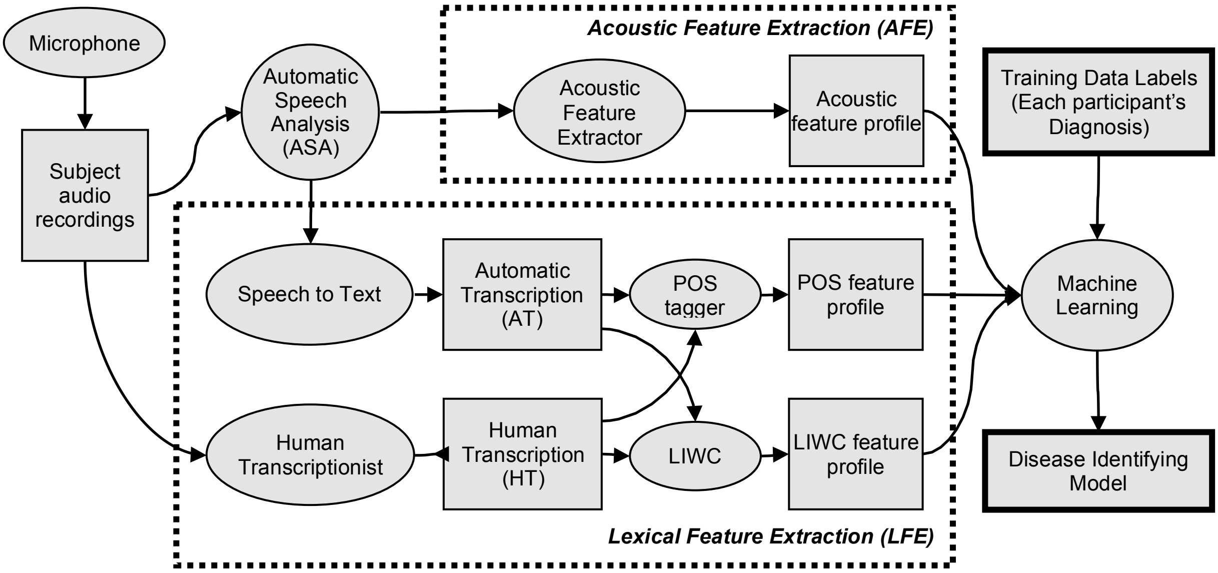
Microphone
Subject audio recordings
Automatic Speech Analysis (ASA)
Acoustic Feature Extraction (AFE)
Acoustic Feature Extractor
Acoustic feature profile
Training Data Labels (Each participant's Diagnosis)
Speech to Text
Automatic Transcription (AT)
POS tagger
POS feature profile
Machine Learning
Human Transcriptionist
Human Transcription (HT)
LIWC
LIWC feature profile
Disease Identifying Model
Lexical Feature Extraction (LFE)

Figure 1. System Information Flow and Evaluation. Participant speech is subjected to automatic speech analysis of two kinds: Acoustic Feature Extraction (AFE) and Lexical Feature Extraction (LFE). Feature selection (not shown) is explained in Sects 2.3 and 2.6. Each machine learning algorithm produces a classification model based on labeled training data. All models used both acoustic and lexical features. Each such disease identifying model is evaluated against held-out training data (not shown). To measure sensitivity to ASR error, half of these models were based on lexical features derived from automatic transcription (AT), the other half from human transcription (HT).

ternational and were paid $10 for their participation.

2.2 Speech Samples

Speech samples were recordings of Part 1 of the Western Aphasia Battery (Kertesz, 1980). Participants are administered a semi-structured interview (e.g., questions such as “How are you?”) and asked to describe a drawn picture of a picnic scene. The resulting 3 to 5 minutes of speech was recorded via wireless lapel microphones. Controls were recorded via digital audio recorder sampling at 48 kHz, 16 bit PCM, and later down-sampled to 16 kHz for use with the speech recognizer. Digital audio was down-sampled at 16 kHz, 16 bit PCM. Recordings were manually segmented in order to separate the interviewee’s voice from the interviewer’s. Only patient speech segments were subject to analysis

2.3 Procedure

To tackle speech-based diagnosis of AD, bvFTD, SD and PNFA, we employ several types of computer-based analyses (see Figure 1). Audio recordings were processed via the Meeting Understanding system (Stolcke at el., 2007), which was custom-tailored to recognize speaker-independent, multi-person speech. First, using this system we perform acoustic-level feature extraction (**AFE**), which obtains measures the duration of consonants, vowels, pauses, and other acoustic-phonetic categories. In parallel, we perform a lexical feature extraction (**LFE**) on transcripts of participant speech producing profiles of each speaker’s language use. This profile characterizes frequencies of different types of words – e.g. frequency of nouns, verbs, function words, words about emotion, etc. – present in a language sample along ~100 dimensions.

Next, The AFE and LFE profiles are combined to form one large vector of features that collectively characterize the speaker. Feature selection is applied to select the most informative features. For feature selection, we performed a one-way ANOVA on each extracted feature to determine which features were significantly related to a diagnostic category using the Benjamini-Hochberg adjustment for multiple comparisons.

The vector of selected features for the speech samples in the *training set* is taken as input to machine learning. Based on these data machine learning automatically induces a diagnostic model that should predict any speaker’s diagnosis based the AFE and LFE profiles of his or her speech sample.

The performance of a learned diagnostic model is measured in terms of ability to generalize to cases that it has not been trained on is measured by feeding *test set* cases – i.e. cases that have not been a part of the *training set*. We compared the accuracy of the machine learning induced algorithms with accuracy studies of traditional diagnostic methods in the literature.

In addition to the above, as part of a desire to achieve insight into the way these models were functioning, we sought verification that a differences in feature profiles as a function of diagnostic group correspond meaningfully to existing expectations derived from the literature. To do so, we formed and tested several predictions about specific feature differences based on the clinical literature (see Hypotheses below).

Finally, we wanted to determine how sensitive the feature differences and classification models were to speech recognition error. To do so we tested each hypothesis on both the human and automatic transcriptions. In addition, we learned a set of models based on automatic transcriptions and a second set of models based on human transcriptions and compared accuracies.

2.4 Acoustic Feature Extraction

We used the automatic speech recognition (**ASR**) system to extract a set of acoustic-level features corresponding to the overall rate, plus the mean and standard deviation of (a) pause lengths and (b) hypothesized phoneme durations. For each speech sample, the speech rate as well as the mean and standard deviation of the duration of pauses, vowels, and consonants were computed. The SRI speech processing system also further identified consonant classes based on manner features (e.g., fricative, stop, etc. …) voicing features (voiced, voiceless) and measured the mean and standard deviation of the duration of these classes. Our Automatic Speech Analysis system produced 41 different duration-based measures extracted from the speech stream.

2.5 Lexical Feature Extraction (LFE)

For each transcript we performed two types of computer-based lexical analysis. The first determined frequencies of 14 different parts of speech (e.g. nouns, verbs, pronouns etc.) using an automatic part-of-speech (**POS**) tagger. The second involved Dr. Pennebaker's *Linguistic Inquiry and Word Count* (**LIWC)** software (Pennebaker, et al 2001), which determines word frequencies organized into 81 categories, such as psychological processes (e.g., emotional or cognitive) and linguistic dimensions (e.g. function words, verb tenses, negations).

To measure sensitivity to speech to text error, each ANOVA was performed twice, once for the "ground truth" human transcriptions (HT) and once for the automatic transcriptions (AT). During hypothesis testing, statistical significance of each pair of AT versus HT based LFEs (i.e., "ground truth") was compared. Additionally different models were learned, half using HT the other half using AT. To test for lexical-level differences between diagnostic categories, we performed a one-way ANOVA for each of the 95 LFE features (e.g. frequency of nouns) in which diagnosis was the independent variable and the given feature's frequency was the dependent variable.

2.6 Machine Learning

We assessed how well a variety of machine learning algorithms predicted a patient's diagnosis, using his or her combined AFE and LFE profile. Evaluation was conducted using five-fold cross-validation over the set of patients, with each "fold" consisting of two phases: a *training phase,* where the feature profiles and diagnoses from 4/5ths of the subjects are used to select features and then train the given learning algorithm, and a *test phase* where the trained learner is given just the feature profiles of the remaining patients, and attempts to predict their diagnoses. This procedure is executed five times, each time using different sets of subjects for the train and test phases, with overall accuracy being the average performance on the test subjects, across all five folds. We applied three learning methods, (1) logistic regression, a statistical learning technique for determining categorical outcomes, (2) Multi-Layered Perceptrons, an artificial intelligence (AI) learning method that roughly mimics biological neural networks, and (3) decision trees, another AI technique which induces sets of rules used to predict outcomes. All three are commonly used machine learning techniques, and for this study we used implementations available in Weka, an off-the-shelf machine learning toolkit (Witten and Frank, 2005).

2.7 Hypotheses

Machine learned classification models can be difficult to understand and often used merely as black boxes. To address this issue, we tried to draw a meaningful link between certain features and diagnosis. In particular, we formed and tested several hypotheses based on expectations derived from clinical literature. We used all the data (rather than one of the training folds) to test these hypotheses.

The hypotheses about the lexical features are as follows. First, based on (Forbes at el., 2002) we predicted that AD patients use more pronouns, verbs, and adjectives and fewer nouns than controls (**H1**).

In SD, one sees decreased lexical access to concrete concepts, so patients tend to use fewer nouns (**H2**). To compensate for such difficulties with word retrieval, they also use more pronouns (**H3**). This gives the impression of empty or circumlocutory speech. For example, rather than saying “The boy is flying a kite,” a SD patient would be more prone to say “*He* is flying *that*.” (Grossman and Ash, 2004).

In PNFA, one sees fewer verbs (**H4**) (Grossman and Ash, 2004). In addition, PNFA patients often exhibit *agrammatism*. Such speech is simplified and ungrammatical and involves fewer function words, for example “give cupcake” or “water now”. Thus (**H5**) is that the speech of patients with PNFA will have fewer function words (**H5**) (Saffran at el., 1989). These hypotheses, along with whether each was supported by our analyses, are listed in Table 2 in Results.

The first acoustic hypothesis about acoustic features (**H6**) is related to the Neary criteria (Neary et al., 1998), which notes that PNFA is characterized by non-fluent spontaneous speech (among other required features). Additionally, patients in this group have significant *apraxia of speech* (Gorno-Tempini at el., 2004). Signs of this condition difficulty include articulatory groping – i.e. where the mouth searches for the correct configurations. Such trial and error speech often sounds “robotic” and can involve sounds that may be held out longer. Thus, given the duration features that are generally associated with apraxia of speech (Samuel at el., 1996; Edythe at el., 1996; Ballard and Robin, 2002), we hypothesize that PNFA patients would exhibit significantly longer vowel and consonant durations than controls (**H6**).

The second acoustic feature hypothesis (**H7**) is based on the fact that in the Neary criteria (Neary at el., 1998) *pressured speech* is a supportive (but not a core) diagnostic feature of both SD and bvFTD. In pressured speech one sees rapid

Hypothesis and source	**Supported in LFE of HT?**	**Supported in LFE or AFE of AT?**	**Figures (see Supplementary Materials)**
H1. AD patients use more pronouns, verbs, and adjectives and fewer nouns than controls (Forbes at el., 2002)	Yes, but only significant for nouns	Yes, significant for nouns, pronouns, and adjectives	Figure 3
H2. SD patients use fewer nouns (Grossman and Ash, 2004)	Yes	Yes, but not significant vs PNFA	Figure 3
H3. SD patients use more pronouns (Grossman and Ash, 2004)	Yes	Partial: SD sig. > CNTRL only	Figure 3
H4. Lower verb frequency in PNFA (Grossman and Ash, 2004)	Yes, but only significant vs. SD	No	Figure 3
H5. Fewer function words in PNFA (Saffran at el., 1989)	Yes	Yes, but only significant vs SD	Figure 3
H6. PNFA patients would exhibit longer vowel and consonant durations	N/A	Yes	Figure 2
H7. SD and bvFTD patients have shorter pauses than controls.	N/A	Yes	Figure 2

Table 2. Hypotheses extracted from literature and whether our measures—based on human transcripts (HT) and automatic transcripts (AT)—support them [Hypotheses 1-5 relate to Lexical Feature Extraction; Hypotheses 6-7 relate to Acoustic Level Analyses]

"flight of ideas" speech. We would thus expect *some* patients in these conditions to exhibit press of speech, and so hypothesize that the mean duration of pauses should be significantly less than controls (**H7**).

3 RESULTS

Results suggest that analyses at the lexical and acoustic levels are capable of detecting differences in accordance with expectations of prior research. Additionally, machine-learning algorithms predict clinical diagnosis surprisingly well.

3.1 Results: Acoustic-Level Hypotheses

For each measure, we performed an ANOVA with respect to diagnosis and found that 25 out of 41 measures were significant at the (Benjamini-Hochberg multiple comparison adjusted) 0.05 level. Hypotheses 7 and 8 in Table 2 and Figure 2 in Supplementary Materials deal specifically with AFE measures. These show that PNFA patients do exhibit significantly longer vowel and consonant durations, as the literature linking PNFA with apraxia of speech would predict. Furthermore, SD and bvFTD patients have significantly shorter pauses than controls, which is consistent with the hypothesis that some patients with these diagnoses exhibit press of speech.

3.2 Results: Lexical-Level Hypotheses

There were several lexical-level differences between diagnostic groups. We checked for significant differences (hereafter, "significant features") with respect to diagnosis while using the Benjamini-Hochberg test for multiple comparisons (Benjamini and Hochberg, 1995). (We use this adjustment for all multiple comparisons). There were several more lexical level differences based on the HTs than one would predict by chance. For example, 11 of the 14 POS features were significant ($p \leq .05$) including verbs, nouns, adjectives and adverbs. For LIWC features, 22 of 81 features were statistically significant at the p

		(A)	**(B)**	**(C)**	**(D)**
		FTLD vs AD vs Controls	**AD vs SD vs PNFA vs bvFTD vs Control**	**FTLD vs AD**	**AD vs Controls**
1.	Random diagnosis	33%	20%	50%	50%
2.	Naïve learner (always picks largest class in training set)	63%	27%	77%	50%
3.	**Our method**	**80%**	**61%**	**88%** **Sens/Spec AD** **.58/0.77** **Sens/Spec FTLD** **.95/.89**	**88%** **$\kappa = .64$ /Spec AD .83/.90** **Sens/Spec Controls** **.92/.86**
4.	Radiologists in Klöppel at el. (2008) using MRI data			69% Sens/Spec AD .64/.71	89% Sensi/Spec AD .88/.90
5.	Frontal Behavioral Inventory in Blair at el. (2007)			75%	
6.	Neuropsychiatric inventory in Blair at el. (2007).			54%	
7.	NINCDS-ARDA criteria in Lopez at el. (1990)				$\kappa = .36 - .65$
8.	DSM-III criteria in Kukull at el. (1990)				$\kappa = .55$
9.	NINCDS criteria in Kukull at el. (1990)				$\kappa = .64$
10.	ECRDC criteria in Kukull at el. (1990)				$\kappa = .37$

Table 3. Accuracy, sensitivity and specificity for Layered Perceptron learned models for FTLD subtypes. (Accuracy of a random and naïve learner id 33% and 43% respectively)

<= 0.05 level, with p ≤ 0.005 for 17 of them. As to the question of whether the profile differences correspond meaningfully to existing literature, Table 2 shows which literature-generated hypotheses were supported. See Figure 3 in Supplementary Materials which show the means and standard error for each diagnostic class on a particular feature.

3.3 Machine Learning Results

Using cross-validation, we tested the ability of machine learning methods to produce algorithms that could synthesize lexical-level and acoustic-level profiles and then identify the clinician diagnosis.

We tried several different machine-learning algorithms and found that performance was roughly the same. See Table 3 for the performance of the Multi-layered Perceptron algorithm, which was slightly superior. Performance was measured across several different diagnostic problems (e.g., FTLD vs AD vs Controls (Column A), AD vs Controls (Column D), etc.). For purposes of rough comparison, Table 3 also provides diagnostic performance of other methods, including radiologists using MRI data.

In evaluating machine learning results, we wished to compare model performance against various benchmarks. The two easiest such benchmark are random guessing (see Table 3 Row 1: given N diagnostic alternatives, one has a 1 / N chance of correctly guessing) and *naïve learner guessing*, (see Table 3 Row 2) which always chooses the most frequent (i.e., modal) diagnosis found in the training sample. The row labeled "Our method" corresponds to the accuracy of models generated from lexical and acoustic features using AT. For this case, HT results differs from AT in accuracy by only 2-3% for all prediction problems. Note that our method is at least equal to the accuracies, sensitivities, specificities, and *kappa*'s of the other clinical benchmarks in most cases. See Table 4, which shows the performance on distinguishing FTLD subtypes. For more detail on machine learning results see Peintner et al (2008).

4 DISCUSSION

The accuracy of the best machine learned diagnostic model was 88% in the binary classifications of AD versus FTLD, and AD versus Controls (Table 3). Acoustic and lexical level differences are detectable despite the present level of ASA inaccuracy. Although diagnosis should never be made on the basis of one source of information, our pilot data show that automatic computer-based analyses of spontaneous speech show promise as diagnostic aids by detecting the at times subtle differences that characterize these neurodegenerative disorders.

Inferences drawn from these results are subject to a variety of assumptions and limitations. Perhaps the biggest limitation is the small number of research participants. Larger samples will be needed in order to make valid generalizations to the population. Small samples increase the probability of Type I and II Errors and decrease power in testing for normality. That said, many of our hypothesized linguistic differences based on prior research were confirmed. Additionally, low N in each group entailed that test sets in each fold were small. Though it is remarkable in our pilot study that we obtained classification accuracy on par with clinical judgment, a larger sample size is required to make a rigorously valid claim about on par accuracy.

Statistically minded readers may question our use of parametric statistics (ANOVA) in feature selection because we have not tested the normality assumption. There are too few observations in each group to test for normality of residuals with any power. In future work with a larger sample we should perform such a test. Alternatively, on the present data we could use the non-parametric Kruskal-Wallis test as a stand in for ANOVA.

Additionally, such readers may question our use of the Benjamini-Hochberg (BH) adjustment which controls false discovery rate over a more stringent correction for familywise error rate such as Bonferoni or Holm. Our rationale was that an occasional false positive (5% if we have a 5% false positive rate) among our total set of positives isn't a big concern. As our focal aim was machine learning, scientific discovery, was a secondary concern. Thus, we were less interested in the question "was there *any* difference between the groups". We were more interested in *which* features showed a difference. Better to have a small proportion of false positives than to miss true positives. In addition, because the false negative rate criterion is less stringent about false positives, the BH procedure tends to have greater power than multiple comparison approaches that control the familywise error rate.

The success of our methods is surprising given (1) we have performed no customization of "off the shelf" LFE and machine learning techniques; (2) models were trained on a relatively small number of subjects; (3) speech samples were short (3-5 minutes). Larger speech samples, larger N and more tailored tools (e.g. language models) will enable lower word error rate, higher accuracy and finer discrimination amongst and within diagnostic types. It also suggests that this can be accomplished without training the system to the voice of each subject.

The results also draw significance because the overall approach may be applied to other neurological or psychological disorders. Many such disorders have characteristic lexical or acoustic profiles. For example, Jarrold (2011) and Stirman et al (2001) have shown that depression is associated with high frequencies of first person words (I, me, I've) and lower frequencies of social and second person words (us,we). Sanchez et al (2011) and Keskinpala (2007) have shown acoustic prosodic features indicative of depression or suicide risk. Our results suggest a very similar study design can be applied to detect these kinds of depression related lexical and acoustic/prosodic profiles.

Our results suggest we may be able train the models to assess specific highly diagnostic language symptoms – such as fluency, circumlocution, and apraxia of speech. This can be particularly important where the inter-rater reliability of given symptoms is poor. We believe that poor inter-rater reliability is mainly caused by the inability to precisely delineate the objective characteristics of these symptoms. Assuming we can get a range of values that characterize a given symptom, we can apply machine learning to identify symptoms in addition to diagnosis.

We view the methods described as analogous to EKG. The EKG trace affords a more quantitative and objective picture of cardiac functioning which complements the stethoscope. Analogously, if scaled-up studies can demonstrate adequate diagnostic accuracy results, then computationally extracted lexico-acoustic profiles may someday augment information provided by current speech and language diagnostic methods which are currently based substantially on subjective clinical judgment. As modern EKG's provide automatic interpretation, our analysis suggests that classification of speech as AD-like or FTLD-like may be possible. The competent physician never relies only the automated diagnosis provided by EKG but also interprets a profile of measures in the context of clinical observation. Our assumption is that the methods outlined above should be used in a way analogously to the EKG.

Accuracy	bvFTD (Sens/Specif)	PNFA	SD
63%	.51 / .58	.54 / .72	.76 / .62

Table 4. Accuracy, sensitivity and specificity for Lay-ered Perceptron learned models for FTLD subtypes. (Accuracy of a random and naïve learner id 33% and 43% respectively)

The results of our hypothesis testing show that differences in feature profiles are generally consistent with what we would expect from the clinical literature. This may be the first of several steps required to provide assurance to clinicians who would prefer to trust a model that had somewhat transparent features to the opaque "black box" models that are often learned. Establishing trust of clinicians is required for wide scale adoption and future work should build on these results.

Our pilot data suggest this approach provides diagnoses of comparable accuracy to other more time intensive or more invasive methods (e.g. neuropsychological testing or imaging). This is a fast, inexpensive, and non-invasive means of obtaining diagnostically useful information. Thus the tool may show most promise as a screening tool to decide which patients need deeper evaluation. Additionally, it may provide objective and quantifiable measures of speech and language symptomatology – a kind of symptomatology for which there are few objective, quantifiable measures.

5 Conclusion

Clinical speech analytics applied to spontaneous speech can detect distinguish between AD, bvFTD, SD PNFA and healthy control groups via lexico-acoustic profiles. Diagnostic accuracy is comparable to other clinical data sources despite speech sample brevity. Accuracy levels suggest the approach offers promise as an additional, objective and easily obtained source of diagnostic information for clinicians.

Reference

Varma A.R., Snowden J.S., Lloyd J.J., Talbot P.R., Mann D.M.A., Neary D. 1999. *Evaluation of the NINCDS-ADRDA criteria in the differentiation of Alzheimer's disease and frontotemporal dementia*, Journal of Neurology, Neurosurgery and Psychiatry, 66: 184-188.

Klöppel, S., Stonnington, C.M., Barnes, J., Chen, F., Chu, C., Good, C.D., Mader, I., Mitchell, L.A., Patel, A.C., Roberts, C.C., Fox, N.C., Jack, R. Jr, Ashburner, J., Frackowiak , RS. 2008. *Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method*, Brain, 131(11): 2969–2974.

S. Pestell, M. Shanks, J. Warrington, and A. Venneri. 2000. *Quality of spelling breakdown in Alzheimer's disease is independent of disease progression*, Journal of Clinical and Experimental Neuropsychology, volume 22, pages 599–612.

H. Platel, J. Lambert, F. Eustache, B. Cadet, M. Dary, F. Viader, and B. Lechevalier. 1993. *Characterstics and evolution of writing impairment in Alzheimer's disease*, Journal of Clinical and Experimental Neuropsychology, volume 22, pages 599–612.

K. Forbes, A. Venneri, and M. Shanks. 2002. *Distinct patterns of spontaneous speech deterioration: an early predictor of Alzheimer's disease*, Brain and Cognition, volume 48(2-3): 356–61.

C. Crockford and R. Lesser. 1994. *Assessing functional communication in aphasia: Clinical utility and time demands of three methods*, European Journal of Disorders of Communication, volume 29: 165–182.

Thomas, V., Keselj, N., Cercone, K., Rockwood, E. 2005. *Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech*, IEEE International Conference on Mechatronics and Automation.

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

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. 1998. *Frontotemporal lobar degeneration: A consensus on clinical diagnostic criteria,* Neurology, 51(6): 1546–54.

Davies, R.R., Hodges, J.R., Kril, J.J., et al. 2005. *The pathological basis of semantic dementia.* Brain, 128(9): 1984–95.

Josephs, K.A., Duffy, J.R., Strand, E.A., et al. 2006. *Clinicopathological and imaging correlates of progressive aphasia and apraxia of speech*. Brain, 129(6): 1385–98.

Bright, P., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. 2008. *Longitudinal studies of semantic dementia: The relationship between structural and functional changes over time*, Neuropsychologia, 46: 2177-2188.

S. Singh, R. Bucks, and J. Cuerden. 2001. *Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech*, Aphasiology, volume 15(6): 571–584.

Stefanie Abel, Walter Huber, Gary S. Dell. 2009. *Connectionist diagnosis of lexical disorders in aphasia*, Aphasiology, volume 23.

Dilek Hakkani-Tür, Dimitra Vergyri, Gökhan Tür. 2010. *Speech-based automated cognitive status assessment*. Interspeech 2010: pages 258-261.

Maider Lehr, Emily T. Prud'hommeaux, Izhak Shafran and Brian Roark. 2012. *Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment*. In Proceedings of Interspeech.

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

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

Pennebaker, J.W., Francis, M.E., Booth, R.J. 2001. *Linguistic Inquiry and Word Count (LIWC): LIWC2001*, Mahwah, NJ: Erlbaum Publishers.

Toutanova, K., Klein, D., Manning, C., Singer, Y. 2003. *Feature-rich part-of-speech tagging with a cyclic dependency network*, in Proceedings of HLT-NAACL 2003, pages 252–259.

Grossman, M., Ash, S. 2004. *Primary Progressive Aphasia: A Review,* Neurocase, 10(1): 3–18.

Gorno-Tempini, M.L, Dronkers, N.F., Rankin, K.P., Ogar, J.M., La Phengrasamy, B.A., Rosen, H.J., Johnson, J.K., Weiner, M.W., Miller, B.L, *Cognition and Anatomy in three variants of primary progressive aphasia*, Annals of Neurology, 2004. 55: 335–346.

Samuel A. K. Seddoh, Donald A. Robin, Hyun-Sub Sim, Carlin Hageman, Jerald B. Moon, John W. Folkins. 1996. *Speech Timing in Apraxia of Speech versus Conduction Aphasia,* Journal of Speech and Hearing Research, 39: 590–603.

Edythe A. Strand, E.A., McNeil, M.R. 1996. *Effects of Length and Linguistic Complexity on Temporal Acoustic Measures in Apraxia of Speech,* Journal of Speech and Hearing Research, 39: 1018–33.

Kirrie J. Ballarrd, Ph.D., and Donald A. Robin. 2002. *Assessment of AOS for Treatment Planning,* Seminars in Speech and Language, 23(4): 281–291.

Witten, I.H., Frank, E. 2005. *Data mining: Practical machine learning tools and techniques*, San Francisco: Morgan Kaufmann. Second edition.

Benjamini, Y., Hochberg Y. 1995. *Controlling the False Discover Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society. Series B *(Methodological)*, 57(1): 289–300.

Saffran, E.M., Berndt, R.S., Schwartz, M.F. 1989. *The quantitative analysis of agrammatic production: procedure and data, Brain and Language*, 37(3): 440–79.

Blair, M., Kertesz, A., Davis-Faroque, N., Hsiung, G.Y.R., Black, S.E., Bouchard, R.W., Gauthier, S., Guzman, D.A., Hogan, D.B., Rockwood, K., Feldman. H. 2007. *Behavioural Measures in Frontotemporal Lobar Dementia and Other Dementias: The Utility of the Frontal Behavioural Inventory and the Neuropsychiatric Inventory in a National Cohort Study,* Dementia and Geriatric Cognitive Disorder, 23: 406-15

Lopez, O. L., Swihart, A. A., Becker, J. T., Reinmuth, O. M., Reynolds, C. F., Rezek, D. L., Daly, F. L. 1990. *Reliability of NINCDS-ADRDA clinical criteria for the diagnosis of Alzheimer's disease,* Neurology, 40: 1517

Kukull, W. A., Larson, E. B., Reifler, B. V., Lampe, T. H., Yerby, M., Hughes, J. 1990. *Interrater reliability of Alzheimer's disease diagnosis*, Neurology, 40(2): 257-60

Peintner, B., Jarrold, W, Vergyri, D., Richey, C., Gorno Tempini, M., and Ogar, J. 2008. *Learning Diagnostic Models Using Speech and Language Measures*, 30th Annual International IEEE EMBS Conference, August 20-24, Vancouver, British Columbia, Canada.

Jarrold, W., Javitz, H.S., Krasnow, R., Peintner, B., Yeh E., Swan, G.E. (2011) *Depression and Self-Focused Language in Structured Interviews with Older Adults* Psychological Reports Oct;109(2):686-700.

Stirman, S.W., & Pennebaker, J.W. (2001). *Word use in the poetry of suicidal and non-suicidal poets.* Psychosomatic Medicine 63, 517-522.

Michelle Hewlett Sanchez, Dimitra Vergyri, Luciana Ferrer,,Colleen Richey, Pablo Garcia, Bruce Knoth, William Jarrold: *Using Prosodic and Spectral Features in Detecting Depression in Elderly-Males.* INTERSPEECH 2011: 3001-3004

H. Kaymaz Keskinpala, T. Yingthawornsuk, D. Mitchell Wilkes, Richard G. Shiavi, R. M. Salomon: Distinguishing high risk suicidal subjects among depressed subjects using mel-frequency cepstrum coefficients and cross validation technique. MAVEBA 2007: 157-160

Supplementary Materials

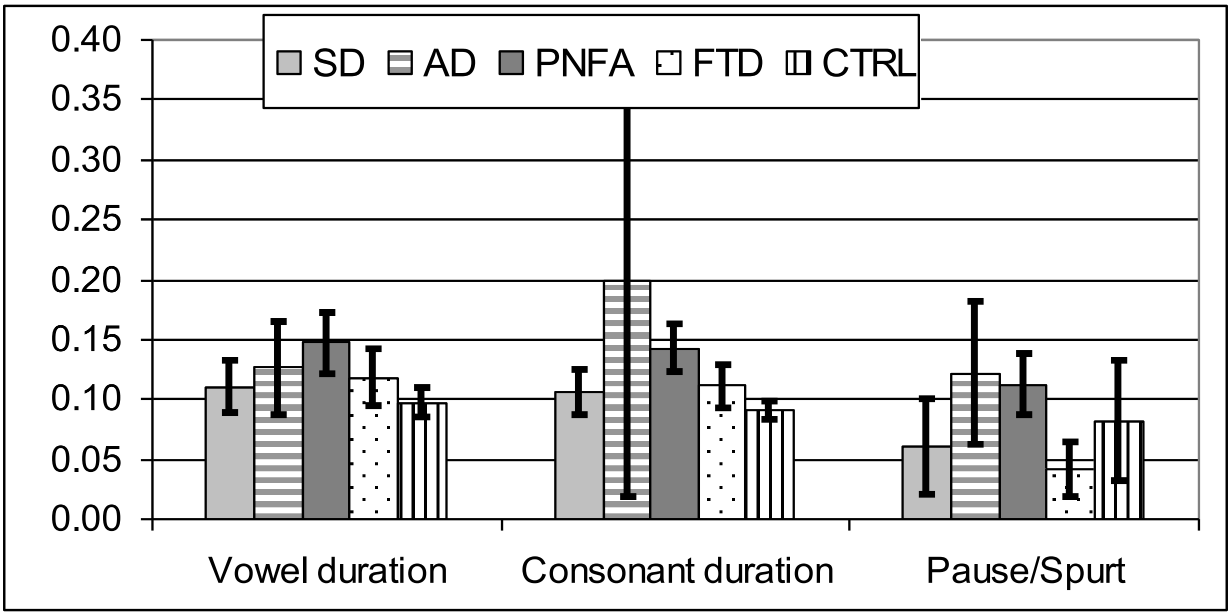

Figure 2. Vowel, consonant, and pause

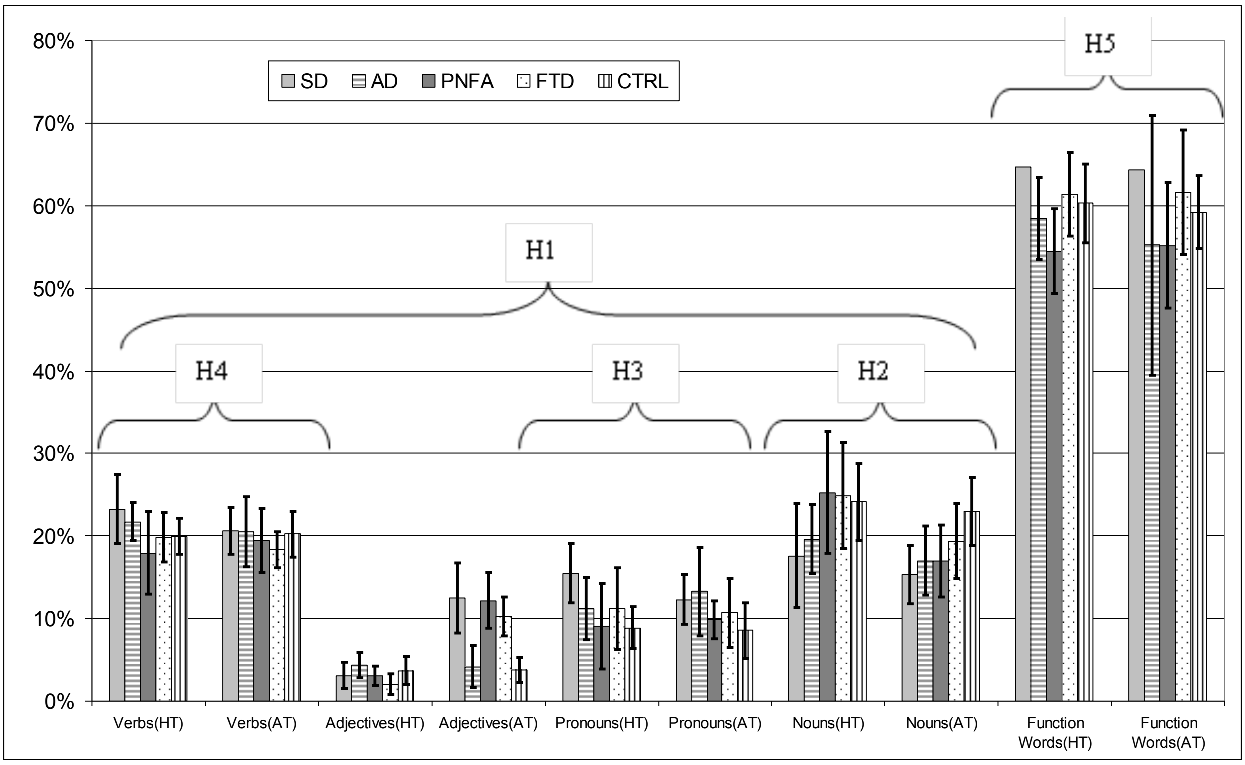

Figure 3. Verb, adjective, pronoun, noun and function word frequencies (H1, H2, H3, H4, H5)