Using text and acoustic features to diagnose progressive aphasia and its subtypes

Kathleen C. Fraser[1], *Frank Rudzicz*[1,2], *Elizabeth Rochon*[2,3]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2] Toronto Rehabilitation Institute, Toronto, Canada
[3] Department of Speech-Language Pathology, University of Toronto, Toronto, Canada
kfraser@cs.toronto.edu, frank@cs.toronto.edu, Elizabeth.Rochon@utoronto.ca

Abstract

This paper presents experiments in automatically diagnosing primary progressive aphasia (PPA) and two of its subtypes, semantic dementia (SD) and progressive nonfluent aphasia (PNFA), from the acoustics of recorded narratives and textual analysis of the resultant transcripts. In order to train each of three types of classifier (naïve Bayes, support vector machine, random forest), a large set of 81 available features must be reduced in size. Two methods of feature selection are therefore compared – one based on statistical significance and the other based on minimum-redundancy-maximum-relevance. After classifier optimization, PPA (or absence thereof) is correctly diagnosed across 87.4% of conditions, and the two subtypes of PPA are correctly classified 75.6% of the time.
Index Terms: aphasia, classification, feature selection

1. Introduction

Primary progressive aphasia (PPA) is a neurodegenerative disease which primarily affects the language areas of the brain. It has two main subtypes: semantic dementia (SD), which is characterized by word-finding difficulties and vague but relatively fluent speech, and progressive nonfluent aphasia (PNFA), which is characterized by slow, hesitant speech and grammatical impairments, but relatively spared single word comprehension [1]. A third subtype, logopenic progressive aphasia, has been identified in recent years but is not considered here (although see Machulda et al. [2] for a recent analysis of that subtype).

Typically, a PPA diagnosis can occur only after a series of tests for cognitive and language function. Usually at least one of these tests involves the production of narrative speech, either in a picture description or a story-telling task. A narrative speech sample can contain rich information about the speaker's ability to choose appropriate content and function words, construct sentences, and convey meaning. However, analysis of narrative speech is typically done by hand and can be prohibitively time-consuming and expensive. Indeed, there is evidence that frontotemporal lobar degeneration of this type can go undiagnosed for up to 7 years, often being misdiagnosed by human assessors as psychiatric disorder or Alzheimer's disease [3]. These problems will increase with the age of populations across many nations. It is therefore important to develop inexpensive and accurate automatic analysis of narrative speech, particularly using features that are relevant to clinical diagnosis.

Our previous work used textual features extracted from transcripts of speech to classify between SD, PNFA, and healthy controls [4]. That work achieved relatively high accuracies between patient groups and controls, but the accuracies were reduced when attempting to distinguish between the two PPA subtypes. By contrast, this paper analyzes acoustic features of patient and control speech, and augments text-based classifiers with these features.

2. Background

Pakhomov et al. [5] extracted a number of different features from the audio files and corresponding transcripts of 38 patients with frontotemporal lobar degeneration (FTLD). They examined the differences between SD and PNFA (which are subtypes of both FTLD and PPA), and behavioural variant FTLD (which is a subtype of FTLD but not of PPA). However, they did not attempt to classify the subtypes based on the extracted features. Peintner et al. [6] analyzed data from 30 FTLD patients and 9 controls. They used a large number of phoneme and linguistic content features to train machine learning classifiers, but did not report which features were selected for classification.

Other related work has applied similar techniques to different clinical groups. Roark et al. [7] tested the ability of a classifier to distinguish patients with mild cognitive impairment from healthy controls based on speech and language measures. Tsanas et al. [8] used speech features to discriminate individuals with Parkinson's disease from healthy controls. A common feature of these studies is the relatively small number of participants. This can present a problem for machine learning approaches, for which a large quantity of training data is preferred. However, clinical data can be expensive and time-consuming to collect. In this domain, appropriate methods for reducing the dimensionality of the data are essential. In this study, we explore two methods of feature selection to reduce the dimensionality of the data: a simple filter method, and minimum-redundancy-maximum-relevance (mRMR).

3. Data

Our data set comprises speech samples from 24 patients with PPA and 16 age- and education-matched controls. Of the 24 PPA patients, 10 were diagnosed with SD and 14 with PNFA. The speech samples were collected as part of a longitudinal study on language impairment in PPA in the Department of Speech-Language Pathology at the University of Toronto. Narrative speech samples were elicited using a standard story-telling procedure (see, for example, Saffran et al. [9]). Participants were given a wordless picture book of the well-known fairy tale "Cinderella", and given a chance to look through the book. The book was then taken away, and participants were asked to tell the story in their own words.

The narrative samples were recorded on a digital audio recorder, and transcribed at the word level by trained research assistants. Transcriptions include filled pauses, repetitions, false starts, and total speech time. Sentence boundaries were marked according to semantic, syntactic, and prosodic cues.

3.1. Features

For each participant, we have two sources of information: the transcript and the audio sample. We extract 58 lexical and syntactic features from the transcript and an additional 23 acoustic features from the audio file, for a total of 81 possible features.

3.1.1. Text features

To examine the syntactic properties of the participants' speech, we use Lu's L2 Syntactic Complexity Analyzer, which counts the number of clauses, dependent clauses, T-units[1], and other syntactic structures [10]. Although this tool was developed to analyze the syntactic complexity of written language, it has also been used to measure the syntactic complexity of speech [11]. We also evaluate syntactic complexity by measuring the height of the parse trees generated by the Stanford parser [12], as well as the maximum, mean, and total Yngve depths[2] [13].

A number of additional features are based on the part-of-speech (POS) tags assigned by the Stanford tagger [14]. SD patients have been observed to produce proportionally fewer nouns and more verbs and pronouns, while PNFA patients tend to produce more nouns and fewer verbs [15, 16]. PNFA patients also tend to omit function words, such as determiners or auxiliaries [17, 16].

We find the frequency of each word in the SUBTL norms, which are derived from a large corpus of subtitles from film and television [18]. We calculate the average frequency over all words as well as specifically for nouns and verbs. Similarly, we calculate the average familiarity, imageability, and age of acquisition of the words in each transcript using the combined Bristol norms and Gilhooly-Logie norms [19, 20]. Each word in these psycholinguistic databases has been ranked according to human perception of how familiar the word is, the approximate age at which a word is learned, and how easily the word evokes an image in the mind. Previous studies have found that SD patients tend to use words which are higher in frequency and familiarity [21], and in some cases lower in imageability [22].

From the transcripts we also measure such quantities as the average length of the words and the type-token ratio, as well as measures of fluency such as the number of filled pauses produced and the rate of speech, or verbal rate. We measure the combined occurrence of all filled pauses, as well as the individual counts for "um" and "uh", since it has been suggested that they may indicate different types of hesitation [23].

3.1.2. Acoustic features

We follow the work of Pakhomov et al. and measure pause-to-word ratio (i.e., the ratio of non-silent segments, excluding filled pauses, to silent segments longer than 150 ms), mean fundamental frequency (F0) and variance, total duration of speech, long pause (> 0.4 ms) count, and short pause (> 0.15 ms) count [5]. To this we add mean pause duration and phonation rate (the amount of the recording spent in voiced speech) [7], as well as the mean and variance for the first 3 formants ($F1, F2, F3$), mean instantaneous power, mean and maximum first autocorrelation function, skewness, kurtosis, zero-crossing rate, mean recurrence period density entropy (a method for measuring the periodicity of a signal, which has been applied to pathological speech generally [24]), jitter [25], and shimmer.

Slow, effortful speech is a core symptom of PNFA, and apraxia of speech is often an early feature of the disease [1]. PNFA patients may make speech sound errors and exhibit disordered prosody [1, 26]. Indeed, atypical F0 range and variance have been shown to be indicative of articulatory neuropathologies within the context of speech recognition [27]. In contrast, speech production is generally spared in SD, although SD patients may produce long pauses as they search for words [16].

3.2. Feature selection

To avoid overfitting, we reduce the dimensionality of our data to be bounded above by the minimum number of data points available for a classification task; since there are 24 speakers with either PNFA or SD, we reduce our feature space from 81 to at most 20. We compare two methods of performing this feature selection. In the first, we calculate Welch's t-test for each feature to calculate the significance of the difference in that feature between the two groups. We then rank each feature by its p-value, and for a feature set of size n we consider only the top n features from the ranked list. This method does not take into account any correlations between variables, but it does offer some insight into which individual features are most strongly indicative of one diagnosis (class) or the other. Feature selection methods based on p-value have been used in previous studies on machine learning classification of frontotemporal lobar degeneration [6] and mild cognitive impairment [7].

The second method we consider is minimum-redundancy-maximum-relevance (mRMR) feature selection in which a set of features is selected such that redundancy (i.e., the average mutual information between features) is minimized and the relevance (i.e., the mutual information between the given features and the class) is maximized [28]. Specifically, for feature f_i in set S and class c, mRMR selects the feature set S^* such that

$$S^* = \arg\max_S \left[\frac{1}{\|S\|} \sum_{f_i \in S} I(f_i; c) - \frac{1}{\|S\|^2} \sum_{f_i, f_j \in S} I(f_i; f_j) \right],$$

where $I(X;Y)$ is the mutual information between X and Y.

Table 1 shows the top $n = 10$ features selected by both the mRMR and p-value methods. Both mRMR and the p-value method select more textual features than acoustic features from all available features; 7/10 features are textual in all cases except for the p-value selection of features for PNFA-vs-SD, in which case 9/10 of the selected features are textual. This might be tempered to some extent by the fact that our acoustic features are more highly correlated (average $r = 0.16$ ($\sigma = 0.47$) among acoustic features and $r = 0.05$ ($\sigma = 0.37$) among textual features). In general, the mRMR and p-value methods are in greater agreement on the PPA-vs-CTRL task (with 6, 7, and 6 features in common across feature sets) than on the PNFA-vs-SD task (with 5, 4, and 4 features in common). Also, the PPA and CTRL classes are more significantly differentiated by the associated top $n = 10$ features than the PNFA and SD classes; all features selected across feature sets in PPA-vs-CTRL are significant at $\alpha \leq 0.05$ (mean $p = 0.003$, ($\sigma = 0.008$)) but fewer than half of the features across feature sets in PNFA-vs-SD are significant (mean $p = 0.131$, ($\sigma = 0.156$)).

[1] T-units are minimally terminable units consisting of a main clause and its dependent clauses.

[2] Yngve depth measures the proportion of left-branching to right-branching in parse tree structures.

		mRMR	p-value
PPA vs. CTRL	text	***frequency***, ***verbalRate***, totalDepth, nounFrequency, verbFrequency, verbs, aveLengthWord, *demonstratives*, totalWords, *familiarity*	verbalRate‡, frequency‡, aveLengthWord‡, demonstratives‡, nounFamiliarity‡, nounFrequency‡, familiarity‡, verbFrequency‡, nouns‡, pronounRatio†
	acous.	***phonationRate***, *meanF2*, *meanRPDE*, *skewness* pause:wordRatio, meanDurationOfPauses, meanInstantaneousPower, F0variance, longPauseCount, meanF0	phonationRate‡, meanRPDE‡, longPauseCount‡, short-PauseCount‡, meanDurationOfPauses‡, meanInstantaneousPower‡, shimmer‡, skewness†, kurtosis*, pauseWordRatio*
	all	***phonationRate***, familiarity, F2variance, *verbalRate*, nounFrequency, verbFrequency, *meanRPDE*, *nounRatio*, TUnitsPerSentence, demonstratives	phonationRate‡, verbalRate‡, frequency‡, aveLengthWord‡, demonstratives‡, meanRPDE‡, nounFamiliarity‡, longPauseCount‡, nounFrequency‡, familiarity‡
PNFA vs. SD	text	nounFamiliarity, *verbalRate*, ***imageability***, *nounFrequency*, adjectives, ***familiarity***, determiners, dependentClauses, nounAOA, S	familiarity‡, nounFamiliarity‡, nounFrequency‡, dependent-ClausesPerClause*, um*, complexTUnits, dependentClauses, verbFamiliarity, demonstratives, determiners
	acous.	***ZCR***, ***shortPauseCount***, *skewness*, totalDurationOfSpeech, *F0variance*, *jitter*, shimmer, meanF0, meanRPDE, phonationRate	meanFirstAutocorrFunc*, jitter, totalDurationOfSpeech, maxFirstAutocorrFunc, pauseWordRatio, F3variance, F2variance, meanF2, meanF0, longPauseCount
	all	*imageability*, ***familiarity***, **jitter**, *verbalRate*, *nounFrequency*, clausesPerTUnit, nounFamiliarity, shortPauseCount, ZCR, demonstratives	familiarity‡, nounFamiliarity‡, nounFrequency‡, dependent-ClausesPerClause*, um*, meanFirstAutocorrFunc*, complex-TUnits, dependentClauses, verbFamiliarity, demonstratives

Table 1: Selected features ($n = 10$) for each task and feature set using the mRMR and p-value methods. Features in **bold** and *italic* appear in the associated selected feature sets for $n = 2$ and $n = 5$ of the mRMR method, respectively; features marked with ‡, † and * represent features on which the given classes are significantly different at $\alpha = 0.005$, $\alpha = 0.01$, and $\alpha = 0.05$, respectively.

4. Experiments in diagnosis

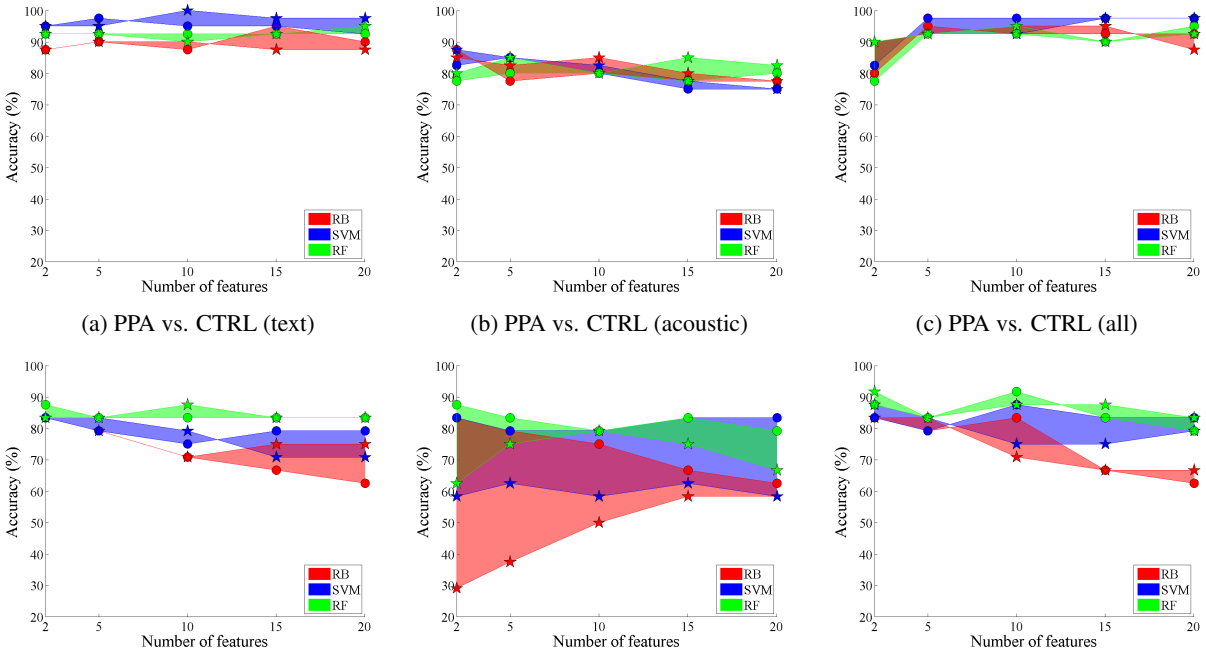
Our experiments compare diagnostic accuracy across a number of empirical variables, namely the task (PPA-vs-CTRL or SD-vs-PNFA), feature set ('Feat. set': text-only, acoustic-only, all), classifier (naïve Bayes (NB), support vector machine with sequential minimal optimization (SVM), and random forests (RF)), number of features considered for classification ('Num. feat.': 2, 5, 10, 15, 20), and the method of feature selection used to derive these reduced sets ('Feat. select', described in section 3.2). The naïve Bayes classifier assumes conditional independence of its features, the SVM is a parametric binary classifier that provides highly non-linear decision boundaries given particular kernels, and the random forest is an ensemble classifier that returns the mode of the class predictions of several decision trees. We optimize the SVM classifier over several combinations of kernel (polynomial of degree 1 or 2, or radial basis function) and complexity ($c = \{0.01, 0.1, 1, 10, 100\}$). We optimize RF for the number of trees ($I = \{5, 10, 15, 20\}$) and the random seed ($S = \{1, 2, 3, 4, 5\}$). Accuracies of diagnostic classification were obtained for each of the possible permutations of our empirical parameters using stratified leave-one-out cross-validation.

Table 2 shows the results of a multi-way analysis of variance (ANOVA) across each of our empirical variables and their two-way interactions. Interestingly, each empirical parameter contributes significantly to the variance *except* for the number of features. Table 3 partitions rates of accurate diagnosis across tasks, feature sets, classifiers, and methods of feature selection. As expected, classifying PPA from CTRL is significantly easier than among sub-types of PPA (heteroscedastic one-tailed $t(268) = 10.354, p < 0.0001, CI = [9.92, \infty]$. Interestingly, although the effect size of considering *all* possible features rather than *only textual* features is small (Cohen's $d = 0.1363$), the difference in accuracy is significant (paired one-tailed $t(89) = -1.798, p < 0.05, CI = [-\infty, -0.09]$. If we consider each task separately, adding acoustics to textual features always increases accuracy, but not significantly (PPA-vs-CTRL: $\mu_{text} = 90.4\%$, $\mu_{all} = 91.2\%$, $t(88) = -0.70, p = 0.24$; SD-vs-PNFA: $\mu_{text} = 78.8\%$, $\mu_{all} = 80.4\%$, $t(88) = -0.96, p = 0.17$.

	Mean sq.	F	$p > F$
Task	7132.24	196.26	$2.27E^{-32}$
Feat. select.	679.69	18.70	$2.31E^{-5}$
Feat. set	2700.66	74.31	$1.96E^{-25}$
Num. feat.	50.59	1.39	0.24
Classifier	985.84	27.13	$2.88E^{-11}$
Task× Feat. select.	437.10	12.03	$6.29E^{-4}$
Task×Feat. set	24.17	0.67	0.52
Task×Num. feat.	103.24	2.84	0.03
Task×Classifier	873.22	24.03	$3.58E^{-10}$
Feat. select.×Feat. set	314.60	8.66	$2.40E^{-4}$
Feat. select.×Num. feat.	31.81	0.88	0.48
Feat. select.×Classifier	96.88	2.67	0.07
Feat. set×Num. feat.	45.20	1.24	0.27
Feat. set×Classifier	39.49	1.09	0.36
Num. feat.×Classifier	51.46	1.42	0.19

Table 2: Multi-way ANOVA (**F** statistics and p values) on accuracy across task, feature selection method, feature set, number of features, classifier, and two-way interactions. Statistically significant ($\alpha = 0.05$) results are in **bold**

Figure 1 shows graphs of accuracy over the number of features in feature sets for each type of feature set (text, acoustic, all) and each task (PPA-vs-CTRL and SD-vs-PNFA).



(a) PPA vs. CTRL (text) (b) PPA vs. CTRL (acoustic) (c) PPA vs. CTRL (all)

(d) SD vs. PNFA (text) (e) SD vs. PNFA (acoustic) (f) SD vs. PNFA (all)

Figure 1: Accuracies across task (PPA-vs-CTRL, SD-vs-PNFA) and feature set (text, acoustic, all) for NB (red), SVM (blue), and RF (green). Star-shaped points represent accuracies obtained using mRMR; circles represent those obtained using the p-value method.

Variable	Value	Accuracy (%)
Task	PPA vs CTRL	$\mu = 87.39, (\sigma = 6.79)$
	SD vs PNFA	$\mu = 75.59, (\sigma = 11.37)$
Feat. set	text	$\mu = 84.62, (\sigma = 8.51)$
	acoustic	$\mu = 74.05, (\sigma = 11.59)$
	all	$\mu = 85.80, (\sigma = 8.84)$
Classifier	NB	$\mu = 77.48, (\sigma = 12.87)$
	SVM	$\mu = 82.13, (\sigma = 11.30)$
	RF	$\mu = 84.85, (\sigma = 6.95)$
Feat. select.	pvalue	$\mu = 83.73, (\sigma = 8.28)$
	mRMR	$\mu = 80.37, (\sigma = 12.08)$

Table 3: Average accuracy μ (and standard deviation σ) of accuracies across experiments partitioned by task, feature set, classifier, and method of feature selection.

5. Discussion

The naïve Bayes method performed surprisingly well here, although generative approaches can sometimes outperform discriminative ones when the available data is less vast [29]. The feature selection method can affect the accuracy of NB, as illustrated in Figure 1e. For a feature set of size two, the NB classifier achieved an accuracy of 83% using the p-value method, and only 29% using mRMR. We hypothesize that this is because the p-value filter, which chooses features on the basis of their mean and variance, is choosing exactly those features which would best distinguish the groups in a Gaussian framework. Indeed, the top two features chosen by mRMR (ZCR and short pause count) both have bimodal (non-Gaussian) distributions.

Using acoustic features as well as text features had a significant effect on the classification accuracies. For the task of classifying PPA versus controls, it is unsurprising that features like phonation rate and short and long pause counts are informative, reflecting the language difficulties experienced by the PPA patients. Other features which were significantly different between the groups, including mean RPDE, mean instantaneous power, shimmer, skewness, and kurtosis, have not to our knowledge been previously reported for PPA.

Only one acoustic feature, mean first autocorrelation function, significantly differentiated the SD and PNFA groups. This is somewhat unexpected, as SD patients are often described as more "fluent" than PNFA patients, suggesting they could be distinguished on the basis of characteristics such as phonation rate and number of pauses. However, fluency may be a poor marker for subtyping PPA, in part because patients without PNFA may show "intermittent dysfluency" due to word-finding difficulties [30]. Wilson et al. [16] also found reduced average speech rate for both PNFA and SD, and suggested that measuring maximum speech rate might be more useful for distinguishing them.

In future work we hope to determine whether there are other acoustic features which may better differentiate the patient groups, and whether it would be more effective to approach the problem as a three-class classification (SD vs. PNFA vs. controls). Future work will also examine the effect (or lack thereof) of the number of features across classifiers, and the interaction between the diagnostic task and the feature set.

6. Acknowledgments

This work was supported by the Canadian Institutes of Health Research (CIHR), Grant #MOP-82744, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

7. References

[1] M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, B. F. Boeve, F. Manes, N. F. Dronkers, R. Vandenberghe, K. Rascovsky, K. Patterson, B. L. Miller, D. S. Knopman, J. R. Hodges, M. M. Mesulam, and M. Grossman, "Classification of primary progressive aphasia and its variants," *Neurology*, vol. 76, pp. 1006–1014, 2011.

[2] M. M. Machulda, J. L. Whitwell, J. R. Duffy, E. A. Strand, P. M. Dean, M. L. Senjem, C. R. Jack Jr, and K. A. Josephs, "Identification of an atypical variant of logopenic progressive aphasia," *Brain and language*, 2013.

[3] V. S. Bahia, "Underdiagnosis of frontotemporal lobar degeneration in Brazil," *Dementia & Neuropsychologia*, vol. 1, no. 4, pp. 261–365, 2007.

[4] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon, "Automated classification of primary progressive aphasia subtypes from narrative speech transcripts," *Cortex*, 2012.

[5] S. V. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman, "Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration," *Cognitive and Behavioral Neurology*, vol. 23, pp. 165–177, 2010.

[6] 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.

[7] 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.

[8] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *Biomedical Engineering, IEEE Transactions on*, vol. 59, no. 5, pp. 1264–1271, 2012.

[9] E. M. Saffran, R. S. Berndt, and M. F. Schwartz, "The quantitative analysis of agrammatic production: procedure and data," *Brain and Language*, vol. 37, pp. 440–479, 1989.

[10] X. Lu, "Automatic analysis of syntactic complexity in second language writing," *International Journal of Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.

[11] M. Chen and K. Zechner, "Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 722–731.

[12] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423–430.

[13] V. Yngve, "A model and hypothesis for language structure," *Proceedings of the American Physical Society*, vol. 104, pp. 444–466, 1960.

[14] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2003, pp. 252–259.

[15] A. E. Hillis, S. Oh, and L. Ken, "Deterioration of naming nouns versus verbs in primary progressive aphasia," *Annals of Neurology*, vol. 55, no. 2, pp. 268–275, 2004.

[16] S. M. Wilson, M. L. Henry, M. Besbris, J. M. Ogar, N. F. Dronkers, W. Jarrold, B. L. Miller, and M. L. Gorno-Tempini, "Connected speech production in three variants of primary progressive aphasia," *Brain*, vol. 133, pp. 2069–2088, 2010.

[17] S. Ash, P. Moore, L. Vesely, D. Gunawardena, C. McMillan, C. Anderson, B. Avants, and M. Grossman, "Non-fluent speech in frontotemporal lobar degeneration," *Journal of Neurolinguistics*, vol. 22, no. 4, pp. 370–383, 2009.

[18] 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.

[19] H. Stadthagen-Gonzalez and C. J. Davis, "The Bristol norms for age of acquisition, imageability, and familiarity," *Behavior Research Methods*, vol. 38, no. 4, pp. 598–605, 2006.

[20] K. Gilhooly and R. Logie, "Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words," *Behavior Research Methods*, vol. 12, pp. 395–427, 1980.

[21] L. Meteyard and K. Patterson, "The relation between content and structure in language production: an analysis of speech errors in semantic dementia," *Brain and Language*, vol. 110, no. 3, pp. 121–134, 2009.

[22] H. Bird, M. A. Lambon Ralph, K. Patterson, and J. R. Hodges, "The rise and fall of frequency and imageability: Noun and verb production in semantic dementia," *Brain and Language*, vol. 73, pp. 17–49, 2000.

[23] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[24] M. Little, P. McSharry, I. Moroz, and S. Roberts, "Nonlinear, biophysically-informed speech pathology detection," in *Proceedings of ICASSP 2006*, Toulouse, France, 2006, pp. 1080–1083.

[25] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–9, 2009.

[26] M. Grossman, "Primary progressive aphasia: clinicopathological correlations," *Nature Reviews Neurology*, vol. 6, pp. 88–97, 2010.

[27] K. Mengistu, F. Rudzicz, and T. Falk, "Using acoustic measures to predict automatic speech recognition performance for dysarthric speakers," in *Proceedings of the 7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications at INTERSPEECH 2011*, Firenze Italy, August 2011.

[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[29] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems (NIPS) 14*, vol. 2, 2002, pp. 841–848.

[30] C. K. Thompson, S. Cho, C.-J. Hsu, C. Wieneke, A. Rademaker, B. B. Weitner, M. M. Mesulam, and S. Weintraub, "Dissociations between fluency and agrammatism in primary progressive aphasia," *Aphasiology*, vol. 26, no. 1, pp. 20–43, 2012.