Automatic speech recognition in the diagnosis of primary progressive aphasia

Kathleen Fraser[1], *Frank Rudzicz*[1,2], *Naida Graham*[2,3], *Elizabeth Rochon*[2,3]

[1]Department of Computer Science, University of Toronto; [2] Toronto Rehabilitation Institute
[3] Department of Speech-Language Pathology, University of Toronto

kfraser@cs.toronto.edu, frank@cs.toronto.edu,
naida.graham@utoronto.ca, elizabeth.rochon@utoronto.ca

Abstract

Narrative speech can provide a valuable source of information about an individual's linguistic abilities across lexical, syntactic, and pragmatic levels. However, analysis of narrative speech is typically done by hand, and is therefore extremely time-consuming. Use of automatic speech recognition (ASR) software could make this type of analysis more efficient and widely available. In this paper, we present the results of an initial attempt to use ASR technology to generate transcripts of spoken narratives from participants with semantic dementia (SD), progressive nonfluent aphasia (PNFA), and healthy controls. We extract text features from the transcripts and use these features, alone and in combination with acoustic features from the speech signals, to classify transcripts as patient versus control, and SD versus PNFA. Additionally, we generate artificially noisy transcripts by applying insertions, substitutions, and deletions to manually-transcribed data, allowing experiments to be conducted across a wider range of noise levels than are produced by a tuned ASR system. We find that reasonably good classification accuracies can be achieved by selecting appropriate features from the noisy transcripts. We also find that the choice of using ASR data or manually transcribed data as the training set can have a strong effect on the accuracy of the classifiers.

Index Terms: automatic speech recognition, classification, progressive aphasia

1. Introduction

Primary progressive aphasia (PPA) is a neurodegenerative disorder in which language is the most affected aspect of cognitive functioning. There are two main variants of PPA: progressive nonfluent aphasia (PNFA), in which speech is hesitant and effortful, and semantic dementia (SD), in which speech is fluent but with severe word findings difficulties [1]. A third subtype, logopenic progressive aphasia, has been identified in recent years but is not considered here.

The features of narrative speech in each variant of PPA have been characterized to some extent, but they are not yet fully understood. Evaluation of spoken output is an important part of diagnosis of PPA and in identification of the variant. From a clinical perspective, analysis of narrative speech has the advantage that it can provide a lot of information from a relatively brief assessment. 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. Systematic analysis of narrative speech is typically done manually, which is time-consuming and may be prohibitively expensive. The automated approach evaluated here has several advantages. For example, this method enables simultaneous consideration of multiple aspects of speech. Also, it should ultimately provide greater sensitivity to changes occurring in the earliest stages of disease, thereby facilitating early diagnosis. Similarly, it should provide objective measures of changes over time in language production, thereby enabling more accurate assessment of disease progression; this is important for patients and their families, as well as for evaluation of efficacy in drug trials (as potentially disease modifying drugs become available).

Fully automated analysis of narrative speech will require automatic speech recognition (ASR) in order to extract lexical and syntactic features from acoustic signals. Despite major improvements in ASR technology over the past few decades, accuracy for unrestricted (i.e., 'dictation-style') speech remains decidedly imperfect, as described in the next section. In order to estimate how effective a classifier of PPA and its subtypes might be when given textual transcripts derived from ASR, a wide range of potential system performances must be considered, to account for real-world variation. This research approximates various levels of ASR performance by randomly corrupting human transcripts according to pre-defined levels of error and compares these results against actual output from a leading commercial dictation system. Error levels are quantified by word-error rate (WER), which is the total number of erroneous insertions, deletions, and substitutions of words in an ASR transcript, divided by the total number of words in a reference transcript[1]. Simulated ASR errors have been used in various contexts, such as training dialogue systems [2] and for testing the safety of dictation systems for use in automobiles [3].

2. Related Work

In general, the accuracy of ASR systems on elderly voices tends to decrease with the age of the speaker [4]. Elderly voices typically have increased breathiness, jitter, shimmer, and a decreased rate of speech [4]. Older speakers may also exhibit articulation difficulties, changes in fundamental frequency, and decreased voice intensity [5]. These factors can result in speech that is less intelligible to both human listeners and ASR systems. For example, Hakkani-Tur *et al.* [6] 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.

[1]If the number of insertions is large, it can overwhelm the total number of words in the reference transcript, therefore allowing for WERs above 100%.

Effective speech recognition can be further challenged by the presence of linguistic impairments such as those occurring in PPA. To our knowledge, there has only been one previous study on automatic speech recognition of PPA speakers. Peintner et al. [7] analyzed speech from patients with PNFA and SD as well as patients with a dementia affecting behavior and deportment, but not language. They achieved a WER of 37% for SD and 61% for PNFA. They also tested a control group, who had an average WER of 20%.

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 normal or has a subtype of PPA. Peintner et al. [7] also attempted this task, although they did not report how the high error rates affected the lexical features studied or their classification accuracy. Other studies in this area have used manually transcribed transcripts [8]. One strategy which combines ASR technology with manual transcripts is to use forced-alignment with manual transcripts to measure acoustic features such as rate of speech and length of pauses [9, 10]. 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.

3. Data

3.1. Narrative samples

Our data set comprises speech samples from 24 patients with PPA and 16 age- and education-matched controls. Of the 24 PPA patients, 14 were diagnosed with PNFA and 10 with SD. 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. See Table 1 for demographic information about the participants.

Narrative speech samples were elicited following the procedure described by Saffran et al. [11]. Participants were given a wordless picture book of the well-known fairy tale "Cinderella", and were asked to look through the book. The book was then removed, and participants were asked to tell the story in their own words.

The narrative samples were recorded on a digital audio recorder, and transcribed by trained research assistants. The manual transcriptions include filled pauses, repetitions, and false starts. Sentence boundaries were marked according to semantic, syntactic, and prosodic cues. The SD patients produced an average of 380 words and 20 sentences, the PNFA patients produced an average of 302 words and 16 sentences, and the control group produced an average of 403 words and 16 sentences.

	SD ($n = 10$)	PNFA ($n = 14$)	Controls ($n = 16$)
Age	65.6 (7.4)	64.9 (10.1)	67.8 (8.2)
Years of education	17.5 (6.1)	14.3 (3.6)	16.8 (4.3)
Sex	3 F	6 F	7 F

Table 1: Demographic information for each participant group. Averages (and standard deviations) are given for age and years of education.

3.2. Features

Two types of features are extracted for each participant individually, namely textual transcripts and acoustic samples. From these, we derive 31 lexical/syntactic features from the text transcripts and 23 features from the acoustics, giving a total of 54 available features, described below.

3.2.1. Text features

A number of features can be extracted from the text transcripts. Some of our features are based on the part-of-speech (POS) tags assigned by the Stanford tagger [12]. 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 [13, 14, 15]. PNFA patients also tend to omit function words, such as determiners or auxiliaries [13, 16].

We look up the frequency of each word in the SUBTL norms, which are derived from a large corpus of subtitles from film and television [17]. We calculate the average frequency over all words as well as specically 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 [18, 19]. Each word in these psycholinguistic databases has been ranked according to human perception of how familiar the word is, how easily the word evokes an image in the mind, and the approximate age at which a word is learned. Frequency, familiarity, imageability, and age of acquisition have all been found to influence speech production in aphasia [14, 20, 21, 22, 23]. The coverage of these norms on our data is variable. The frequency norms have excellent coverage – between 0.92 and 0.95 across the three groups on the manually transcribed data. The coverage for the familiarity, imageability, and age of acquisition norms is not as good, possibly due to the fact that the authors of the norms specifically excluded high frequency words [18]. The coverage for those norms ranges from 0.25 to 0.31 for all content words across the three groups for the manual transcripts.

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. 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 [24].

In previous work using manual transcripts, researchers have also examined measures which can be derived from parse trees, such as Yngve depth, or the number and length of different syntactic constructions [8, 9]. However, such parse trees will depend on the location of the sentence boundaries in the transcript, the placement of which can be a difficult task for ASR systems [25]. Indeed, the Nuance system used here does not place punctuation except by explicit command. For the purposes of this preliminary study, we avoid using features which depend on accurate sentence boundaries.

3.2.2. Acoustic features

We follow the work of Pakhomov et al. [10] and measure pause-to-word ratio (i.e., the ratio of non-silent segments to silent segments longer than 150 ms), mean fundamental frequency (F0) and variance, total duration of speech, long pause count (> 0.4 ms), and short pause count (> 0.15 ms and < 0.4 ms). To this we add mean pause duration and phonation rate (the amount of the recording spent in voiced speech) [9], 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 [26]), jitter [27], and shimmer.

Slow, effortful speech is one of the core symptoms of PNFA, and apraxia of speech can be an early feature [1]. PNFA patients may make speech sound errors and exhibit disordered prosody [1, 28]. Similarly, typical F0 range and variance have been shown to be indicative of articulatory neuropathologies within the context of speech recognition [29, 30]. In contrast, speech production is generally spared in SD, although SD patients may produce long pauses as they search for words [13].

4. Methods

4.1. ASR and simulated errors

We use two methods to produce errorful textual transcripts. The first method represents the current leader in commercial dictation software, Nuance Dragon NaturallySpeaking Premium; here, audio files are transcribed by Nuance's desktop dictation software. The second method corrupts human-produced transcripts according to pre-defined levels of WER; this method allows for an indirect approximation of the performance given a wide range of potential alternative ASR systems.

The Nuance Dragon NaturallySpeaking 12.5 Premium for 64-bit Windows dictation system (hereafter, 'Nuance') is based on traditional hidden Markov modeling of acoustics and, historically, on trigram language modeling [31]. This system is initialized with the default 'older voice' model suitable for individuals 65 years of age and older. The default vocabulary consists of 150,478 words, plus additional control phrases for use during normal desktop dictation (e.g., "*new paragraph*", "*end of sentence*"); this feature cannot be deactivated. The core vocabulary, however, can be changed. In order to get a more restricted vocabulary, all words used in our manually transcribed Cinderella data set plus all words used in a selection of 9 stories about Cinderella from the Gutenberg project (totalling 22,168 word tokens) were combined to form a reduced vocabulary of 2633 word types. Restricted vocabularies, by their nature, have higher random baselines and less phonemically confusable word pairs, usually resulting in proportionally higher accuracies in ASR. The Nuance system scales the language model to the reduced vocabulary.

For the simulated ASR transcripts, each word in the manual transcript is modified with a probability equal to the desired WER. In this set of experiments, we use a language model obtained from the Gigaword corpus [32], since the Nuance language model is proprietary and not accessible to the user. A word w can be modified in one of three ways:

- Substitution – w is replaced with a new word w_S.
- Insertion – w is followed be a new word w_I.
- Deletion – w is removed.

In the case of insertion, the word to be inserted is chosen randomly according to the bigram distribution of the language model. That is, words that frequently occur after w are more likely to be chosen as w_I. If w is not found in the Gigaword vocabulary, then w_I is chosen randomly according to the unigram distribution of the language model. In the case of substitution, the new word is randomly chosen from a ranked list of words with minimal phonemic edit distance from the given word, as computed by the Levenshtein algorithm.

Once it has been determined that a word will be modified, it is assigned one of the above modifications according to a predefined distribution. Different ASR systems may tend towards different distributions of insertion errors (IE), substitution errors (SE), and deletion errors (DE). We create data noise according to three distributions, each of which favours one type of error over the others: [60% IE, 20% SE, 20% DE], [20% IE, 60% SE, 20% DE], and [20% IE, 20% SE, 60% DE]. We then also adjust these proportions according to proportions observed in Nuance output, as described in Section 5.

4.2. Classification

We use stratified leave-one-out cross-validation to test our diagnostic classifiers. For each fold, one transcript is removed as test data. We then apply a simple feature selection algorithm to the remaining transcripts: we calculate a Welch's t-test for each feature individually and determine the significance of the difference between the groups on that feature. We then rank each feature by increasing p-value, and include as input to the classifier only the top ten most significant features in the list. For each fold, different training data is used and therefore different features may be prioritized in this manner. Similar methods for feature selection have been used in previous studies on the classification of dementia subtypes [7, 9, 33].

Once the features have been selected, we train three types of classifier: naïve Bayes (NB), support vector machine with sequential minimal optimization (SVM), and random forests (RF). The classifiers are then tested with the same subset of features derived from the held-out transcript. This procedure is repeated for every transcript in the data set, and the average accuracy is computed.

We consider two classification tasks, PPA-vs.-control and SD-vs.-PNFA, since these binary tasks allow for less confusion than a trinary classification task and can be cascaded. For each task, there are two possible feature sets: text features only, or a combination of text and acoustic features. There are also two possible training sets for each task: i) the classifiers can be trained on the human-transcribed data and tested on the ASR data[2], and ii) the classifiers are both trained and tested on the noisy ASR (or simulated ASR) data. We test our classifiers on each combination of these variables.

5. Results

5.1. Features and feature selection

First, we examine whether the feature selection method selects different types of features depending on the WER. It might be expected that as the WER increases, the text features will become less significant. Figure 1 shows the p-values, averaged across folds, for the text and acoustic features selected at each WER for each noise distribution. Note that the values of the acoustic features do not change with the noise levels, but the average p-value will change as different features are selected in each case, depending on the values of the text features. For the case of PPA versus controls, a mix of text and acoustic features are chosen, and the features tend to be significant at $p < 0.05$, even when the error rate is high. A combination of text and acoustic features are also selected for SD versus PNFA at all

[2]This represents the scenario in which researchers have access to a corpus of manual transcriptions for training purposes

noise levels; however in this case the mean p-values are often not significant, suggesting that the features are not as discriminative between these groups. This effect is reflected in the lower classification accuracies for the SD versus PNFA task reported below. So, Figure 1 does not support the hypothesis that text features become irrelevant at the highest noise levels, but rather suggests that the transcripts still contain some information which is at least as valuable as the acoustic information in the speech signal.

	p-value	PPA mean	Control mean
Nuance default vocabulary			
verb imageability	0.0006	401	354
noun frequency	0.002	3.51	3.26
noun familiarity	0.04	575	558
Nuance reduced vocabulary			
average word length	0.003	5.44	6.21
noun frequency	0.006	3.13	2.77
noun imageability	0.01	487	554
noun familiarity	0.02	558	531
frequency	0.04	3.60	3.20

Table 2: Significant text features ($p < 0.05$) for PPA vs. Controls using the Nuance system with default and reduced vocabularies.

	p-value	SD mean	PNFA mean
Nuance default vocabulary			
noun familiarity	0.002	596	560
familiarity	0.002	594	568
Nuance reduced vocabulary			
None	N/A	N/A	N/A

Table 3: Significant text features ($p < 0.05$) for SD vs. PNFA using the Nuance system with default and reduced vocabularies.

Some text features are still significant in the Nuance data as well, despite the high WER. Table 2 shows the text features that were significant ($p < 0.05$) when comparing PPA and controls using the two Nuance models. As before, since the feature set changes with each fold in the cross-validation, the p-value is an average across folds. The means for the two groups are also shown to indicate the direction of the difference. Using the default vocabulary, there are three significant text features: verb imageability, noun frequency, and noun familiarity. These three features are all significant in the manually-transcribed data as well, and with the same direction. For the system trained on the reduced vocabulary, there are five significant text features, as indicated, only one of which (noun imageability) is not significant in the manual transcripts. All five features show differences in the same direction. Table 3 shows that only noun familiarity and overall familiarity are significant in the SD vs. PNFA case using the default vocabulary system, as they are in the manually transcribed data, with the difference in the same direction. There are no significant text features using the reduced vocabulary system.

The significant acoustic features for each classification task are shown in Tables 4 and 5. These features remain the same regardless of the transcription method. For a complete discussion of the acoustic features of this data set, see [33].

	p-value	PPA mean	Control mean
phonation rate	0.0000006	0.733	0.920
mean duration of pauses	0.00002	37 800	14 500
mean recurrence period density entropy	0.00002	0.549	0.477
long pause count	0.0006	34.7	10.6
skewness	0.0006	-0.0733	-0.532
mean instantaneous power	0.0003	-26.1	-22.1
short pause count	0.002	49.9	22.1
kurtosis	0.005	20.4	14.1
shimmer	0.05	0.00560	0.00748

Table 4: Significant acoustic features ($p < 0.05$) for PPA vs. Controls.

	p-value	SD mean	PNFA mean
mean first autocorrelation function	0.02	0.848	0.730

Table 5: Significant acoustic features ($p < 0.05$) for SD vs. PNFA.

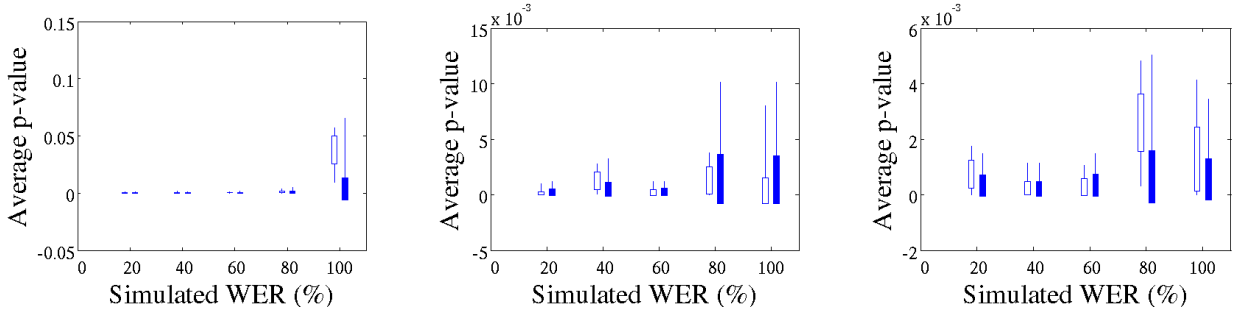
5.2. Recognizing PPA speech

Table 6 shows the WER of the Nuance system across populations and vocabularies. Somewhat surprisingly, using the reduced vocabulary reduces accuracy considerably, despite all words in the test set being present in the vocabulary. A possible explanation may be found in the distribution of error types across the uses of both vocabularies, which is shown in table 7. In particular, when using the reduced vocabulary, Nuance makes significantly more deletion errors, which may be attributed to a lower confidence assigned to its word sequence hypotheses which in turn may be attributed to a language model that is not adapted to non-default vocabularies. A general language model may assign a high lexical probability to a series of words that are phonemically similar to an utterance but if those words are not in the reduced vocabulary, a more domain-specific sequence of words may be assigned a low lexical probability and therefore a low confidence. When confidence in a hypothesis is below some threshold, that hypothesis may not be returned, resulting in an increase in deletion errors. Not having access to these internals of the Nuance engine prohibits modification at this level.

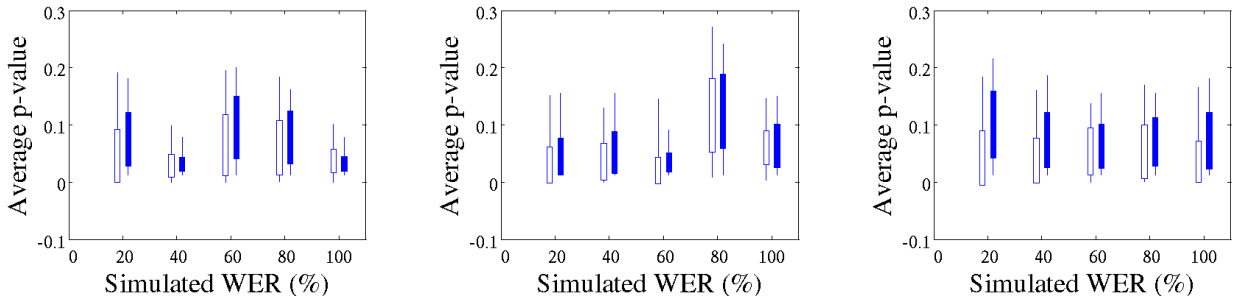
Another point to highlight is that, given Nuance's default vocabulary, there is no significant difference between the WER obtained with the control and PNFA groups ($t(26.78) = -0.62, p = 0.54, CI = [-0.16, 0.08]$), nor with the con-

	Default Vocabulary	Reduced Vocabulary
SD	73.1	98.1
PNFA	67.7	97.3
Control	64.0	97.1
All	67.5	97.5

Table 6: Mean word error rates for the Nuance systems on each of the participant groups.



(a) PPA vs. control, 0.2 IE, 0.2 SE, 0.6 DE (b) PPA vs. control, 0.2 IE, 0.6 SE, 0.2 DE (c) PPA vs. control, 0.6 IE, 0.2 SE, 0.2 DE



(d) SD vs. PNFA, 0.2 IE, 0.2 SE, 0.6 DE (e) SD vs. PNFA, 0.2 IE, 0.6 SE, 0.2 DE (f) SD vs. PNFA, 0.6 IE, 0.2 SE, 0.2 DE

Figure 1: Acoustic features (filled bars) and text features (empty bars) selected for the feature sets at each WER for each distribution of insertion errors (IE), substitution errors (SE), and deletion errors (DE). Each bar represents one standard deviation from the mean, and the lines indicate the minimum and maximum values.

	Default Vocabulary	Reduced Vocabulary
Insertion errors	0.00602	0.00008
Substitution errors	0.39999	0.11186
Deletion errors	0.59398	0.88804

Table 7: Distribution of error types for the Nuance systems.

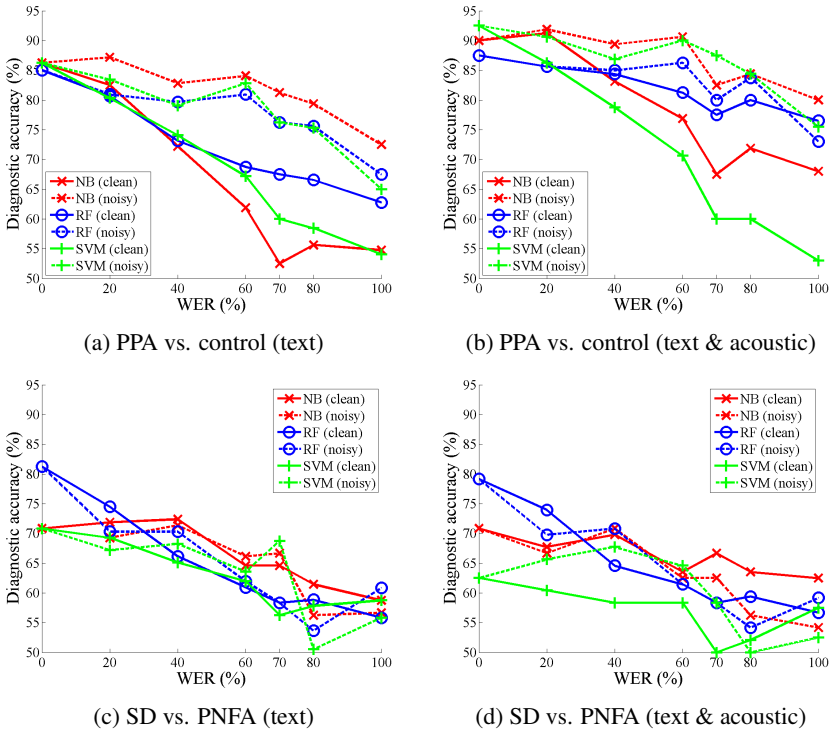
trol and SD groups ($t(23.77) = -1.47, p = 0.16, CI = [-0.22, 0.04]$), although the differences in Table 7 might seem large.

5.3. Diagnosing PPA and its subtypes

We evaluate the accuracy of diagnosing PPA and its subtypes based on the selected features across the three classification methods using the simulated ASR method. In practice, classification models might be trained on data that have been manually transcribed by humans (clinicians or otherwise). However, as the amount of data increases, this becomes less practical and it may become necessary to train these models from transcripts that were automatically generated from ASR. We replicate our experiments once on data that have been manually transcribed and once on the same data, but with transcripts corrupted by synthetic word errors (in which case the training data and test data have the same WER). Classifiers trained on human-produced transcripts have an average accuracy of 65.71% ($\sigma = 12.42$) and those trained on 'noisy' transcripts have an average accuracy of 70.72% ($\sigma = 13.89$), which is significant at heteroscedastic $t(543) = -4.47, p < 0.00001, CI = [-0.072, -0.028]$. These differences can be observed in Figure 2. Interestingly, the classifiers trained with 'noisy' transcripts outperform those trained with 'clean' transcripts fairly consistently in the PPA vs. control task, but this is far less pronounced (and to some extent reversed) in the SD vs. PNFA task. This may be partially explained by a significant three-way interaction between WER, the task (i.e., the participant groups), and the training set (i.e., 'noisy' vs. 'clean') on a followup ANOVA ($F(6) = 2.43, p < 0.05$).

This trend is also apparent when the classifiers are tested using the Nuance transcripts. Figure 3 shows the classification accuracies for each classifier on each diagnostic task using the data generated using the default and reduced vocabularies. When classifiying PPA versus controls, training on the 'noisy' Nuance data always leads to equal or greater accuracies than training on the 'clean' (human-transcribed) data. For SD versus PNFA, the results are mixed, although the results from the reduced vocabulary suggest the opposite trend.

We compare the diagnostic accuracies across all classifiers given transcripts from Nuance using the reduced vocabulary with the accuracies of the synthetic WER method using the nearest WER (100%) and the associated error type distribution (i.e., 10% substitutions, 90% deletions, over all errors). We find no difference between results obtained with Nuance data and those obtained with the synthetic method ($t(44.25) = 1.1072, p = 0.27, CI = [-0.04, 0.13]$). We repeat this analysis with the default Nuance vocabulary and its equivalent synthetic WER (70%) and distribution (i.e., 40% substitution, 60% deletion) and again find no significant difference ($t(44.61) = 1.46, p = 0.15, CI = [-0.02, 0.11]$). Here, distributions of WER are approximately Gaussian over the various parameterizations of the systems. The lack of apparent difference in diagnosis when using the Nuance ASR and the synthetic method supports the use of the latter in these experi-



(a) PPA vs. control (text)

(b) PPA vs. control (text & acoustic)

(c) SD vs. PNFA (text)

(d) SD vs. PNFA (text & acoustic)

Figure 2: Accuracy in diagnosing the indicated classes given features derived from potentially error-full textual transcriptions alone and in combination with features derived directly from the acoustics. Lines marked with x's, circles, and pluses indicate the use of the naïve Bayes, random forest, and support vector machine classifiers. Solid lines indicate those trained with human-transcribed (clean) data and dashed lines indicate those trained with corrupted data.

ments.

Among the simulated ASR data, an n-ary ANOVA reveals significant main effects for each of the classification problems (PPA-vs.-control or PNFA-vs.-SD; $F(1) = 124.19, p = 0$), WER ($F(5) = 31.69, p = 0$), error distribution (proportions of IE, SE, and DE; $F(4) = 6.32, p < 0.0005$), and training set ('noisy' or 'clean'; $F(1) = 35.41, p = 0$) on the accuracy of classification; there is no effect of the classifier, however ($F(2) = 2.27, p = 0.1039$). There were significant interaction effects between WER and the classification problem ($F(5) = 5.18, p < 0.0005$), error distribution ($F(12) = 2.2, p < 0.05$), and the training set ($F(5) = 4.95, p < 0.0005$), but not with the data subset (text or text with acoustics; $F(5) = 1.42, p = 0.2146$), or the classifier ($F(10) = 0.49, p = 0.8993$).

6. Discussion

Our goal is to provide assistive technologies, including diagnostic software, to various populations with pathological speech and language, including those with PPA. This study represents an initial step towards ASR for this population. One main result of this research is that fairly accurate diagnosis of PPA and of its subtypes can remain relatively accurate, even at very high levels of WER, by selecting appropriate features from the data at training time. Acoustic features are valuable, as they remain constant as the WER increases. However, our data suggest that some features from the text can still be informative, even when the transcripts are very noisy.

One important direction for future work is to improve ASR for clinical populations. Clearly, modern speech recognition has greater difficulty in recognizing PPA speech relative to speech the general elderly population, especially for individuals with SD. While more appropriate acoustic models built for older-adult voices will be important (based on available data), a focus on improving language modeling and the pruning of the lattices produced by hidden Markov models may be more fruitful if the cause of the pathology is semantic or lexical.

Another limitation of our approach is that the t-test method for feature selection does not consider interactions between features. In the future we would like to examine these interactions, particularly between text and acoustic features.

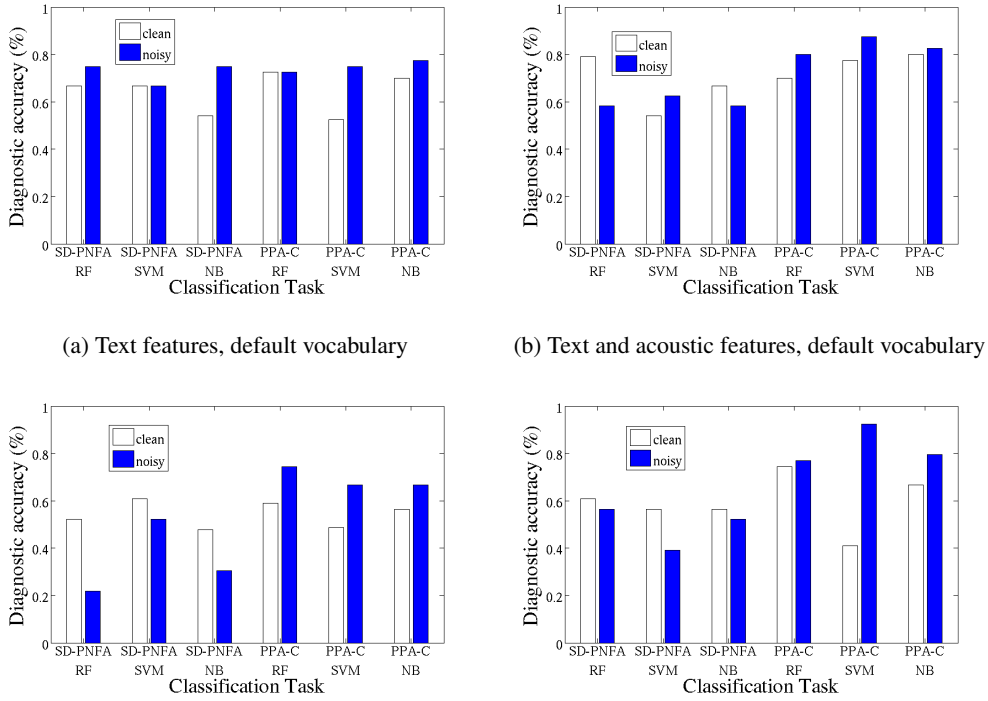
In this study we did not take into account any syntactic features, although agrammatism and/or syntactic simplification are characteristic of PNFA. Presumably, including information of this type could increase the classification accuracy. One approach would be to apply a sentence boundary detection algorithm to the ASR transcripts and extract traditional syntactic complexity measures (e.g. Yngve depth). Another approach would be to explore localized complexity metrics which do not depend on full sentence parses.

7. Acknowledgements

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).

8. 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.



(a) Text features, default vocabulary

(b) Text and acoustic features, default vocabulary

(c) Text features, reduced vocabulary

(d) Text and acoustic features, reduced vocabulary

Figure 3: Classification accuracies using transcripts from the Nuance system with the default and reduced vocabularies, for random forest (RF), support vector machine (SVM), and naïve Bayes (NB) classifiers. Empty bars indicate the accuracy achieved when training on the clean, human-transcribed data, while filled bars indicate the accuracy when training on the noisy ASR data.

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] J. Schatzmann, B. Thomson, and S. Young, "Error simulation for training statistical dialogue systems," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 526–531.

[3] M. Labský, J. Cuřín, T. Macek, J. Kleindienst, L. Kunc, H. Young, A. Thyme-Gobbel, and H. Quast, "Impact of word error rate on driving performance while dictating short texts," in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ser. AutomotiveUI '12. ACM, 2012, pp. 179–182.

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

[5] 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.

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

[7] 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.

[8] 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*, 2013.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] P. Hoffman, R. W. Jones, and A. Lambon Ralph, Matthew, "Be concrete to be comprehended: Consistent imageability effects in semantic dementia for nouns, verbs, synonyms and associates," *Cortex*, vol. 49, no. 5, pp. 1206–1218, 2013.

[21] D. Crepaldi, C. Ingignoli, R. Verga, A. Contardi, C. Semenza, and C. Luzzatti, "On nouns, verbs, lexemes, and lemmas: Evidence from the spontaneous speech of seven aphasic patients," *Aphasiology*, vol. 25, no. 1, pp. 71–92, 2011.

[22] F. Cuetos, C. Rosci, M. Laiacona, and E. Capitani, "Different variables predict anomia in different subjects: A longitudinal study of two Alzheimer's patients," *Neuropsychologia*, vol. 46, no. 1, pp. 249–260, 2008.

[23] M. A. Lambon Ralph, K. S. Graham, A. W. Ellis, and J. R. Hodges, "Naming in semantic dementia – What matters?" *Neuropsychologia*, vol. 36, no. 8, pp. 775–784, 1998.

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

[25] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1526–1540, 2006.

[26] 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.

[27] 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.

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

[29] 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.

[30] R. D. Kent and Y.-J. Kim, "Toward an acoustic typology of motor speech disorders," *Clinical linguistics & phonetics*, vol. 17, no. 6, pp. 427–445, 2003.

[31] K. Francois, "The comprehensive dragon naturallyspeaking guide," in *Inclusive Learning Technologies Conference*, 2008.

[32] D. Graff and C. Cieri, *English Gigaword Corpus*. Linguistic Data Consortium, 2003.

[33] K. C. Fraser, F. Rudzicz, and E. Rochon, "Using text and acoustic features to diagnose progressive aphasia and its subtypes," in *Proceedings of Interspeech*, 2013.