Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation

Duc Le and Emily Mower Provost

University of Michigan
Computer Science and Engineering, Ann Arbor, MI 48109
{ducle, emilykmp}@umich.edu

Abstract

Patients with aphasia often have impaired speech-language production skills, resulting in tremendous difficulties in tasks that require verbal communication. To facilitate rehabilitation outside of therapy, we are collaborating with the University of Michigan Aphasia Program (UMAP) to develop an automated system capable of providing feedback regarding the patient's verbal output. In this paper we introduce a robust method for extracting rhythm and intonation features from aphasic speech based on template matching. These features, combined with Goodness of Pronunciation (GOP) scores and our previous feature set, help our system achieve human-level performance in classifying the quality of speech produced by patients attending UMAP. The results presented in this work demonstrate the efficacy of our technique and the potential of this system for handling natural speech data recorded in non-ideal conditions as well as the unpredictability in aphasic speech patterns.

Index Terms: aphasia, speech-language disorder, clinical application, speech quality analysis, machine learning

1. Introduction

Individuals who acquire aphasia may have impairments in various language-related areas, including speech production, reading, writing, word finding, and auditory comprehension. Because of these deficits, patients with aphasia often have tremendous communication difficulties and frequently experience social isolation [1]. Individual therapy with Speech-Language Pathologists (SLPs) is the traditional form of treatment and has been shown to be most effective at high frequency and intensity [2]. Outside of therapy, many patients practice using commercial software exclusively designed for aphasia treatment. These programs, which are developed for personal computers and, more recently, mobile devices, have exhibited positive effects on aphasia rehabilitation [3–6]. However, most commercial applications lack the ability to provide users with feedback regarding their verbal output. This hinders the effectiveness of in-home exercises and may lead patients, especially those with severely impaired language comprehension, to unknowingly acquire bad habits. To resolve this shortcoming, we are working with the University of Michigan Aphasia Program (UMAP) to develop a system that can perform automatic quality assessment of a patient's speech and provide guidance during an exercise [7]. The system has the potential to improve the efficacy of in-home practice and reinforce traditional therapy.

We currently focus on the speech quality assessment aspect of the application. This task is challenging due to the subjective nature of human scoring as well as the high variability in severity and types of the patient's impairments. Previous works on pathological speech assessment used word recognition rates of automatic recognizers to estimate a speaker's intelligibility [8, 9]. This method is difficult to apply to our problem because individuals with aphasia often produce grammatically incorrect utterances containing different amounts of jargon, mispronunciations, made-up words, and erratic pauses, making unconstrained speech recognition challenging. Other studies estimated intelligibility by extracting phonemic and phonological features from speech [10, 11]. This approach requires a phonetically diverse set of utterances from a single speaker and therefore has limited applicability to our task which performs utterance-level analysis. In other aphasia-related works, Abad et al. [12] developed an application to recognize phrases spoken by aphasia patients during word naming exercises. They targeted users who have word-finding difficulties but no auditory comprehension or speech production impairment. In contrast, our users may have difficulties in both. Fraser et al. [13] obtained good accuracies in classifying subtypes of primary progressive aphasia by doing feature selection. While their work focused on diagnosis, ours targets rehabilitation.

In this paper we present feature extraction methods that enable our system to perform at least as well as the average human evaluator in classifying the quality of aphasic speech. Specifically, we leverage existing measures of pronunciation, rhythm, and intonation from the language learning literature, motivated by the observation that many aphasia rehabilitation activities resemble language learning exercises. These measures compare the patient's speech against a reference database of healthy speech, thus complementing our existing acoustic features derived only from the patient's data [7]. We introduce a novel reference alignment extraction algorithm to overcome the unpredictability of aphasic speech and facilitate the computation of rhythm and intonation scores. We find that these new features yield significant improvement in classification accuracies when combined with our previous feature set. In addition, experiments using subsets of the features provide insights into the factors that influence human perception. The novelty of this work lies in our system's application to aphasia and a robust algorithm for extracting reference alignments, a prerequisite for computing rhythm and intonation estimates.

2. Data Collection

2.1. Aphasic Speech

Our aphasic speech corpus consists of utterances produced by six patients attending UMAP who do not have cognitive impairment and exhibit varying severity and types of deficits. The corpus is heterogeneous: while some subjects produced mainly pronunciation errors and few prosodic abnormalities, others

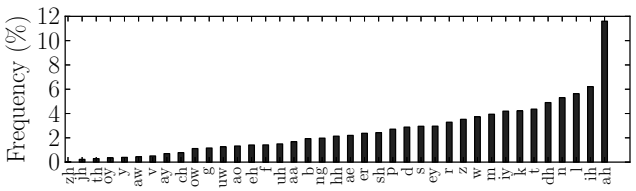

Fig. 1: Phone histogram of the non-aphasic speech corpus.

showed an opposite pattern. Our mobile application, based on Verb Network Strengthening Treatment [14], is the primary tool for data collection. Patients are shown a picture stimulus and asked to describe it verbally either using predefined options or in their own words. The application operates at the sentence level, which is suggested to be more beneficial than word-level exercises for recovering communication skills in highly routine conversational tasks [3, 15]. The dataset contains 1,047 utterances or 2.13 hours of speech. See [7] for more detail about the corpus, the application, and the data collection process.

2.2. Non-aphasic Speech

Our primary goal in this paper is to compare the patients' recorded speech patterns to those present in healthy speech. To enable this comparison, we collected speech recordings from 11 native speakers (5 males and 6 females) who do not have any speech-language impairment. The data is recorded using the built-in microphone on the same type of tablet used for data collection. We do not control the recording environment of this corpus to simulate the condition under which speech data would be obtained in actual application usage. All speakers were asked to take the tablet and find a relatively quiet space to perform the recordings in their own time. Consequently, the recording environment may be varied both across and within speakers. The speech prompts are extracted from our mobile application and have significant repetitions of common words such as "he", "she", and "the", making the dataset phonetically unbalanced (Figure 1). The corpus contains over 8 hours of speech, 13,767 utterances, and 67,889 instances of 521 unique words.

3. Data Annotation

3.1. Transcriptions

We obtain two types of transcripts for each utterance in the aphasic speech corpus: *Free-form* and *Context-based*. For *Free-form*, annotators are asked to transcribe the utterances exactly, including silences, fillers, and speech regions that cannot be reliably decoded. For *Context-based*, we ask annotators to perform the same task while using the speech prompts shown to patients at recording time to transcribe unintelligible regions.

Both types of transcripts can be converted to a high-level representation by clustering annotated labels into four broad categories: (1) *Non-Speech* refers to sounds not produced by patients such as silence and background noise, (2) *Filler* consists of filled pauses such as "um" and "eh", (3) *Vague-Speech* contains patients' speech activities that are unintelligible, and (4) *Clear-Speech* indicates speech segments that can be easily understood. This kind of labeling is motivated by the fact that exact transcription of aphasic speech can be challenging due to the patient's impairment, thus we want annotation labels that can be extracted with high consistency. On a common set of 60 sentences, 10 from each patient, the mean Cohen's kappa score for our four annotators with respect to these high-level labels is 0.92 for *Free-form* and 0.94 for *Context-based*.

We find that more speech can be decoded and the resulting transcripts are closer to the original prompts given the prompts as context. Moving from *Free-form* to *Context-based*, the percentage of speech regions that can be understood increases from 84.13% to 98.24%. In addition, the average substitution error with respect to the prompts decreases from 1.42 to 0.75 words per utterance. It should be noted that our current reliance on these manual transcripts is only temporary. In future work we will explore region classification and, motivated by the effect of prompt knowledge on human annotation, constrained context-based speech recognition to automate their creation.

3.2. Qualitative Scores

As described in [7], we collaborated with the SLPs in UMAP to create four criteria to assess the patient's quality of speech: *Clarity*, *Fluidity*, *Effort*, and *Prosody*. Four human annotators, all of whom undergraduate students who received communication training from the UMAP staff, labeled these criteria on a scale of 1 to 2 for *Prosody* and 1 to 4 for the other three, where a higher score denotes better quality and vice versa. We limit *Prosody* to 2 classes because our preliminary study indicated that the subjective perception of this criterion varies more than the others, thus increasing the number of scores would result in chance level agreement. *Clarity*, *Fluidity*, and *Effort* are further clustered to create *2-* and *3-class* problems, under the hypothesis that the original *4-class* scheme might not be optimal for classification. Following [8, 16], the average scores across all evaluators are rounded and used as ground-truths.

4. Method

4.1. Acoustic Modeling

We train a standard speaker-independent cross-word triphone acoustic model with two Gaussian mixtures on our non-aphasic speech corpus using data from 10 speakers. The number of mixtures is determined by the word error rate (WER) on the held-out data of the 11th speaker (5.85%). Forced alignment via this acoustic model is the foundation on which our new features are extracted. Previous works have demonstrated that measures derived from forced alignment output correlate well with the degree of intelligibility in pathological speech [10, 17], thus motivating our approach. We will explore word-level modeling in future research, a technique which has been shown to be beneficial for small-vocabulary dysarthric speech recognition [18], given that our application has a relatively limited vocabulary.

4.2. Feature Extraction

4.2.1. Transcript and Acoustic Features

Our previous features were made up of two sets: transcript- and acoustic-based [7]. Features in the former set were extracted for each utterance from its high-level *Free-form* transcript described in Section 3.1. They contain the duration of *Non-Speech*, *Filler*, *Vague-Speech*, and *Clear-Speech*, total duration, voiced duration, speech duration, start time of first speech activity, *Clear-Speech* rate, long (> 0.4s) and short (> 0.15s, ≤ 0.4s) pause count [19], along with phonation rate and mean pause duration [20]. Acoustic features in the latter set are extracted from each voiced region over the boundary defined in the transcript. They consist of the mean and variance of intensity, jitter [21], mean and variance of fundamental frequency (F0) [19], mean and variance of the first three formants (F1, F2, F3), mean instantaneous power, mean and maximum first autocorrelation function, skewness, kurtosis, zero-crossing rate, and shimmer. Our previous results indicated that transcript features have more

impact on classification accuracies, echoing the finding in [13].

4.2.2. Pronunciation Scores

Our pronunciation scores are based on Goodness of Pronunciation (GOP), a commonly used metric first introduced by Witt and Young [22]. The idea behind GOP is to calculate the difference between the average acoustic log-likelihood of a force-aligned phoneme and that of an unconstrained phone loop. If this number is close to 0, the pronunciation of this phone is more likely to be correct and vice versa. Originally defined to compute the pronunciation score of a single phoneme, GOP can be modified to accommodate an arbitrary phone sequence:

$$GOP(\mathbf{p}) = \frac{1}{N} \log \frac{P(O|\mathbf{p})}{P(O|PL)} \quad (1)$$

where $\mathbf{p}$ is the sequence of phones, O is the acoustic, N is the number of frames, and PL is the unconstrained phone loop. To obtain GOP for a word, we force align its speech over all possible pronunciations of that word to find the best phone sequence $\mathbf{p}$, which will allow us to compute $P(O|\mathbf{p})$.

We extract GOP scores for all words in an utterance by force aligning the speech to its *Context-based* transcript. We then weigh the GOP scores by word durations, hypothesizing that longer words have more impact on the perception of the entire sentence. Finally, we extract the mean, standard deviation, median, minimum, and maximum word-level GOP scores to use as features for the utterance. We also extract a similar set of features at the phone level, under the assumption that they will provide complementary information.

4.2.3. Extracting Reference Alignments

A prerequisite for computing rhythm and intonation scores in this work is the ability to, given the transcript of an aphasic speech utterance, extract corresponding alignment profiles from a reference database of non-aphasic speech. For example, suppose the patient says "The people clapped", we would want to find the same sentence spoken by someone without aphasia and analyze how the two utterances' durations and pitch contours differ. This will allow us to compute rhythm and intonation scores, respectively. However, extracting an identical sentence is impractical for our application because of two reasons. Firstly, it is not possible to anticipate all the sentences and words that individuals with aphasia will produce. When interacting with our mobile application, the patients' speech-language deficits often caused them to verbalize the sentences differently from what was asked. Only 16.63% of the utterances in our aphasic speech corpus completely match their given prompts, and 13.33% of the words are out-of-vocabulary. Secondly, as the application grows and new sentences are added, it is impractical to maintain a matching reference database.

We hypothesize that the characteristics of an acoustic unit (word, syllable, or phone) are influenced by its immediate neighbors, motivated by [11] and the ability of triphones to capture coarticulation. Based on this idea, we developed an algorithm to search for a reference alignment of any target utterance by gradually increasing the level of granularity until a match is found. The utterance is first broken into triwords, defined as the words with their left and right neighbors. The algorithm finds occurrences of each triword in the reference database which match, in decreasing preference, both left and right contexts, only left or right context, or no context. If the word cannot be found, the triword is broken into syllables according to a pronunciation dictionary and the search continues for each trisyllable. Similarly, if the syllable is not found, it is broken into phones and the search continues for each triphone. The process is guaranteed to succeed if the reference database contains instances of all phones. Table 1 shows a sample alignment for the target sentence "The people clapped". Since "people" is out-of-vocabulary, the search breaks it down to two syllables "p iy" and "p ah l". The second syllable is also missing, so the search breaks it down further into individual phones. Our algorithm currently assumes that a match without context at a higher level (word, syllable) is better than a match with context at a lower level (syllable, phone), which we will investigate in future work.

Target	Level	Reference	Context	Instances
the	WORD	the	L	6,436
people	SYL.	p iy	L	65
-	PHONE	p	L + R	12
-	PHONE	ah	L + R	22
-	PHONE	l	L + R	139
clapped	WORD	clapped	R	20

Table 1: A reference alignment extracted for the target sentence "The people clapped". Because "people" is out-of-vocabulary, the search must descend into the syllable and phone level.

Each unit in the reference alignment is augmented with its duration and pitch contour to facilitate the computation of rhythm and intonation scores. Details on how to adapt existing measures of rhythm and intonation to aphasic speech through this alignment process are covered in the next two sections.

4.2.4. Rhythm Scores

Earlier work on rhythmic analysis proposed features computed from the target speech such as %V (average proportion of vocalic intervals), ΔC and ΔV (average standard deviations of consonantal and vocalic intervals) [23], and normalized Pairwise Variability Index (PVI) [24]. The efficacy of these metrics has been demonstrated, but they are less suitable for our tasks because of two reasons. Firstly, the above features are typically computed at the speaker level and may not be stable enough for short utterances that contain considerably less data. Secondly, speech patterns of patients with aphasia have very high variability, thus computing statistics on their speech alone might not be conducive to generalization. More recently, Tepperman et al. introduced Pairwise Variability Error (PVE), a metric that directly compares two speakers' rhythms [25]. Given duration profiles of a target and reference utterance, denoted as $\{t_1, t_2, ..., t_N\}$ and $\{r_1, r_2, ..., r_N\}$ respectively, where each element is the duration of an acoustic unit (word, syllable, or phone), PVE computes the difference of these two profiles:

$$PVE = \frac{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |(t_i - t_{i-m}) - (r_i - r_{i-m})|}{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |t_i - t_{i-m}| + |r_i - r_{i-m}|} \quad (2)$$

where M is a hyperparameter specifying the maximum distance between a pair of units that can be considered for comparison.

Finding references was not a problem in their work because the target non-native learners would always attempt to reproduce the given prompts. However, the issue of prompt and speech mismatch discussed above makes this a challenge for us, which can be overcome with the reference alignment algorithm. The target duration profile is first obtained by force-aligning the speech to its *Context-based* transcript. The reference duration profile can then be constructed by querying the non-aphasic speech database for an alignment and computing the average duration of instances in the same unit. We do not perform linear scaling on the reference durations as in [25] to

retain information about speaking rates and to avoid durational distortion caused by long pauses in aphasic speech. For each utterance we compute four PVE scores with M ranging from 1 to 4 (same as [25]), constituting the utterance's rhythm features.

4.2.5. Intonation Scores

Previous studies suggested that pitch contours in patients with aphasia may exhibit anomalies in sentence-length utterances [26, 27]. We are therefore interested in systematically comparing the contours of aphasic speech to those of healthy speech. This comparison is driven by Dynamic Time Warping (DTW), a method previously used to measure the similarity of pitch contours with differing lengths [28]. To extract intonation features for a target utterance, we first obtain its reference alignment using the algorithm presented in Section 4.2.3. As a result, every target unit (word, syllable, or phone) is matched with a set of reference units, each of which contains a pitch contour. For each unit, we compute the mean and relative standard deviation of the DTW distances between the target and all reference contours. Prior to computation, the reference contours are shifted to have the same mean as the target; this accounts for pitch differences across speakers. We then weigh the mean DTW distance of each unit by its duration and inverse of the aforementioned relative standard deviation, under the hypothesis that longer units have more impact on human perception and those with high variance should not be weighed as heavily. Finally, we extract the mean, standard deviation, median, minimum, and maximum unit-level distances to use as intonation features for the utterance.

4.3. Classification

For comparison, we follow the same setup in our previous work [7]. We perform leave-one-subject-out cross-validation on the aphasic speech corpus, applying minimum-redundancy-maximum-relevance (mRMR) feature selection [29] on each fold prior to training. We evaluate every fold using five commonly-used classifiers: C4.5 Decision Tree, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine. We use the default settings in the WEKA toolkit for all algorithms [30]. Results of the best classifiers will be reported.

5. Results and Discussion

As the scores are unbalanced, we evaluate a classifier's performance by its unweighted average recall (UAR), the mean per-class accuracy. We are interested in how the new system compares to both our previous work and the average human performance, defined as the mean UAR of our four annotators with respect to the ground-truths [7]. Statistical significance is determined by binomial tests. We treat each utterance as an independent random experiment; the average human or the previous system's UAR denotes the hypothetical chance of success, while the new UAR estimates the observed number of successes. For a two-tailed binomial test, the prediction results are deemed not significantly different if $p > 0.05$. Otherwise, we say one of the results is significantly better than the other since the binomial distribution is symmetric (one-tailed test, $p = 0.025$). Table 2 summarizes the results of our new system. Compared to the average human performance, the system performs significantly better in 2-class *Clarity*, 3-class *Fluidity*, *Prosody*, and equally well in the 7 remaining tasks. Compared to our previous work using only acoustic and transcript features, the system performs significantly better in 7 out of 10 categories. The performance gain is more significant in problems with more than 2 classes, suggesting that our new features are more useful for finer-grained classification.

	2-class	3-class	4-class
Clarity	75.8^{*} (RF)	$62.7_{\natural}$ (NB)	$53.8_{\natural}$ (NB)
Fluidity	$73.0_{\natural}$ (NB)	$70.2^{*}_{\natural}$ (NB)	$52.1_{\natural}$ (NB)
Effort	78.9 (LR)	$66.4_{\natural}$ (NB)	$53.8_{\natural}$ (NB)
Prosody	66.3^{*} (RF)	N/A	

$*$ = sig. better than human | $\natural$ = sig. better than previous
RF: Random Forest | NB: Naive Bayes | LR: Log. Regression

Table 2: UAR (%) of the best classifier for each category. Numbers without significance markers are not significantly different from the average human or the previous system's performance.

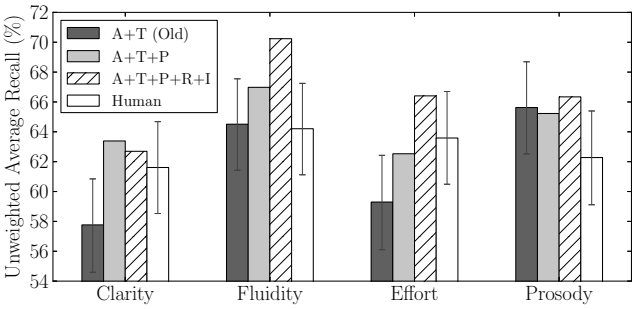

Fig. 2: Effect of **P**ronunciation and **R**hythm+**I**ntonation on classification accuracies when added to the old **A**coustic+**T**ext feature set. Results on 3-class problems and *Prosody* are shown. The vertical bar indicates the region within which the number is deemed not significantly different from the mean.

We also investigate the effect of individual feature sets on classification and human perception. For 3-class problems, adding pronunciation scores to the existing acoustic and transcript features results in improvements across *Clarity*, *Fluidity*, and *Effort*, with the first category receiving the most gain (Figure 2). Adding rhythm and intonation scores further improves *Fluidity* and *Effort* but not *Clarity* and *Prosody*. The performance gain is more pronounced when considering these two sets of features jointly compared to their individual impacts. We observe a similar trend in 2- and 4-class problems (graphs not shown). These observations suggest that our targeted speech qualities are not perceived in isolation. Rather, changes in the perception of one category will also influence that of others.

6. Conclusion and Future Work

In this work we present our techniques for extracting pronunciation, rhythm, and intonation features from aphasic speech, along with an algorithm to robustly extract reference alignments for a target utterance. Our experimental results indicate that it is possible for an automated system to achieve human-level accuracies in classifying the quality of speech produced by patients with aphasia. We show that pronunciation features influence the perception of not only *Clarity* but also *Fluidity* and *Effort*, while rhythm and intonation features mainly affect the latter two.

In future work we will investigate methods to lift the dependence on human-labeled transcripts through automatic segment categorization and constrained context-based speech recognition. Our rhythm features can be further improved through envelope-based analysis which incorporates different dimensions than our current interval-based approach [31]. Finally, modeling phonological features appears promising not only for classification but also for providing automatic feedback [10,11].

7. References

[1] L. R. Cherney, A. S. Halper, A. L. Holland, and R. Cole, "Computerized Script Training for Aphasia: Preliminary Results," *American Journal of Speech-Language Pathology (AJSLP)*, vol. 17, no. 1, pp. 19–34, Feb 2008.

[2] S. Bhogal, R. Teasell, M. Speechley, and M. L. Albert, "Intensity of Aphasia Therapy, Impact on Recovery," *Stroke*, vol. 34, no. 4, pp. 987–993, Apr. 2003.

[3] L. M. Manheim, A. S. Halper, and L. Cherney, "Patient-Reported Changes in Communication After Computer-Based Script Training for Aphasia," *Archives of Physical Medicine and Rehabilitation*, vol. 90, no. 4, pp. 623–627, Apr 2009.

[4] R. C. Katz, "Computers in the treatment of chronic aphasia," *Seminars in Speech and Language*, vol. 31, no. 1, pp. 34–41, Feb 2010.

[5] R. Nobis-Bosch, L. Springer, I. Radermacher, and W. Huber, "Supervised Home Training of Dialogue Skills in Chronic Aphasia: A Randomized Parallel Group Study," *Journal of Speech, Language, and Hearing Research (JSLHR)*, Dec. 2010.

[6] L. Allen, S. Mehta, J. A. McClure, and R. Teasell, "Therapeutic interventions for aphasia initiated more than six months post stroke: A review of the evidence," *Topics in Stroke Rehabilitation*, vol. 19, no. 6, pp. 523–535, 2012.

[7] D. Le, K. Licata, E. Mercado, C. Persad, and E. Mower Provost, "Automatic Analysis of Speech Quality for Aphasia Treatment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[8] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic Scoring of the Intelligibility in Patients with Cancer of the Oral Cavity," in *Proc. of the 8th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Antwerp, Belgium, 2007, pp. 1206–1209.

[9] K. Riedhammer, G. Stemmer, T. Haderlein, M. Schuster, F. Rosanowski, E. Noth, and A. Maier, "Towards Robust Automatic Evaluation of Pathologic Telephone Speech," in *Automatic Speech Recognition and Understanding (ASRU)*, Kyoto, Japan, Dec 2007, pp. 717–722.

[10] G. Van Nuffelen, C. Middag, M. De Bodt, and J. P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International Journal of Language and Communication Disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[11] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated Intelligibility Assessment of Pathological Speech Using Phonological Features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 3:1–3:9, Jan. 2009.

[12] A. Abad, A. Pompili, A. Costa, and I. Trancoso, "Automatic word naming recognition for treatment and assessment of aphasia," in *Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[13] K. Fraser, F. Rudzicz, and E. Rochon, "Using text and acoustic features to diagnose progressive aphasia and its subtypes," in *Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2013.

[14] L. A. Edmonds, S. E. Nadeau, and S. Kiran, "Effect of Verb Network Strengthening Treatment (VNeST) on Lexical Retrieval of Content Words in Sentences in Persons with Aphasia," *Aphasiology*, vol. 23, no. 3, pp. 402–424, Mar 2009.

[15] L. R. Cherney, A. S. Halper, and R. C. Kaye, "Computer-based script training for aphasia: Emerging themes from post-treatment interviews," *Journal of Communication Disorders*, vol. 44, no. 4, pp. 493–501, 2011.

[16] M. P. Black, J. Tepperman, and S. S. Narayanan, "Automatic Prediction of Children's Reading Ability for High-Level Literacy Assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1015–1028, May 2011.

[17] J. Carmichael and P. Green, "Revisiting Dysarthria Assessment Intelligibility Metrics," in *Proc. of the 8th International Conference on Spoken Language Processing (INTERSPEECH)*, Jeju Island, Korea, 2004.

[18] H. V. Sharma, M. Hasegawa-Johnson, J. Gunderson, and A. Perlman, "Universal Access: Preliminary Experiments in Dysarthric Speech Recognition," in *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, United Kingdom, 2009.

[19] 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, no. 3, pp. 165–177, Sep 2010.

[20] B. Roark, M. Mitchell, J. 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.

[21] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter Estimation Algorithms for Detection of Pathological Voices," *EURASIP Journal on Advances in Signal Processing*, pp. 9:1–9:9, Jan. 2009.

[22] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 23, pp. 95 – 108, 2000.

[23] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 75, no. 1, pp. AD3 – AD30, 2000.

[24] E. Grabe and E. L. Low, "Durational Variability in Speech and the Rhythm Class Hypothesis," *Laboratory Phonology VII*, pp. 515–546, 2002.

[25] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing Suprasegmental English Through Parroting," in *Proc. of Speech Prosody*, Chicago, IL, USA, 2010.

[26] M. Danly and B. Shapiro, "Speech prosody in Broca's aphasia," *Brain and Language*, vol. 16, no. 2, pp. 171 – 190, 1982.

[27] J. Gandour, S. H. Petty, and R. Dardarananda, "Dysprosody in Broca's aphasia: A case study," *Brain and Language*, vol. 37, no. 2, pp. 232 – 257, 1989.

[28] A. Rilliard, A. Allauzen, and P. B. de Mareil, "Using Dynamic Time Warping to Compute Prosodic Similarity Measures," in *Proc. of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy, 2011, pp. 2021–2024.

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

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

[31] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages," *Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, Jul 2013.