AUTOMATIC ANALYSIS OF SPEECH QUALITY FOR APHASIA TREATMENT

Duc Le[†]*, Keli Licata*[‡]*, Elizabeth Mercado*[‡]*, Carol Persad*[‡]*, and Emily Mower Provost*[†]

University of Michigan, Ann Arbor, MI 48109, USA
[†]Computer Science and Engineering, [‡]University Center for Development of Language and Literacy
{ducle, klicata, eolszews, cpersad, emilykmp}@umich.edu

ABSTRACT

Aphasia is a common language disorder which can severely affect an individual's ability to communicate with others. Aphasia rehabilitation requires intensive practice accompanied by appropriate feedback, the latter of which is difficult to satisfy outside of therapy. In this paper we take a first step towards developing an intelligent system capable of providing feedback to patients with aphasia through the automation of two typical therapeutic exercises, sentence building and picture description. We describe the natural speech corpus collected from our interaction with clients in the University of Michigan Aphasia Program (UMAP). We develop classifiers to automatically estimate speech quality based on human perceptual judgment. Our automatic prediction yields accuracies comparable to the average human evaluator. Our feature selection process gives insights into the factors that influence human evaluation. The results presented in this work provide support for the feasibility of this type of system.

***Index Terms*—** aphasia, speech-language disorder, machine learning, clinical application

1. INTRODUCTION

In the US, there are approximately one million people with aphasia and more than 100,000 acquire it every year due to brain injury, most commonly from a stroke [1]. Individuals with aphasia exhibit high variability in their specific impairments. Those with non-fluent aphasia produce slow, halting, and effortful speech. In contrast, those with fluent aphasia can speak effortlessly but their sentences contain jargon and are often void of meaning. Some may have problems with word-finding (anomia) or motor speech production (apraxia). In general, most aphasia patients have speech production impairments and some form of language comprehension deficit, making social interaction difficult and frustrating.

Traditional treatment for aphasia involves individual therapy with trained Speech-Language Pathologists (SLPs). Individual therapy has been shown to be most effective in helping patients regain language skills when carried out at high frequency and intensity [1]. However, not all patients with aphasia can achieve the optimal practice frequency and intensity through therapy alone due to financial limitations and scheduling constraints. In addition to therapy, patients often practice on their own using commercial software programs specifically designed for aphasia treatment. While the effect of computer use in aphasia rehabilitation is generally positive [2–4], most programs do not provide meaningful feedback to patients regarding their verbal output during the course of an exercise. For example, some programs let patients play back their own speech but do not provide any qualitative analysis of said speech. This lack of feedback may cause patients to develop bad habits over time, which will be further exacerbated by infrequent interaction with the SLPs.

We aim to address this issue by developing an intelligent system capable of providing automatic feedback to patients about their verbal output during practice, thus improving the effectiveness of in-home exercises to support traditional therapy as needed. As a first step, we partnered with the University of Michigan Aphasia Program (UMAP) to develop a mobile application that includes the therapeutic exercises of sentence building and picture description. We collected over two hours of aphasic speech from six UMAP clients while they interacted with our application. After data collection, we extracted speech features for each utterance and trained automatic classifiers to estimate the quality of speech based on scores assigned by human evaluators. The automatic prediction achieved comparable accuracies with human scoring. In addition, feature selection results give insights into the factors that influenced human evaluation. The novelty of this work lies in the development of new assistive technology for aphasia rehabilitation, which includes data recording, a set of criteria for assessing aphasic speech quality, and the analysis of how automatic evaluation differs from that of humans.

2. RELATED WORK

Processing aphasic speech for therapeutic and diagnostic purposes has been the subject of several previous works. Abad et al. [5] used keyword spotting to recognize phrases spoken by aphasia patients during word naming exercises. Their work differs from ours in two ways. Firstly, their targeted users are individuals with aphasia who have anomia but no difficulties

[1] http://aphasia.org/?q=content/aphasia-faq

with auditory comprehension or speech-language production. In contrast, our targeted users may have difficulties in both. Secondly, their work aims to recognize spoken words, while ours attempts to estimate speech quality. Other works focused on the medical diagnosis for subtypes of aphasia and related disorders [6–8], unlike ours which targets rehabilitation.

This work adopts techniques used in [8–10]. Black et al. [9] extracted speech features to predict children's high-level reading ability. Wang et al. [10] used transcript-based features to predict the severity of mispronunciation. Both works utilized fine-grained information about phonetics and common letter-to-sound mistakes. This technique is challenging to apply to our target population whose pronunciation errors are not systematic. Fraser et al. [8] used feature selection to classify two subtypes of primary progressive aphasia. Their text features capture the complexity of patients' narratives, which is not applicable to our work because exact transcriptions are difficult to obtain and the utterances are simple in structure.

3. DATA

3.1. Mobile Application

Our mobile application is the primary tool for data collection. Designed to run on Android tablet devices, it features exercises modeled after picture description, an activity commonly administered by SLPs in UMAP during individual therapy. The activity helps patients practice expressive communication skills, such as word-finding, sentence construction, use of appropriate verb tenses, and articulation of target words. In the application, patients are presented with a picture stimulus and asked to describe it verbally either using predefined options or in their own words. Figure 1 shows a sample exercise with predefined options. We allow users to adjust the difficulty level for every exercise. We also utilize text-to-speech to provide auditory feedback in addition to visual and textual information as our users may have difficulties with reading and/or auditory comprehension.

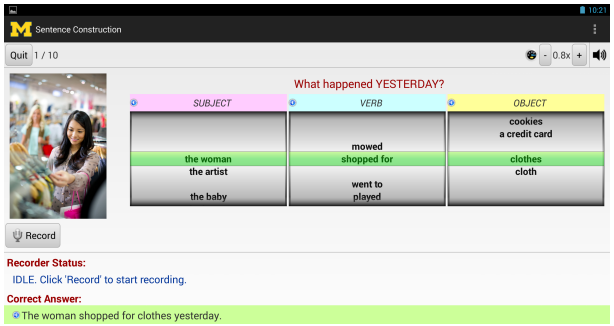

Fig. 1. Screenshot of the Sentence Construction exercise.

3.2. Data Collection

For this pilot study we recruited six individuals attending UMAP who have aphasia and do not have cognitive impairment. UMAP offers an intensive therapy program which, for full-time clients, typically involves 40 hours of speech-language therapy a week for four weeks. A team of research staff consisting of four undergraduate students sat individually with the patients during lunchtime and provided support, as needed, while they completed the exercise on the mobile application. The research team received training from UMAP staff regarding how to assist individuals with aphasia. Our goal was to collect natural speech recordings that best resemble the type of data the application would have seen had the patients used it on their own. We adjusted the difficulty based on recommendations from the SLPs and used the tablet's built-in microphone for all recordings. Table 1 lists the age and gender information, diagnosis, amount of recorded data, and Aphasia Quotient (AQ) for each subject in our dataset. The Aphasia Quotient [11] is a commonly used measure for the severity of aphasia. In total the dataset contains over two hours of aphasic speech. Collecting this kind of data is particularly challenging because individuals with aphasia generally require more time and effort to produce a sentence.

Age, Gender	Diagnosis	Data	AQ
50, Female	expressive	122 utterances (16.92 mins)	91.6 (mild)
70, Male	fluent, anomia	375 utterances (41.17 mins)	75.7 (mild)
49, Male	non-fluent, anomia	146 utterances (23.82 mins)	65.6 (moderate)
49, Male	fluent, apraxia	112 utterances (13.79 mins)	49.9 (severe)
68, Male	expressive, receptive	81 utterances (10.15 mins)	43.9 (severe)
50, Male	non-fluent, apraxia	211 utterances (21.75 mins)	42.6 (severe)

Table 1. Subject breakdown of the dataset.

3.3. Data Annotation

3.3.1. Transcription

Each utterance was transcribed by one of the four members of the research staff into time segments belonging to four broad categories: (1) *Non-Speech* consists of sounds not spoken by the patient such as silence and audible background noise, (2) *Filler* consists of filled pauses such as "um" and "eh", (3) *Vague-Speech* consists of patient's speech activities which are unclear, and (4) *Clear-Speech* consists of speech segments that can be easily understood. Exact transcription of aphasic speech is challenging due to the patient's speech-language impairment. We are interested in annotation labels that can be extracted with relative ease and result in high agreement across annotators to enable autonomous tablet-based interaction. In this work we are only concerned with which category a sound segment belongs to, not its exact speech content. We

estimated the reliability of annotation by having all evaluators transcribe a common set of 60 sentences, 10 from each patient. The mean Cohen's kappa score with respect to the category labels is 0.92, a very high agreement level.

3.3.2. Qualitative Scores

In order to provide feedback, the system must be able to estimate the quality of speech produced by patients. With guidance from the SLPs, we created four criteria for evaluating a patient's speech: *Clarity*, *Fluidity*, *Effort*, and *Prosody*. We asked each of the four evaluators to rate every utterance on a scale from 1 to 2 for *Prosody* and 1 to 4 for the other three criteria. A lower score denotes lower quality and vice versa. A score of 0 may be assigned to utterances without enough speech activity for analysis. We clustered the scores for *Clarity*, *Fluidity*, and *Effort* in two additional ways, under the hypothesis that the *4-class* system might not be optimal for automatic classification. In the *2-class* scheme, $\{1,2\}$ and $\{3,4\}$ are collapsed into two groups, thus categorizing an utterance as low- or high-quality. In the *3-class* scheme, $\{2,3\}$ becomes one category while 1 and 4 remain the same, thus separating low- and high-quality utterances from ambiguous ones.

Following the work of Black et al. [9], we constructed a "de-noised" set of ground-truth labels by averaging the scores across each evaluator and rounding to the closest integer. These ground-truth scores represent the collective opinions and are used to train our automatic classifiers. Utterances with an average score of 0 are not used because they indicate insufficient speech activity for analysis. As the scores are unbalanced, we evaluate a classifier's performance based on its unweighted average recall (UAR), defined as the mean per-class accuracy. The goal of the classifiers is to achieve the same level of performance as that of individual evaluators, shown in Table 2. This accounts for the challenges associated with the perceptual judgment of speech quality.

	2-class	**3-class**	**4-class**
Clarity	71.0 ± 5.3	61.6 ± 9.0	51.2 ± 8.2
Fluidity	75.2 ± 3.7	64.2 ± 4.8	54.6 ± 3.5
Effort	76.7 ± 5.3	63.6 ± 4.5	55.5 ± 4.0
Prosody	62.3 ± 7.5	N/A	

Table 2. Mean and standard deviation of unweighted average recall (%) of human evaluators. The average scores are used as ground truths and each evaluator is viewed as a classifier.

4. METHOD

4.1. Feature Extraction

4.1.1. Transcript Features

We hypothesize that the transcripts encode information about the four aspects of speech quality that we seek to model. For instance, we expect a high number of pauses to represent disconnected speech and thus correspond to low *Fluidity* scores. For each utterance we extracted the following features from its transcript: duration of *Non-Speech*, *Filler*, *Vague-Speech*, and *Clear-Speech*, total duration, voiced duration, speech duration, start time of first speech activity, and fraction of *Clear-Speech* over speech duration. We also extracted long pause ($>$ 0.4s) and short pause ($>$ 0.15s, $\leq$ 0.4s) count [7], along with phonation rate and mean pause duration [12].

4.1.2. Acoustic Features

Acoustic features contain lower-level information about the sound properties. For each voiced segment in an utterance's transcript, we extracted the mean and variance of intensity, jitter [13], mean and variance of fundamental frequency (F0) [7], 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 [8]. The segments are weighed by duration and their weighted average yields the features for the entire utterance. We expect some features to correlate directly with our targeted speech qualities. For example, high jitter is associated with a harsher voice [14] and may equate to low *Effort* scores.

4.2. Classification

We partitioned the dataset using leave-one-subject-out cross-validation, motivated by the assumption that our application must be able to generalize beyond individual speakers. To avoid overfitting, we performed feature selection on the training set of each fold using the minimum-redundancy-maximum-relevance (mRMR) method, which outputs the subset of features that correlate well with the class label but not with each other [15]. mRMR was also used in Fraser et al. [8], yielding good results in classifying subtypes of primary progressive aphasia. We evaluated each fold using several commonly-used classifiers, including C4.5 Decision Tree, Logistic Regression, Naive Bayes, Random Forest, and Support Vector Machine. Since our dataset is relatively small, we did not do model selection and instead used the default settings specified in the Weka toolkit [16]. As our dataset grows, we will explore model selection and speaker-dependent adaptation to improve classification performance.

5. RESULTS AND DISCUSSION

5.1. Selected Features

We ran the mRMR feature selection algorithm on the complete dataset to identify the most representative features for each score category. Table 3 lists the selected features across all scoring schemes (*2*, *3*, and *4-class*). The algorithm selected mostly transcript features in all categories. This trend, which was also observed in [8], suggests that human evaluators rely more on the high-level features captured by the transcripts. Only two or fewer acoustic features were selected for *Clarity*, *Fluidity*, and *Effort*, but five were chosen for *Prosody*.

Clarity	fillerDuration, clearSpeechDuration, clearSpeechRate, phonationRate
Fluidity	totalDuration, longPauseCount, clearSpeechDuration, clearSpeechRate, phonationRate, vagueSpeechDuration, nonSpeechDuration, *meanIntensity*
Effort	fillerDuration, totalDuration, longPauseCount, clearSpeechDuration, vagueSpeechDuration, speechDuration, nonSpeechDuration, voicedDuration, *meanIntensity*, *zeroCrossingRate*
Prosody	fillerDuration, speechDuration, clearSpeechDuration, nonSpeechDuration, clearSpeechRate, voicedDuration, phonationRate, *meanF1*, *skewness*, *meanF0*, *stdDevF0*, *meanIntensity*

Table 3. Features selected by mRMR across all grouping schemes (*2*, *3*, and *4-class*). *Italic* denotes acoustic features.

This suggests that our high-level transcript features did not capture *Prosody* as strongly, forcing human evaluators to fall back on the low-level acoustic features.

The specific features selected provide insights into how human evaluators assessed quality of speech. The duration of *Clear-Speech*, which denotes the amount of speech perceived as easily understood, is present in all four categories. Long pause count is present in *Fluidity* and *Effort*, suggesting that connected speech tends to be perceived as more fluid and effortless. The mean and standard deviation of the fundamental frequency (F0) only appear in *Prosody*, hinting a connection to intonation, i.e. pitch variation. These results can be potentially communicated back to patients in real-time on an utterance basis, thus guiding their practice sessions. In future work we will explore ways to realize this idea.

5.2. Classification Results

Our classifiers' performance is assessed in relation to that of the average human evaluator. We use binomial tests to determine if the automatic classification is significantly different from human scoring. Specifically, we treat each utterance as an independent random experiment; the average human UAR denotes the hypothetical chance of success for each trial and the classifier's UAR is used to estimate the observed number of successes. For a two-tailed binomial test, if $p > 0.05$, the prediction result is deemed not significantly different from human scoring. Otherwise, because the binomial distribution is symmetric, we can conclude that the automatic prediction is significantly better (or worse) than human (one-tailed test, $p = 0.025$). Table 4 lists the UAR of the best classifier for each category along with its relation to the average evaluator.

The classifiers are better at classifying *Clarity* and *Prosody* than *Fluidity* and *Effort*. This suggests that our feature set does not capture as much information about the latter two properties. The results also show that automatic classification achieves better performance when there are fewer classes. We hypothesize that this improvement is caused by the need to make fewer fine distinctions as well as the increase in per-class training data, a result of collapsing scores into fewer categories. This hypothesis is supported by the fact that Naive Bayes, which has been shown to work well on smaller datasets [17], is the predominant method of choice in *3-class* and *4-class*. In contrast, the best *2-class* results are all achieved by Random Forest, an ensemble classifier that aggregates results from multiple decision trees. Interestingly, the way Random Forest obtains its final prediction roughly corresponds to how our ground-truths were computed from human evaluators. Moreover, Random Forest performs better when there is higher variance in human UARs. This is consistent with the results in Breiman [18], which showed that the algorithm works best when the trees are uncorrelated.

	2-class	**3-class**	**4-class**
Clarity (size: 991)	73.7* (RF)	57.8$_\dagger$ (NB)	50.0* (NB)
Fluidity (size: 980)	69.2$_\dagger$ (RF)	64.5* (NB)	48.3$_\dagger$ (RF)
Effort (size: 982)	77.9* (RF)	59.3$_\dagger$ (NB)	50.2$_\dagger$ (NB)
Prosody (size: 971)	65.6$_\natural$ (RF)	N/A	

$*$ = not sig. different $\natural$ = sig. higher $\dagger$ = sig. lower
RF = Random Forest NB = Naive Bayes

Table 4. UAR (%) of the best classifier for each category and how it compares to the average human evaluator.

6. CONCLUSION AND FUTURE WORK

Our results indicate that it is possible to construct an automatic classifier comparable to the average human for estimating each of these four aspects of speech quality. Feature selection suggests that humans rely more on high-level transcript features during evaluation, and that acoustic features capture *Prosody* more strongly than *Clarity*, *Fluidity*, and *Effort*.

In future work we will lift the dependency on manually labeled transcripts through the automatic categorization of time segments. As our dataset grows, we will explore model selection and speaker-dependent adaptation to improve classification performance. Lastly, we will investigate ways to provide patients with concrete feedback based on the output of feature selection and automatic classifiers.

7. ACKNOWLEDGMENTS

We would like to thank Patrick Shin, Yoolim Jung, Kelly Karpus, Carly Swiftney, and the UMAP staff for their assistance in application development, data collection, and annotation.

8. REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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, ISCA.

[6] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. Gorno Tempini, and J. Ogar, "Learning diagnostic models using speech and language measures," in *Proc of the 30th Annual International IEEE EMBS Conference*, Vancouver, British Columbia, Canada, 2008.

[7] 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.

[8] 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.

[9] 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.

[10] H. Wang, Z. Wu, S. Zhang, and H. Meng, "Predicting gradation of l2 english mispronunciations using crowdsourced ratings and phonological rules," in *Proc. of Speech and Language Technology in Education (SLaTE)*, Grenoble, France, 2013.

[11] A. Kertesz and E. Poole, "The aphasia quotient: the taxonomic approach to measurement of aphasic disability," *Canadian Journal of Neurological Sciences*, vol. 1, no. 1, pp. 7–16, Feb 1974.

[12] 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.

[13] 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. 9:1–9:9, Jan. 2009.

[14] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, no. 2, 1990.

[15] 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.

[16] 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.

[17] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Proc. of the 15th Annual Conference on Neural Information Processing Systems (NIPS)*. 2001, pp. 841–848, MIT Press.

[18] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.