Speech Communication 100 (2018) 1–12

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic quantitative analysis of spontaneous aphasic speech

Duc Le[*,a,b], Keli Licata[a,c], Emily Mower Provost[a,b]

[a] University of Michigan, Ann Arbor, MI 48109, USA
[b] Computer Science and Engineering, USA
[c] University Center for Language and Literacy, USA

ARTICLE INFO

Keywords:
Aphasia
Clinical application
Quantitative analysis
Disordered speech recognition
Aphasia quotient estimation

ABSTRACT

Spontaneous speech analysis plays an important role in the study and treatment of aphasia, but can be difficult to perform manually due to the time consuming nature of speech transcription and coding. Techniques in automatic speech recognition and assessment can potentially alleviate this problem by allowing clinicians to quickly process large amount of speech data. However, automatic analysis of spontaneous aphasic speech has been relatively under-explored in the engineering literature, partly due to the limited amount of available data and difficulties associated with aphasic speech processing. In this work, we perform one of the first large-scale quantitative analysis of spontaneous aphasic speech based on automatic speech recognition (ASR) output. We describe our acoustic modeling method that sets a new recognition benchmark on AphasiaBank, a large-scale aphasic speech corpus. We propose a set of clinically-relevant quantitative measures that are shown to be highly robust to automatic transcription errors. Finally, we demonstrate that these measures can be used to accurately predict the revised Western Aphasia Battery (WAB-R) Aphasia Quotient (AQ) without the need for manual transcripts. The results and techniques presented in our work will help advance the state-of-the-art in aphasic speech processing and make ASR-based technology for aphasia treatment more feasible in real-world clinical applications.

1. Introduction

Aphasia, a common speech-language disorder typically caused by focal brain damage, currently affects over two million people in the US and more than 180,000 acquire it every year (Association, 2016). Aphasia may cause impairments in both expressive and receptive language skills, including speaking, writing, reading, and listening (Basso, 2003; Davis, 2006; Helm-Estabrooks et al., 2013). Persons with aphasia (PWAs) often face significant communication difficulties, which may lead to feelings of frustration, loneliness, loss of autonomy, and social isolation, among others (Simons-Mackie et al., 2010).

Spontaneous speech (e.g., answering an open-ended interview question, retelling a story, describing a picture) plays a prominent role in a PWA's everyday interaction and is widely regarded in the aphasia literature as one of the most important modalities to analyze (Mayer and Murray, 2003; Prins and Bastiaanse, 2004; Fox et al., 2009; Jaecks et al., 2012). Example applications of spontaneous speech analysis include aphasia classification (Goodglass et al., 2000), treatment planning (Prins and Bastiaanse, 2004), recovery tracking (Grande et al., 2008), and diagnosis of residual aphasia post onset (Jaecks et al., 2012).

Analysis of spontaneous aphasic speech is typically carried out in clinical settings by Speech-Language Pathologists (SLPs) and often confined to a relatively small amount of speech samples with manually coded transcripts, which can be very time consuming to complete (Prins and Bastiaanse, 2004). Furthermore, the analysis itself often requires a SLP's expert knowledge of aphasia and linguistics. As a result, only the small percentage of PWAs who have frequent interaction with SLPs can access and benefit from spontaneous speech analysis, the results of which carry important implications for a PWA's everyday interaction and future treatment plans. At the same time, SLPs in many settings have high productivity expectations and limited time outside of direct patient contact, thus restricting them from conducting such analysis regularly.

Techniques in automatic speech processing can potentially help SLPs perform this type of analysis more efficiently, thereby making its results and findings more commonly available to PWAs. However, previous works in the area of aphasic speech processing have two major limitations that prevent the development of fully automated systems capable of analyzing spontaneous aphasic speech. First, they often assume the availability of expertly produced speech transcripts, which are very time consuming to complete manually (Fraser et al., 2014, 2013b;

* Corresponding author.
E-mail addresses: ducle@umich.edu (D. Le), klicata@umich.edu (K. Licata), emilykmp@umich.edu (E. Mower Provost).

https://doi.org/10.1016/j.specom.2018.04.001
Received 10 August 2017; Received in revised form 11 March 2018; Accepted 4 April 2018
Available online 06 April 2018

Lee et al., 2013b, 2015) and difficult to generate automatically (Fraser et al., 2013a; Peintner et al., 2008). Second, they typically target speech with known prompts (Le et al., 2014; Le and Provost, 2014; Le et al., 2016; Abad et al., 2012, 2013). This removes the need for unconstrained automatic speech recognition (ASR) and greatly simplifies transcript generation, which can be achieved by variants of forced alignment (Le et al., 2014; Le and Provost, 2014; Le et al., 2016) or keyword spotting (Abad et al., 2012, 2013). However, the reliance of this type of system on known prompts makes it inapplicable to spontaneous speech.

It is evident that ASR is a major bottleneck for spontaneous aphasic speech analysis. ASR performance must be sufficiently accurate such that the results and findings are not significantly affected by transcription mismatches. In addition, the features derived from ASR output must be relatively robust to recognition errors. However, the performance of ASR on aphasic speech and robustness of features against transcription errors have been under-explored in the literature. Our work helps bridge this gap by performing one of the first large-scale studies on ASR-based spontaneous aphasic speech analysis.

We present the paper in three sequential components. First, we describe our method for training acoustic models, which leads to significant improvement in aphasic speech recognition accuracy, achieving **37.37%** Word Error Rate (WER) on AphasiaBank, a large-scale aphasic speech corpus (Macwhinney et al., 2011). Next, we discuss various clinically relevant quantitative measures that can be extracted from the resulting transcripts. We show that with our feature calibration method, the majority of these measures are highly robust to ASR errors and can reliably be used for clinical diagnosis. Finally, we demonstrate that these measures can be leveraged to accurately predict the revised Western Aphasia Battery (WAB-R) Aphasia Quotient (AQ), one of the most widely used metrics for aphasia assessment (Kertesz, 2006). Our system achieves **9.18** Mean Absolute Error (MAE) and **.799** correlation in predicting WAB-R AQs without the need for manual transcripts. A high-level overview of the system is shown in Fig. 1.

The technical novelty of this work lies in our proposed calibration method for correcting ASR-based quantitative measures and our modeling approach which combines free speech and semi-spontaneous speech features. The techniques and results presented in this work will help advance the state-of-the-art in aphasic speech processing, as well as make automated spontaneous aphasic speech analysis more feasible in real-world clinical applications.

2. Background

2.1. Linguistic analysis of spontaneous aphasic speech

Linguistic analysis of aphasic speech can be divided into two types, qualitative and quantitative (Prins and Bastiaanse, 2004). The former assesses PWAs' speech based on a qualitative rating scale, such as the Boston Diagnostic Aphasia Examination (Goodglass et al., 2000) or Aachen Aphasia Test (Miller et al., 2000), both of which have a significant portion dedicated to spontaneous speech. The advantage of qualitative analysis is that it is relatively simple and efficient to perform (Katz et al., 2000). However, qualitative rating scales often have difficulties in measuring a PWA's improvement (Prins and Bastiaanse, 2004) and may lack sensitivity (Grande et al., 2008). By contrast, quantitative analysis typically involves the investigation of objective and quantifiable measures that can directly indicate changes in aphasia. However, these quantitative measures are often time consuming to obtain and can require significantly deeper consideration of various linguistic features as well as specialized training in aphasiology to complete and interpret (Prins and Bastiaanse, 2004).

Quantitative analysis of spontaneous aphasic speech has a wide range of applications and is extensively studied in the clinical literature. For example, Grande et al. proposed a set of five basic parameters to measure changes in spontaneous aphasic speech (Grande et al., 2008). This parameter set, which captures lexical and semantic content, syntactic completeness, linguistic complexity, and mean utterance length, was shown to be more sensitive to change compared to qualitative rating scales. Fergadiotis and Wright showed that lexical diversity measures extracted from spontaneous speech can differentiate between PWAs and healthy controls (Fergadiotis and Wright, 2011). Finally, Jaecks et al. were able to diagnose residual aphasia using a set of variables spanning information density, syntactic variability, linguistic errors, and cohesion (Jaecks et al., 2012). The measures proposed in these works form the basis of our feature set (Section 5).

2.2. Automated speech-based methods for aphasia assessment

Automatic analysis of aphasic speech has also been studied in the engineering community. Lee et al. proposed the use of forced alignment in conjunction with manually labeled transcripts to analyze large amount of Cantonese aphasic speech (Lee et al., 2013b; 2015). They found that compared to healthy speech, aphasic speech contains fewer words, longer pauses, and higher numbers of continuous chunks, with fewer words per chunk (Lee et al., 2013b). Further, aphasic speech exhibits different intonation patterns (Lee et al., 2015). Fraser et al. tackled automatic classification of different subtypes of primary progressive aphasia (PPA) based on narrative speech, utilizing a combination of text and acoustic features (Fraser et al., 2014, 2013b). While they achieved good prediction accuracy on these tasks, their proposed feature set relied on intricate transcripts produced manually by trained research assistants. Their follow-up work attempted to evaluate the proposed approach on transcripts generated with an off-the-shelf ASR system (Fraser et al., 2013a). However, the ASR performance was relatively poor, attaining WER between 67.7% and 73.1%. As a result,

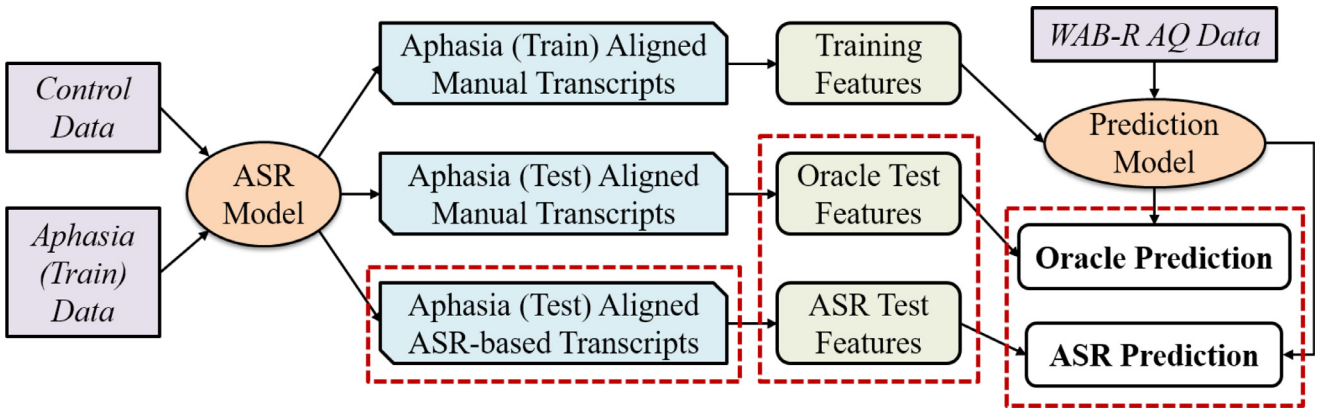

Fig. 1. High-level overview of our proposed system. The red boxes denote components that will be the focus of our analysis.

Table 1
Summary of AphasiaBank data used in this paper. The speakers are split into two groups, those who have aphasia (*Aphasia*) and healthy controls (*Control*).

		Aphasia	Control
Demographics	*Gender*	238 M, 163 F	85 M, 102 F
	Age	62 ± 12	63 ± 17
Speech Data	*Duration*	89.2 h	41.7 h
	Utterances	64,748	38,186
	Words	458,138	371,975
Utterance Type	*Free*	28,157	16,465
	Semi	36,591	21,721

their analysis was limited to simulated ASR output with preset WER levels, and the robustness of their feature set to realistic recognition errors remained unclear.

Our previous work proposed a system to automatically estimate qualitative aspects of read aphasic speech through transcript, pronunciation, rhythm, and intonation features (Le et al., 2014; Le and Provost, 2014; Le et al., 2016). We showed that by using modified forced alignment for automatic transcription, our system could achieve results comparable to those using manual transcripts. Our approach took advantage of the fact that the speech prompt was known ahead of time, thus significantly constraining the space of possible utterances. However, this is an unrealistic assumption for spontaneous speech. Peintner et al. proposed a set of speech and language features extracted from ASR output to distinguish between three types of frontotemporal lobar degeneration, including progressive non-fluent aphasia (Peintner et al., 2008). While their work showed promising results, it was performed on a relatively small dataset and there was no analysis regarding the reliability of ASR-based features. By contrast, the work presented in this paper is conducted on a large-scale aphasic speech corpus with detailed discussion regarding the robustness of ASR-derived quantitative measures.

2.3. Disordered speech recognition

There has been extensive work in the related field of dysarthric speech recognition (Sharma and Hasegawa-Johnson, 2010; Aniol, 2012; Christensen et al., 2012; Sharma and Hasegawa-Johnson, 2013; Christensen et al., 2013a,b, 2014). ASR for dysarthric and disordered speech in general is faced with abnormal speech patterns (Mengistu and Rudzicz, 2011), high speaker variability (Mustafa et al., 2015), and data scarcity (Christensen et al., 2012). Methods for alleviating these problems include speaker-dependent Gaussian Mixture Model (GMM) adaptation (Sharma and Hasegawa-Johnson, 2010; Christensen et al., 2012; Sharma and Hasegawa-Johnson, 2013), generation of auxiliary acoustic features used within tandem-based systems (Aniol, 2012; Christensen et al., 2013a), learning systematic speaker-specific pronunciation errors (Christensen et al., 2013b), and similarity-based speaker selection for acoustic modeling (Christensen et al., 2014).

There has been relatively little work on ASR for aphasic speech. Most existing works were limited to using mismatched healthy acoustic models for recognition (Peintner et al., 2008; Abad et al., 2013; Lee et al., 2016). Our previous work established the first ASR baseline on AphasiaBank, a large-scale aphasic speech corpus (Macwhinney et al., 2011), showing that adding utterance-level i-vectors to the input drastically improves the recognition accuracy of Deep Neural Network (DNN) acoustic models (Le and Mower Provost, 2016). More recently, we utilized out-of-domain adaptation and deep Bidirectional Long-Short Term Memory Recurrent Neural Network (BLSTM-RNN) acoustic model to automatically transcribe and detect paraphasias in read aphasic speech (Le et al., 2017). In this paper, we combine the techniques used in these works and analyze aphasic speech recognition results in more detail.

3. Data

3.1. Speech data

All experiments in this work are carried out on AphasiaBank, a large-scale audiovisual dataset primarily used by clinical researchers to study aphasia (Forbes et al., 2012; Macwhinney et al., 2011). AphasiaBank is a growing collection of sub-datasets contributed by different research groups under various elicitation protocols. For this work, we select English sub-datasets that have at least four speakers and are collected with the core AphasiaBank protocol, a series of open-ended questions designed to gather verbal discourse samples. These inclusion criteria result in 401 PWAs and 187 control speakers without aphasia, spanning 19 sub-datasets and 130.9 h of speech. Utterances in AphasiaBank can be categorized based on their applied elicitation method, which can be either free speech (e.g., open interview, conversational speech) or semi-spontaneous (e.g., storytelling, picture description) (Prins and Bastiaanse, 2004). The speech-language patterns of the same PWA may be different across these two categories (Prins and Bastiaanse, 2004; Jaecks et al., 2012), thus it may be beneficial to analyze them separately. Table 1 describes the dataset in more detail.

3.2. Transcripts

Utterances in AphasiaBank were transcribed using the CHAT format (MacWhinney, 2000). The transcriptions contain a variety of special codes to aid with language sample analysis, such as word-level errors, sound fragments, repetitions, non-verbal actions, among others. The first row of Table 2 shows an example transcript containing a sound fragment **&uh**, a semantic error **bit** with known target **get**, a real-word phonological error **pea** with known target **the**, and two non-word phonological errors with known targets, **peanut** and **butter**. The actual pronunciations of these non-word phonological errors are transcribed in the International Phonetic Alphabet (IPA) format, arked with the **@u** trailing symbol. CHAT transcripts contain a rich source of information about a PWA's speech-language patterns that enable various forms of manual analyses. However, they are not suitable targets for standard ASR and thus need to be simplified. We propose two ways to process raw CHAT transcripts, one preserving the original pronunciations (***cleaned***) and one emphasizing the language usage patterns (***target***).

For *cleaned* transcripts, we replace all sound fragments and interjections, in addition to **um** and **uh**, with a generic filler token, denoted by < **FLR** > . Other special tokens include < **SPN** > (spoken noise, e.g., onomatopoeia, babbling), < **LAU** > (laughter), and < **BRTH** > (breathing sounds). In order to preserve the original pronunciations, we retain all word-level errors. We convert IPA strings to special hashed tokens such that the same IPA pronunciations map to the same hash. The second row of Table 2 shows an example *cleaned* transcript, in which **&uh** is mapped to < **FLR** > , semantic error **bit** and real-word phonological error **pea** are retained, and the two non-word phonological errors **pinək@u** and **bʌðə@u** are replaced with hashed tokens < **U1** > and < **U2** > .

While *cleaned* transcripts preserve the original pronunciations and are therefore suitable targets for acoustic modeling, they are difficult for an ASR system to produce because of the retained word-level errors.

Table 2
Example transcript and its processed forms. *Cleaned* transcripts preserve the original pronunciation of each word. *Target* transcripts replace all word-level errors, excluding semantic errors, with their known targets (if available).

Original	And I &uh bit [: get] [* s:ur] out pea [: the] [* p:w] pinək@u [: peanut] [* p:n] bʌðə@u [: butter] [* p:n].
Cleaned	And I < FLR > bit out pea < U1 > < U2 > .
Target	And I < FLR > bit out the peanut butter.

We showed in previous work that phonological and especially neologistic errors are challenging to recognize automatically due to their mismatched pronunciations and lack of representation in the lexicon (Le et al., 2017). We address this problem by producing ***target*** transcripts in which all word-level errors, excluding semantic, are replaced with their known targets. An example is shown in Table 2, where **pea** **< U1 >** **< U2 >** is replaced with **the peanut butter**. Semantic errors are retained because they may have completely different pronunciations than their targets, thus replacing them will cause significant difficulties for ASR. These transcripts better reflect a PWA's language usage patterns and will be used for language modeling as well as ASR evaluation.

3.3. Lexicon preparation

Our lexicon is based on the CMU dictionary[1], containing 39 regular phones, plus five phones representing special tokens: silence, **< FLR >**, **< LAU >**, **< SPN >**, and **< BRTH >**. Each IPA pronunciation is heuristically mapped to a sequence of CMU phones. For example, **pinək@u** and **bʌðə@u** are converted to **p iy n ah k** and **b ah dh er**, respectively. Finally, we estimate the pronunciations of the remaining OOV words using the LOGIOS lexicon tool[2], which makes use of normalization, inflection, and letter-to-sound rules.

3.4. Speaker-level ratings and assessment

AphasiaBank contains a number of speaker-level test results, including WAB-R AQ, AphasiaBank Repetition Test, Boston Naming Test–Short Form, Northwestern Verb Naming Test, Complex Ideational Material–Short Form, and Philadelphia Sentence Comprehension Test. Among these tests, WAB-R AQ is the most commonly administered, with test data available for 355 PWAs (out of 401). The other tests are conducted on fewer PWAs and/or not as widely used outside the scope of AphasiaBank. Since an increase in AQ can signify improvement in a PWA's language capabilities, reliable automatic AQ estimation may play an important role in monitoring a PWA's recovery progress over time. We are interested in seeing how well AQ can be estimated for each PWA in our dataset.

WAB-R AQ is an aggregated score ranging from 0 to 100 that measures a PWA's overall language capabilities (Kertesz, 2006). It consists of four separate subtests, Spontaneous Speech, Auditory Comprehension, Repetition, and Naming/Word Finding. The severity of aphasia can be roughly categorized according to this score: mild (76–100), moderate (51–75), severe (26–50), and very severe (0–25). The PWAs have a mean AQ of 71.1 and a standard deviation of 19.5, with the majority classified as mild (174), followed by moderate (131), severe (38), and very severe (12). Fig. 2 plots the histogram of the available AQ scores in our dataset.

3.5. Experimental setup

An automated system for aphasic speech analysis must be able to handle previously unseen speakers. We adopt a speaker-independent 4-fold cross-validation scheme, similar to that used in our previous work (Le and Mower Provost, 2016). For each fold, we withhold 25% of speakers from each sub-dataset in the **Aphasia** set to form a test set. The remaining data and all **Control** speakers are used for training. Test results from all folds will be pooled together for analysis. The amount of per-fold training data, including **Control** speakers, ranges from 106.8 to 110.5 h.

[1] http://www.speech.cs.cmu.edu/cgi-bin/cmudict.
[2] http://www.speech.cs.cmu.edu/tools/lextool.html.

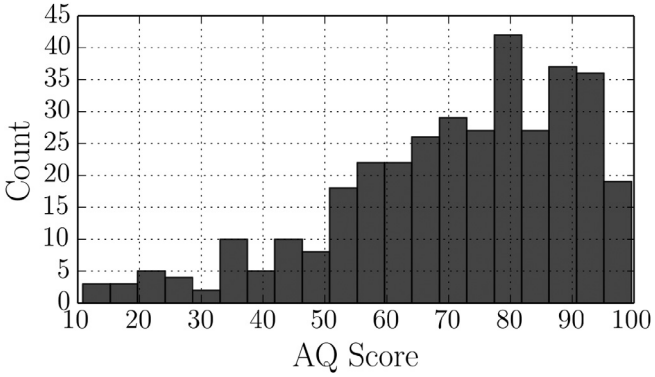

Fig. 2. Histogram of WAB-R AQ scores.

4. Automatic transcription

The first step of spontaneous aphasic speech analysis is to obtain a detailed transcript for each utterance, including precise alignments of words and phones. These transcripts are time consuming to create manually; an alternative is to utilize ASR to generate them automatically. In this section, we describe the components of our ASR system that allows the transcription of spontaneous aphasic speech.

4.1. Acoustic modeling

In this work, we train a deep multi-task BLSTM-RNN on log Mel filterbank coefficient (MFB) features augmented with utterance-level i-vectors (Fig. 3). Our previous work showed that utterance-level i-vectors significantly improved ASR performance on AphasiaBank when used in conjunction with a DNN acoustic model and Mel-frequency cepstral coefficient (MFCC) features (Le and Mower Provost, 2016). Our work on automatic paraphasia detection made use of a deep multi-task BLSTM-RNN architecture trained on MFBs that predicts the senone and monophone labels simultaneously (Le et al., 2017). We will analyze the effect of BLSTM-RNN acoustic model, i-vectors, and multi-task learning in Section 7.1.

4.1.1. Frontend

We resample all audio files to 16kHz and use Kaldi (Povey et al., 2011) to extract two sets of frame-level acoustic features: (a) 12-dimensional MFCCs plus energy, along with the first and second order derivatives, and (b) 40-dimensional MFBs. We use a 25 ms window and 10 ms frame shift for both feature types. Finally, we z-normalize the features at the speaker level.

4.1.2. Development data

We withhold 15% of training **Aphasia** speakers from each sub-dataset to form a held-out development set. We will refer to the remaining training data as the training set in this section.

4.1.3. Frame-level labels

We obtain senone and monophone labels for each frame through forced alignment using a bootstrap context-dependent tied-state triphone HMM-GMM system trained with Maximum Likelihood. The number of senones across different folds ranges from 4,472 to 4,563. Forced alignment is carried out on *cleaned* transcripts because they preserve all word-level pronunciations and will likely yield more accurate frame labels.

4.1.4. I-vector extraction

We adopt a similar i-vector extraction technique to the one used in Le and Mower Provost (2016). For the frontend, we splice 9 consecutive MFCC frames together, followed by 40-dimensional Linear Discriminant

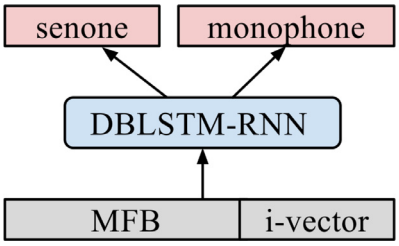

Fig. 3. Deep multi-task BLSTM-RNN acoustic model.

Table 3
13 applied statistics.

quartiles 1–3
3 inter-quartile ranges
1% percentile (≈ min), 99% percentile (≈ max)
percentile range 1% – 99%
mean, standard deviation
skewness, kurtosis

Analysis (LDA) with senones as class labels. We model these LDA-transformed features using a Universal Background Model (UBM) with 1024 Gaussians. Finally, we train an extractor to generate a 32-dimensional i-vector for every utterance in our dataset. The LDA matrix, UBM, and i-vector extractor are trained using only the training set. The resulting utterance-level i-vectors encode speaker as well as channel information, and will be used as auxiliary input features for our acoustic model.

4.1.5. Multi-task BLSTM-RNN training

Following Le et al. (2017), we augment each MFB frame with five left and five right neighbors[3] in addition to the corresponding utterance-level i-vector, resulting in 472 dimensions per frame. We model these features with a stacked BLSTM-RNN comprising four hidden layers, each with 1200 units (600 for forward, 600 for backward). The model has two parallel softmax output layers corresponding to the senone and monophone labels.

We train the model with the Adam optimizer (Kingma and Ba, 2014) and total Cross Entropy (CE) loss weighted equally across the two tasks. We utilize full Backpropagation Through Time (BPTT), limited to utterances that are shorter than 25 s. Only less than 0.5% of training utterances are longer than 25 s, many of which have badly aligned transcripts that may negatively affect model training. Therefore, we hypothesize that excluding these utterances will have minimal impact on acoustic model performance.

We use 0.4 dropout and an initial learning rate of 0.001. We perform early stopping based on the development senone Frame Error Rate (FER) and step-decay learning schedule (Le and Mower Provost, 2016). If the senone FER on the development set increases at the end of a training epoch, we halve the current learning rate and restore the previous model parameters. The training process finishes once the learning rate drops below 0.00005.

4.1.6. Baseline acoustic models

In order to investigate the effect of i-vectors and multi-task learning, we train three additional BLSTM-RNN acoustic models using the same method. These include a multi-task BLSTM-RNN model trained on MFBs without utterance-level i-vectors and two single-task BLSTM-RNNs with only senone output, one trained on MFBs plus i-vectors, and one trained on just MFBs.

Finally, we train baseline DNN acoustic models using the method described in Le and Mower Provost (2016) to better evaluate the effectiveness of BLSTM-RNN. The model consists of four hidden layers with 2048 units each and one senone output layer. The input features are 27-frame context windows of MFBs with or without augmented utterance-level i-vectors.

[3] Input features to RNN acoustic models are traditionally single acoustic frames. In our work, we found that using single and multiple frames (context windows) as input features gives very similar recognition rates (less than 0.4% relative difference).

4.2. Language modeling and decoding

Our preliminary results indicate that while *cleaned* transcripts are more appropriate for obtaining frame-level labels through forced alignment, *target* transcripts better reflect the language usage patterns of PWAs and are more suitable ground-truths for ASR. We use SRILM (Stolcke et al., 2011) to train a trigram language model (LM) with backoff on the training *target* transcripts. We tune the decoder's LM weight {9, 10, ..., 20} and word insertion penalty {0.0, 0.5, 1.0} based on the development WER. Our decoder outputs the hypothesized transcripts as well as the word- and phone-level alignments.

5. Quantitative analysis

In the context of this work, the goal of quantitative analysis is to produce a set of quantifiable measures (i.e., features) for each speaker that are characteristic of aphasic speech, compatible with ASR output, and robust to recognition errors. We consider adopting and extending existing measures that have been proposed in the engineering literature for disordered speech assessment. In addition, we aim to operationalize quantitative measures that have traditionally been used only in clinical studies. We focus specifically on measures that can separate different severity levels of aphasia and/or distinguish between PWAs and healthy controls. The extracted features (Table 4) are organized into six groups, each of which captures a specific speech-language aspect of a PWA. The extraction of these features relies on speech transcripts, which may be either time aligned manual transcripts or ASR-generated output (Fig. 1).

5.1. Information density (DEN)

This group of features captures the amount of information conveyed in a PWA's speech, under the hypothesis that those with milder aphasia produce relatively denser information content. Features 1–2 capture a PWA's speech rate, which has been shown in previous work to be useful for assessing the quality of aphasic speech (Le et al., 2014; Le and Provost, 2014; Le et al., 2016) as well as distinguishing between subjects with PPA and healthy controls (Fraser et al., 2014, 2013b).

Features 3–4 are adopted from a set of basic parameters proposed by Grande et al. to objectively measure changes in spontaneous aphasic speech (Grande et al., 2008). Following their work, we define interjections to be fillers (< **FLR** >) and the particles *yes, yeah*, and *no*. Open class words are nouns, verbs, adjectives, and derivative adverbs (heuristically determined as those ending with $-ly$). Closed class words comprise determiners, pronouns, conjunctions, and genuine (i.e., non-derivative) adverbs. We generate Part of Speech (POS) tags for all words in our transcripts using NLTK (Bird et al., 2009) and the universal tag set. Percentage words (*W*) is expected to capture word-finding difficulties since it decreases with more frequent use of interjections. Meanwhile, percentage open-class words (*OCW*) characterizes agrammatism, in which PWAs produce mainly content words and relatively few function words (Grande et al., 2008).

Features 5–6 are based on mean length of utterances in words, a widely used measure in spontaneous aphasic speech analysis (Prins and Bastiaanse, 2004; Grande et al., 2008; Jaecks et al., 2012). We extend this measure by computing a more comprehensive set of statistics over the collection of utterance lengths, using the 13 statistics listed in Table 3. We also consider utterance length measured in the number of phones instead of words as they may capture a PWA's speech production ability more accurately. We expect more severe PWAs to produce

Table 4
Extracted quantitative measures for each speaker. {} denotes a collection of numbers summarized into speaker-level measures using the statistics listed in Table 3.

Information Density (DEN)		
1	Words/min	Words / Total duration (minutes)
2	Phones/min	Phones / Total duration (minutes)
3	W	Words / (Words + Interjections)
4	OCW	Open class / Open + closed class
5	{Words/utt}	Words spoken per utterance
6	{Phones/utt}	Phones spoken per utterance
7	Nouns	Nouns / Words
8	Verbs	Verbs / Words
9	Nouns/verb	Nouns / Verbs
10	Noun ratio	Nouns / (Nouns + Verbs)
11	Light verbs	Light verbs / Verbs
12	Determiners	Determiners / Words
13	Demonstratives	Demonstratives / Words
14	Prepositions	Prepositions / Words
15	Adjectives	Adjectives / Words
16	Adverbs	Adverbs / Words
17	Pronoun ratio	Pronouns / (Nouns + Pronouns)
18	Function words	Function words / Words
Dysfluency (DYS)		
19	Fillers/min	Fillers / Total duration (minutes)
20	Fillers/word	Fillers / Words
21	Fillers/phone	Fillers / Phones
22	Pauses/min	Pauses / Total duration (minutes)
23	Long pauses/min	Long pauses / Total duration (minutes)
24	Short pauses/min	Short pauses / Total duration (minutes)
25	Pauses/word	Pauses / Words
26	Long pauses/word	Long pauses / Words
27	Short pauses/word	Short pauses / Words
28	{Seconds/pause}	Duration of pauses in seconds
Lexical Diversity and Complexity (LEX)		
29	Type–token ratio	Unique words / Words (open class)
30	{Freq/word}	Word frequency score
31	{Img/word}	Word imageability score
32	{AoA/word}	Word age of acquisition score
33	{Fam/word}	Word familiarity score
34	{Phones/word}	Number of phones per word
Part-of-Speech Language Model (POS-LM)		
35	{Bigram CE/utt}	POS bigram Cross Entropy per utterance
36	{Trigram CE/utt}	POS trigram Cross Entropy per utterance
Pairwise Variability Error (PVE)		
37	{PVE_1/utt}	Utterance PVE score ($M = 1$)
38	{PVE_2/utt}	Utterance PVE score ($M = 2$)
39	{PVE_3/utt}	Utterance PVE score ($M = 3$)
40	{PVE_4/utt}	Utterance PVE score ($M = 4$)
Posteriorgram-Based Dynamic Time Warping (DTW)		
41	{Raw dist/word}	Raw DTW distance per word
42	{Norm dist/word}	Normalized DTW distance per word
43	{Segment/word}	Longest horizontal/vertical aligned segment per word

shorter utterances on average while having less varied utterance lengths.

Features 7–18 characterize a PWA's POS usage patterns, which have been shown to be important for residual aphasia (Jaecks et al., 2012) and PPA (Fraser et al., 2014, 2013b) diagnosis. Following Breedin et al. (1998) and Fraser et al. (2014), we classify verbs as *light* or *heavy* depending on their semantic complexity. A verb is considered *light* if its lemmatized form is *be, have, come, go, give, take, make, do, get, move*, or *put*; otherwise, the verb is categorized as *heavy*. Function words include determiners, pronouns, prepositions, conjunctions, particles, and modals; they are expected to occur more frequently in milder PWAs (Grande et al., 2008).

5.2. Dysfluency (DYS)

Dysfluency is an important aspect of aphasic speech which has been used in qualitative analysis (Le et al., 2014; Le and Provost, 2014; Le et al., 2016) and PPA diagnosis (Fraser et al., 2013b). Features 19–28 capture the amount of dysfluency (i.e., fillers and pauses) in each PWA's speech. Following Pakhomov et al. (2010), we define pauses as regions of silence between words that are longer than 150 ms; these are further categorized as short ($\leq$ 400 ms) or long ($>$ 400 ms). We extract the occurrence frequency of fillers and pauses normalized by speech duration, total words, and total phones. Finally, we extract the statistics over all pause durations. We expect milder PWAs to exhibit less dysfluency and vice versa.

5.3. Lexical diversity and complexity (LEX)

Lexical diversity, defined as the range of vocabulary employed by a speaker, has been shown to be significantly different between PWAs and healthy controls (Fergadiotis and Wright, 2011). A standard measure that captures lexical diversity is the ratio between the number of unique words (*types*) and total words (*tokens*), commonly referred to as type–token ratio (TTR). Following Fergadiotis and Wright (2011), we extract TTR (feature 29) using only lemmatized open class words to remove the influence of grammars on lexical diversity. PWAs with mild aphasia tend to have less word-retrieval difficulties; as a result, we expect them to have relatively higher TTR compared to more severe PWAs.

The complexity of a speaker's vocabulary is also an important measure of aphasic speech. We hypothesize that PWAs with mild aphasia tend to use words that are longer and less frequently used compared to those with severe aphasia. Brysbaert and New introduced the SUBTL norms, a mapping from words to their frequencies in American English based on an analysis of film and television subtitles (Brysbaert and New, 2009). In addition, the combined work of Stadthagen-Gonzalez and Davis (2006) and Gilhooly and Logie (1980) produced a database of word-level imageability, age of acquisition, and familiarity scores, which can be used to estimate a word's complexity. In this work, we extract statistics over all word-level frequency, imageability, age of acquisition, and familiarity scores for each speaker, resulting in features 30–33. Similar measures were used by Fraser et al. for PPA diagnosis (Fraser et al., 2014; 2013b). However, they only extracted the mean scores, whereas we consider a more comprehensive set of statistics (Table 3). Finally, feature 34 approximates the pronunciation complexity of a PWA's vocabulary based on the number of phones present in a word.

5.4. Part of speech language model (POS-LM)

The degree of a PWA's syntactic deviation from that of healthy controls may help separate subjects with different severity levels. We model the syntactic structure present in healthy speech by training bigram and trigram LMs with backoff on the POS transcripts of **Control** speakers. Given a POS LM $\mathscr{M}$, the Cross Entropy (CE) of a POS sequence $p_1 p_2 \cdots p_N$ denotes how closely it adheres to the model:

$$\mathscr{H}(p_1 p_2 \cdots p_N | \mathscr{M}) = \frac{\log P(p_1 p_2 \cdots p_N | \mathscr{M})}{N} \quad (1)$$

PWAs with milder language impairment are expected to produce more standard POS sequences, thus resulting in higher CE on average. Features 35–36 capture this idea through the statistics of utterance-level bigram and trigram CE scores. A similar approach was used by Roark et al. to detect mild cognitive impairment (Roark et al., 2011).

5.5. Pairwise variability error (PVE)

Speech rhythm was shown in our previous work to be helpful for estimating qualitative aspects of aphasic speech (Le and Provost, 2014; Le et al., 2016). In the context of this work, we expect the rhythmic patterns of less severe PWAs to be more similar to **Control** speakers and vice versa. We quantify rhythmic deviations using Pairwise Variability Error (PVE), a metric first proposed by Tepperman et al. (2010) to compare the rhythms of a candidate (**Aphasia**) and reference (**Control**)

speaker. Given duration profiles of a candidate and reference utterance, denoted as $\{c_1, c_2, ..., c_N\}$ 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:

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

where M is a hyperparameter specifying the maximum distance between a pair of units considered for comparison. PVE scores range from 0 to 1, where values closer to 0 denote higher similarity between the candidate and reference rhythms.

The candidate and reference duration profiles for an utterance are generated using the Reference Alignment algorithm proposed in our previous work (Le and Provost, 2014). This algorithm aligns a candidate utterance to a prototypical reference utterance, accounting for OOV words by breaking them down into finer granularity levels. Features 37–40 comprise statistics of utterance-level PVE scores with context parameter M varying from 1 to 4, the same range used in Tepperman et al. (2010), Le and Provost (2014) and Le et al. (2016).

5.6. Posteriorgram-based dynamic time warping (DTW)

Our final feature group is based on the observation that PWAs with more severe aphasia tend to have worse pronunciations. The monophone output of our multi-task BLSTM-RNN acoustic model can be viewed as a compact representation of each speech frame. Combined with the aligned transcripts, we can represent each word as a multi-dimensional time series (i.e., posteriorgram), where each point in the series is a probability distribution over 44 phones. Intuitively, words that are pronounced correctly will have posteriorgrams that are similar to those associated with **Control** speakers. We showed in our previous work that Dynamic Time Warping (DTW) can be used to detect paraphasias through posteriorgram comparison (Le et al., 2017). The DTW-based features, inspired by Lee and Glass (2012), Lee et al. (2013a) and Lee and Glass (2013), were shown to outperform Goodness of Pronunciation (GOP) (Witt and Young, 2000) and phone edit distance features. We will therefore adopt DTW-based features in this work.

As a first step to feature extraction, we represent the correct pronunciation of each word as a collection of posteriorgrams extracted from **Control** speakers, which we will refer to as the ***reference*** set. For efficiency reasons, we limit the maximum number of reference posteriorgrams per word to 100, randomly subsampled if necessary. We can then compare a pair of candidate and reference posteriorgrams using DTW, where the distance between two frames c_i and r_j is defined as their inner product distance:

$$D(c_i, r_j) = -\log(c_i \cdot r_j) \quad (3)$$

We extract the following features for each word in our dataset by comparing its posteriorgram with the reference set: mean raw DTW distance, mean DTW distance normalized by aligned path length, and mean length of the longest horizontal/vertical aligned segment normalized by aligned path length. Special tokens and words with fewer than five reference posteriorgrams are skipped. Finally, we calculate the statistics over these word-level measures, producing features 41–43.

5.7. Feature calibration

A desirable property of automatic quantitative analysis is that features extracted with ASR-generated transcripts should accurately reflect a PWA's true measures, i.e., features extracted with manual (oracle) transcripts. We observe that there often exists a systematic deviation between these two sets. For example, ASR-based *words/min* features are typically smaller than their oracle counterparts due to deletion errors; however, they still have very high correlation with one another (Fig. 4). This relationship can be exploited to calibrate ASR-based features to become more similar to oracle features, i.e., closer to a PWA's true measures.

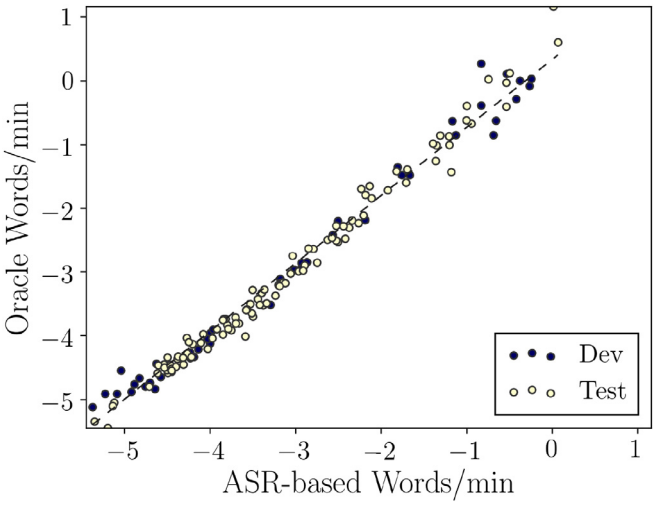

Fig. 4. Example calibration of *words/min* feature. A linear transformation model is trained on development speakers ($y = 1.07x + .33$) and applied to test speakers. Feature values are z-normalized using statistics extracted from healthy controls.

We consider performing calibration for every individual feature by training a linear transformation model on development speakers and applying it to test speakers (e.g., Fig. 4). To ensure that feature calibration is effective, we apply the transformation only if the oracle and ASR-based development features are: (1) statistically significantly different before calibration (two-tailed paired t-test of equal means, $p = .05$), and (2) not statistically significantly different after calibration. If condition (1) is not met, it implies that the feature is already well calibrated and no further action is required. If condition (2) is not met, calibration will likely be ineffective, hence we do not apply the transformation. We will analyze the impact of feature calibration in Section 7.2.

6. WAB-R AQ Prediction

The system's goal is to automatically predict WAB-R AQ, an assessment score closely tied to aphasia severity (Kertesz, 2006). This provides an output that has clinical utility, one that does not require thorough knowledge of linguistics and aphasiology to understand, and one that can be quickly interpreted given the significant time constraints present in clinical settings. The automatic estimation of AQ from spontaneous speech has many potential benefits. For example, it will enable progress monitoring without necessitating frequent repeats of the WAB-R assessment procedures, thus saving time for more important treatment activities. In addition, because the WAB-R cannot be administered repeatedly in a short time period due to the practice effect, reliable automatic AQ estimation independent of the WAB-R can help provide a more complete and robust picture of a PWA's recovery trajectory.

6.1. Experimental setup

We frame WAB-R AQ prediction as a regression problem, with the proposed quantitative measures as features and AQ scores as the target labels. For this set of experiments, we select PWAs who have recorded AQ scores as well as speech samples in both the free speech and semi-spontaneous categories. 348 out of 401 PWAs meet these requirements. We maintain the same speaker-independent four-fold cross-validation split described in Section 3.5, where 25% of speakers are held out from each fold as test data.

We z-normalize all features using statistics computed on **Control** speakers. This aids in model training and enables easy interpretation of the resulting features. For example, a negative *words/min* feature means that the subject speaks more slowly than the typical healthy control, whereas a positive *OCW* feature indicates a relatively less frequent use of function words. Finally, this strategy ensures that features extracted across different folds are comparable and can be analyzed together, since the z-normalization statistics remain the same.

Our preliminary results indicate that Support Vector Regression (SVR) performs favorably in this task compared to Linear Regression, k-Nearest Neighbor, and Multi-Layer Perceptron. We use scikit-learn (Pedregosa et al., 2011) to train SVR on training features extracted from time aligned manual transcripts. We first perform hyperparameter selection using 10-fold cross-validation with MAE as the metric. Our hyperparameter ranges are as follows: penalty term C $\{1.0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, slack parameter ϵ $\{1.0, 10^{-1}, 10^{-2}, 10^{-3}\}$, kernel type {*RBF, linear*}, and shrinking heuristic {*true, false*}. We train the final model on the full training set using the cross-validated hyperparameters.

We perform prediction using three sets of test features:

- ***Oracle***: features extracted with manual transcripts.
- ***Auto***: features extracted with ASR-generated transcripts.
- ***Calibrated****: Auto* features after calibration.

These three sets of results represent our system's performance given perfect and imperfect ASR. Our objective is to achieve good prediction results while minimizing the impact of ASR errors. We post-process the model outputs, clipping them within [0, 100], the known range of WAB-R AQ.

It is worth noting that our regression model is trained on oracle features, and the same model is used with both oracle and ASR-based test features for prediction. Alternatively, we can use ASR-based features for both training and testing, which Fraser et al. found to be beneficial for PPA classification (Fraser et al., 2013a). We do not adopt this approach in our work because it will mask the effect of ASR errors on prediction performance, which we plan to analyze in Section 7.3. Investigation of this modeling approach will be left for future work.

6.2. Feature extraction protocols

Research in aphasiology suggested that the speech-language patterns of a PWA may be different across free speech and semi-spontaneous speech (Prins and Bastiaanse, 2004; Jaecks et al., 2012). As a result, features extracted on these two categories may exhibit different and possibly complementary characteristics. We investigate four variations of our feature set based on this observation:

- ***All***: features extracted on all available utterances.
- ***Free***: features extracted on free speech utterances.
- ***Semi***: features extracted on semi-spontaneous utterances.
- ***Combined***: concatenation of *Free* and *Semi* features.

Analyzing the relative performance of these feature protocols will help indicate the type of aphasic speech most suitable for automatic AQ prediction. In addition, *Combined* features may improve performance if free speech and semi-spontaneous speech do indeed provide complementary information.

7. Results and discussion

7.1. ASR Performance

Table 5 summarizes the performance of our ASR systems, broken down by acoustic model and input feature type. The best performance is obtained with the proposed multi-task BLSTM-RNN acoustic model trained on MFBs augmented with utterance-level i-vectors. Adding i-vectors to the input reduces WER by 2.2% (DNN), 4.3% (single-task BLSTM-RNN), and 4.1% (multi-task BLSTM-RNN) relative to their counterparts without i-vectors. This confirms our finding in previous work (Le and Mower Provost, 2016) and demonstrates the efficacy of i-vectors in speaker-independent acoustic modeling for aphasic speech. Single-task BLSTM-RNN greatly outperforms DNN, reducing the WER by 14.9% (without i-vectors) and 16.7% (with i-vectors). While the relative improvement in recognition rate attributed to multi-task learning is small (around 1%), this method enables the production of low-dimensional monophone posteriorgrams and the extraction of DTW features. While it is theoretically possible to use senone posteriorgrams for DTW feature extraction, their high dimensionality will make this process prohibitively expensive.

Table 5
Aphasic speech WER under different input feature and acoustic model configurations.

Features	DNN	BLSTM-RNN	
		Single-Task	*Multi-Task*
MFB	46.26	39.37	38.95
MFB + i-vectors	45.26	37.69	**37.37**

We further analyze the WER breakdown according to utterance type (Table 6). Semi-spontaneous utterances are generally easier to recognize compared to free speech. A possible explanation is that the former are more constrained in terms of vocabulary range and syntactic structure, and are therefore more compatible with our language model. This suggests that applications requiring highly accurate ASR should work exclusively with semi-spontaneous speech.

WER also varies based on the severity of aphasia defined by the WAB-R AQ (Table 6). Speech of more severe PWAs tend to be more difficult to recognize and vice versa, possibly due to the speech-language impairments present in this population, which result in irregular language patterns, high amount of dysfluency, and word-level pronunciation errors. However, the speaker-level WERs have only a moderate Pearson's correlation of − .545 with WAB-R AQ. This suggests that AQ scores can be used to estimate ASR performance for a given PWA, but only to a limited extent. Nevertheless, these results indicate that those with severe aphasia will have significant difficulties with applications that are highly reliant on ASR output.

Finally, we investigate the error rates of individual words, defined as the sum of insertion, deletion, and substitution errors made on a word divided by the total number of occurrences of that word. We limit the analysis to words that occur at least 100 times in the transcripts. Table 7 lists the words with the highest and lowest errors. It can be observed that words with high error rates are generally short and conversational in nature, while those with low errors tend to be longer content words. This suggests that ASR is more suitable for non-conversational aphasic speech and explains why semi-spontaneous utterances are easier to recognize.

Given these analyses, it is possible that WER can be further reduced by personalizing the acoustic and language models for one utterance type and/or severity group. Moreover, speakers who have similar error patterns can be grouped together for more fine-grained acoustic and language model training. We will explore model personalization

Table 6
WER breakdown by utterance type and aphasia severity defined by WAB-R AQ.

Utterance Type		Aphasia Severity			
Free	*Semi*	*Mild*	*Moderate*	*Severe*	*V. Severe*
38.79	36.24	33.68	41.11	49.21	63.17

Table 7
Words with the highest and lowest error rates.

High Errors			Low Errors		
Word	*Count*	*Error*	*Word*	*Count*	*Error*
hm	210	1.0	happy	274	0.12
mhm	656	0.99	window	593	0.11
I'd	168	0.96	house	457	0.11
yep	101	0.86	stepmother	133	0.11
let	128	0.84	speech	317	0.10
we're	124	0.81	castle	108	0.09
< SPN >	1321	0.81	hospital	416	0.09
I've	249	0.79	people	544	0.08
< BRTH >	1345	0.73	beautiful	262	0.08
am	153	0.73	weeks	131	0.08

methods in future work.

7.2. Feature robustness to ASR errors

One of the most important requirements of ASR-driven quantitative analysis is that the extracted measures must be sufficiently robust to recognition errors. We say a feature is ***robust*** if its values derived from ASR-generated output are not statistically significantly different from those based on manual transcripts. For regular features, we employ a two-tailed paired t-test of equal means, $p = .05$. For features involving the 13 statistics in Table 3, we use a two-way repeated measures Analysis of Variance (ANOVA) with Greenhouse-Geisser correction, $p = .05$, to study the effect of statistic (*1st quartile, 2nd quartile, ..., skewness, kurtosis*) and transcript type (*manual, ASR-based*). The feature is considered robust if the effect of transcript type is not statistically significant.

It can be seen from Table 8 that our proposed calibration method has a positive impact on improving feature robustness. Many features that are not originally robust, such as *words/min, determiners*, and *fillers/min*, become robust after calibration. Meanwhile, the vast majority of features that are already robust before calibration, such as *adverbs, pronoun ratio*, and *{bigram CE/utt}*, remain so after calibration. This suggests that even though ASR errors may lead to feature extraction mismatch, this mismatch is often systematically biased and can be corrected with linear transformation. The remaining analysis will therefore focus on calibrated features.

Several quantitative measures are consistently robust across all three feature extraction protocols (*All, Free*, and *Semi*). These include *words/min, phones/min, {words/utt}, {phones/utt}, determiners, demonstratives, adverbs, pronoun ratio* (***DEN***), *fillers/min, fillers/phone, pauses/min* (***DYS***), *{img/word}, {AoA/word}, {fam/word}* (***LEX***), *{bigram CE/utt}, {trigram CE/utt}* (***POS-LM***), *{$PVE_{1,2,4}$/word}* (***PVE***), and *{raw dist/word}* (***DTW***). These measures, especially those in the ***DEN*** and ***LEX*** feature groups, have been demonstrated to be clinically useful for the analysis of aphasia (Grande et al., 2008; Fergadiotis and Wright, 2011; Jaecks et al., 2012). The fact that such quantitative measures can be reliably extracted based on ASR output is promising. SLPs can use them to assist with clinical diagnosis and treatment planning without having to extract them manually, which is often prohibitively time consuming. This technology will help SLPs form a more complete picture of a PWA's speech-language profile, which can potentially result in more suitable treatment approaches.

The robustness of other features may vary depending on the type of speech from which they are extracted. For example, *nouns, prepositions*, and *function words* can be extracted reliably from free speech but not semi-spontaneous speech; the opposite is true for *verbs*. We have not found a simple explanation as to why some features are robust in one speech type but not the other. This is likely due to the combination and complex interaction of several factors, including ASR error patterns and differences in language use.

Table 8
Comparison of oracle and ASR-based quantitative measures, using two-tailed paired t-test of equal means for regular features and two-way repeated measures ANOVA for statistics features, both with $p = .05$ (▲: not sig. different before calibration; ▼: not sig. different after calibration).

			All	Free	Semi
DEN	1	Words/min	▼	▼	▼
	2	Phones/min	▼	▼	▼
	3	W	▼		
	4	OCW	▼	▲▼	
	5	{Words/utt}	▼	▼	▼
	6	{Phones/utt}	▼	▼	▼
	7	Nouns		▲▼	
	8	Verbs			▲▼
	9	Nouns/verb			
	10	Noun ratio			▲▼
	11	Light verbs			
	12	Determiners	▼	▼	▼
	13	Demonstratives	▲▼	▼	▲▼
	14	Prepositions		▲▼	
	15	Adjectives		▼	▲▼
	16	Adverbs	▲▼	▲▼	▲▼
	17	Pronoun ratio	▲▼	▲▼	▲▼
	18	Function words		▼	
DYS	19	Fillers/min	▼	▼	▼
	20	Fillers/word			▼
	21	Fillers/phone	▼	▼	▼
	22	Pauses/min	▲▼	▼	▼
	23	Long pauses/min		▼	▼
	24	Short pauses/min	▲▼		▼
	25	Pauses/word			
	26	Long pauses/word			
	27	Short pauses/word	▼	▲▼	
	28	{Seconds/pause}		▲▼	
LEX	29	Type–token ratio	▼	▼	
	30	{Freq/word}	▲	▲▼	▲▼
	31	{Img/word}	▲▼	▲▼	▲▼
	32	{AoA/word}	▲▼	▼	▲▼
	33	{Fam/word}	▲▼	▲▼	▲▼
	34	{Phones/word}			
POS-LM	35	{Bigram CE/utt}	▲▼	▲▼	▲▼
	36	{Trigram CE/utt}	▲▼	▼	▲▼
PVE	37	{PVE_1/utt}	▲▼	▼	▲▼
	38	{PVE_2/utt}	▲▼	▼	▲▼
	39	{PVE_3/utt}	▲▼		▲▼
	40	{PVE_4/utt}	▲▼	▼	▲▼
DTW	41	{Raw dist/word}	▲▼	▼	▲▼
	42	{Norm dist/word}	▼		
	43	{Segment/word}		▲▼	

Table 9
WAB-R AQ prediction results measured in Mean Absolute Error (MAE) and Pearson's correlation, broken down by **transcript type** (*Oracle, Auto, Calibrated*) and **feature extraction protocol** (*All, Free, Semi, Combined*). These two factors specify how the features are extracted (Section 6).

Protocol	MAE (Pearson's correlation)		
	Oracle	*Auto*	*Calibrated*
All	9.54 (.787)	9.90 (.776)	9.82 (.769)
Free	10.95 (.675)	11.89 (.625)	12.06 (.602)
Semi	9.00 (.799)	9.26 (.792)	**9.21 (.788)**
Combined	**8.86 (.801)**	**9.18 (.799)**	9.24 (.786)

7.3. WAB-R AQ Prediction

The WAB-R AQ prediction results, measured in MAE and Pearson's correlation, are summarized in Table 9. The results are partitioned based on two factors. First, **transcript type** (*Oracle, Auto, Calibrated*) specifies the source from which features are computed (Section 6.1). Second, **feature extraction protocol** (*All, Free, Semi, Combined*) indicates the type of speech used for feature extraction (Section 6.2). As

expected, *Oracle* features (i.e., those extracted from manual transcripts) result in more accurate predictions than *Auto* and *Calibrated* (i.e., ASR-based features). The best performance is obtained with the *Combined* and *Semi* protocols, suggesting that quantitative measures should be extracted for free and semi-spontaneous speech separately.

We perform two-way repeated measures ANOVA with Greenhouse-Geisser correction ($p = .05$) to further analyze the effect of transcript type and feature extraction protocol, using speaker-level prediction errors as the response variable. There is no statistically significant interaction between these two factors, $F(3.119, 1082.428) = 2.312$, $p = .072$. The effect of transcript type is significant, $F(1.300, 451.168) = 5.016$, $p = .017$. Using post-hoc multiple paired t-tests with Bonferroni correction ($p = .05$), we find that *Oracle* results in significantly lower errors than *Auto*, $t(347) = -3.208$, $p = .004$, as well as *Calibrated*, $t(347) = -3.362$, $p = .002$. This suggests that further improvement in aphasic speech recognition is needed to fully bridge the performance gap caused by ASR errors. Feature calibration helps bring ASR-derived measures closer to their oracle counterpart; however, it does not have significant impact on automatic prediction. Results obtained with *Calibrated* are not significantly different from *Auto*, $t(347) = .375$, $p = 1.0$. A possible explanation for this observation is that the change in feature magnitude resulting from calibration is relatively small, thus the final predictions remain largely unaffected. The effect of feature extraction protocol is also significant, $F(1.694, 587.911) = 25.770$, $p < .001$. Follow-up comparisons reveal that *Combined* and *Semi* results are not significantly different, $t(347) = -.455$, $p = 1.0$; meanwhile, these two significantly outperform *All* and *Free* ($p < .001$). Finally, we find that using only free speech for feature extraction performs significantly worse than all other protocols ($p < .001$), possibly due to the unstructured nature and relatively high WER associated with free speech.

Table 10 lists the prediction results of individual **feature groups** (*DEN, DYS, LEX, POS-LM, PVE, DTW*) under the *Combined* protocol. The best and worst features for AQ prediction are *LEX* and *DYS*, respectively. Similar to above, we use a two-way repeated measures ANOVA with Greenhouse-Geisser correction ($p = .05$) to analyze the effect of feature group and transcript type. There is no statistically significant interaction between these two factors, $F(5.550, 1925.784) = .687$, $p = .648$. While the effect of feature group is significant, $F(4.431, 1537.617) = 16.718$, $p < .001$, it is not so for transcript type, $F(1.206, 418.377) = 2.099$, $p = .144$. Post-hoc multiple paired t-tests with Bonferroni correction ($p = .05$) further show that *LEX* significantly outperforms all other features ($p < .05$), while *DYS* is significantly worse than the remaining groups ($p < .001$). Finally, we observe that the combination of all proposed measures is significantly better than any individual feature group ($p < .001$), suggesting that it is crucial to consider multiple aspects of a PWA's speech-language patterns to reliably predict WAB-R AQ.

The remaining analyses focus on the results of our best automated system (*Auto* transcript type and *Combined* protocol). Specifically, we are interested in speaker-level properties that can separate PWAs with low and high AQ prediction errors. Fig. 5 plots the ground-truth AQs against corresponding predicted labels. Intuitively, we expect the system to perform better on PWAs who have more accurate transcripts (i.e., lower WER) and less severe aphasia (i.e., higher AQ). However, we found limited evidence to support these hypotheses. Speaker-level prediction errors have a relatively weak Pearson's correlation of .162 with WERs and − .180 with AQs. Another hypothesis is that AQ values that are more representative of the training set are easier to predict. We measure the representativeness of an arbitrary AQ score based on its distance to the mean AQ of all training speakers (i.e., *label distance*). Lower label distance denotes higher representativeness and vice versa. The correlation between label distances and prediction errors is .302, which is higher compared to WER and AQ, but still does not indicate a clear relationship.

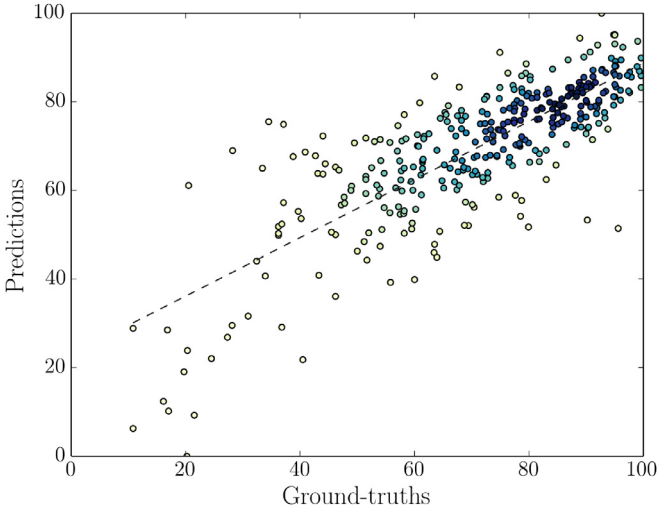

Fig. 5. AQ prediction plot. Darker color means higher density.

Individual characteristics are not correlated with system performance. However, we can partition PWAs into groups based on AQ prediction error (defined by MAE) to understand the general characteristics associated with accurate system performance. We divide the speakers into two groups, one with low MAE and one with high MAE, based on a predefined threshold. We then identify properties that are statistically significantly different across these two groups (Welch's t-test of equal means, $p = .05$). These properties could be used in the future as a preliminary screen to identify PWAs who will benefit from this type of system. We define our threshold based on AQ test-retest reliability. Researchers demonstrated that the average deviation in AQ when rescoring PWAs who were stable at the time of initial testing is 5.316 (Shewan and Kertesz, 1980). In other words, automatic AQ prediction can be considered satisfactory if the MAE does not exceed this value. As shown in Fig. 6, this threshold results in 237 PWAs in the *Low Errors* group with a MAE of 5.30 ± 3.11, and 111 PWAs in the *High Errors* group with a MAE of 17.46 ± 6.86. Further analysis reveals that PWAs in the *Low Errors* group have significantly lower WER, higher AQ, and smaller label distance (Table 11). This suggests that we can roughly estimate the range of prediction errors given a PWA's WER level and/or current AQ score.

Table 10
Performance breakdown of individual feature groups (Section 5) under the *Combined* protocol.

Group	MAE (Pearson's correlation)		
	Oracle	*Auto*	*Calibrated*
DEN	11.06 (.676)	11.46 (.623)	11.47 (.626)
DYS	14.16 (.429)	14.45 (.422)	14.31 (.429)
LEX	**10.11 (.744)**	**10.44 (.733)**	**10.57 (.722)**
POS-LM	11.71 (.629)	11.72 (.645)	11.75 (.645)
PVE	11.73 (.615)	11.94 (.591)	11.96 (.587)
DTW	12.43 (.583)	12.45 (.547)	12.17 (.569)

8. Conclusion and future work

In this work, we perform one of the first large-scale studies on automatic quantitative analysis of spontaneous aphasic speech. Our acoustic modeling method based on deep BLSTM-RNN and utterance-level i-vectors sets a new benchmark for aphasic speech recognition on AphasiaBank. We show that with the help of feature calibration, our proposed quantitative measures are robust against ASR errors and can potentially be used to assist with clinical diagnosis and/or progress monitoring. Finally, we demonstrate the efficacy of these measures by

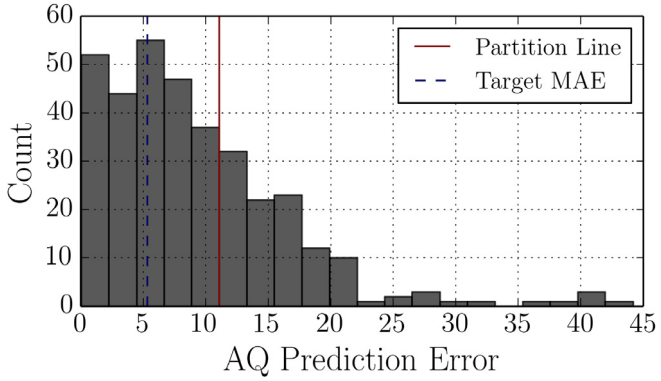

Fig. 6. Histogram of WAB-R AQ prediction errors. The partition line divides PWAs into two groups, *Low Errors* and *High Errors*. The Mean Absolute Error (MAE) of the first group is approximately 5.316, the natural AQ variation for the same subject between different testing sessions.

Table 11
Comparison of PWAs with low and high AQ prediction errors. Values shown are mean (standard deviation). *t*-statistics and two-tailed *p*-values are reported for Welch's t-test of equal means.

Property	Low Errors	High Errors	*t*	*p*
WER	38.78 (14.09)	44.06 (17.09)	− 2.824	.005
AQ	72.85 (17.03)	67.50 (23.27)	2.160	.032
Label Distance	20.16 (7.21)	24.64 (9.10)	− 4.542	< .001

using them to predict WAB-R AQ with promising accuracy. The results and techniques presented in our work will help make automated spontaneous speech analysis for aphasia more feasible, enabling SLPs to quickly and reliably analyze a large amount of speech data that would otherwise be too time consuming to inspect manually.

For future work, we plan to investigate acoustic and language model personalization methods to further improve ASR performance on aphasic speech. This will help increase the reliability of ASR-based quantitative measures as well as reduce the gap between oracle and automatic performance in WAB-R AQ estimation. We also plan to test and further refine our system in realistic clinical applications to determine the full extent of automated aphasic speech assessment.

References

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

Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., Farrajota, L., Martins, I.P., 2013. Automatic word naming recognition for an on-line aphasia treatment system. Comput. Speech Lang. 27 (6), 1235–1248.

Aniol, M., 2012. Tandem features for dysarthric speech recognition. Master's thesis Edinburgh University, United Kingdom.

Association, N. A., 2016. Aphasia. http://www.aphasia.org/. Accessed: 2016-11-12.

Basso, A., 2003. Aphasia and Its Therapy. Oxford University Press.

Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. O'Reilly Media.

Breedin, S.D., Saffran, E.M., Schwartz, M.F., 1998. Semantic factors in verb retrieval: an effect of complexity. Brain Lang. 63 (1), 1–31.

Brysbaert, M., New, B., 2009. 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. Behav. Res. Methods 41 (4), 977–990.

Christensen, H., Aniol, M.B., Bell, P., Green, P., Hain, T., King, S., Swietojanski, P., 2013. Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France.

Christensen, H., Casanueva, I., Cunningham, S., Green, P., Hain, T., 2014. Automatic selection of speakers for improved acoustic modelling: recognition of disordered speech with sparse data. IEEE Workshop on Spoken Language Technology (SLT). South Lake Tahoe, NV, USA.

Christensen, H., Cunningham, S., Fox, C., Green, P., Hain, T., 2012. A comparative study of adaptive, automatic recognition of disordered speech. Proc. of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH). Portland, OR, USA.

Christensen, H., Green, P., Hain, T., 2013. Learning speaker-specific pronunciations of disordered speech. Proc. of the 14th Annual Conference of the International Speech Communication Association (INTERSPEECH). Lyon, France.

Davis, G.A., 2006. Aphasiology: Disorders and Clinical Practice, 2. Pearson.

Fergadiotis, G., Wright, H., 2011. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. Aphasiology 25 (11), 1414–1430.

Forbes, M.M., Fromm, D., MacWhinney, B., 2012. Aphasiabank: a resource for clinicians. Seminars in Speech and Language. Vol. 33. NIH Public Access, pp. 217.

Fox, S., Armstrong, E., Boles, L., 2009. Conversational treatment in mild aphasia: a case study. Aphasiology 23 (7–8), 951–964.

Fraser, K., Rudzicz, F., Graham, N., Rochon, E., 2013. Automatic speech recognition in the diagnosis of primary progressive aphasia. Proc. of the 4th Workshop on Speech and Language Processing for Assistive Technologies. Grenoble, France.

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

Fraser, K.C., Meltzer, J.A., Graham, N.L., Leonard, C., Hirst, G., Black, S.E., Rochon, E., 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. Cortex 55, 43–60.

Gilhooly, K.J., Logie, R.H., 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. Behav. Res. Methods Instrum. 12 (4), 395–427.

Goodglass, H., Kaplan, E., Barresi, B., 2000. Boston Diagnostic Aphasia Examination, 3. Philadelphia: Lippincott, Williams & Wilkins.

Grande, M., Hussmann, K., Bay, E., Christoph, S., Piefke, M., Willmes, K., Huber, W., 2008. Basic parameters of spontaneous speech as a sensitive method for measuring change during the course of aphasia. Int. J. Lang.Commun. Disord. 43 (4), 408–426.

Helm-Estabrooks, N., Albert, M.L., Nicholas, M., 2013. Manual of Aphasia and Aphasia Therapy, 3. Pro-Ed.

Jaecks, P., Hielscher-Fastabend, M., Stenneken, P., 2012. Diagnosing residual aphasia using spontaneous speech analysis. Aphasiology 26 (7), 953–970.

Katz, R.C., Hallowell, B., Code, C., Armstrong, E., Roberts, P., Pound, C., Katz, L., 2000. A multinational comparison of aphasia management practices. Int. J. Lang.Commun. Disord. 35 (2), 303–314.

Kertesz, A., 2006. The Western Aphasia Battery - Revised. Texas: Harcourt Assessments.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR).

Le, D., Licata, K., Mercado, E., Persad, C., Mower Provost, E., 2014. Automatic analysis of speech quality for aphasia treatment. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy.

Le, D., Licata, K., Mower Provost, E., 2017. Automatic paraphasia detection from aphasic speech: apreliminary study. Interspeech. Stockholm, Sweden.

Le, D., Licata, K., Persad, C., Mower Provost, E., 2016. Automatic assessment of speech intelligibility for individuals with aphasia. IEEE Trans. Audio, Speech, Lang. 24, 2187–2199.

Le, D., Mower Provost, E., 2016. Improving automatic recognition of aphasic speech with AphasiaBank. Interspeech. San Francisco, USA.

Le, D., Provost, E.M., 2014. Modeling pronunciation, rhythm, and intonation for automatic assessment of speech quality in aphasia rehabilitation. Proc. of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore.

Lee, A., Glass, J., 2012. A comparison-based approach to mispronunciation detection. IEEE Spoken Language Technology Workshop (SLT). pp. 382–387.

Lee, A., Glass, J.R., 2013. Pronunciation assessment via a comparison-based system. ISCA International Workshop on Speech and Language Technology in Education (SLaTE). Grenoble, France, pp. 122–126.

Lee, A., Zhang, Y., Glass, J.R., 2013. Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Vancouver, BC, Canada, pp. 8227–8231.

Lee, T., Kong, A., Chan, V., Wang, H., 2013. Analysis of auto-aligned and auto-segmented oral discourse by speakers with aphasia: apreliminary study on the acoustic parameter of duration. Procedia - Social and Behavioral Sciences. Vol. 94. pp. 71–72.

Lee, T., Kong, A., Lam, W., 2015. Measuring prosodic deficits in oral discourse by speakers with fluent aphasia. Front. Psychol.(47).

Lee, T., Liu, Y., Huang, P., Chien, J., Lam, W., Yeung, Y., Law, T., Lee, K., Kong, A., Law, S., 2016. Automatic speech recognition for acoustical analysis and assessment of cantonese pathological voice and speech. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China.

MacWhinney, B., 2000. The Childes Project: Tools for Analyzing Talk: Vol. II: The Database. Mahwah.

Macwhinney, B., Fromm, D., Forbes, M., Holland, A., 2011. Aphasiabank: methods for studying discourse. Aphasiology 25 (11), 1286–1307.

Mayer, J., Murray, L., 2003. Functional measures of naming in aphasia: word retrieval in confrontation naming versus connected speech. Aphasiology 17 (5), 481–497.

Mengistu, K.T., Rudzicz, F., 2011. Comparing humans and automatic speech recognition systems in recognizing dysarthric speech. Proceedings of the 24th Canadian Conference on Artificial Intelligence. St. John's, Canada, pp. 291–300.

Miller, N., Willmes, K., De Bleser, R., 2000. The psychometric properties of the english language version of the Aachen aphasia test (EAAT). Aphasiology 14 (7), 683–722.

Mustafa, M.B., Rosdi, F., Salim, S.S., Mughal, M.U., 2015. Exploring the influence of

general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. Expert. Syst. Appl. 42 (8), 3924–3932.
Pakhomov, S.V., Smith, G.E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., Knopman, D.S., 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. Cognit. Behav. Neurol. 23 (3), 165–177.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Peintner, B., Jarrold, W., Vergyri, D., Richey, C., Tempini, M.G., Ogar, J., 2008. Learning diagnostic models using speech and language measures. Proc of the 30th Annual International IEEE EMBS Conference. Vancouver, British Columbia, Canada.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The kaldi speech recognition toolkit. Automatic Speech Recognition and Understanding (ASRU). Hawaii, USA.
Prins, R., Bastiaanse, R., 2004. Analysing the spontaneous speech of aphasic speakers. Aphasiology 18 (12), 1075–1091.
Roark, B., Mitchell, M., Hosom, J., Hollingshead, K., Kaye, J., 2011. Spoken language derived measures for detecting mild cognitive impairment. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2081–2090. http://dx.doi.org/10.1109/TASL.2011.2112351.
Sharma, H.V., Hasegawa-Johnson, M., 2010. State-transition interpolation and MAP adaptation for HMM-based dysarthric speech recognition. HLT/NAACL Workshop on Speech and Language Processing for Assistive Technologies (SLPAT). Los Angeles, CA, USA, pp. 72–79.
Sharma, H.V., Hasegawa-Johnson, M., 2013. Acoustic model adaptation using in-domain background models for dysarthric speech recognition. Comput. Speech Lang. 27 (6), 1147–1162.
Shewan, C.M., Kertesz, A., 1980. Reliability and validity characteristics of the western aphasia battery (WAB). J. Speech Hearing Disord. 45 (3), 308–324.
Simons-Mackie, N., Raymer, A., Armstrong, E., Holland, A., Cherney, L., 2010. Communication partner training in aphasia: a systematic review. Arch. Phys. Med. Rehabil. 91 (12), 1814–1837.
Stadthagen-Gonzalez, H., Davis, C.J., 2006. The bristol norms for age of acquisition, imageability, and familiarity. Behav. Res. Methods 38 (4), 598–605.
Stolcke, A., Zheng, J., Wang, W., Abrash, V., 2011. SRILM at sixteen: update and outlook. Automatic Speech Recognition and Understanding (ASRU).
Tepperman, J., Stanley, T., Hacioglu, K., Pellom, B., 2010. Testing suprasegmental english through parroting. Proc. of Speech Prosody. Chicago, IL, USA.
Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun. 30 (23), 95–108. http://dx.doi.org/10.1016/S0167-6393(99)00044-8.