Towards Automatic Speech-Language Assessment for Aphasia Rehabilitation

by

Duc Le

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2017

Doctoral Committee:

Assistant Professor Emily Kaplan Mower Provost, Chair
Dr. Christian Fügen, Facebook Inc.
Professor Alfred O Hero III
Associate Professor Honglak Lee
Associate Professor Carol Catherine Persad

Duc Le

ducle@umich.edu

ORCID iD: 0000-0001-9490-2563

I dedicate this dissertation to my parents, who have always believed in me and supported me every step of the way.

ACKNOWLEDGMENTS

I woud like to thank my advisor, Dr. Emily Mower Provost, for her guidance and support. She has helped me grow tremendously during my time in graduate school, both as a person and a researcher.

Many thanks to my committee members, Dr. Fügen, Dr. Hero, Dr. Lee, and Dr. Persad, for their helpful comments and suggestions.

Thanks also go to members of the University of Michigan Aphasia Program for introducing me to aphasia and helping me find my research direction. Special thanks go to Keli Licata, who has been an amazing collaborator with consistently insightful comments and observations.

Finally, I would like to thank my friends and colleagues in the CHAI Lab, the University of Michigan Table Tennis Club, and many more, for making my time here so interesting and enjoyable.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Aphasia is a common neurological disorder that can severely impact a person's communication abilities. Speech-based technology has the potential to reinforce traditional aphasia therapy through the development of automatic speech-language assessment systems. Such systems can provide clinicians with supplementary information to assist with progress monitoring and treatment planning, and can provide support for on-demand auxiliary treatment. However, current technology cannot support this type of application due to two major limitations. First, the majority of speech-language assessment techniques assume the availability of manually labeled transcripts, which are time consuming to obtain and typically not available in real-world clinical applications. Second, automatic speech recognition (ASR) traditionally has poor performance on aphasic speech, resulting in inaccurate transcripts that prevent the automation of these techniques.

The focus of this dissertation is on the development of computational methods that can accurately assess aphasic speech across a range of clinically-relevant dimensions without the need for manual transcripts. The dissertation is organized into three parts:

- **Part I:** The first part focuses on novel techniques for assessing qualitative aspects of intelligibility in *constrained* aphasic speech. In this problem setup, speech production occurs in controlled environments, lexical content is restricted, and the target prompt for each utterance is known. While the speech-language impairments associated with aphasia often prevent exact verbalization of the prompts, this constraint greatly simplifies ASR and allows for more accurate transcript generation. We show that transcripts for constrained aphasic speech can be generated automatically with modified forced alignment in place of traditional ASR. These transcripts, combined with novel features that capture a speaker's pronunciation, rhythm, and intonation patterns, yield prediction results that are comparable to those of human evaluators.

- **Part II:** The methods presented previously rely on the prior availability of target prompts. This assumption limits the applicability of these methods to *unconstrained* speech, in which target prompts are not available. The majority of speech produced in normal everyday interaction is unconstrained, thus necessitating the development of robust assessment techniques for this type of speech. Automatic assessment of unconstrained speech is often reliant on ASR output; at the same time, ASR performance on aphasic speech is traditionally poor. Based on this need, the second part of this dissertation improves speech recognition accuracy for speakers with aphasia to lay the foundation for automated assessment of unconstrained aphasic speech. The proposed acoustic modeling techniques, which focus on adapting pre-trained acoustic models to small datasets and leveraging auxiliary input features to mitigate speaker variability, lead to significant improvement in aphasic speech recognition.

- **Part III:** The final part of the dissertation investigates the efficacy of ASR-based analysis across a range of clinically-relevant tasks, including automatic paraphasia (naming error) detection, extraction of clinically-motivated quantitative measures, and estimation of Aphasia Quotient, a standard measure of aphasia severity, from unconstrained aphasic speech. We propose a calibration method that enables information density, dysfluency, and lexical features, many of which have important clinical implications, to be reliably extracted from ASR output. We demonstrate that these ASR-based features can be used to accurately predict Aphasia Quotient.

The unification of the methods and results presented in this work helps enable robust automated technologies for accurately recognizing and assessing aphasic speech without human intervention. We conclude the dissertation with future directions that target the development of specialized ASR models for aphasia and the deployment of our proposed techniques in real-world clinical applications.

CHAPTER 1

Introduction

Aphasia is an acquired chronic language disorder resulting in a loss of language skills that generally arises from focal brain damage to the left cerebral hemisphere [13]. In the US, there are approximately two million people with aphasia and more than 180,000 acquire it every year due to brain injury, most commonly from a stroke [5]. The type and severity of language deficits depends on the size and location of the brain lesion. Individuals typically exhibit expressive and/or receptive language deficits. Those with expressive (non-fluent) aphasia typically have difficulties producing language, with minimal word production (referred to as telegraphic speech), while generally retaining the ability to comprehend most spoken language. They may exhibit difficulties with comprehension of more complex language. Others with receptive (fluent) aphasia typically speak fluently, but often with little content or meaning, while demonstrating difficulties with spoken language comprehension. Common verbal expression deficits in both types of aphasia include phonemic errors and speech dysfluencies [19, 160]. All persons with aphasia (PWAs) have problems with word-finding (anomia) to some degree, and most also have reading and writing impairments. Further, many individuals with aphasia experience motor speech production deficits, such as apraxia and/or dysarthria, which complicate recovery [6]. A PWA's verbal output may appear impaired due to language problems such as word retrieval and sentence formulation difficulties, or speech production issues caused by apraxia, dysarthria, or both. The speech-language deficits associated with aphasia impact one's ability to communicate effectively, making social interaction difficult and frustrating. This results in feelings of social isolation, loss of autonomy, and loneliness, among others [23, 150].

The most effective forms of aphasia treatment are long-term intensive targeted therapies with Speech-Language Pathologists (SLPs) [12, 13, 103]. Previous research suggests that high-intensity treatments are more beneficial than low-intensity ones [12, 62, 113]. In addition, treatments must meet a minimum level of frequency and intensity to yield positive effects [131]. Significant improvements from aphasia are typically observed in the acute post-onset period; however, recovery can continue indefinitely with appropriate treatment and/or dynamic interactions with one's environment [62]. Unfortunately, many do not have consistent access to individualized speech-language therapy services due to the high cost burden, lack of available long-term options, and/or lack of local treatment options [120]. As a result, many PWAs only participate in short-term and/or low-intensity therapeutic care, often administered in hospital environments, and they do not receive sufficient treatment for long-term progress [104]. These factors highlight a need for auxiliary sources of treatment and increased efficiency in assessment procedures for SLPs.

Speech-based technology has the potential to fill these gaps by administering clinically-relevant speech-language feedback to PWAs automatically, as well as providing SLPs with diagnostic and progress monitoring tools to help guide the treatment process. The market has recognized this need. In the last several years, the number of commercially available aphasia programs and applications has increased. These software tools allow individuals to practice their speech-language skills, but often do not provide the feedback necessary for self-assessment and error correction [62]. Practice without feedback may reinforce errors rather than facilitating improvement. Some of these applications allow PWAs to send their speech recordings to SLPs for further analysis. However, SLPs in many settings have high productivity expectations and limited time outside of direct patient contact to manually examine and analyze a large amount of speech data.

The long-term vision of this dissertation is to develop systems that can accurately assess aphasic speech across a range of clinically-relevant dimensions and deliver meaningful feedback that will help maintain and track the recovery progress of PWAs over time. Such systems will help facilitate more efficient assessment pipelines for SLPs through the ability to quickly process large amounts of speech samples that would otherwise be prohibitively time consuming to perform manually. These systems have the potential to improve the well-being of PWAs by complementing and extending traditional aphasia therapy.

1.1 Problem Statement and Methods

We argue that a successful speech-language assessment system for aphasia requires two primary abilities: (1) to transcribe speech content without human intervention and (2) to accurately estimate clinically-relevant measures from aphasic speech. This dissertation focuses on the development of novel computational methods to automate these abilities, with an encompassing goal of enabling reliable fully automated speech-based technology to support aphasia rehabilitation.

1.1.1 Aphasic Speech Transcription

Automatic transcription refers to the process of estimating the lexical content of a given speech sample, as well as identifying the precise timing of acoustic units (i.e., words, syllables, and phones). An essential component of a speech-language assessment system is the ability to accurately extract clinically-relevant measures (i.e., features) from a PWA's speech to support diagnosis and progress monitoring. Transcripts enable the extraction of a set of lexical and linguistic features that play an important role in the study of aphasia, such as vowel duration, filler frequency, part-of-speech usage patterns, lexical diversity, and vocabulary range.

Traditional techniques in ASR relied on a combination of Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). The field has experienced major breakthroughs in recent years, primarily due to massive datasets and advances in deep learning [30, 31, 54, 64, 110, 135, 136, 139]. However, disordered speech recognition is still mostly constrained to the traditional HMM-GMM model. This is due to a variety of factors including data scarcity, atypical speech input, and high speaker variability. These factors severely impact ASR performance on disordered speech in general and aphasic speech in particular, except in applications with highly constrained lexical content. In this work, we hypothesize that:

1. Deep learning techniques can achieve significant improvement over traditional HMM-GMM approaches, even given limited training data for aphasic speech, by adapting models learned on large external corpora to a smaller targeted dataset.

2. Aphasic speech recognition will benefit from speaker-independent adaptation, methods which help the model generalize better to unseen speakers, due to the high degree of speaker variability associated with aphasia.

We present experiments that evaluate these hypotheses in Chapter 4. We first investigate the efficacy of out-of-domain adaptation, in which an acoustic model initially trained on a large amount of data is adapted to a smaller corpus. We show that with this technique, deep learning-based models can significantly outperform HMM-GMMs on a small dataset with only two hours of speech. In addition, we demonstrate that ASR performance on aphasic speech is greatly improved with utterance-level i-vectors, an auxiliary input feature that captures speaker and other sources of variations.

1.1.2 Estimation of Clinically-Relevant Measures

The high-level goal of automated speech assessment is to estimate the characteristics of aphasic speech that are relevant to aphasia rehabilitation. These may include qualitative assessment of human evaluators regarding a PWA's speech, such as measures of pronunciation, fluidity, and intonation. Accurate prediction of these properties will provide PWAs with feedback and potentially increase the efficacy of unsupervised speech-language exercises. Other targets for assessment include objective quantitative measures that can be used by SLPs to better understand the recovery progress of PWAs, such as rate of speech, lexical diversity, and mean length of utterances. These measures, which are often time consuming to produce manually, will provide SLPs with additional information for treatment planning. The primary challenges in developing a speech-language assessment system are the handling of potentially incorrect transcripts, especially those generated automatically, and the engineering of features that capture the target qualitative measures. We hypothesize that:

1. Given human-labeled transcripts and novel feature engineering, it is possible to achieve human-level performance in a range of assessment tasks on aphasic speech.

2. Automatic transcription can replace manual transcripts in some of these tasks with minimal impact on system performance.

We evaluate these hypotheses across various assessment tasks and speech types in Chapter 3, 5, and 6. Chapter 3 investigates novel computational methods for assessing qualitative aspects of intelligibility in constrained aphasic speech. Chapter 5 tackles the problem of automatic paraphasia (naming error) detection. Finally, Chapter 6 tests these hypotheses through the extraction of clinically-relevant quantitative measures and the estimation of aphasia severity from unconstrained aphasic speech.

1.2 Background and Related Work

1.2.1 Methods for Studying and Assessing Aphasia

Early work on the acoustical analysis of aphasic speech was limited to manual inspection of speech waveforms and spectrograms on a small number of short utterances [160]. Lee et al. proposed the use of HMM-based forced alignment to speed up the transcription process and enable the analysis of larger amount of Cantonese aphasic speech [87,88]. They found that compared to healthy speech, aphasic speech contains fewer words, longer pauses, and higher number of continuous chunks, with fewer words per chunk [87]. Further, aphasic speech exhibits different intonation patterns [88]. The limitation of their works lies in the requirement for manual transcripts and the mismatched acoustic model.

Several previous works have tackled the problem of processing aphasic speech for therapeutic and diagnostic purposes [1, 2, 41–43, 69, 122, 141]. Abad et al. [1, 2] used keyword spotting to recognize phrases spoken by PWAs during word naming exercises. Their targeted users are individuals with aphasia who have word-finding problems but no difficulties with auditory comprehension or speech-language production. In contrast, the typical PWA tend to have difficulties in both. Further, their work targeted a relatively restricted type of speech (single words) with limited applicability outside of their proposed application. Fraser et al. combined transcript and low-level acoustic features to classify between two subtypes of primary progressive aphasia (PPA) [42, 43]. Their work relied on fine-grained expertly labeled transcripts, which are expensive and time-consuming to create. Their follow-up work attempted to generate these transcripts with ASR; however, the poor recognition performance limited their analysis to simulated ASR output with preset error levels [41]. Peintner et al. proposed speech and language features to distinguish between three types of frontotemporal lobar degeneration, including progressive non-fluent aphasia [122]. They used an ASR system to automatically transcribe speakers' spontaneous responses to the Western Aphasia Battery assessment test [72]. Their ASR system was trained only on healthy speech with mismatched demographics, which led to high recognition error. In addition, they did not analyze the effect of ASR errors on feature extraction.

1.2.2 Methods for Studying and Treating Apraxic Speech

Apraxia of Speech (AOS) results from impairments to motor networks, while aphasia is related to impairments in language networks. AOS frequently co-occurs with aphasia. AOS is characterized by errors at the phoneme-level, which impact both consonants and vowels

[19, 55]. It is also characterized by sound substitutions, impaired fluency, atypical prosody, and sound distortions [34, 159]. AOS causes the production of speech described as trial-and-error groping, resulting in frequent restarts and repetitions of sounds and syllables [56]. AOS also commonly affects the temporal prosody of speech, resulting in slow speech with prolonged vowels and consonants [118].

There has been limited work exploring quantitative approaches to understanding the diagnosis and assessment of AOS. Haley and colleagues investigated the validity and reliability of two different quantification strategies to characterize the type and severity of errors seen in AOS [56]. In particular, the authors were interested in comparing clinician rated scales that rely more on clinical judgment to produce operationalized-based approaches that focus on the quantification of specifically defined errors. A subset of the metrics that they introduced include: segmental substitution (phone-level substitution errors), segmental distortions (mispronunciations of phones, not substitutions), revision and repetition of sounds, and unit durations [56]. Results showed more consistent and reliable coding using the operationally based approach. Given the high co-occurrence rate of aphasia and AOS, capturing these metrics automatically will be beneficial for the analysis and assessment of aphasia. However, additional metrics must be developed to better capture the language impairments associated with aphasia.

1.2.3 Methods for Studying and Treating Dysarthric Speech

Research in automatic modeling of disordered speech has historically focused on dysarthric speech, commonly seen in Parkinson's Disease, Stroke, Cerebral Palsy, and Amyotrophic Lateral Sclerosis (ALS) [157]. Some types of aphasia may be accompanied by dysarthria [62], further emphasizing the importance of this research. Dysarthria is a motor speech disorder, often caused by neurological injury [25], which affects muscles involved in speech production, such as the lips, tongue and vocal folds. There have been many studies investigating how ASR technologies can be adapted for use by individuals with dysarthria [27,28,58,60,132]. Christensen et al. introduced techniques to model dysarthric speech by bootstrapping models with healthy speech, collected from the AMI Meeting corpus and TED Talk dataset [25]. Research has demonstrated that finite state transducers could be effectively used to correct pronunciation errors in dysarthric speech [111,145].

There has also been work investigating how ASR technologies can be used to provide speech feedback and training [76]. Saz et al. demonstrated that ASR-based technologies could be used for speech and language therapy, focusing on children and young adults with neuromuscular disorders [141]. Their system obtained comparable performance to human experts. Hawley and colleagues introduced methods to provide speech training using ASR technologies [59]. They provided software to five individuals with dysarthria and found that three of the participants showed improvement over the three-week trial period. However, they also found that one of the challenges with ASR-based technologies is that the technology tended to emphasize longer-duration phonemes (e.g., vowels) as a source for potential improvement, rather than the production of challenging and rapidly transitioning consonants, areas in which an individual may actually experience the most deficits [61]. Research has demonstrated that automatic speech processing tools can be employed to assess the pronunciation and intelligibility of disordered speech. Yin and colleagues demonstrated techniques to identify pronunciation errors given a constrained target sentence using confidence measures [164], whereas Christensen et al. demonstrated methodologies to automatically learn mispronunciations of dysarthric speech [25]. Ferrier et al. demonstrated the link between the intelligibility of speech and a subject's ability to use conventional speech modeling tools [38]. The primary challenge in adapting these methods to the targeted application domain is that they primarily focus on speech production issues, whereas aphasia is first and foremost a language disorder with possible speech production impairments due to concomitant motor control disorders.

1.2.4 Methods in Pathological Speech Assessment

In this section, we review prominent methods in pathological speech assessment that are not tied to specific disorders. Previous works in this area used Word Error Rate (WER) from an ASR system evaluated against predefined speech prompts to estimate a speaker's intelligibility [97, 128]. The primary challenge of applying this method is the requirement of the target prompts, which are not available for spontaneous speech. Other studies estimated intelligibility by extracting speaker-level phonemic and phonological features from a phonetically diverse set of utterances [106, 157]. Kim et al. used sentence-level prosody, voice quality, and pronunciation features for intelligibility classification [74]. Finally, Berisha et al. proposed a method to select acoustic features that correlate with SLPs' ordinal ratings of dysarthric speech [11]. A common drawback of these works is that they typically assume the availability of manually labeled transcripts, an unrealistic requirement in most clinical applications. Some works focus exclusively on acoustic features and therefore do not require transcripts; however, such approaches prevent the extraction and analysis of language features, which are crucial for aphasia.

1.2.5 Methods in Computer Aided Language Learning

Methods in Computer Aided Language Learning (CALL) focus on quantifying the differences between native and non-native speech, which can be useful for separating healthy and aphasic speech. Previous work in CALL mostly targeted pronunciation modeling, utilizing variants of Goodness of Pronunciation (GOP) [66, 67, 163], template-based comparison [84, 86, 117], extension of traditional ASR [91, 126], among other methods [162]. The GOP metric, first proposed by Witt and Young, is derived from the log posterior score of a HMM-GMM acoustic model [163]. Follow-up work investigated GOP extraction using HMM-DNN [66] and optimizing GOP with a discriminative training objective function [67]. Template-based methods involve comparison of word-level posteriorgrams extracted by a HMM-DNN acoustic model [84, 86]. Nicolao et al. extended these methods to enable phone-level pronunciation error detection [117]. Qian et al. proposed to augment canonical recognition networks to detect and diagnose mispronunciation [126]. Li et al. developed a unified framework for detecting and diagnosing mispronunciation using DNNs [91]. Similar to [126], their proposed system is based on ASR, but is simpler and more flexible. Other work in this area focused on high-level assessment of a subject's overall reading ability instead of token-level assessment [15–18]. Finally, Tepperman et al. proposed Pairwise Variability Error (PVE), a metric for highlighting the differences in rhythm between native and non-native speakers [154]. The majority of existing research in CALL focus on modeling a speaker's pronunciation, which by itself does not fully characterize the speech-language characteristics associated with aphasia. In addition, methods in this area typically assume that the speaker always reproduces the target prompts perfectly. This is often not true for aphasic speech, in which various speech-language impairments may lead to deviations from the target prompts. As a result, modifications to these methods are required to account for the prompt mismatches.

1.2.6 ASR Overview

In ASR, the acoustic signal of an utterance is represented by a feature vector (i.e., *observation*) $\mathbf{o} = (o_1, \ldots, o_T)$, a sequence of T observations. A potential transcript is denoted as $\mathbf{w} = (w_1, \ldots, w_K)$, a sequence of K words. The goal of ASR is to find the optimal transcript $\mathbf{w}^*$ that maximizes the probability $P(\mathbf{w}|\mathbf{o})$:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} P(\mathbf{w}|\mathbf{o}) = \arg\max_{\mathbf{w}} p(\mathbf{o}|\mathbf{w})P(\mathbf{w}) \tag{1.1}$$

Here, $p(\mathbf{o}|\mathbf{w})$ is determined by an *acoustic model* and $P(\mathbf{w})$ is estimated by a *language model* (e.g., n-gram). In practice, the acoustic model is not normalized and the recognition problem is typically formulated as:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \log p(\mathbf{o}|\mathbf{w}) + \alpha \log P(\mathbf{w}) + \beta|\mathbf{w}| \tag{1.2}$$

where α and β are empirically determined constants denoting the language model weight and word insertion penalty, respectively.

A standard modeling assumption in ASR is that each word w can be represented as a sequence of basic sounds (i.e., *phones*) $\mathbf{q}^{(w)} = (q_1, \ldots, q_n)$. Let $\mathbf{q}$ be a possible phone sequence for the word sequence, $\mathbf{w}$. We can then rewrite $p(\mathbf{o}|\mathbf{w})$ as:

$$p(\mathbf{o}|\mathbf{w}) = \sum_{\mathbf{q}} p(\mathbf{o}|\mathbf{q}) P(\mathbf{q}|\mathbf{w}) \tag{1.3}$$

where $P(\mathbf{q}|\mathbf{w})$ is given by a *pronunciation model*. In monophone modeling, each phone q is represented by a HMM, typically a linear left-to-right model with 3–5 hidden states[1]. Under this model, each observation, o_i, is emitted by a HMM state, s_i, where the emission probability $p(o_i|s_i)$ is governed by the output observation distribution $b_{s_i}(o_i)$, and the transition between states $P(s_i|s_j)$ is determined by the transition probability $a_{s_i s_j}$.

Let $\mathbf{s} = (s_0, s_1, \ldots, s_T, s_{T+1})$ be a possible state sequence obtained from the composite HMM for the phone sequence $\mathbf{q}$ and observation sequence $\mathbf{o}$, where s_0 and s_{T+1} are the non-emitting start and end states, respectively. $p(\mathbf{o}|\mathbf{q})$ can now be computed as:

$$p(\mathbf{o}|\mathbf{q}) = \sum_{\mathbf{s}} a_{s_0 s_1} \prod_{t=1}^{T} b_{s_t}(o_t) a_{s_t s_{t+1}} \tag{1.4}$$

The performance of an ASR system is determined in a large part by how the emission probability $b_{s_t}(o_t)$ is calculated. In this section, we review three prominent methods for modeling emission probabilities.

[1] In large-vocabulary speech recognition, a phone is commonly represented by a set of HMMs accounting for different left and right context. This method, typically referred to as triphone modeling, helps capture co-articulation and usually gives better performance than monophone models. Hidden states in triphone HMMs are called *senones*. The basic mathematical formulations of these two methods are largely similar.

1.2.6.1 Gaussian Mixture Model (GMM)

Emission probabilities are traditionally modeled with GMMs. Under this model, each HMM hidden state, s_t, is associated with a mixture of multivariate Gaussian densities:

$$b_{s_t}(o_t) = \sum_{m=1}^{M^{(s_t)}} c_m^{(s_t)} \mathcal{N}(o_t; \mu_m^{(s_t)}, \Sigma_m^{(s_t)}) \tag{1.5}$$

where $M^{(s_t)}$ is the number of mixture components, $c_m^{(s_t)}$ is the weight of the m-th component, $1 \leq m \leq M^{(s_t)}$, and $\mathcal{N}(\cdot; \mu_m^{(s_t)}, \Sigma_m^{(s_t)})$ is a multivariate Gaussian with mean $\mu_m^{(s_t)}$ and covariance matrix $\Sigma_m^{(s_t)}$.

GMMs can model probability distributions to an arbitrary level of accuracy given enough components, and are fairly easy to train using Expectation-Maximization (EM). However, a major disadvantage of GMMs is that they cannot effectively capture information over a large number of consecutive feature frames [64]. In addition, GMMs typically assume diagonal covariance matrices due to computational issues. This necessitates uncorrelated input features, which prevent GMMs from modeling feature interaction.

1.2.6.2 Deep Neural Network (DNN)

DNNs recently emerged as an alternative to GMMs that are capable of modeling reasonably large windows of frames as well as feature interaction [30, 31, 64, 110]. The application of DNN to acoustic modeling is based on the reformulation of the emission probability $p(o_t|s_t)$ according to Bayes' rule:

$$p(o_t|s_t) \propto \frac{P(s_t|o_t)}{P(s_t)} \tag{1.6}$$

where $P(s_t|o_t)$ is the *posterior probability* and $P(s_t)$ is the *prior probability* of state s_t.

Instead of estimating the emission probability directly, DNNs estimate the posterior probability using a conventional multilayer perceptron (MLP) with several hidden layers.

For a DNN with $L+1$ layers, where layer 0 is the input layer, layers 1 to $L-1$ are the hidden layers, and layer L is the output layer, the output of the first L layers can be computed as:

$$v^l = f(z^l) = f(W^l v^{l-1} + b^l), \text{ for } 0 < l < L \tag{1.7}$$

where v^l, z^l, W^l, and b^l are the output vector, excitation vector, weight matrix, and bias vector at layer l, respectively. $f(\cdot)$ is an element-wise activation function; common choices for this function are sigmoid, hyperbolic tangent, and rectified linear unit (ReLU).

The last DNN layer consists of S outputs, in which the i-th output corresponds to the posterior probability of the i-th HMM hidden state given the input observation o:

$$v_i^L = P(i|o) = \text{softmax}_i(z^L) = \frac{e^{z_i^L}}{\sum_{j=1}^{S} e^{z_j^L}} \tag{1.8}$$

where z_i^L is the i-th element of the excitation vector z^L and S is the number of HMM states.

Unlike GMMs, DNNs take as input a context window of multiple consecutive frames, typically spanning 110ms to 270ms [64, 110]. This ability to model large context windows is the key advantage of DNNs compared to GMMs. However, a limitation of DNNs is that they can only model data within fixed-size context windows and are not suited for handling long-term dependencies [137].

1.2.6.3 Recurrent Neural Network (RNN)

More recently, RNN-based acoustic models have achieved state-of-the-art results on various ASR benchmarks [54, 135, 136, 139]. Similar to DNNs, RNNs estimate the posterior probability $P(s_t|o_t)$ instead of the emission probability $p(o_t|s_t)$. The main advantage of RNNs over DNNs lies in their ability to model long-range temporal dependencies without relying on fixed-size context windows. A standard RNN layer receives an input vector sequence $\mathbf{x} = (x_1, \ldots, x_T)$ and produces a hidden vector sequence $\mathbf{h} = (h_1, \ldots, h_T)$:

$$(h_t, c_t) = \mathcal{H}(x_t, h_{t-1}, c_{t-1}) \tag{1.9}$$

where h_t and c_t are the hidden and cell activation vectors at time step t, and $\mathcal{H}$ is the activation function. A popular choice for $\mathcal{H}$ is the Long-Short Term Memory (LSTM) activation function, a special type of unit designed to better find and exploit long-range context [65]. A RNN layer with LSTM activation function is commonly referred to as a LSTM-RNN layer.

Bidirectional LSTM-RNN (BLSTM-RNN) is an extension to this architecture, which adds a parallel LSTM-RNN layer that processes the input sequence backward:

$$(\overrightarrow{h}_t, \overrightarrow{c}_t) = \overrightarrow{\mathcal{H}}(x_t, \overrightarrow{h}_{t-1}, \overrightarrow{c}_{t-1}) \tag{1.10}$$

$$(\overleftarrow{h}_t, \overleftarrow{c}_t) = \overleftarrow{\mathcal{H}}(x_t, \overleftarrow{h}_{t+1}, \overleftarrow{c}_{t+1}) \tag{1.11}$$

The output of a BLSTM-RNN layer is the concatenated hidden vector $h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$. Multiple BLSTM-RNN layers can be stacked on top of each other to create a deep BLSTM-RNN architecture. Finally, an output layer can be added:

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{1.12}$$

where $W_{\overrightarrow{h}y}$ and $W_{\overleftarrow{h}y}$ are the hidden-output weight matrices and b_y is the bias vector. Similar to DNNs, softmax normalization is applied to the output vector y_t, resulting in a distribution over HMM states given an input observation.

1.2.7 ASR for Disordered Speech

There has been extensive work in the related field of dysarthric speech recognition [3, 25–28, 146, 147]. ASR for dysarthric and disordered speech in general is faced with abnormal speech patterns [105], high speaker variability [112], and data scarcity [27]. Methods for alleviating these problems include speaker-dependent GMM adaptation [27, 146, 147], generation of auxiliary acoustic features used within tandem-based systems [3, 25], learning systematic speaker-specific pronunciation errors [28], and similarity-based speaker selection for acoustic modeling [26]. Most of these works focused on single-word recognition, whereas our work targets disordered continuous speech, which has remained relatively under-explored in the literature. In addition, the application of deep learning-based acoustic models to this area has remained fairly limited.

There has been comparatively little work on ASR for aphasic speech. Existing works are limited to using healthy acoustic models to recognize aphasic speech [2, 89]. Further, aphasia and dysarthria have several key differences. A PWA's verbal expression is modulated by language impairment and co-occurring motor control disorders, which often include AOS and dysarthria itself. AOS can make the speech produced by PWAs inconsistent, thus increasing intra-speaker variability. Verbal output and language usage patterns of different PWAs may vary depending on the aphasia type, such as fluent and non-fluent aphasia. It is unclear if techniques that work for dysarthria will also translate to aphasia.

1.3 Contributions

Our work presents novel computational methods to enable reliable speech-language assessment, with the long-term goal of transforming therapeutic care for PWAs by providing individualized on-demand therapy and speech-based progress monitoring tools. This necessitates advances in automatic disordered speech recognition and assessment. The research contributions of this dissertation are as follows:

- Aphasic Speech Intelligibility Assessment
  - Introduced the UMAP dataset and provided baseline intelligibility classification results using transcript and acoustic features [78].
  - Introduced a novel feature set that captures the pronunciation, rhythm, and intonation of PWAs by comparing with healthy speech patterns [83].
  - Created a fully automated intelligibility assessment system by removing the dependence on human-labeled transcripts using variants of forced alignment. Introduced new clinically-motivated features and demonstrated that the system can achieve competitive performance with human evaluators [81].
- Aphasic Speech Recognition
  - Established the first large-vocabulary continuous speech recognition (LVCSR) baseline on AphasiaBank, a large corpus traditionally used by clinical researchers to study aphasia. Showed that i-vectors can be used to compensate for the variability in speech patterns among PWAs. Proposed adaptation methods to improve recognition performance on a small aphasic speech corpus [82].
  - Introduced new training methods that significantly improved recognition accuracy on AphasiaBank. The proposed ASR system formed the basis for subsequent works targeting ASR-driven analysis of aphasic speech [80].
- Quantitative Analysis of Aphasic Speech
  - Established the first evaluation framework and baseline feature set for automatic paraphasia detection. Showed that speaker-level phonemic paraphasia production rate can be estimated with reasonable accuracy using ASR output [79].
  - Proposed a feature calibration method that allows clinically-relevant quantitative measures to be extracted reliably from ASR-generated transcripts. Showed

that ASR-based features can be used to accurately predict Aphasia Quotient, a standard measure of aphasia severity [80].

The unification of these works will enable an automated system capable of capturing clinically-relevant characteristics of either read or spontaneous aphasic speech without the need for manually labeled transcripts. The output from this system can be used as direct feedback for PWAs, as well as complementary information to assist SLPs with progress monitoring and treatment planning.

1.4 Dissertation Outline

The dissertation is organized as follows. Chapter 2 provides an overview of the datasets used in this work, including the development, collection, and annotation of the University of Michigan Aphasia Program (UMAP) dataset. Chapter 3 covers our work on automated speech intelligibility assessment for utterances with well-defined prompts. Chapter 4 describes methods to improve ASR performance on aphasic speech in order to enable automated analysis of unconstrained speech. Chapter 5 highlights our work on automatic paraphasia detection. Chapter 6 focuses on the relationship between feature robustness and transcription errors, as well as aphasia severity estimation. Finally, Chapter 7 summarizes the dissertation and discusses potential directions for future work.

CHAPTER 2

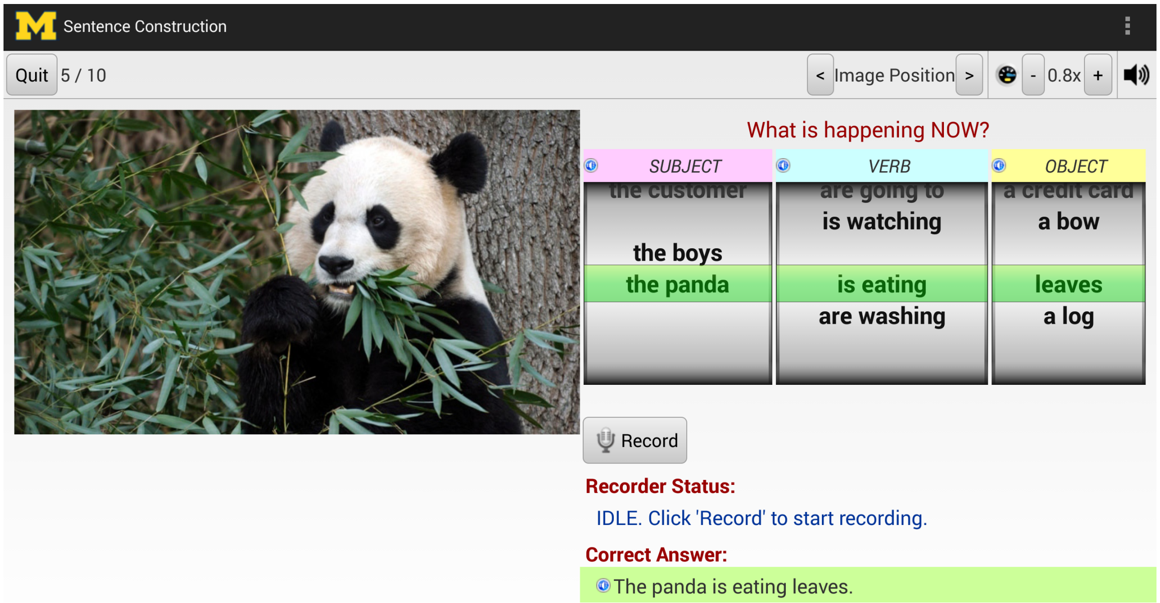
Datasets

2.1 UMAP Dataset

One of the long-term objectives of this dissertation is to develop an intelligent system capable of providing automatic speech-language feedback to persons with aphasia (PWAs). One of the major challenges to achieving this objective is the lack of a publicly available dataset containing speech data collected in the context of a therapeutic application. To this end, we partnered with the University of Michigan Aphasia Program (UMAP) to develop a mobile application that includes therapeutic exercises of sentence building and picture description. We collected approximately five hours of aphasic speech from 17 UMAP clients while they interacted with the application. Human annotators transcribed and evaluated each utterance across three aspects of speech intelligibility: *Clarity*, *Fluidity*, and *Prosody*. In addition, we collected 10.5 hours of speech from non-aphasic controls to allow for a comparison between the speech-language patterns of these two populations. This dataset forms the basis of our work on automatic speech intelligibility assessment (Chapter 3).

2.1.1 Speech Intelligibility Ratings

An important problem in this work involves constructing ground-truth labels for speech intelligibility from human evaluators. This task is traditionally performed by expert listeners, such as Speech-Language Pathologists (SLPs). However, previous work has shown that



What is happening NOW?

Recorder Status:

IDLE. Click 'Record' to start recording.

Correct Answer:

The panda is eating leaves.

Figure 2.1: Screenshot of an exercise with predefined options.

with appropriate elicitation techniques, untrained listeners can estimate speech intelligibility with close to expert-level judgment [101]. Further, it has been suggested that SLPs may be overly familiar with disordered speech and may assign higher scores than non-expert listeners, a phenomenon commonly referred to as the "familiarity effect" [77, 107, 170].

One popular metric for measuring speech intelligibility is Ease of Listening (EOL). In EOL, a 5-point Likert scale is employed to elicit perceptual measures of dysarthric speech intelligibility from naïve listeners [77, 108]. Alternative approaches include using continuous [29] or similarity [11] labels. We adopt the Likert scale in this work because it is more in line with SLP scoring practices and is still the most common choice for human perceptual studies.

2.1.2 Aphasic Speech Corpus

2.1.2.1 Mobile Application

We developed a mobile application designed to run on Android tablet devices for the purpose of data collection and speech-language therapy. The application was designed using an iterative process in which feedback from PWAs and SLPs at UMAP shaped the interface and functioning of the system. In the application, users are presented with a picture stimulus, along with optional predefined word options, and asked to verbally produce a sentence to describe the picture. The sentence must contain a subject, verb, and object. Sentences of this form can be thought of as Main Concepts of the picture being presented [115]. The application features exercises primarily targeting sentence formulation while also allowing users to work on word-finding, use of verb tenses, and repetition and articulation of target words and phrases, to ultimately facilitate expressive communication. It is intended to be used by PWAs for home practice, as well as by SLPs and PWAs together in therapy sessions as stimuli for speech-language activities using functional, evidence-based treatment techniques such as Verb Network Strengthening Treatment (VNeST) [35].

All speech output is recorded using the tablet's built-in microphone, sampled at 44.1 kHz. Figure 2.1 shows a sample exercise with predefined word options. The application operates at the sentence level, which was suggested to be more beneficial than word-level exercises for recovering communication skills in highly routine conversational tasks [24, 99]. The difficulty level can be adjusted through the application interface. We also utilize text-to-speech with configurable speech rate to provide auditory feedback in addition to visual and textual information as the users may have difficulties with reading and/or auditory comprehension. The information gathered from this application is beneficial for both the PWAs in self-monitoring and the SLPs in determining the appropriate course of treatment. It is also a valuable data source for aphasic speech modeling since the collected dataset contains not only speech samples but also their recording context.

2.1.2.2 Collection Methodology

We recruited 17 individuals attending UMAP who have aphasia and do not have cognitive impairment for this study. UMAP offers an intensive therapy program which, for full-time clients, typically involves 24 hours of speech-language therapy a week for four weeks. Each study subject was screened and recommended by the assigned primary SLP in UMAP. A team of research staff interacted with each individual for an average of 30 minutes a day, three days a week for up to three weeks. During these sessions, the research staff provided support, as needed, while the participants completed the exercises on our mobile application. The research team consisted of undergraduate and graduate students who received training from UMAP staff regarding how to assist individuals with aphasia.

Our goal was to collect speech recordings that best resemble the type of data the application would have received from natural interaction with the PWAs. Recordings were made in one of the three classrooms at UMAP, depending on what was available at the time. We adjusted the difficulty based on recommendations from the SLPs and used the tablet's built-in microphone for all recordings. We collected two types of recordings based on the PWAs' severity and personal preference. The first is read speech, in which PWAs assemble a sentence using predefined word options (Figure 2.1), and then read the sentence out loud. The second is free-form speech, in which PWAs describe the picture in their own words.

It should be noted that for the reading task, PWAs often do not reproduce the target sentence exactly. This may be caused by difficulties initiating speech, word finding problems, repetition, and various types of paraphasias. Our work on intelligibility assessment (Chapter 3) focuses on read speech because we can systematically make use of the target prompts, which constrain the recognition problem and make automatic transcription more feasible. Recognizing and assessing free-form speech will be left for future work.

2.1.2.3 Detailed Analysis

Table 2.1 lists the age, sex, diagnosis, amount of recorded data, and Aphasia Quotient (AQ) before and after treatment at UMAP for each subject in the dataset. AQ is a summary score which measures the severity of aphasia [73]. AQ is obtained using the Western Aphasia Battery-Revised (WAB-R) assessment test, which comprises of individual subtests targeting a PWA's functional communication, repetition, word finding, and auditory com-

ID	**Age**	**Sex**	**Diagnosis**		**# of Recordings**		**Aphasia Quotient**	
			Type	***MCD***	***Read***	***Free-form***	***Before***	***After***
RM	65	F	FLU	DYS	122	–	94.6	95.6
JF	50	F	NFL	–	30	92	91.6	–
CC	33	M	NFL	–	131	–	78.3	87.2
JR	70	M	FLU	–	105	270	75.7	87.8
RK	60	M	NFL	AOS	170	–	75.4	78.6
CH	79	M	FLU	AOS	133	–	69.1	80.3
AN	86	M	FLU	AOS	69	104	68.6	76.2
TP	48	M	FLU	AOS	89	–	67.9	77.0
MH	35	F	FLU	–	84	–	62.5	66.5
JE	70	F	NFL	AOS	121	68	61.2	71.2
KH	50	M	NFL	–	72	–	59.5	71.9
DD	49	M	NFL	–	58	88	58.9	72.1
BW	66	F	NFL	AOS	49	–	53.1	62.3
PT	55	F	FLU	AOS	112	–	51.0	82.7
DB	68	M	NFL	–	81	–	43.9	94.2
JT	49	M	FLU	AOS	112	–	40.4	59.4
TL	50	M	NFL	AOS	147	64	34.6	50.6

AQ Class: 0-25 (**very severe**), 26-50 (**severe**), 51-75 (**moderate**), 76-100 (**mild**) [72]
Aphasia Type: fluent (**FLU**), non-fluent (**NFL**)
Motor Control Disorder: dysarthria (**DYS**), apraxia of speech (**AOS**)

Table 2.1: Subject breakdown of the aphasic speech corpus.

prehension ability [72]. According to the WAB-R's AQ classification, most PWAs in the dataset have moderate to mild aphasia. The AQs and diagnoses are shown to demonstrate the heterogeneity of the dataset.

The corpus contains 1,685 read and 686 free-form utterances, totaling approximately 5 hours of data from 11 male and 6 female PWAs with an average age of 58 ± 14. AOS was manifested in 9 out of 17 speakers, while only one had dysarthria. The speech patterns produced by the speakers differ greatly. Some subjects have relatively intact pronunciation but exhibit disrupted rhythm and prosody. Others display highly fluent speech but impaired articulation. Many participants have trouble pronouncing uncommon and phonetically complex words, and/or producing verb tenses other than present continuous. Each utterance contains on average 0.238 fillers (e.g. "um", "eh") and 0.085 false starts (e.g. "ddog", "yes-yesterday"). To summarize, the dataset contains a diverse collection of speakers who have different impairments and exhibit a wide range of speech-language patterns.

2.1.3 Healthy Speech Corpus

We hypothesize that comparing and contrasting how aphasic and healthy speech differ will lend insights into and enhance the process of modeling speech intelligibility. For this purpose, we collected speech recordings from 14 native speakers (7 males and 7 females) of American English who have no speech-language impairment. The age range of this population is 20 to 32, which is significantly lower than the subjects in the aphasic dataset. In future work, we will collect healthy speech data that better match the demographics of the aphasic corpus.

The data were recorded with the same device type and recording algorithm used in data collection. We did not control the recording environment of this corpus to simulate the condition under which speech data would be obtained in actual application usage. All speakers were asked to take the tablet and find a relatively quiet space to perform the recordings in their own time. Consequently, the recording environment may be varied both across and within speakers. Healthy speakers were presented with the same speech prompts given to PWAs, but the accompanied picture and word options were not shown. These prompts have significant repetitions of common words such as "he", "she", and "the", making the dataset phonetically unbalanced. The corpus contains 10.5 hours of speech, 17,559 utterances, and 86,596 instances of 527 unique words.

2.1.4 Data Annotation

In this section, we describe how utterances in the aphasic corpus are annotated. The annotation process consists of two tasks: transcription and scoring. The first task provides word-level labels for automatic speech recognition (ASR) training. The second task produces qualitative sentence-level scores for modeling speech intelligibility.

2.1.4.1 Transcriptions

Each utterance (both read and free-form) is transcribed by a member of the research staff. Transcription of each utterance progresses in two passes. In the first pass, transcribers annotate each utterance at the word level based only on audio data. Transcribers are asked to mark sub-regions of the utterance as vague when the speech content is not clear enough to reliably decode and provide a guess for their content if possible. In the second pass, transcribers go through each utterance again, but this time using context information to help refine their guesses for sub-regions previously marked as vague. The provided context information includes the speech prompts and word options shown to PWAs at recording time. Previous research has suggested that constraining the transcription process, as was done in the second-pass transcripts, can help resolve subtle differences in pronunciation among less intelligible speakers, whereas the first-pass transcripts better approximate regular everyday interaction [107].

The first-pass transcripts are used to extract training data for acoustic modeling in ASR. Our preliminary experiments indicate that excluding noisy data, i.e. speech regions that humans cannot reliably decode, helps improve the recognition accuracy of the acoustic model. Because transcribers do not have access to context information during this stage, they must rely more on acoustic data as opposed to word priming to make a distinction between vague and clear segments.

The second-pass transcripts are the targets for the ASR system. These transcripts have the ability to capture what the PWAs tried to say over speech regions with poor intelligibility. This type of information provides the potential to model intelligibility at a greater depth. For example, if we know that the PWA attempted to say “is playing” in a given segment, we can compare its pronunciation, rhythm, and prosody to those of the same segment spoken by a healthy control.

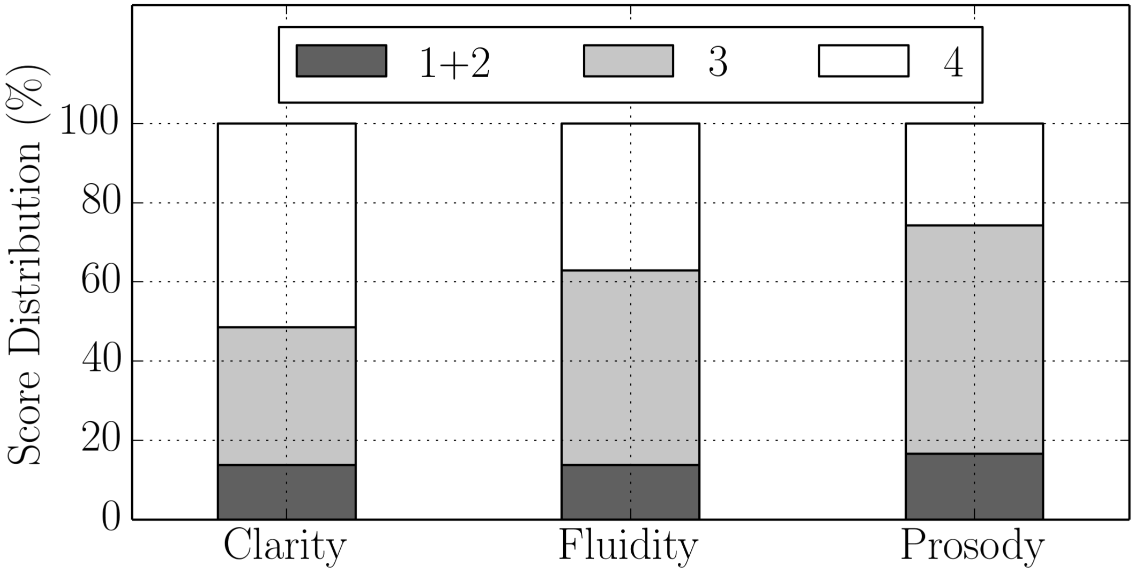


Figure 2.2: Distribution of speech intelligibility scores.

2.1.4.2 Qualitative Scores

The goal of this annotation step is to obtain ground-truth labels for speech intelligibility. With guidance from the SLPs, we created three criteria for evaluating an utterance's intelligibility: *Clarity*, *Fluidity*, and *Prosody*. These criteria capture the quality of pronunciation, the degree of fluidity, and the monotonicity of speech, respectively. Each utterance in the aphasic speech corpus is scored by at least three members of the research staff, all of whom are native speakers of American English without any auditory comprehension deficit. The annotators only have access to the audio data of an utterance; they do not see the identity of the speaker to account for biases in judgment. Each category is scored on a Likert scale of 1 to 4, where a higher score denotes better quality. Annotators may assign a special label, "Not enough data" if they deem that the utterance does not have enough data for analysis. 169 out of 1,672 utterances were marked as such by at least half of the annotators and are removed from the dataset. During this process, annotators are provided with utterances in random order and asked to rate one of the three scoring criteria. They also have access to prototypical examples for each intelligibility score, defined as utterances that have perfect score agreement drawn from the smaller dataset used in our earlier work [78, 83]. Similar techniques have been used in other work to help annotators calibrate their ratings and yield higher inter-rater agreement level [107, 161]. There is also evidence in the literature that a group of non-expert listeners can deliver close to expert-level assessments regarding speech intelligibility [77, 101, 107, 170].

Following [18, 97], we construct a "de-noised" set of ground-truth labels by averaging the scores across all evaluators and rounding to the closest integer. These ground-truth scores represent the collective opinions and are used to train the automatic classifiers. Post-hoc investigation of the ground-truths reveals that the score "1" constitutes only 2.80%, 2.13%, and 3.13% of *Clarity*, *Fluidity*, and *Prosody* labels, respectively. As a result, we merge "1" and "2" together to make a new label category. Figure 2.2 shows the distribution of scores in the aphasic speech corpus after merging.

We evaluate system performance using unweighted average recall (UAR), i.e. the mean per-class accuracy, to account for class imbalance. We can establish a target performance and estimate the degree of human agreement by treating each evaluator as a classifier and computing its UAR with respect to the ground-truths. Our ultimate goal is to achieve human-level UAR with automatic classification.

We observe that the label "3" consistently has lower human agreement across all scoring categories as it is often confused with the other two labels, more so with "1+2". We therefore investigate an additional labeling scheme by merging "1+2" and "3" into a new category, in order to investigate the trade-off between label granularity and classification accuracy. This also results in a more balanced dataset and higher inter-rater agreement than merging "3" and "4". The resulting *2-class* problem is simpler in nature (separating low- and high-quality utterances) and has more reliable ground-truths. A similar merging approach was done in [74] to reduce a label from five to two classes. Table 2.2 summarizes the degree of agreement between human annotators in terms of UAR and Cohen's kappa. As can be seen, *Clarity* has the highest agreement level, followed by *Fluidity* and *Prosody*.

		Clarity	**Fluidity**	**Prosody**
UAR	*3-class*	75.4 ± 4.5	70.9 ± 3.4	68.3 ± 5.5
	2-class	82.3 ± 4.9	80.6 ± 3.6	78.2 ± 4.6
Kappa	*3-class*	0.66 ± 0.09	0.56 ± 0.07	0.50 ± 0.07
	2-class	0.64 ± 0.10	0.62 ± 0.10	0.51 ± 0.10

Table 2.2: Degree of human agreement in speech scoring w.r.t. the ground-truths, measured by average and standard deviation of Unweighted Average Recall (%) and linearly weighted Cohen's kappa.

2.1.4.3 Annotator Instructions

The instructions for the scoring process were given orally to individual annotators. They followed a predefined format without a fixed script. We informally described each scoring category and stepped through all possible answer choices, along with their associated prototypical examples. Below is a summary of the three scoring categories.

Clarity is defined as the degree to which a sentence can be understood. It is intended to capture the overall pronunciation quality of a sentence. The elicitation question for this category is: *How clear is the pronunciation?* The possible answer choices, from low to high quality, are: *Very Unclear*, *Mostly Unclear*, *Mostly Clear*, and *Very Clear*.

Fluidity is defined as the degree to which a sentence can be uttered at an appropriate speed and without pauses or hesitation. The elicitation question for this category is: *How fluid is the speech?* The possible answer choices, from low to high quality, are: *Very Broken*, *Mostly Broken*, *Mostly Fluid*, and *Very Fluid*.

Prosody is arguably the most difficult category to define. We define it broadly as the correctness of intonation. Utterances that are overly monotonous or have widely varying pitch are both considered incorrect. We found that this definition of *Prosody* resulted in higher human agreement compared to directly quantifying the degree of monotonicity. The elicitation question for this category is: *Is the intonation correct?* The possible answer choices, from low to high quality, are: *Very Incorrect*, *Mostly Incorrect*, *Mostly Correct*, and *Very Correct*.

All scoring categories have one additional answer choice, *Not Enough Data*, which is reserved for utterances that the annotators deem to have insufficient data for analysis.

2.2 AphasiaBank Dataset

AphasiaBank is a large-scale audiovisual dataset containing interactions in several languages between PWAs and research investigators [39, 96]. It is primarily used by clinical researchers to study aphasia and has recently been introduced to the engineering community [82, 89]. AphasiaBank data are organized according to their elicitation protocols. Data associated with a specific protocol contain a number of *sub-datasets* collected by different research groups under various recording conditions. In this dissertation, we consider data of native English speakers collected with the *AphasiaBank* and *Scripts* protocols.

2.2.1 AphasiaBank Protocol

This is the core protocol of AphasiaBank, which involves open-ended questions designed to elicit verbal discourse samples. Example questions include: "*How do you think your speech is these days?*," "*Tell me as much of the story of Cinderella as you can*," and "*Describe for me what you see in this picture*." The type of speech collected under this protocol is spontaneous. Utterances can be further divided into two categories based on their applied

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

Table 2.3: Summary of the core AphasiaBank dataset. The speakers are split into two groups, those who have aphasia (*Aphasia*) and healthy controls (*Control*).

And I &uh bit [: get] [* s:ur] out pea [: the] [* p:w] pinək@u [: peanut] [* p:n] bʌðə@u [: butter] [* p:n].

[: get] [* s:ur]: unrelated semantic error with known target *get*
[: the] [* p:w]: real-word phonological error with known target *the*
[: peanut] [* p:n]: non-word phonological error with known target *peanut*
[: butter] [* p:n]: non-word phonological error with known target *butter*

Table 2.4: Example AphasiaBank transcript with semantic and phonological word errors.

elicitation methods, namely *free speech* (e.g., open interview, conversational speech) and *semi-spontaneous speech* (e.g., storytelling, picture description) [125].

We identify sub-datasets under this protocol that contain at least four speakers. This results in 19 sub-datasets with 401 PWAs and 187 healthy controls (323 males, 265 females, age 63 ± 14)[1]. The distribution in aphasia severity of the 401 PWAs, defined by the revised Western Aphasia Battery Aphasia Quotient [72], is 43.4% mild, 32.7% moderate, 9.5% severe, 3.0% very severe, and 11.4% unknown. We discard less than 1% of the utterances whose transcripts include unintelligible or overlapping speech, which makes them incompatible with automatic speech processing. The final dataset contains approximately 130.9 hours of speech (89.2 hours of aphasic speech, 41.7 hours of healthy speech), 102,934 utterances, and 830,193 words. Table 2.3 summarizes the speaker demographics, amount of speech data, and number of utterances for each spontaneous speech category in the dataset.

Utterances in AphasiaBank are transcribed using the CHAT format [95]. The transcriptions contain a variety of special codes to aid with language sample analysis, such as word-level and utterance-level errors [4], sound fragments, repetitions, non-verbal actions, among others. An example is shown in Table 2.4 where the transcript contains one semantic, one real-word phonological, and two non-word phonological errors. The actual pronunciations for the non-word phonological errors are transcribed in the International Phonetic Alphabet (IPA) format, denoted by the @u trailing symbol. Word-level errors occur relatively infrequently in this dataset, accounting for 2.53% and 0.03% of all words

[1] Based on available data at the time this was written. New data are continually added to AphasiaBank.

	Fridriksson	**Adler**
Gender	8 M, 4 F	5 M, 1 F
Duration	3.1 hours	1.1 hours
Utterances	990	349
Words	9,310	3,886
Word Errors	22.6%	5.1%

Table 2.5: Summary of speech data and speaker demographics under the *Scripts* protocol.

in the *Aphasia* and *Control* partitions, respectively.

This collection of data, which we will refer to as the **core AphasiaBank dataset**, is used in our work on large-vocabulary aphasic speech recognition (Chapter 4) and automatic quantitative analysis of spontaneous aphasic speech (Chapter 6).

2.2.2 Scripts Protocol

In contrast to the *AphasiaBank* protocol which elicits spontaneous speech, the *Scripts* protocol is intended to collect read speech samples from predefined scripts. Two sub-datasets fall under this elicitation protocol, *Fridriksson* and *Adler*.

The *Fridriksson* sub-dataset employs four identical scripts for every participant. These include *advocacy* (addressing communication challenges due to aphasia), *eggs* (how to make scrambled eggs), *vast* (use of video assisted speech therapy), and *weather* (describing the weather in southern United States). The sub-dataset comprises 3.1 hours of speech from 12 PWAs (8 males, 4 females), totaling 990 utterances and 9,310 words[1].

Instead of using fixed scripts, the *Adler* sub-dataset utilizes 1–3 personalized scripts for each speaker. Topics of these scripts include aphasia recovery, family members, and memorable personal events, among others. The sub-dataset contains 1.1 hours of speech from 6 PWAs (5 males, 1 female), totaling 349 utterances and 3,886 words[1].

Table 2.5 summarizes the data under the *Scripts* protocol. Word-level errors occur much more frequently in this dataset compared to the core AphasiaBank data, especially for the *Fridriksson* sub-dataset. Due to the high occurrence rate of word errors, *Fridriksson* data will be the target of our work on automatic paraphasia detection (Chapter 5).

2.3 Work Published

The work presented in this chapter was published in the following articles:

1. **Duc Le**, Keli Licata, Elizabeth Mercado, Carol Persad, and Emily Mower Provost. "Automatic Analysis of Speech Quality for Aphasia Treatment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy. May, 2014.

2. **Duc Le**, Keli Licata, Carol Persad, and Emily Mower Provost, "Automatic Assessment of Speech Intelligibility for Individuals With Aphasia," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 24:11(2187-2199). November, 2016.

CHAPTER 3

Automatic Assessment of Aphasic Speech Intelligibility

3.1 Introduction

In this chapter, the term ***aphasic speech*** denotes a PWA's verbal output, which can be modulated by motor control disorders, for example, apraxia and dysarthria. ***Speech intelligibility*** refers to the perceptual quality of aphasic speech, which can be affected by both language and speech impairments. We aim to apply speech assessment and machine learning techniques on the University of Michigan Aphasia Program (UMAP) dataset (Section 2.1) to automatically quantify multiple aspects of aphasic speech intelligibility. These objective measures will let persons with aphasia (PWAs) self-monitor their verbal output as well as help Speech-Language Pathologists (SLPs) decide on appropriate therapy choices. The system has the potential to enable effective auxiliary in-home practice and assist with traditional therapy as needed.

Figure 3.1 shows an overview of the system for speech intelligibility assessment. Utterances from PWAs are first processed by a forced alignment component using either human-labeled transcripts or predefined speech prompts. The former represents an oracle approach and helps set the performance ceiling for the system, while the latter completely automates the forced alignment pipeline. The output of this system serves as a preprocessing step for feature extraction. Our novel feature set consists of clinically-motivated features that

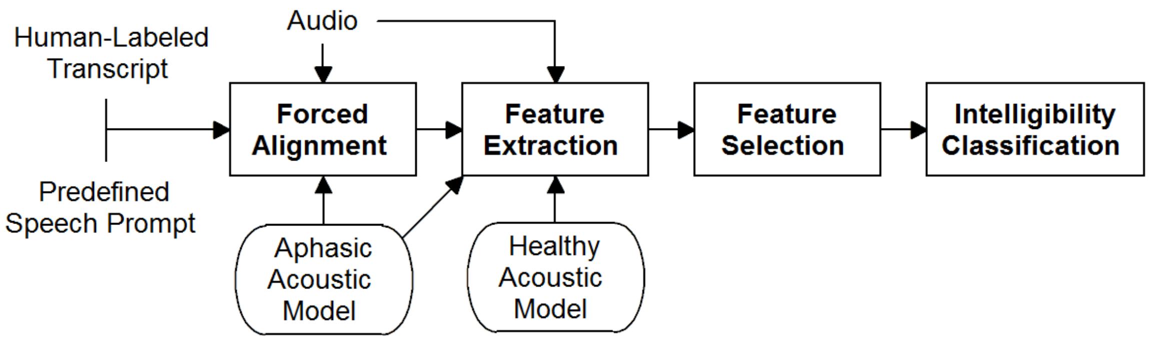


Figure 3.1: System diagram for estimating speech intelligibility.

capture various aspects of speech intelligibility, including pronunciation, rhythm, and intonation. We demonstrate that the system can perform close to the level of human evaluators in estimating intelligibility scores.

The contributions of this work are four-fold. Firstly, the system's novel application to aphasia has the potential to greatly benefit the well-being of PWAs by enabling self-directed practice with automatic feedback. Secondly, we introduce forced-alignment-based techniques for automatic transcript generation that perform well on aphasic speech in spite of limited data and atypical speech input. Thirdly, we describe a novel feature set specifically engineered to capture speech intelligibility. Lastly, the detailed analysis of classification performance and feature relevance uncovers the research problems that need to be solved in order to bridge the gap between human and automatic intelligibility assessment, along with possible approaches to tackle them.

3.2 Oracle Forced Alignment

The goal of this step is to use forced alignment to obtain a detailed transcript of what was spoken by the PWA in a given utterance, including precise alignments of words, syllables, and phones. This transcript is an important prerequisite for extracting features relevant to speech intelligibility classification (Section 3.4). Forced alignment requires as input a preliminary word-level transcript without timing. We initially make use of the oracle transcripts labeled manually by human annotators. Section 3.3 will discuss methods to completely automate forced alignment.

3.2.1 Speech Preprocessing

For each utterance in the healthy and aphasic speech datasets, we downsample the audio to 16 kHz and extract 13-dimensional Mel-frequency Cepstral Coefficients (MFCCs) using a 25ms Hamming window with 10ms frame step. Each MFCC frame is augmented with the first and second temporal derivatives, resulting in a 39-dimensional feature vector. Finally, the features are z-normalized at the speaker level.

3.2.2 Acoustic Modeling

In this work, there is much more healthy speech compared to aphasic speech for acoustic modeling. Counting only speech frames from clearly understood segments, healthy speech amounts to 9.1 hours whereas aphasic speech comprises only 1.7 hours. Furthermore, we are interested in a speaker-independent acoustic model to better understand how the system will perform on an unknown speaker. Data from the aphasic corpus will be further reduced as a result of leave-one-speaker-out cross-validation. We adopt the out-of-domain adaptation approach, in which a model initially trained on the more abundant out-of-domain (healthy) speech is adapted using a smaller amount of in-domain (aphasic) data. Out-of-domain adaptation on disordered speech has been employed successfully in [25], where the adapted models outperform those trained only on the in-domain data. Intuitively, this method helps alleviate the data sparsity issue when training on aphasic speech alone.

State-of-the-art acoustic models typically involve a number of Hidden Markov Models (HMMs), one for each phone, where the emission probabilities are estimated using a Deep Neural Network (DNN). Training data for the DNN is usually obtained by using an initial acoustic model based on Gaussian Mixture Model (GMM) to perform forced alignment. In large-vocabulary continuous speech recognition (LVCSR) systems, the HMMs model context-dependent tied-state triphones instead of individual monophones. However, when there is limited speech data and the vocabulary is relatively constrained such as the case for our data, using monophone acoustic models may be more appropriate. Our preliminary experiments indicate that there is little improvement in performance when using triphone models, while the system complexity and training time dramatically increase. As a result, we only train monophone acoustic models in this work.

We follow the recipe in [110] to train a DNN acoustic model on healthy speech, bootstrapped from a standard HMM-GMM model trained with Maximum Likelihood, 3-state left-to-right HMMs representing the 40 phones defined in the CMU lexicon[1], and 64 diagonal covariance Gaussians per state. We augment each MFCC frame with 13 neighbors from both sides, padding out-of-boundary elements with the nearest frame. We first generatively pretrain a Restricted Boltzmann Machine (RBM) with two layers, 1024 units per layer, and sigmoid activation. Similar to [110], we use a batch size of 128, L2 regularization weight of 0.0002, and learning rates of 0.002 for the first Gaussian-binary layer and 0.02 for the second binary-binary layer. For each RBM layer, pretraining terminates when the relative change in reconstruction cost is less than 0.1%. RBM has been shown to be effective for ASR when the dataset and/or the network is relatively small [166].

We subsequently add a softmax output layer with 120 units, corresponding to the HMM states of the 40 phones, to the RBM. The network is finetuned using stochastic gradient descent to predict the HMM state label for each acoustic frame. We use a batch size of 256, 0.5 momentum, and a small L2 regularization weight of 0.00002. An early stopping approach is employed for finetuning. 10% of acoustic frames from each training speaker are randomly withheld to form a validation set. The initial learning rate starts at 0.1. Once the change in validation error falls below 0.05% absolute, the learning rate decays by half after every epoch. Finetuning will terminate when the change in validation error once again falls below 0.05%. The context window size (13), number of hidden layers (2), and

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

Method	**1-best**	**2-best**	**3-best**
GMM (Healthy)	63.30	51.68	45.67
GMM (MAP)	55.62	43.01	36.60
DNN (Healthy)	48.48	35.78	29.89
DNN (Aphasic)	37.86	26.51	21.61
DNN (Adapted)	**37.11**	**25.75**	**20.61**

Table 3.1: Single word recognition Word Error Rate (%) on the aphasic speech dataset using a uniform word language model over all 592 unique words in the vocabulary. The total number of words in the dataset is 8,564.

L2 regularization weight (0.00002) were selected with cross-validation. Specifically, this combination of hyperparameters minimizes the average leave-one-speaker-out Phone Error Rate (PER) on healthy speech.

Finally, we adapt the healthy DNN acoustic model by retraining the network on aphasic speech using a similar finetuning recipe with more conservative parameters. This can be viewed as a form of discriminative pretraining [166]. We adapt the model for each speaker in the aphasic corpus using data from all other speakers, in concordance with the leave-one-speaker-out validation scheme. We use a batch size of 256, no momentum, and a L2 regularization weight of 0.0001. A similar early stopping approach is adopted. 15% of data from each speaker is withheld to form a validation set. The initial learning rate is smaller, starting at 0.05. The initial learning rate (0.05), momentum (0), and L2 regularization weight (0.0001) were again selected with cross-validation to minimize the average leave-one-speaker-out PER on aphasic speech.

We perform isolated word recognition (IWR) on 8,564 word-level segments extracted from the human-labeled transcripts to evaluate the acoustic models. A unigram language model with identical probability over all 592 words present in the transcripts is used for decoding. The 1-best, 2-best, and 3-best WER for the unadapted and adapted acoustic models are summarized in Table 3.1. For reference, we also include the performance of the original GMM model trained on healthy speech, the GMM model adapted to aphasic speech using Maximum a Posteriori (MAP) adaptation, and the DNN model trained only on aphasic speech. The adapted DNN model clearly outperforms both GMM models and the unadapted (healthy) DNN. While the adapted DNN is not statistically significantly better than the aphasic DNN for 1-best (paired t-test with per-speaker WERs, p = 0.068), it is significantly better for 2-best (p = 0.014) and 3-best (p = 0.008). A more prominent gain may be achieved with a larger out-of-domain dataset that better matches the demographics of our aphasic corpus.

3.3 Automatic Forced Alignment

We need a method to automatically generate coarse word-level transcripts to remove the dependence on human annotators. Unconstrained ASR may not yield sufficient recognition accuracy due to the small size of the dataset and atypical speech input. However, all utterances considered in this work are constrained by the provided speech prompts. Each prompt consists of three distinct parts: subject, verb, and object. A verb may be in one of three tenses: present continuous ("he is driving a car"), simple past ("he drove a car"), or simple future ("he will drive a car"). Due to the language impairments associated with aphasia, PWAs may not reproduce the target prompts perfectly. Speech-language errors may include phonemic errors (e.g., sound distortion, substitution, omission), lexical errors (e.g., word repetition, substitution, omission), insertion of fillers, and false starts. Nevertheless, the prompts help greatly constrain the search space of potential utterances.

We convert the speech prompt of each utterance into a *simple* recognition network by connecting all words in the prompt linearly, with optional silence in between. This network enables forced alignment on the speech input, which provides precise timing information for words, syllables, and phones. This method is not suitable for all utterances because PWAs often do not reproduce the prompts accurately. Our analysis of the collected transcripts show that there are three prominent types of errors made by PWAs during speech production. First, 48.1% of sentences in simple past and future tenses include the

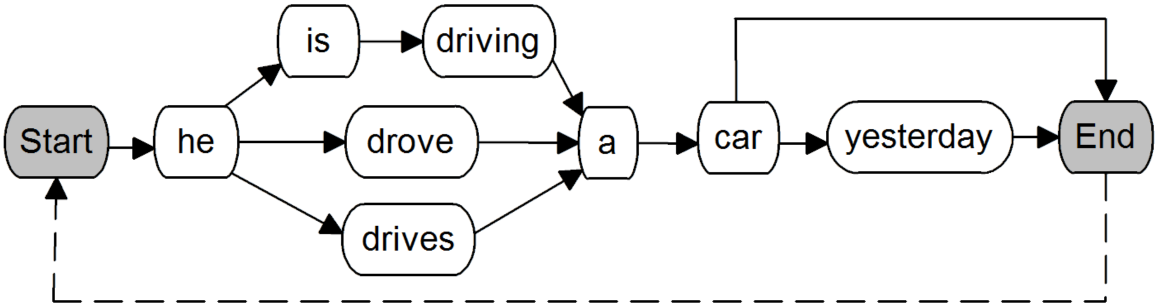


Figure 3.2: Example extended recognition network generated for the prompt "he drove a car." Optional silence can be inserted in between words. All outgoing edges have identical weights. The dashed edge is optional and can be traversed at most once.

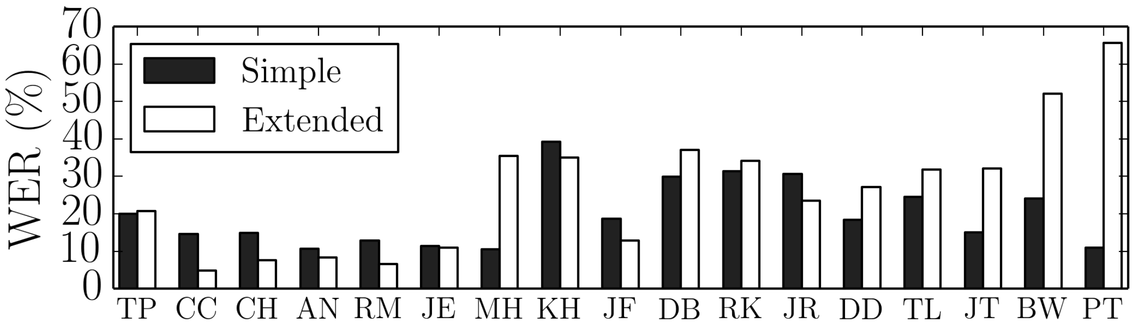


Figure 3.3: Per-speaker Word Error Rate (WER) using simple and extended forced alignment. Speakers are sorted by increasing isolated word recognition WER.

adverb "yesterday" and "tomorrow," respectively. These adverbs serve as indicators for making the intended tenses clear and are not part of the prompt. However, many PWAs prefer to include them in their speech as the adverbs help them select the correct verb tense more easily. This can be captured by extending the network to allow for optional insertion of the adverb depending on the target verb tense. Second, many PWAs consistently produce simple present and continuous tenses in place of the requested simple past and/or future tenses. We can capture this type of error, which accounts for 13.4% of all utterances, by ensuring that the former tenses are always included in the recognition network. Finally, 5.1% of utterances contain repetitions of two or more parts, which can be accounted for by allowing the utterance to be repeated at most once. We encode these three observations in the *extended* recognition network. Figure 3.2 shows an example extended network generated for the prompt “he drove a car.”

Our preliminary experiments indicate that this forced-alignment-based approach, though restricted, performs better than n-gram language models due to the lack of aphasic acoustic data and atypical speech events, such as fillers, false starts, and audible background noise. N-gram models will be used in later chapters when working with unconstrained speech. Figure 3.3 shows the per-speaker WER using simple and extended forced alignment (SFA and EFA, respectively). Compared to SFA, EFA typically leads to WER reduction for more intelligible speakers, but can produce significantly worse results for less intelligible PWAs. This suggests that it might be possible to systematically select the type of forced alignment to use for each PWA based on diagnosis or a simple word pronunciation test conducted beforehand. Future work will explore the related problem of language model personalization based on speaker diagnoses.

This method is also reasonably suitable for handling atypical speech events, which are difficult to recognize directly due to lack of data. Using EFA, 76.1% of these events get absorbed by the silence model. This is possible because the number of non-silence tokens in the recognition network is finite, and the likelihood obtained from matching these tokens to the correct words is higher than matching them to atypical speech sounds.

3.4 Feature Extraction

Given the detailed transcripts generated from *Oracle*, *Simple*, or *Extended* forced alignment, we now discuss feature extraction methods to capture speech intelligibility. The features considered in this work are grouped into four sets: *Transcript*, *Pronunciation*, *Rhythm*, and *Intonation*. These sets extract high-level information related to different aspects of speech intelligibility. All four sets rely on an utterance’s associated transcript, which means that the accuracy of the transcript directly affects the quality of the extracted features. Low-level acoustic features such as intensity, jitter, shimmer, and zero-crossing rate, which were investigated in our earlier work [78, 83], as well as the widely used GeMAPS feature set [36], did not work well on this dataset, possibly due to the uncontrolled recording conditions. For example, the distance between a speaker and the tablet microphone may vary between utterances, which directly affects intensity and pitch tracking.

3.4.1 Transcript Features

We hypothesize that the transcripts (both human-labeled and automatic) encode information about the aspects of speech intelligibility targeted in this work. The transcripts include detailed timing information of each token (word, syllable, or phone). Transcript tokens can be broadly divided into three groups. **Clear speech** denotes speech regions that can be clearly understood. **Non-speech** corresponds to non-verbal regions in the utterance, such as silence and background noise. Finally, **vague speech** includes fillers and pronunciations that are determined to be unclear by human annotators. Because automatic (*Simple* and *Extended*) forced alignment does not detect fillers and unclear speech, this token category is only available in *Oracle* forced alignment.

For each utterance, we first extract duration-based measures from its transcript to characterize the distribution of different token categories. Specifically, we compute the duration of non-speech, vague speech, and clear speech, total duration, and voiced duration, defined as the total duration of both clear and vague speech. To normalize for utterance length, we extract the fraction of clear speech over total duration and voiced duration. The next set of transcript features roughly capture dysfluency in a PWA's speech. We measure the start time of first speech activity, which may denote utterance initiation difficulty. We also extract long pause ($>$ 0.4s) and short pause ($>$ 0.15s, $\leq$ 0.4s) count [119], as well as the mean [130], median, minimum, maximum, and standard deviation of pause durations. It should be noted that features involving vague speech are only relevant when using *Oracle* transcripts. For automatic (*Simple* and *Extended*) transcripts, these features are set to zero and will always be eliminated by feature selection (Section 3.5). In total, 16 transcript features are extracted for each utterance.

The next set of features is inspired by diagnostic measures collected by UMAP SLPs to analyze a PWA's speech. Specifically, we extract the number of words and syllables produced overall and per minute. In addition, we extract a similar set of features for content words only to account for the possibility that non-content words (heuristically defined to include “is”, “was”, “are”, “were”, “the”, “a”, “will”) have lower impact on the perception of speech intelligibility, given that they carry relatively less meaning. Post-hoc analysis indicates that features for content words are complementary and help improve classification performance. 8 new features are added in total.

3.4.2 Pronunciation Features

Our first set of pronunciation features are based on Goodness of Pronunciation (GOP), a commonly used metric first introduced by Witt and Young [163]. The idea behind GOP is to calculate the difference between the average acoustic log-likelihood of a force-aligned phoneme and that of an unconstrained phone loop. If this number is close to 0, the pronunciation of this phone is more likely to be correct and vice versa. Originally defined to compute the pronunciation score of a single phoneme, GOP can be modified to accommodate an arbitrary phone sequence:

$$GOP(\mathbf{p}) = \frac{1}{N} \log \frac{P(O|\mathbf{p})}{P(O|PL)} \tag{3.1}$$

where $\mathbf{p}$ is the sequence of phones, O is the acoustic observation, N is the number of frames, and PL is the unconstrained phone loop. To obtain GOP for a word, we force align its speech over all possible pronunciations to find the best phone sequence $\mathbf{p}$. $P(O|\mathbf{p})$ and $P(O|PL)$ can be rewritten as a product of HMM transition probabilities and acoustic likelihoods, where the latter are obtained by dividing the DNN posteriors by state priors.

We extract GOP scores for all words in an utterance by force aligning the speech to its associated transcript using the DNN acoustic model trained on healthy speech. Preliminary experiments indicated that the healthy acoustic model is better suited for GOP computation than the adapted model, as evidenced by improved classification performance and information gain with respect to the *Clarity* ground-truth labels. We then weight the GOP scores by word durations, given the early observations that duration-weighted features perform better in classification, possibly because longer words have more impact on the perception of the entire sentence. Finally, we extract the mean, standard deviation, median, minimum, and maximum word-level GOP scores to use as features for the utterance. We repeat this process once more for content words only, based on the idea that non-content words contribute less to the overall perception of speech intelligibility. Another similar set of features is extracted at the phone level, which was previously shown to provide complementary information [83]. In total, 15 GOP features are extracted for each utterance.

In addition to GOP, we extract additional metrics based on how well the speaker's speech sample fits into the acoustic model, motivated by similar features used in [97, 128] to assess speaker-level intelligibility. We first extract isolated word segments for each utterance based on the timing information encoded in its associated transcript. We then report the 1-, 2-, and 3-best WER when performing IWR on these segments using the adapted DNN acoustic model (Section 3.2.2). We additionally compute the error rates weighted by word duration, which may provide complementary information if longer words have more impact on human perception. Post-hoc feature analysis confirms this hypothesis. We expect that the error rates will negatively correlate with *Clarity* scores. 6 new features are added, increasing the size of the pronunciation feature set to 21.

3.4.3 Reference Alignment

A prerequisite for computing rhythm and intonation scores in this work is the ability to, given the transcript of an aphasic speech utterance, extract corresponding alignment profiles from a reference database of non-aphasic speech [83]. For example, suppose the PWA says

"The people clapped", the goal is to find the same sentence spoken by a healthy control and analyze how the two utterances' durations and pitch/intensity contours differ. This will allow us to compute rhythm and intonation scores, respectively. However, extracting an identical sentence is impractical for two reasons. Firstly, it is not possible to anticipate all the sentences and words that PWAs will produce. When interacting with the mobile application, the PWAs' speech-language deficits often caused them to verbalize the sentences differently from what was asked. Secondly, as the application grows and new sentences are added, it is impractical to maintain a matching reference database.

We hypothesize that the characteristics of an acoustic unit (word, syllable, or phone) are influenced by its immediate neighbors, motivated by [106] and the ability of triphones to capture coarticulation. Based on this idea, we developed an algorithm to search for a reference alignment of any target utterance by gradually increasing the level of granularity until a match is found. The utterance is first broken into triwords, defined as the words with their left and right neighbors. The algorithm finds occurrences of each triword in the reference database which match, in decreasing preference, both left and right contexts, only left or right context, or no context. If the word cannot be found, the triword is broken into syllables according to a pronunciation dictionary and the search continues for each trisyllable. Similarly, if the syllable is not found, it is broken into phones and the search continues for each triphone. The process is guaranteed to succeed if the reference database contains instances of all phones. Table 3.2 shows a sample alignment for the target sentence

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

Table 3.2: Reference alignment for the target sentence "The people clapped." The search must descend into the syllable and phone level for the out-of-vocabulary word "people."

“The people clapped”. Since “people” is out-of-vocabulary, the search breaks it down to two syllables “p iy” and “p ah l”. The second syllable is also missing, so the search breaks it down further into individual phones. The algorithm currently assumes that a match without context at a higher level (word, syllable) is better than a match with context at a lower level (syllable, phone), which will be investigated further in future work.

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

3.4.4 Rhythm Features

Earlier work on rhythmic analysis for language classification proposed features computed from the target speech such as %V (average proportion of vocalic intervals), ΔC and ΔV (average standard deviations of consonantal and vocalic intervals) [127], and normalized Pairwise Variability Index (PVI) [52]. The efficacy of these metrics has been demonstrated, but they are less suitable for this work because of two reasons. First, these features are typically computed at the speaker level and may not be stable enough for short utterances that contain considerably less data. Second, speech patterns across different PWAs are highly variable, thus computing statistics on their speech alone might not be conducive to generalization. More recently, Tepperman et al. introduced Pairwise Variability Error (PVE), a metric that directly compares two speakers’ rhythms [154]. Given duration profiles of a target and reference utterance, denoted as $\{t_1, t_2, ..., t_N\}$ and $\{r_1, r_2, ..., r_N\}$ respectively, where each element is the duration of an acoustic unit (word, syllable, or phone), PVE computes the difference of these two profiles:

$$PVE = \frac{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |(t_i - t_{i-m}) - (r_i - r_{i-m})|}{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |t_i - t_{i-m}| + |r_i - r_{i-m}|} \tag{3.2}$$

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

The target duration profile is first obtained by force-aligning the speech to its transcript using the adapted DNN acoustic model (Section 3.2.2). The reference duration profile can then be constructed by querying the non-aphasic speech corpus using the Reference Alignment algorithm. We do not perform linear scaling on the reference durations as in [154] to retain information about speaking rates and to avoid durational distortion caused by long pauses in aphasic speech. For each utterance, we compute four PVE scores with M ranging from 1 to 4 (same as [154]), constituting the utterance's rhythm features.

3.4.5 Intonation Features

Previous studies suggested that pitch contours in PWAs may exhibit anomalies in sentence-length utterances [32, 45]. Methods for labeling speech prosody involve the inspection of pitch contours of phrases and syllables [21, 149]. We compare the contours of aphasic speech to those of healthy speech using Dynamic Time Warping (DTW), a method previously used to measure the similarity of pitch contours with differing lengths [129]. Similar to above, we first obtain a reference and target alignment for each utterance using the Reference Alignment algorithm. We compute the average DTW distance between each *target* word produced by the PWAs and the same *reference* words spoken by the healthy controls. Prior to computation, the reference contours are shifted to have the same mean as the target; this accounts for pitch differences across speakers. We also compensate for different speaking styles by only using reference words of the healthy speaker that yields the minimum average DTW distance. We then weight the DTW distance of each unit by its duration, under the hypothesis that longer units have more impact on human perception. The final step is to extract the mean, standard deviation, median, minimum, and maximum unit-level distances to use as utterance-level intonation features.

We extract a similar set of features for intensity contours, based on the idea that they also influence the perception of speech intelligibility by modulating emphasis patterns. The final intonation feature set contains 10 features in total, 5 for pitch and 5 for intensity.

3.5 Classification Methods

We partition the dataset using leave-one-subject-out cross-validation, motivated by the design goal that the application must generalize beyond individual speakers. Features are globally z-normalized using statistics from the training set. To avoid overfitting, feature selection is performed 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 [123]. mRMR was used in Fraser et al. [42], yielding good results in classifying subtypes of primary progressive aphasia. We evaluate each fold using several commonly-used classifiers, including C4.5 Decision Tree, Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine. Since our dataset is relatively small (1,503 data points), we did not do model selection and instead used the default settings specified in the Weka toolkit [57]. We will explore model selection and speaker-dependent adaptation to improve classification for larger datasets.

3.6 Results and Discussion

3.6.1 GeMAPS Baseline

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) is a collection of acoustic features commonly used for speech emotion recognition (SER) [36]. GeMAPS may contain useful features for our classification tasks since they share certain similarities with SER.

Table 3.3 shows the UARs achieved on this feature set using the proposed classification pipeline. All results are statistically significantly worse than those achieved using the proposed features (paired t-test, $p < 0.001$). Further, we found that adding GeMAPS features

	Clarity	**Fluidity**	**Prosody**
3-class	32.8 (S)	50.1 (L)	44.6 (N)
2-class	58.0 (L)	64.2 (L)	55.8 (L)

S: Support Vector Machine | N: Naïve Bayes | L: Logistic Regression

Table 3.3: Classification Unweighted Average Recall (%) using the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) features.

to the existing feature set does not improve performance; they actually worsened the results in certain cases. This may be caused by the non-ideal recording conditions of the dataset. Further, the proposed high-level acoustic features may already capture relevant information encoded by GeMAPS features, thus making them redundant. These observations help emphasize the importance of feature engineering in this work.

3.6.2 Classification Performance

Table 3.4 summarizes the classification results and the best performing classifier for methods using human-labeled (*Oracle*) and automatic (*Simple*, *Extended*) transcripts. We also explore a fourth method, *Merged*, which involves combining features of *Simple* and *Extended* transcripts. These two transcript types may offer complementary information, as suggested by the WER results in Figure 3.3.

Across the three scoring categories, *Clarity* has the highest degree of agreement in humans, followed by *Fluidity* and *Prosody*, respectively. However, the trend is different for automatic classification, where *Fluidity* is the easiest to classify, followed by *Clarity* and *Prosody*. Similar to our previous works, *Prosody* remains the most challenging task for both humans and automatic classifiers [78, 83].

A classification UAR is considered comparable to human performance if it is better or within one standard deviation from the average human UAR. We can see that *Fluidity* can be estimated very reliably, with results comparable to human across all labeling schemes and transcript types. The results for *Clarity* suggest that the proposed feature set can capture

		Oracle	**Simple**	**Extended**	**Merged**
3-class	*Clarity*	67.3 (N)	59.3 (N)	61.9 (N)	64.3 (N)
	Fluidity	76.1*(N)	73.3*(N)	73.1*(N)	74.1*(N)
	Prosody	66.9*(N)	65.3*(N)	63.5*(N)	65.0*(N)
2-class	*Clarity*	79.1*(L)	75.6 (L)	77.9*(L)	78.8*(L)
	Fluidity	86.5*(L)	81.9*(L)	83.4*(L)	83.2*(S)
	Prosody	72.5 (N)	72.6 (N)	70.4 (N)	72.3 (N)

N: Naïve Bayes | L: Logistic Regression | S: Support Vector Machine

$*$ = higher than or within one std. deviation from avg. human UAR

Table 3.4: Classification Unweighted Average Recall (%) of our speech intelligibility assessment systems. *Oracle* denotes results using human-labeled transcripts. *Simple*, *Extended*, and *Merged* indicate results using automated transcripts.

this category more effectively at the coarser *2-class* level, producing results comparable to human, but is not sensitive enough for the *3-class* case. The opposite is true for *Prosody*, where *3-class* results are closer to human performance. These observations suggest that labeling schemes should be adjusted accordingly depending on the target scoring category.

We compare the classification performance of the four transcript types using repeated measures Analysis of Variance (ANOVA) followed by paired t-test, both with a significance level of 0.05. We use speaker-level accuracies as observations and repeat the test for each combination of labeling scheme and scoring category. These tests show that 3-class *Clarity* using *Oracle* transcripts is statistically significantly better (s.s.b.) than using *Simple* (p = 0.035) and *Extended* (p = 0.008) transcripts, but not *Merged* (p = 0.079). Similarly, 2-class *Clarity* with *Oracle* transcripts is s.s.b. than with *Simple* (p = 0.021), but not *Extended* (p = 0.109) and *Merged* (p = 0.329). 2-class *Fluidity* is significantly better for *Oracle* compared to all other methods, *Simple* (p = 0.012), *Extended* (p = 0.001), and *Merged* (p = 0.017). Of the three automatic methods, *Merged* performs the best and is the closest to *Oracle*, confirming the intuition that the two transcript types, *Simple* and *Extended*, offer complementary information. *Clarity* performance with *Merged* is s.s.b. than with *Extended* for 3-class (p = 0.006) and *Simple* for 2-class (p = 0.038). *Merged* generally performs better than or comparable to both *Simple* and *Extended* for the other categories, but the differences are not statistically significant.

In all *3-class* classification tasks, Naïve Bayes yields the highest UAR. On the other hand, Logistic Regression and SVM outperform Naïve Bayes in the *2-class* version of *Clarity* and *Fluidity*. One possible explanation is the lack of per-class training data in the *3-class* labeling scheme. Naïve Bayes is known to perform well when training data are scarce [114]. More complicated algorithms, specifically Logistic Regression and SVM in this case, begin to outperform Naïve Bayes as the amount of per-class data increases when switching to *2-class*. The same line of reasoning can also explain why *Prosody* is an exception to this phenomenon. The label "4" only constitutes 25.7% of the ground-truths and will suffer from training data scarcity in both labeling schemes. We use all classifiers with default hyperparameters in this work. In future work, we will look into model selection, which can be beneficial for methods with many hyperparameters.

Finally, we analyze the correlation between speaker-level classification accuracies and various speaker characteristics available in the dataset. We limit the analysis to *Oracle* transcripts to eliminate variations caused by automatic transcription errors. A statistically significant correlation is found between age and 3-class *Fluidity* ($r = 0.60$, $p = 0.011$), as well as AOS and 2-class *Clarity* ($r = 0.61$, $p = 0.009$). This implies that 3-class *Fluidity* can be predicted more reliably for more elderly speakers, and 2-class *Clarity* is easier to estimate for those with AOS. These results may help us personalize the model using readily available speaker properties.

3.6.3 Feature Analysis

The goal of this section is to identify the most relevant features for each scoring category, as well as how their relevance changes when moving from human-labeled (*Oracle*) to automatic (*Merged*) transcripts. We first perform mRMR on the entire dataset to partly eliminate features with high correlation. These features are grouped into six sets:

- *ROS*: rate of speech features, e.g., number of syllables and words spoken per minute,

	Clarity		**Fluidity**		**Prosody**	
	Oracle	***Merged***	***Oracle***	***Merged***	***Oracle***	***Merged***
ROS	0.17 (2)	0.12 (3)	0.37 (4)	0.31 (6)	0.14 (2)	0.13 (2)
DUR	0.08 (2)	0.17 (1)	0.32 (3)	0.32 (3)	0.10 (5)	0.14 (4)
GOP	0.21 (8)	0.17 (15)	0.11 (2)	0.13 (4)	0.06 (3)	0.06 (3)
IWR	0.17 (3)	0.18 (5)	N/A (0)	N/A (0)	0.02 (1)	0.05 (1)
PVE	0.21 (2)	0.15 (3)	0.40 (4)	0.25 (4)	0.16 (3)	0.12 (2)
DTW	0.08 (1)	0.03 (1)	0.10 (4)	0.07 (4)	0.04 (4)	0.04 (3)

Table 3.5: Mean Information Gain for different feature sets across scoring categories (2-class) and transcript types. Numbers inside the parentheses denote the number of features from each set selected by minimum-redundancy-maximum-relevance (mRMR).

phonation rate (Section 3.4.1)

- *DUR*: duration features, e.g., duration of pauses and filler (Section 3.4.1)

- *GOP*: word and phone GOP scores (Section 3.4.2)

- *IWR*: isolated word recognition features (Section 3.4.2)

- *PVE*: PVE features (Section 3.4.4)

- *DTW*: pitch and intensity DTW features (3.4.5)

We then compute the mean Information Gain (IG) for each feature set with respect to the *2-class* ground-truth labels (Table 3.5). *3-class* labels are omitted from the analysis given that they produce very similar results.

Pronunciation (GOP, IWR) features are the most prominent indicators of *Clarity* in terms of the number of features selected and IG. On the other hand, *Fluidity* and *Prosody* are mostly dominated by transcript (ROS, DUR) and rhythm (PVE) features. Intonation (DTW) features are selected more frequently in *Fluidity* and *Prosody*; however, they contribute relatively little in terms of IG for all cases. The overall IGs for different scoring categories roughly mirror the automatic classification performance, where *Fluidity* has both the highest information gain and UAR, followed by *Clarity* and *Prosody*, respectively. This suggests that there is room for improvement in feature engineering for *Clarity* and *Prosody*.

Comparing the selected features of *Oracle* and *Merged* highlights the impact of automatic transcript generation on the effectiveness of particular features. The general structure of the two lists remains the same, where GOP and IWR figure prominently in *Clarity*, while ROS, DUR, and PVE dominate *Fluidity* and *Prosody*. There are several differences between the two transcript types, however. GOP features have higher IG in the *Oracle* version of *Clarity* than *Merged*, suggesting that GOP is affected by less accurate transcripts to a certain extent. At the same time, IWR can partly compensate for the degradation in GOP when using automatic transcripts, as evidenced by its relatively stable IG. For *Fluidity*, PVE experiences a decrease in IG when moving from *Oracle* to *Merged* and gets displaced by ROS and DUR as the most important features. This suggests that PVE, much like GOP, is more dependent on the transcript quality. Of the two transcript features, ROS is more affected by the associated transcripts, where as DUR is relatively stable.

There are two potential solutions to address the feature differences between *Oracle* and *Merged*. One, the engineering of new features that are more robust toward inaccurate transcripts. Two, improvements in accuracy of automatic transcript generation. We will explore these directions in future work.

3.7 Conclusion

In this chapter, we presented one of the first comprehensive solutions for estimating aspects of aphasic speech intelligibility in a completely automatic manner. We described techniques for automatic transcript generation, including DNN acoustic modeling, out-of-domain adaptation, and forced-alignment-based language modeling. We presented novel features to capture speech intelligibility, some of which are adapted from clinical practice, such as the rate of word and syllable production. The results demonstrated the potential of automatic approaches for classifying speech intelligibility. Most notably, *Fluidity* can be estimated at human-level accuracy using automatically generated transcripts. However, the estimation of *Clarity* and *Prosody* has not yet achieved human-level performance.

Moving forward, there are two separate problems we need to tackle in order to bridge the gap between human and automatic performance in classifying qualitative measures of aphasic speech. We need to make oracle methods approach human-level accuracy by making advances in feature engineering, feature selection, and classification algorithms. Lexical and linguistic features [42,43,53,68,122] may provide further improvement given that aphasia is primarily a language disorder. We investigate these feature types in Chapter 6 for estimating aphasia severity from spontaneous speech. While mRMR generally works well in practice, other feature selection methods that are directly tied to classification performance may result in additional gain. Lastly, we will look into hyperparameter tuning and model selection to better accommodate the test speaker.

The second problem that needs to be solved is to make fully automatic methods approach oracle-level performance by making advances in ASR, specifically acoustic and language modeling. Aphasic speech contains many abnormalities, including fillers, false starts, mispronunciations, word repetition, insertion, substitution, and deletion with respect to the target prompt. We need specialized acoustic and language models to recognize these atypical patterns effectively. Personalized acoustic and language models are promising, given that the corpus contains a heterogeneous set of speakers and a general model may not be the most appropriate [26–28, 146, 147]. Auxiliary input features that capture relevant speaker characteristics can potentially mitigate the high degree of speaker variability in disordered speech [3,25]. We will investigate methods for improving ASR performance on aphasic speech in the next chapter.

Classifying speech intelligibility is only one among many other problems that need to be solved in order to enable effective in-home exercise for aphasia rehabilitation. PWAs require meaningful feedback during the course of an exercise. Additional research is needed to leverage classification results to produce concrete feedback that the PWAs can use to improve their speech. Provided that PWAs may have limited motor control and audiovisual perception impairment, more work in user interface design is required to develop an application that is both easy to use and sufficiently engaging.

3.8 Work Published

The work presented in this chapter was published in the following articles:

1. **Duc Le**, Keli Licata, Elizabeth Mercado, Carol Persad, and Emily Mower Provost. "Automatic Analysis of Speech Quality for Aphasia Treatment," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy. May, 2014.

2. **Duc Le** and Emily Mower Provost. "Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation." *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Singapore. September, 2014.

3. **Duc Le**, Keli Licata, Carol Persad, and Emily Mower Provost, "Automatic Assessment of Speech Intelligibility for Individuals With Aphasia," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 24:11(2187-2199). November, 2016.

CHAPTER 4

Improving Aphasic Speech Recognition

4.1 Introduction

In Chapter 3, we proposed an automated intelligibility assessment system for constrained aphasic speech. Our system used modified forced alignment in place of traditional automatic speech recognition (ASR) for transcript generation. This technique was possible because the target prompts were known and did not deviate significantly from the produced utterances, due to the controlled recording environment and restricted lexical content. However, this is an unrealistic assumption for unconstrained speech, which plays an important role in everyday interaction of persons with aphasia (PWAs). Conventional ASR is therefore required to enable automatic speech-language assessment for this type of speech.

ASR for aphasic speech is considerably challenging for a number of reasons. First, a PWA's pronunciation can be distorted due to co-occurring motor control disorders such as apraxia of speech (AOS) or dysarthria. Second, language impairments may result in halting speech that contains jargon and various types of paraphasias, all of which can potentially induce recognition errors. Third, the size of most aphasic speech datasets is relatively small, partly due to the difficulties involved in collecting this type of data at a large scale. Fourth, the high variability among PWAs makes it difficult for models to generalize to unseen speakers, especially when training data are limited. As a result, the majority of previous works in aphasic speech assessment had to use mismatched acoustic models trained on external datasets, and ASR performance on aphasic speech is not well understood.

In this chapter, we present a two-part study that aims to improve aphasic speech recognition by leveraging data from AphasiaBank [96] (Section 2.2). The initial study establishes the first large-vocabulary continuous speech recognition (LVCSR) baseline on English AphasiaBank using Deep Neural Network (DNN) acoustic models trained on Mel-frequency cepstral coefficient (MFCC) features augmented with utterance-level i-vectors. The results show that appending i-vectors to frame-level acoustic features leads to a **3.1%** to **15.1%** relative reduction in per-speaker Phone Error Rate (PER), with more severe speakers receiving larger improvement. We also investigate out-of-domain adaptation methods to adapt AphasiaBank models to the University of Michigan Aphasia Program (UMAP) dataset (Section 2.1). The proposed discriminative pretraining method results in a mean relative PER reduction of **18.8%** per speaker, with a standard deviation of **9.1%**.

Our follow-up study extends the previous ASR training methods in four ways to further improve aphasic speech recognition performance. First, we replace word-level phonological and neologistic errors with their known targets to make AphasiaBank transcripts, originally transcribed in CHAT format [95], more compatible with traditional ASR systems. Second, we use log Mel filterbank coefficient (MFB) features in place of MFCCs; the former have recently become the standard input for the large majority of state-of-the-art neural network-based acoustic models. Third, we augment the original training set with 41.7 hours of AphasiaBank data collected from healthy controls. Finally, we replace DNN with a multi-task deep Bidirectional Long-Short Term Memory Recurrent Neural Network (BLSTM-RNN) acoustic model that predicts the senone (i.e., hidden state within triphone HMM) and monophone labels simultaneously. Together, these changes result in an overall Word Error Rate (WER) of **37.37%**, a relative improvement of **30.8%** compared to the best DNN system from the first study.

This work helps further the understanding of aphasic speech recognition, provides insights into the types of speakers who would benefit from different adaptation techniques, demonstrates the potential of AphasiaBank in ASR, and suggests that automatic speech-language assessment of unconstrained aphasic speech, which relies heavily on ASR, may be feasible in certain contexts. The ASR systems described in this work will be an important component in the remaining chapters of this dissertation.

4.2 Related Work

4.2.1 Under-Resourced ASR

Disordered speech recognition also shares important similarities with low-resource ASR due to the issue of data scarcity. Common techniques for handling this problem include deep bottleneck [47,167] or posterior-based [8] features used within tandem-based systems [63], and discriminatively pretrained DNN acoustic model using out-of-domain data [156]. A shared theme of these methods is the use of external speech (e.g., multilingual data) for enhancing the performance of in-domain models. In the context of ASR for disordered speech, out-of-domain data usually consist of healthy speech [3, 25]. However, there is an inherent mismatch between healthy and disordered speech [27], suggesting that healthy speech data may not be the most appropriate choice for out-of-domain adaptation. We leverage aphasic speech directly as out-of-domain data in this work.

4.2.2 ASR with i-Vectors

i-Vector front-end analysis [33, 50] has emerged as the state-of-the-art in speaker verification [133] and a variety of other speech processing tasks [7, 48, 134, 165], for example: language detection [48, 165], accent detection [7], age estimation [134], and compensation for gender in speaker recognition [144]. The i-vector technique assumes that there is a difference between a general model, called a universal background model (UBM), and a model associated with a subset of speech data and that this difference lies in a low-dimensional space [133]. This space is called the total variability space because it captures all sources of variability in speech such as speaker identity, language, channel information, age, gender and emotion. A speaker's acoustic profile is formulated as: $M = m + Tw$, where m is a speaker-independent and session-independent supervector, T is a low rank matrix, and w is a random vector (normally distributed $N(0, I)$). T is trained using an Expectation Maximization (EM) algorithm [71]. The components of w form the i-vector, which contains information describing how the speaker is different from a general subject population.

Recent studies have shown that appending i-vectors to frame-level acoustic features leads to significant improvement for DNN-based acoustic models [46, 140, 143, 153]. Systems having i-vectors as auxiliary features can be thought of as performing bias adaptation based on the input data, leading to better generalization. The i-vector approach is promising for handling the high speaker variability present in disordered speech; however, its application to this type of data has been limited.

4.3 Data

4.3.1 ApahasiaBank

In this work, we consider utterances in the core AphasiaBank dataset (Section 2.2.1) spoken by PWAs. We resample all audio files to 16kHz and use Kaldi [124] to extract two sets of frame-level acoustic features: (1) 12-dimensional MFCCs plus energy, along with the first and second order derivatives, and (2) 40-dimensional MFBs. We use a 25ms window and 10ms frame shift for both feature types. The features are z-normalized at the speaker level. Finally, we perform speaker-independent 4-fold partitioning to evaluate ASR performance on unseen speakers. 25% of speakers are withheld from each sub-dataset to form the test set. We further withhold 15% of training speakers from each sub-dataset to form a development set. The test sets across these four folds form a complete partition of the dataset. The amount of per-fold training data ranges from 55.1 to 58.7 hours.

The core AphasiaBank dataset also contains 41.7 hours of healthy speech data. In the second study, we consider augmenting each fold-specific training set with this collection of healthy speech, under the hypothesis that having additional training data will improve recognition performance.

4.3.2 UMAP

The UMAP dataset was used in Chapter 3 for studying automatic speech intelligibility assessment. A major bottleneck in this work was the reliance on predefined speech prompts. Achieving good ASR performance on UMAP will move us closer to deploying the system for real-world usage with spontaneous speech as input. We split each UMAP utterance into continuous segments of intelligible speech, each of which contains on average 2 to 4 words. We will perform ASR evaluation on these segments. The segment-level data contains in total 2.1 hours and 12,661 instances of 1,073 unique words.

We apply an identical feature extraction pipeline used in AphasiaBank. ASR evaluation will be done through leave-one-speaker-out cross-validation, which results in 17 folds where data from one speaker are withheld for testing and the rest are used for training. We further withhold 15% of utterances from each training speaker to form a development set. The size of the per-fold training set ranges from 1.7 to 2 hours.

4.4 Initial Work

4.4.1 AphasiaBank Transcript Preparation

Utterances in AphasiaBank were transcribed using the CHAT format [95]. The transcriptions contain a variety of special codes to aid with language sample analysis, such as word-level and utterance-level errors [4], sound fragments, repetitions, non-verbal actions, among others. The first row of Table 4.1 shows an example *raw* transcript containing

Raw	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>.

Table 4.1: Example AphasiaBank transcript and its cleaned version.

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, marked 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 a method to process CHAT transcripts to be compatible with traditional ASR systems while preserving as much of the original pronunciations as possible. 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). IPA strings are converted to special hashed tokens such that the same IPA pronunciations map to the same hash. The second row of Table 4.1 shows an example *cleaned* transcript, in which **&uh** is mapped to <**FLR**>, and the two non-word phonological errors **pinək@u** and **bʌðə@u** are replaced with hashed tokens <**U1**> and <**U2**>. All ASR-related experiments involving AphasiaBank in this section are conducted on *cleaned* transcripts.

4.4.2 Lexicon Preparation

The lexicon used in this work is based on the CMU dictionary[1], containing 39 regular phones, plus five special phones: silence, <**FLR**>, <**LAU**>, <**SPN**>, and <**BRTH**>. Each IPA pronunciation is heuristically mapped to a sequence of CMU phones. For example,

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

	AphasiaBank	**UMAP**
GMM	CD tied-state triphones trained with Maximum Likelihood.	
	Parameters:	
	25,000 Gaussians; 3,000 senones.	8,000 Gaussians; 700-800 senones.
DNN	5 hidden layers, 1024 units per layer, sigmoid activation, SGD training. 27-frame context windows, HMM-GMM alignments, CE objective.	
	W/o i-vectors: exponential-decay (0.4 initial rate, 0.05% threshold).	
	No regularization.	2×10^{-5} L2 regularization weight.
	W/ i-vectors: step-decay (0.4 initial rate, 0.01 minimum rate).	
	10^{-5} L2 regularization weight.	2×10^{-5} L2 regularization weight.
i-vectors	***UBM***: 1024 Gaussians trained on 9-frame spliced MFCCs + 40-dim LDA on senones. Only voiced frames are used.	
	Type of i-vector:	
	32-dim utterance-level.	32-dim session-level.
Decoding	Continuous phone loop with trigram phone-level language model.	

Table 4.2: Training and decoding methods for intra-dataset automatic speech recognition experiments. See text for description of learning schedule and i-vector type.

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.

4.4.3 Intra-Dataset Speech Recognition

In this section, we outline our experiments for intra-dataset speech recognition, which will result in a speaker-independent cross-validated PER for each dataset. We consider two classes of methods, one based on the traditional context-dependent tied-state triphone HMM-GMM model, and one based on the more modern hybrid HMM-DNN system [64, 110]. Two versions of HMM-DNN are trained, one with and one without i-vectors in the input features. Details about this experiment are summarized in Table 4.2. We use Kaldi [124] for HMM-GMM modeling and i-vector extraction, and Theano [155] for DNN training. Additional data for replicating this work, such as fold selection, transcription, and audio

[2] http://www.speech.cs.cmu.edu/tools/lextool.html

segmentation, are available online[3]. For the remainder of this section, we will elaborate on the hyperparameter choice, learning schedule, and i-vector extraction.

4.4.3.1 Hyperparameter Selection

HMM-GMM and HMM-DNN both require a number of hand-picked hyperparameters, such as the number of Gaussians and tied-states for the former, and the DNN architecture and training recipe for the latter. Hyperparameters for AphasiaBank were selected based on the average PER achieved on the development set across all four cross-validation folds. On the other hand, hyperparameters for UMAP were selected using an oracle method that optimizes for test PER. Doing so helps us obtain the strongest UMAP baseline to compare against out-of-domain adaptation techniques described in later sections.

4.4.3.2 Learning Schedule

Learning schedule refers to the adjustment of learning rate after each DNN stochastic gradient descent epoch. We find that different learning schedules must be used for models with and without i-vectors to achieve optimal results.

Exponential-decay: This schedule first trains the network at a fixed initial learning rate (e.g., 0.4). Once the change in frame-level error on the development set drops below a threshold (e.g., 0.05% absolute), we halve the learning rate after every epoch. The training process terminates once the change in development error once again drops below the threshold. We find that this schedule is appropriate for models without using i-vectors, possibly because it finishes faster and avoids overfitting the network to the training set, which is easier to do without having additional features to model.

Step-decay: This schedule is similar to the one used in [110]. Instead of halving the learning rate after every epoch, it halves the learning rate and restores previous network weights whenever the development error does not improve. The training process terminates

[3] http://www.umich.edu/ ducle/IS16appendix

once the learning rate drops below a minimum value (e.g., 0.01). We find that this schedule is appropriate for less stable learning process, such as when i-vectors are used.

4.4.3.3 i-Vector Extraction

i-Vectors are typically extracted at the speaker level over a relatively large amount of data. However, this approach requires all data from a speaker to be available before decoding, which is often not possible in ASR. The alternative is to extract utterance-level i-vectors, which have been shown to improve ASR performance [143]. In this work, AphasiaBank i-vectors are extracted on a per-utterance basis. On the other hand, this type of i-vector does not work well on UMAP, possibly because the utterances in UMAP are excessively short and do not contain sufficient distinguishing information. We instead use session i-vectors, which are extracted from speech data produced by the PWA in one single recording session[4]. There are 125 sessions, each containing 1 minute of speech on average.

The input features are transformed before UBM training and i-vector extraction to better reflect variability in the phoneme space, following [140]. We first perform energy-based Voice Activity Detection (VAD) to discard silent frames. Next, nine consecutive MFCC frames are spliced and projected down to 40 dimensions using Linear Discriminant Analysis (LDA), with triphone states as the target class labels. The i-vector dimension is set to 32 for both datasets, based on preliminary experiments and the system described in [46].

4.4.3.4 Results

Table 4.3 summarizes the mean and standard deviation of speaker-level PERs on AphasiaBank, where the speakers are grouped by the level of severity defined by WAB-R AQ.

We first turn attention to the relatively high PERs achieved on this dataset. This may be caused by the abnormal speech patterns associated with aphasia that are difficult to capture with conventional ASR techniques. Speech data in AphasiaBank were recorded

[4] Session-level i-vectors are not the same as speaker-level i-vectors since a speaker typically has multiple recording sessions.

Severity	**No i-vectors**	**With i-vectors**
mild	48.95 ± 11.55	47.41 ± 10.46
moderate	57.04 ± 13.22	52.79 ± 10.37
severe	65.44 ± 18.65	61.00 ± 13.20
v. severe	89.27 ± 29.14	75.81 ± 18.65
unknown	60.36 ± 29.75	54.35 ± 18.64

Table 4.3: AphasiaBank per-speaker Phone Error Rate (PER), grouped by severity.

using video cameras situated far away from the speaker. This far-field recording condition is known to significantly reduce recognition performance. Two observations can be made from these results. One, if we want to apply ASR technology to help improve the well-being of PWAs, it is crucial to constrain the recognition problem in some way, such as restricting the vocabulary or task grammar. Aphasic speech may be too challenging for unconstrained LVCSR to achieve an acceptable recognition accuracy. Two, it is important to realize that ASR is only a precursor and not an end goal for speech-based technology aimed toward PWAs. It will be interesting to investigate tasks that can be performed reasonably well given imperfect ASR output. This will help us better understand what kind of ASR-dependent technology is feasible for aphasic speech.

These results also show that both the mean and standard deviation of per-speaker PERs tend to increase as a PWA's aphasia becomes more severe on the WAB-R AQ scale. This is a useful observation as it shows that AQ, despite being a measure of general language skills and not of speech itself, can be a reasonable estimate for the effectiveness of ASR. Being able to predict how well an ASR system will work for a speaker using readily available information such as AQ may help the system adapt to that speaker more quickly and effectively. A natural extension of this observation is to use a speaker's severity level directly as input to the DNN, such as encoding it as a one-hot vector. However, our preliminary experiments indicate that this approach does not yield additional improvement on top of i-vectors. We will explore different methods to augment acoustic modeling with PWAs' diagnoses in future work.

Finally, we note the effectiveness of i-vectors for aphasic speech recognition. Models that use i-vectors in the input features experience a reduction in both the mean and standard deviation of per-speaker PER. While the relative improvement for speakers with mild aphasia is relatively small (3.1%), the improvement is more noticeable for those with moderate to severe (6.8% – 7.5%), and especially very severe aphasia (15.1%). Christensen et al. noted that although their out-of-domain adaptation technique is quite effective, speakers with more severe dysarthria tend to benefit less from adaptation [25]. Our results suggest a complementary method for improving the recognition rate of the more severe population.

4.4.4 Adapting AphasiaBank to UMAP

We consider two methods to use AphasiaBank to improve recognition results on UMAP.

merged: In this method, we merge the full AphasiaBank corpus' training and development set with the UMAP counterparts, and train a new DNN using the same recipe and architecture described in Table 4.2. This method allows the network to directly model UMAP data while also modeling the large amount of speech present in AphasiaBank. A potential disadvantage of this method is that it might not model UMAP data extensively since UMAP contributes only a relatively small fraction of the training data.

dpAB: We investigate **d**iscriminative **p**retraining with **A**phasia**B**ank data inspired by the work of Thomas et al. for low-resource ASR [156]. The authors in [156] proposed retraining only the softmax layer while keeping the lower layers fixed. However, we find that retraining the entire AphasiaBank DNN on the UMAP training set, using the step-decay learning schedule and no regularization, yields better results. This suggests that the high-level representation learned by AphasiaBank DNN does not transfer directly to UMAP data. This indicates a large mismatch between the two datasets, and further suggests that methods which aim to constrain the shift in parameters from the original model by inserting additional layers on top of a fixed network [90] or regularizing the change in output distribution [168] may have limited efficacy. Speaker adaptation on the same dataset,

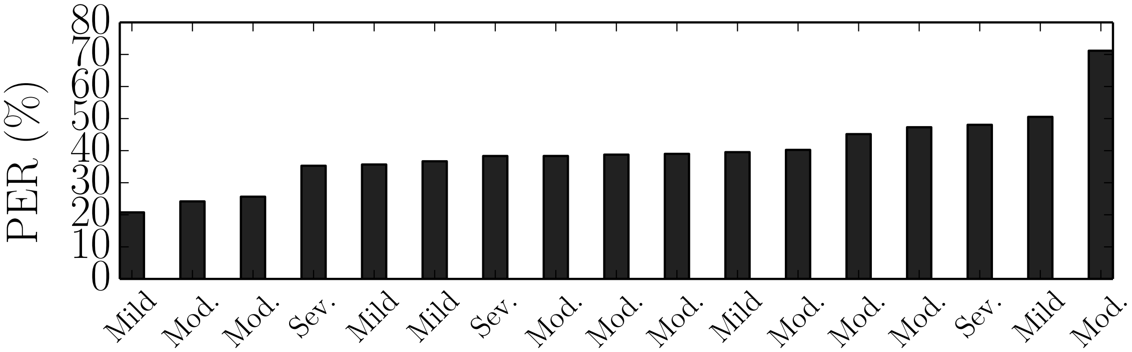


Figure 4.1: UMAP per-speaker Phone Error Rate (PER) using Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) trained on UMAP data. x-axis denotes each subject's severity level according to the revised Western Aphasia Battery Aphasia Quotient.

Model	**No i-vectors**	**With i-vectors**
AB-DNN	4.8 ± 15.1	3.4 ± 15.5
UMAP-DNN	-1.0 ± 7.6	2.9 ± 9.0
merged	-14.7 ± 9.3	**-16.6 ± 8.9**
dpAB	**-18.8 ± 9.1**	-15.9 ± 7.6

Table 4.4: Relative change (%) in UMAP per-speaker Phone Error Rate (PER) compared to the Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) baseline. A negative value means reduced PER. *AB-DNN* and *UMAP-DNN* are Deep Neural Networks (DNNs) trained only on AphasiaBank and UMAP data, respectively.

which does not suffer from such data mismatch, may benefit more from these approaches.

We also considered using deep bottleneck features (DBNFs) generated by AphasiaBank DNN in a tandem-based system. However, our preliminary experiments were not able to outperform the HMM-GMM baseline. Again, this may be due to the high level of mismatch between AphasiaBank and UMAP. As a result, we do not consider DBNFs here.

4.4.4.1 Results

Figure 4.1 shows the PERs for different speakers in the UMAP dataset using the HMM-GMM baseline model. The PERs range from 20.8% to 71.2% (mean 39.7%, std. deviation 11.1%). We will estimate the effectiveness of different adaptation methods based on the resulting change in PER for each speaker. These are summarized in Table 4.4.

The first two rows, *AB-DNN* and *UMAP-DNN*, refer to DNN acoustic models trained only on AphasiaBank and UMAP data, respectively. Compared to the baseline, the resulting PERs for both models improve for some speakers and worsen for others, and there is no clear advantage to using either model. The fact that *UMAP-DNN* was not able to outperform the HMM-GMM baseline reinforces the data scarcity problem in aphasic speech recognition. Looking at individual speakers, *AB-DNN* tends to work better for those who are similar to the typical speakers in AphasiaBank, namely those with mild and fluent aphasia. On the other hand, there is no obvious pattern as to which type of speaker benefits from the *UMAP-DNN* model.

Of the two adaptation methods, the best result (**18.8% ± 9.1%** relative improvement) is achieved with *dpAB*, which uses UMAP data to finetune a DNN that was discriminatively pretrained on AphasiaBank. Speakers with mild severity receive the largest improvement (22.5% ± 7.5%), while those with fluent and non-fluent aphasia experience a similar degree of PER reduction (19.2% ± 9.3% vs. 18.3% ± 9.0%). On the other hand, the *merged* adaptation method provides more benefit to those with fluent aphasia, resulting in 18.7% ± 6.1% relative improvement compared to 14.7% ± 10.5% for non-fluent.

Finally, we analyze the effect of i-vectors on adaptation. Using i-vectors resulted in better performance for *AB-DNN* and *merged*, but worse performance for *UMAP-DNN* and *dpAB*. The common theme among the two methods that were not able to take advantage of i-vectors is that only UMAP i-vectors were used for DNN training. On the other hand, using UMAP i-vectors directly in testing (*AB-DNN*) or training them jointly with AphasiaBank i-vectors (*merged*) proved beneficial. The 125 UMAP session i-vectors are possibly too few in number and too dissimilar for the network to take advantage of in a speaker-independent setup. Additional work is needed to leverage i-vectors in limited-data situations.

Raw	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.

Table 4.5: Example AphasiaBank transcript and its two 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).

4.5 Follow-Up Study

Having established an initial LVCSR baseline on AphasiaBank using DNN acoustic models trained on MFCCs augmented with utterance-level i-vectors, we now aim to further improve recognition accuracy on AphasiaBank. This is achieved with four major changes to the existing ASR training methods, which we describe in detail in the following sections.

4.5.1 Target AphasiaBank Transcripts

We previously proposed a way to clean *raw* AphasiaBank transcripts to make them more compatible with standard ASR systems. While the resulting *cleaned* transcripts preserve the original pronunciations and are therefore suitable targets for acoustic modeling, the retained word-level errors are difficult to recognize for two reasons. First, they are not well captured by the language model given the irregular language patterns associated with word errors. Second, many word-level errors, especially neologistic and non-word phonological errors, are not present in the training lexicon and will therefore be unrecognizable. We mitigate this problem by producing a second set of ***target*** transcripts in which all word-level errors, excluding semantic, are replaced with their known targets. An example is shown in Table 4.5, 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 *target* transcripts better reflect a PWA's language usage patterns and will be used for language modeling as well as ASR evaluation. Acoustic model training will still make use of *cleaned* transcripts.

4.5.2 Frontend

The acoustic models employed in this section will make use of MFB features instead of MFCCs. The former have recently become the standard input in the majority of state-of-the-art neural network-based ASR systems because modern deep learning acoustic models can better exploit the inter-correlation of MFBs compared to the decorrelated MFCCs. Our i-vector extraction system uses GMMs with diagonal covariance matrices and will still make use of LDA-transformed MFCC features.

4.5.3 Control Data

We augment the original training data with AphasiaBank healthy control speech, which amounts to 41.7 hours across 187 speakers. These additional data are used in both acoustic and language model training. We hypothesize that including this set of healthy speech will improve the generalizability and accuracy of our models. AphasiaBank healthy data are collected with similar elicitation protocols as those used for aphasic speech. As a result, their lexical content will resemble that of the test data to a certain extent and help improve language modeling. In addition, speech production difficulties in aphasia may lead to inaccurate frame-level target labels in the original training set and negatively affect acoustic modeling. Adding healthy speech, which is likely to have accurate frame-level target labels with similar distribution as those in the test set (due to the similar lexical content), will help mitigate this problem.

4.5.4 Multi-Task BLSTM-RNN Acoustic Model

We replace DNN with BLSTM-RNN for acoustic modeling, motivated by the fact that the latter has recently achieved state-of-the-art results on various ASR benchmarks [54, 136,

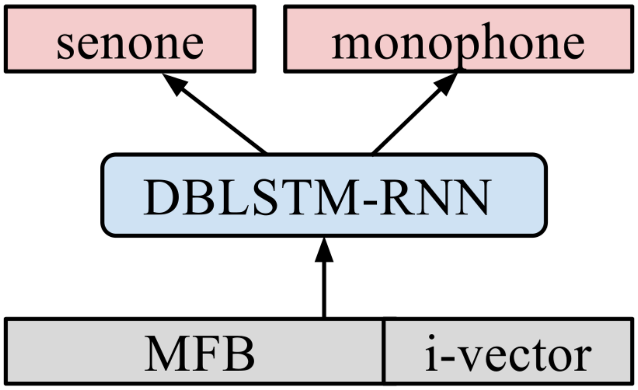


Figure 4.2: Deep multi-task Bidirectional Long-Short Term Memory Recurrent Neural Network (BLSTM-RNN) acoustic model.

138]. Further, we train the model to simultaneously predict both the correct senone and monophone labels, a well-known technique that can improve classification performance due to its regularization effect [9, 10, 142]. In addition to regularizing the network, this multi-task model has two additional advantages when applied to aphasic speech. First, the more robust but less fine-grained monophone labels can act as a correcting signal for senone labels, which may be inaccurate due to the speech production difficulties associated with aphasia. Second, the monophone output induces a multi-dimensional time series that can be used to compactly represent words and phones. These time series, also referred to as *posteriorgrams*, help extract features for paraphasia detection (Chapter 5) and aphasia severity estimation (Chapter 6).

Following [79], we augment each MFB frame with five left and five right neighbors in addition to the corresponding utterance-level i-vector, resulting in 472 dimensions per frame[5]. These features are modeled with a stacked BLSTM-RNN comprising four hidden layers, each with 1,200 units (600 for forward, 600 for backward). The model has two parallel softmax output layers corresponding to the senone and monophone labels (Figure 4.2). The number of senones varies across folds, ranging from 4,472 to 4,563.

[5] 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).

The model is trained using the Adam optimizer [75] 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 seconds. Only less than 0.5% of training utterances are longer than 25 seconds, 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, along with early stopping based on the development senone Frame Error Rate (FER) and step-decay learning schedule [82]. After each training epoch, we halve the current learning rate and restore the previous model parameters if the senone FER on the development set increases. The training process finishes once the learning rate drops below 0.00005.

Single-Task Baseline: To analyze the effect of multi-task learning, we use an identical method to train BLSTM-RNN acoustic models with only a single senone output layer.

DNN Baseline: To better evaluate the effectiveness of BLSTM-RNN, we employ DNN acoustic models trained using the same recipe described in Section 4.4.3. The models consist of four hidden layers with 2048 units each and one senone output layer.

4.5.5 Results

For each evaluation fold, we use SRILM [152] to train a trigram language model (LM) with backoff on the training *target* transcripts. We tune the decoder's language model

Features	**DNN**		**BLSTM-RNN**	
	5×1024	***4×2048***	***ST***	***MT***
MFCC	55.07*	48.61	-	-
MFCC + i-vectors	54.01*	47.14	-	-
MFB	-	46.26	39.37	38.95
MFB + i-vectors	-	45.26	37.69	**37.37**

*: previous baseline | *N*×*L*: N hidden layers with L units each | *ST*: single-task | *MT*: multi-task

Table 4.6: AphasiaBank Word Error Rate (WER) under different input feature and acoustic model configurations.

Utterance Type		**Aphasia Severity (WAB-R AQ)**			
Free Speech	***Semi-Spontaneous***	***Mild***	***Moderate***	***Severe***	***V. Severe***
38.79	36.24	33.68	41.11	49.21	63.17

Table 4.7: AphasiaBank Word Error Rate (WER) by utterance type and aphasia severity according to the revised Western Aphasia Battery Aphasia Quotient (WAB-R AQ).

weight {9, 10, . . . , 20} and word insertion penalty {0.0, 0.5, 1.0} based on the WER of the development set. The decoded test output is aggregated across all four folds and evaluated against the reference *target* transcripts. The results are summarized in Table 4.6. The best performance, **37.37%** WER, is obtained with the proposed multi-task BLSTM-RNN acoustic model trained on MFBs and utterance-level i-vectors. This is a 30.8% relative improvement compared to our previous DNN baseline trained on MFCCs and i-vectors.

Comparing DNN performance on MFCC and MFB features, it is clear that the latter give better results. The relative improvements in terms of WER are 4.8% (without i-vectors) and 4.0% (with i-vectors). This justifies our decision to use MFB in place of MFCC for acoustic model training.

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 previous finding and demonstrates the efficacy of i-vectors in speaker-independent acoustic modeling for aphasic speech.

Single-task BLSTM-RNN greatly outperforms DNN, resulting in a relative WER reduction of 14.9% (without i-vectors) and 16.7% (with i-vectors). While the improvement in recognition rate attributed to multi-task learning is small (around 1% relative), this method enables the production of posteriorgrams, which will be used later for automatic paraphasia detection (Chapter 5) and aphasia severity estimation (Chapter 6).

Table 4.7 breaks down the WER based on utterance type (free vs. semi-spontaneous speech, described in Section 2.2.1). As can be seen, semi-spontaneous speech is generally easier to recognize compared to free speech. A possible explanation is that the former

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>	1,321	0.81	hospital	416	0.09
I've	249	0.79	people	544	0.08
<BRTH>	1,345	0.73	beautiful	262	0.08
am	153	0.73	weeks	131	0.08

Table 4.8: Words with the highest and lowest error rates.

is more constrained in terms of vocabulary range and syntactic structure, and is therefore more compatible with the language model. This suggests that applications requiring highly accurate ASR should focus on semi-spontaneous speech.

WER also varies based on the severity of aphasia defined by the revised Western Aphasia Battery Aphasia Quotient [72] (Table 4.7). 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 loosely estimate ASR performance for a given PWA. Further, these results indicate that those with severe aphasia will likely have significant difficulties with applications that are reliant on ASR.

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 4.8 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. Combined with the previous observation that semi-spontaneous speech has lower WER, this suggests that ASR is more suitable for non-conversational aphasic speech.

Given these analyses, it is possible that WER can be further reduced by personalizing the acoustic and language models for individual utterance types and/or severity groups. Moreover, speakers who have similar error patterns can potentially be grouped together for more fine-grained acoustic and language model training.

4.6 Conclusion

In this study, we established the first LVCSR baseline on English AphasiaBank, and showed that AphasiaBank data can be leveraged to improve the recognition rate on a smaller aphasic speech corpus by a large margin through discriminative pretraining. The analysis suggests that discriminative pretraining provides more benefit to PWAs with lower severity, while i-vector-based adaptation benefits those with higher severity. However, more work is needed to combine the benefit of both approaches.

Our follow-up work expanded upon this initial study to further improve recognition accuracy on AphasiaBank using a multi-task BLSTM-RNN acoustic model trained on MFBs and utterance-level i-vectors. The proposed system achieves an overall WER of **37.37%**, a 30.8% relative improvement compared to the previous baseline. Subsequent analysis shows that semi-spontaneous speech is easier to recognize than free speech, and that there is a moderate correlation between WAB-R AQ and speaker-level WER. Finally, word-level errors suggest that ASR is more suitable for non-conversational aphasic speech.

We plan to extend this work in two major directions. First, we are interested in the extent to which an improved ASR model can replace human-labeled transcripts in analyzing aphasic speech. We investigate this problem in Chapter 5 and 6. Second, we will explore more fine-grained adaptation methods based on diagnoses and other speaker properties. Given the high speaker variability present in aphasic speech, more highly personalized models may result in further gain [26–28, 146, 147].

4.7 Work Published

The work presented in this chapter was published in the following articles:

1. **Duc Le** and Emily Mower Provost. "Improving Automatic Recognition of Aphasic Speech with Aphasia Bank." *17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, USA. September, 2016.

2. **Duc Le**, Keli Licata, and Emily Mower Provost. "Automatic Quantitative Analysis of Spontaneous Aphasic Speech." *Speech Communication*. (in submission)

CHAPTER 5

Automatic Paraphasia Detection

5.1 Introduction

Anomia (word retrieval deficit) is the core symptom of aphasia and is present in virtually all persons with aphasia (PWAs) [62]. Those who have anomia often produce various types of paraphasias (naming errors), the most common of which are *semantic*, *phonemic*, and *neologistic*. In these three categories, respectively, the PWA may substitute the target word (e.g., *harmonica*) with a semantically related word (e.g., *flute*), a phonemically related word (e.g., *karmonica*), or a non-word (e.g., *parokada*). The type and frequency of the produced paraphasias play an important role in estimating the severity of anomia as well as determining an appropriate treatment approach [44,116]. For example, PWAs who produce mainly semantic paraphasias may benefit from treatment approaches focusing on word meaning, while treatment approaches targeting the phonological structure of target words may be more appropriate for PWAs who produce mainly phonemic paraphasias [98,116].

Being able to detect paraphasias automatically from a PWA's speech (e.g., through a computer-based word-finding exercise) would provide SLPs with a useful tool for both diagnostic and progress-monitoring purposes and, as such, would help guide the treatment process. Additionally, it could lead to computer-based activities for in-home practice for PWAs, thereby increasing the intensity of practice and facilitating carry-over of progress from therapy to other environments. It could also serve to increase a PWA's awareness of errors and enhance self-monitoring skills and, thus, promote independence in overall communication. However, the automatic detection of paraphasias has not previously been studied in the literature.

In this chapter, we present a pilot study that investigates the feasibility of detecting phonemic and neologistic paraphasias automatically from aphasic speech. We demonstrate that when the target transcript is known, phonemic and neologistic paraphasias can be successfully distinguished from correctly pronounced words. We also investigate a variant of the problem in which the target transcript needs to be generated automatically. In this setup, the system is able to outperform the naïve baseline in detecting the presence of paraphasias in utterances, and achieve good correlation in estimating the rate of phonemic paraphasia production for each speaker. The results and analyses provided in this work help lay the initial foundation for future work targeting automatic paraphasia detection.

5.2 Related Work

To the best of our knowledge, no existing work has looked at paraphasia detection in aphasic speech from a technical perspective. Previous works primarily tackled utterance-level and speaker-level classification problems for therapeutic and diagnostic purposes [42, 78, 81, 83, 122]. Peintner et al. [122] proposed speech and language features to distinguish between three types of frontotemporal lobar degeneration, including progressive non-fluent aphasia. Fraser et al. [42] combined text and low-level acoustic features to classify primary progressive aphasia (PPA). Our previous work tackled the problem of predicting utterance-level pronunciation, fluidity, and prosody scores given read speech samples of PWAs [78, 81, 83]. The most closely related works are those of Abad et al. [1, 2], which used keyword spotting to recognize phrases spoken by PWAs in word naming exercises. However, they did not consider fine-grained word-level labels such as paraphasias.

In an oracle setting where there is access to a PWA's target transcript, automatic paraphasia detection shares certain similarities with mispronunciation detection, an extensively studied problem in the literature. The task in both cases is to classify each word in the transcript as either correct or containing errors. We adopt techniques proposed by Lee et al. [84–86], which compared a non-native speaker's word- and phone-level pronunciations against those of a native speaker, using Dynamic Time Warping (DTW) features extracted on phoneme posteriorgrams. However, PWAs often do not produce the correct target due to their speech-language impairments. Consequently, target transcriptions may not be available, and reference utterances do not always exist, making it difficult to apply techniques from mispronunciation detection. In this work, we investigate the oracle use case where target transcripts are available, as well as a more realistic scenario in which automatic speech recognition (ASR) is used to generate the transcripts automatically.

5.3 Data

In this work, we focus on the *Fridriksson* sub-dataset of the *Scripts* portion of English AphasiaBank (Chapter 2), which contains recordings of 12 PWAs reading from four predefined scripts (*advocacy*, *eggs*, *vast*, and *weather*). The other *Scripts* sub-dataset, *Adler*, consists of six high-functioning PWAs and very few instances of paraphasias. We therefore exclude the *Adler* sub-dataset from this study.

Each utterance in the dataset was transcribed verbatim with word-level error codings in concordance with the CHAT transcription format [4, 95]. Word-level error codes include semantic, phonemic, and neologistic paraphasias, each of which is accompanied by a target

Target	I have aphasia
P1	I have the aphasia
P2	have æfeziə@u [: aphasia] [* n:k]
P3	I have vəfeʒə@u [: aphasia] [* p:n]

[* n:k]: neologistic paraphasia | [* p:n]: phonemic paraphasia

Table 5.1: Example AphasiaBank *Scripts* transcripts.

Speaker	**Utts**	**Words**	**Phonemic**	**Neologistic**
P1	85	787	90	72
P2	108	879	108	66
P3	109	1060	113	46
P4	88	767	108	75
P5	67	652	101	36
P6	37	262	28	61
P7	103	1118	67	18
P8	104	1076	117	24
P9	93	901	146	53
P10	6	47	2	4
P11	67	607	136	112
P12	123	1154	101	32
Total	**990**	**9310**	**1117**	**599**

Table 5.2: AphasiaBank *Scripts* dataset summary.

word. Table 5.1 shows example transcripts of three PWAs reading the prompt "*I have aphasia*." P1 produced the target without any paraphasia, but added an extra "*the*." P2 and P3 produced neologistic and phonemic paraphasias, respectively, for the target word "*aphasia*." The actual pronunciation was transcribed in IPA format (ending with `@u`).

We target phonemic and neologistic paraphasias in this work. Detecting semantic paraphasias requires a different approach and will be addressed in future work. Table 5.2 summarizes the 12 speakers in the dataset, along with the utterance and word count, as well as the number of phonemic and neologistic paraphasias. In total, phonemic and neologistic paraphasias account for 12.0% and 6.4% of the words, respectively.

All experiments will be performed with leave-one-speaker-out cross-validation in order to assess the system's performance on unseen speakers. We further withhold 10% of utterances from each training speaker to form a development set.

5.4 Paraphasia Detection

5.4.1 With Known Target Transcripts

We first want to determine if it is possible to separate phonemic and neologistic paraphasias from correct words. The ***target*** transcript of an utterance is defined as the original transcript in which all phonemic and neologistic paraphasias are replaced with their corresponding targets. Thus, the target transcripts in Table 5.1 will be: "*I have the aphasia*" (P1), "*have aphasia*" (P2), and "*I have aphasia*" (P3). Assuming that the target transcripts are available, the goal is then to label each word according to the following binary classification schemes:

- *C–pn*: correct (C) vs. phonemic or neologistic (pn).
- *C–p*: correct (C) vs. phonemic (p).
- *C–n*: correct (C) vs. neologistic (n).

where correct words are defined as those without any error code. We exclude words that do not fall under any labeling category (e.g., semantic paraphasias), as well as audible background noise, breathing sounds, fillers, and laughters.

Metric: although the focus of this work is to detect phonemic and/or neologistic paraphasias, detecting correctly produced words is arguably equally important. We therefore utilize the average F1 score across classes for evaluation.

Baseline: no baseline currently exists as this is the first work to tackle paraphasia detection. We adopt a simple approach that labels every word as correct (the majority class).

5.4.2 Without Known Target Transcripts

The target transcripts will not be available in advance for many real-world applications. We propose to transcribe test utterances automatically with ASR to overcome this limitation.

Given the hypothesized transcripts, it is possible to utilize the same classification models in Section 5.4.1 to obtain predicted word labels.

We consider three types of evaluation metrics that measure the system's performance at the word, utterance, and speaker level. These metrics will help determine the system's applicability under different levels of analyses.

Word-Level Metric: the ideal paraphasia detection system should simultaneously generate the correct target transcripts and label each word accurately. We encode this idea by augmenting the hypothesized and reference target transcripts with corresponding word labels. Under the *C–pn* classification scheme, the augmented reference transcripts in Table 4.1 will be: "*I/C have/C the/C aphasia/C*" (P1), "*have/C aphasia/pn*" (P2), and "*I/C have/C aphasia/pn*" (P3). Given an augmented hypothesized transcript, its Word Error Rate (WER) compared to the reference captures both transcription and word labeling errors. This metric will henceforth be referred to as augmented WER (AWER).

Utterance-Level Metric: aphasic speech is known to be difficult to recognize [82], thus achieving good AWER may be challenging. Instead of providing detailed word-level predictions, the system can simply output whether or not an utterance contains paraphasias, i.e., a binary prediction problem. We again adopt average F1 as the evaluation metric.

Speaker-Level Metric: using the same reasoning, the system can be modified to estimate the rate of paraphasia production for a given speaker, which helps indicate anomia severity. This task is evaluated using the Pearson correlation coefficient (r) between the predicted and actual paraphasia occurrence rate per minute for all speakers in the dataset.

5.5 Methods

5.5.1 Acoustic Modeling

Given the small size of the dataset, we adopt an out-of-domain training approach, motivated by previous work in disordered speech recognition [25, 82]. We first train an acoustic model on the core AphasiaBank dataset, which contains approximately 130.9 hours of spontaneous speech elicited through the AphasiaBank protocol. We then adapt (retrain) the model on each training fold in our *Scripts* dataset. These two models are referred to as the out-of-domain (OOD) and in-domain (ID) models, respectively.

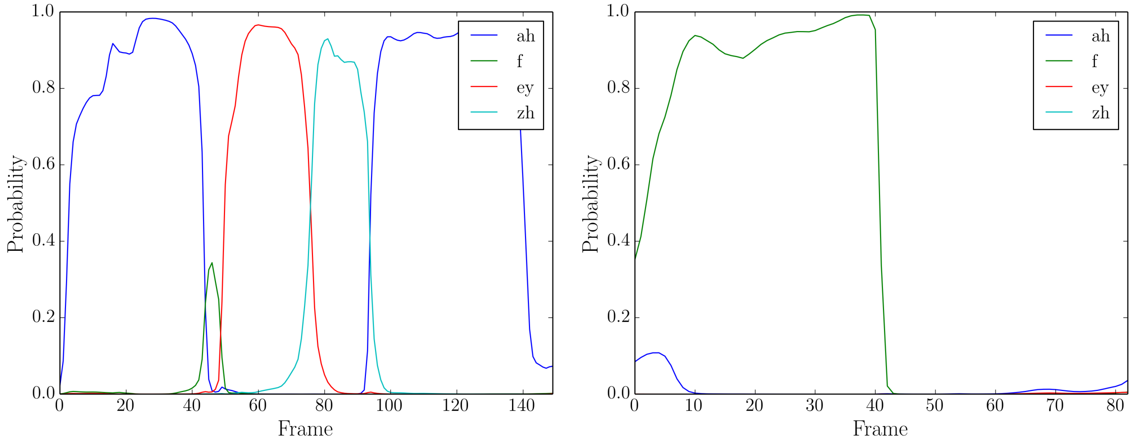
This work employs a multi-task deep Bidirectional Long-Short Term Memory Recurrent Neural Network (BLSTM-RNN) acoustic model that jointly predicts the correct senone as well as monophone labels, similar to the architecture described in Section 4.5.4. The monophone output of the network represents a distribution over phonemes, also referred to as phoneme posteriorgrams. They can be viewed as a low-dimensional representation of each speech frame. Combined with alignment information, each word can then be represented as a sequence of posteriorgrams, i.e., a multi-dimensional time series.

Input Features: 40-dimensional log Mel filterbank coefficients (MFBs) are extracted with Kaldi [124], using a 25ms window and 10ms frame shift. We perform per-speaker z-normalization and augment each feature frame with five left and right neighbors, resulting in 440 dimensions per frame.

Model Architecture: our multi-task BLSTM-RNN consists of four hidden BLSTM layers, each with 1200 units (600 for forward, 600 for backward). The senone and monophone output layers contain 4550 and 46 units, respectively.

OOD Training: We train the network using the Adam optimizer [75], full Backpropagation Through Time (BPTT), Cross Entropy (CE) loss, 0.4 dropout, and an initial learning rate of 0.001. Early stopping is applied based on the development frame error rate (FER) and an exponential-decay learning schedule [82].

ID Adaptation: We adapt the OOD network to the smaller training set using the same strategies as in OOD training, with two modifications. First, we modify the loss function to also minimize the Kullback-Leibler divergence (KLD) between the ID and OOD model outputs, which has been shown to be an effective regularization technique [169]. Second, we employ the step-decay schedule [82] with a 0.00005 minimum learning rate. The KLD



(a) Correctly Produced

(b) Neologistic Paraphasia

Figure 5.1: Example posteriorgrams of a correctly produced word (a) and a neologistic paraphasia (b). The target word in both cases is *aphasia* (ah f ey zh ah).

weight {0.25, 0.5} and dropout rate {0.4, 0.6} are chosen based on the development FER.

5.5.2 Feature Extraction

The ID acoustic model obtained from the previous step can be used to detect word and phone boundaries via forced alignment with the target transcripts. Given this information, our objective is to extract features for each word that can help separate phonemic/neologistic paraphasias from correctly produced words.

The phoneme posteriorgrams produced by the multi-task BLSTM-RNN model provide a compact representation of word and phone segments. Figure 5.1 shows example posteriorgrams of two words with the same target (*aphasia*), one correctly produced and one with neologistic paraphasia. The plots are limited to phones that make up the pronunciation of *aphasia* (ah, f, ey, zh). As can be seen, there are visible differences between the two posteriorgrams. Our proposed feature set will thus focus on quantifying this difference. The features can be divided into the following groups.

Goodness of Pronunciation (GOP): GOP is a widely used metric for assessing pronunciation, first proposed by Witt and Young [163]. It has also been used successfully in our previous work to estimate aphasic speech quality [78, 81]. GOP involves calculating the difference between the average acoustic log-likelihood of a force-aligned word-level segment and that of an unconstrained phone loop. The closer this number is to 0, the more likely that the pronunciation of this word is correct. We extract the GOP as well as the raw forced alignment score for each word. All calculations are performed on our DBLSTM-RNN's phoneme posteriorgram output.

Phone Edit Distance (DIST): both phonemic and neologistic paraphasias involve deviations between the spoken and correct phone sequences. The spoken phone sequence can be estimated from an unconstrained phone loop over the phoneme posteriorgram associated with the word segment, and the correct phone sequence can be obtained from forced alignment results on the target transcript. For each pair of spoken and correct phone sequences, we extract the raw edit distance, edit distance normalized by alignment length, as well as the number of insertions, deletions, and substitutions normalized by alignment length.

Dynamic Time Warping (DTW): the underlying assumption behind these features is that the phoneme posteriorgrams of phonemic and neologistic paraphasias are different from those of correct words. Given a *candidate* word, we can find *references* of this word in the ID training set that are marked as correctly produced, along with their phoneme posteriorgrams. Following Lee et al. [84–86], we compare posteriorgram pairs using DTW, where the distance between two frames c_i and r_j is defined as their inner product distance:

$$D(c_i, r_j) = -\log(c_i \cdot r_j) \tag{5.1}$$

We extract the following features for each candidate-reference posteriorgram pair: raw DTW distance, DTW distance normalized by aligned path length, and length of the longest horizontal/vertical aligned segment normalized by aligned path length. We extract the mean, median, lower and upper quartile, and standard deviation of each feature group to produce word-level features. We extract a similar set of features for all candidate-reference phone pairs within the word, given that they might provide complementary information. If a candidate word has fewer than three references, we use the average features of all correct words in the training set. This accounts for 6–7% of all candidate words.

Duration Measures (DUR): these features are also inspired by Lee et al. [84–86] and extracted similarly to DTW. However, we compare the differences in durations instead of posteriorgrams. For each candidate-reference word/phone pair, we extract the ratio between their durations, the difference in duration normalized by candidate duration, and the difference in duration normalized by reference duration.

As a final post-processing step, we z-normalize all features using statistics computed from correctly produced words in the training set.

5.5.3 Automatic Transcription

Automatic transcription of test utterances can be performed by combining our BLSTM-RNN acoustic model with a language model (LM) for decoding. We experiment with two LM types in this work. Firstly, we use a trigram model estimated on the ID training and development set. We refer to this model as the *global* LM. Secondly, we take advantage of the fact that utterances in the dataset are limited to four predefined scripts with different vocabulary and sentence structures. Therefore, it may be beneficial to use a trigram model estimated on the portion of the training and development set corresponding to the same script as the current test utterance. We refer to this as the *task-specific* LM. In both cases, the LM weight and word insertion penalty are chosen based on the development WER.

Table 5.3 lists the test WERs for different acoustic and language model combinations. As expected, the best performance is obtained with an in-domain acoustic model and task-

	Global LM	**Task LM**
OOD AM	65.82	60.97
ID AM	47.68	**45.11**

Table 5.3: Word Error Rate (WER) with different language and acoustic model types.

	C–pn	**C–p**	**C–n**
Baseline	.442	.461	.484
GOP	.615 (SVM)	.560 (LR)	.590 (SVM)
DIST	.619 (DT)	.556 (DT)	.662 (DT)
DTW	.699 (SVM)	.611 (LR)	.746 (LR)
DUR	.628 (LR)	.556 (DT)	.652 (LR)
All Feats.	**.704 (LR)**	**.632 (LR)**	**.761 (LR)**

SVM: Support Vector Machine | DT: Decision Tree | LR: Logistic Regression

Table 5.4: Paraphasia detection results with known target transcripts, measured in average F1. The best performing classifiers are indicated in parentheses.

specific language model. We will use this system for all experiments involving ASR.

5.6 Results and Discussion

5.6.1 Paraphasia Detection With Known Transcripts

Paraphasia classification results from known transcripts using different feature sets and labeling schemes, measured in average F1 scores, are summarized in Table 5.4. We show results from the classifier that yields the best overall test performance.

All of our systems are able to outperform the naïve baseline, demonstrating that it is feasible to automatically separate correctly produced words and phonemic/neologistic paraphasias. In particular, neologistic paraphasias (*C–n*) are easier to detect than phonemic paraphasias (*C–p*). This is consistent with the clinical definitions of these two paraphasia types. Because neologistic paraphasias, by definition, involve more deviations from the sounds in the target word, they are better characterized by our proposed features.

In all three labeling schemes (*C–pn*, *C–p*, and *C–n*), the best performance is obtained by using all features, with DTW generating the best individual results. This demonstrates the utility of the phoneme posteriorgram representation produced by our multi-task BLSTM-RNN acoustic model. A potential method to further exploit phoneme posteriorgrams is to use them as features in whole-word acoustic modeling, which may lead to better discrimination than template matching techniques. GOP features traditionally perform favorably compared to DTW for mispronunciation detection [86], but not so in our work. A possible way to improve GOP performance in this task is to extract phone-level GOP scores alongside word-level features. Likewise, duration-based features (DUR) may benefit from established measures in rhythm analysis, such as Pairwise Variability Error [154]. We will explore these ideas in future work.

Finally, we observe that different feature sets benefit from different classification algorithms. Logistic regression and SVM work well with primarily continuous features such as GOP, DTW, and DUR. By contrast, decision tree yields better performance on DIST, whose features are largely discrete.

5.6.2 Paraphasia Detection Without Known Transcripts

We are interested in how the best (bolded) models in Table 5.4 perform when target transcripts for test utterances are generated automatically with ASR. Table 5.5 lists the results at the word, utterance, and speaker level, as described in Section 5.4.2.

For word-level, the goal of the system is to simultaneously recognize and label each word. However, our system is unable to outperform the naïve baseline in terms of AWER. As previously discussed, this is challenging because aphasic speech poses significant problems for ASR, and it is difficult to obtain reliable word-level predictions without accurate

	C–pn	**C–p**	**C–n**
Word **[AWER]**	53.46 (53.39)	54.18 (51.48)	47.84 (47.18)
Utterance **[Avg. F1]**	.594 (.412)	.611 (.373)	.604 (.404)
Speaker **[r]**	.479 (N/A)	.749* (N/A)	.057 (N/A)

*statistically significant ($p \approx 0.005$, 2-tailed test)

Table 5.5: Paraphasia detection results without known target transcripts. Naïve baseline performance is in parentheses.

target transcripts. This suggests that aphasic speech ASR performance must be improved before paraphasias can be detected reliably at the word level without known transcripts.

Meanwhile, utterance-level results, which involve detecting the presence of paraphasias in an utterance, appear more promising. Our system outperforms the naïve baseline in all three classification schemes, suggesting that although word-level predictions may be unreliable, clinically-relevant information can still be extracted at a coarser level of analysis.

We also observe positive results for estimating the paraphasia production frequency of a particular speaker, which can be tied to anomia severity. Specifically, we obtain a statistically significant Pearson correlation coefficient of **0.749** ($p \approx 0.005$, 2-tailed test) for estimating the rate of phonemic paraphasia production. However, there is virtually no correlation for neologistic paraphasias. We hypothesize that while neologistic paraphasias are easy to classify from known transcripts, they are difficult to detect in a free-form setting because our ASR system fails to recognize them. This again calls for further improvement in aphasic speech recognition.

5.7 Conclusion

In this chapter, we presented the first study on detecting phonemic and neologistic paraphasias automatically from aphasic speech, utilizing techniques from ASR and mispronunciation detection. We demonstrated the feasibility of detecting paraphasias from known target transcripts. We showed the utility of utterance- and speaker-level analysis when target transcripts are generated automatically with ASR.

For future work, we will investigate additional feature extraction methods to better characterize paraphasias, such as those based on whole-word acoustic models and phonological features [106]. We will experiment with ways to further improve ASR technology to better accommodate aphasic speech with high amounts of paraphasias, such as utilizing personalized and data-driven pronunciation models [28, 92–94, 102, 158]. Finally, we will explore computational approaches for tackling semantic paraphasia detection.

5.8 Work Published

The work presented in this chapter was published in the following article:

1. **Duc Le**, Keli Licata, and Emily Mower Provost. "Automatic Paraphasia Detection from Aphasic Speech: A Preliminary Study." *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden. August, 2017.

CHAPTER 6

Automatic Quantitative Analysis of Spontaneous Aphasic Speech

6.1 Introduction

Spontaneous speech (e.g., answering an open-ended interview question, retelling a story, describing a picture) plays a prominent role in everyday interaction of a person with aphasia (PWA) and is widely regarded in the clinical literature as one of the most important modalities to analyze [40, 68, 100, 125]. Example applications of spontaneous speech analysis include aphasia classification [51], treatment planning [125], recovery tracking [53], and diagnosis of residual aphasia post onset [68].

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 [125]. 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.

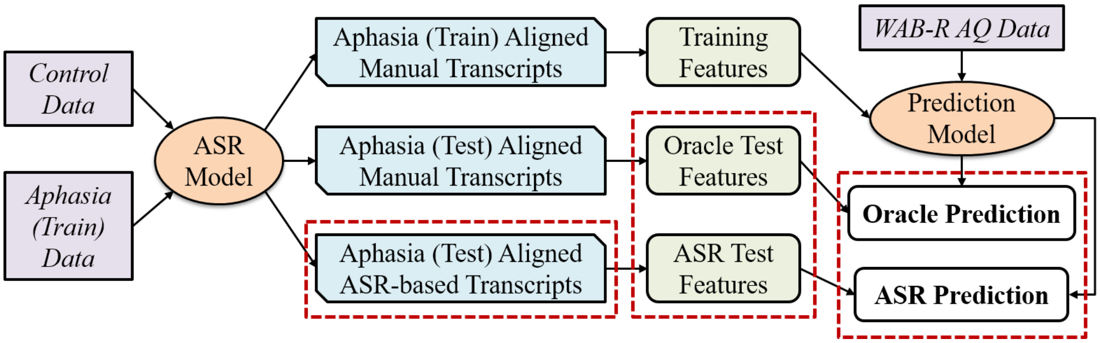


Figure 6.1: High-level overview of our proposed system. The red boxes denote components that will be the focus of our analysis.

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 [42, 43, 87, 88] and difficult to generate automatically [41, 122]. Second, they typically target speech with known prompts [1, 2, 78, 81, 83]. This removes the need for unconstrained automatic speech recognition (ASR) and simplifies transcript generation, which can be achieved by modified forced alignment [78,81,83] or keyword spotting [1,2]. However, the reliance on known prompts makes this type of system 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 robustness of ASR-based features against transcription errors has 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 this work in two sequential components. First, we discuss various clinically relevant quantitative measures that can be extracted from transcripts generated by an ASR system. 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. Second, 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 [72]. 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 Figure 6.1.

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 clinical applications.

6.2 Related Work

6.2.1 Linguistic Analysis of Spontaneous Aphasic Speech

Linguistic analysis of aphasic speech can be divided into two types, qualitative and quantitative [125]. The former assesses PWAs' speech based on a qualitative rating scale, such as the Boston Diagnostic Aphasia Examination [51] or Aachen Aphasia Test [109], 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 [70]. However, qualitative rating scales often have difficulties in measuring a PWA's improvement [125] and may lack sensitivity [53]. 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 [125].

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 [53]. This parameter set, which captures lexical and semantic content, syntactic completeness, linguistic complexity, and mean utterance length, were 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 [37]. Finally, Jaecks et al. were able to diagnose residual aphasia using a set of variables spanning information density, syntactic variability, linguistic errors, and cohesion [68]. These proposed measures form the basis of our feature set (Section 6.5).

6.2.2 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 [87, 88]. 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 [87]. Further, aphasic speech exhibits different intonation patterns [88]. 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 [42, 43]. 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 [41]. However, the ASR performance was relatively poor, attaining word error rate (WER) between 67.7% and 73.1%. As a result, their analysis was limited to simulated ASR output with preset WER levels, and the robustness of their feature set remained unclear.

Our previous work proposed a system to automatically estimate qualitative aspects of read aphasic speech through transcript, pronunciation, rhythm, and intonation features [78,81,83]. 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 [122]. 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 here is conducted on a large-scale aphasic speech corpus with detailed discussion regarding the robustness of ASR-derived quantitative measures.

6.3 Data

6.3.1 Speech Data

All experiments in this work are carried out on AphasiaBank [39, 96]. We select English sub-datasets that have at least four speakers and are collected with the core AphasiaBank

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

Table 6.1: Summary of AphasiaBank data used in this work. The speakers are split into two groups, those who have aphasia (*Aphasia*) and healthy controls (*Control*).

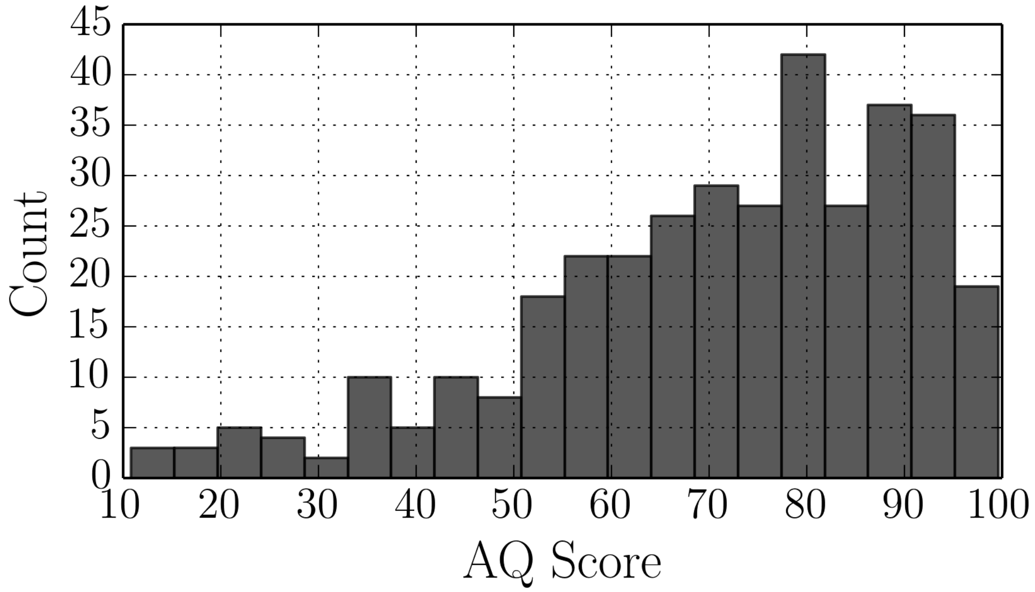


Figure 6.2: Histogram of WAB-R AQ scores.

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 hours of speech. Utterances in AphasiaBank can be categorized based on their applied elicitation method, which is either free speech (e.g., open interview, conversational speech) or semi-spontaneous (e.g., storytelling, picture description) [125]. The speech-language patterns of the same PWA may be different across these two categories [68, 125], thus it may be beneficial to analyze them separately. Table 6.1 describes the dataset in more detail.

6.3.2 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 [72]. 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). Figure 6.2 plots the histogram of the available AQ scores.

6.3.3 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 described in Chapter 4. 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 hours.

6.4 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 work, we employ the ASR system described in Chapter 4, which consists of a multi-task deep Bidirectional Long-Short Term Memory Recurrent Neural Network

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

Table 6.2: 13 applied statistics.

(BLSTM-RNN) acoustic model and trigram word-level language model. Our decoder outputs the hypothesized transcripts as well as the word- and phone-level alignments.

6.5 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 6.3) 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 (Figure 3.1).

Table 6.3: Extracted quantitative measures for each speaker. {} denotes a collection of numbers summarized into speaker-level measures using the statistics listed in Table 6.2.

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 words / Open + closed class words
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

6.5.1 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 [78,81,83] as well as distinguishing between subjects with PPA and healthy controls [42,43].

Features 3–4 are adopted from a set of basic parameters proposed by Grande et al. to objectively measure changes in spontaneous aphasic speech [53]. 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 [14] 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 few function words [53].

Features 5–6 are based on mean length of utterances in words, a widely used measure in spontaneous aphasic speech analysis [53, 68, 125]. 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 6.2. 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 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 [68] and PPA [42,43] diagnosis. Following [20,43], 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 [53].

6.5.2 Dysfluency (DYS)

Dysfluency is an important aspect of aphasic speech which has been used in qualitative analysis [78, 81, 83] and PPA diagnosis [42]. Features 19–28 capture the amount of dysfluency (i.e., fillers and pauses) in each PWA's speech. Following [119], we define pauses as regions of silence between words that are longer than 150ms; these are further categorized as short ($\leq$ 400ms) or long ($>$ 400ms). We extract the occurrence frequency of fillers and pauses normalized by speech duration, total words, and total phones (features 19–27). Finally, we extract the statistics over all pause durations (feature 28). We expect milder PWAs to exhibit less dysfluency and vice versa.

6.5.3 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 [37]. 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 [37], 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 [22]. In addition, the combined work of Stadthagen-Gonzalez and Davis [151] and Gilhooly and Logie [49] 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 [42, 43]. However, they only extracted the mean scores, whereas we consider a more comprehensive set of statistics (Table 6.2). Finally, feature 34 approximates the pronunciation complexity of a PWA's vocabulary based on the number of phones present in a word.

6.5.4 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 $\mathcal{M}$, the Cross Entropy (CE) of a POS sequence $p_1p_2 \ldots p_N$ denotes how closely it adheres to the model:

$$\mathcal{H}(p_1p_2 \ldots p_N|\mathcal{M}) = \frac{\log P(p_1p_2 \ldots p_N|\mathcal{M})}{N} \tag{6.1}$$

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 [130].

6.5.5 Pairwise Variability Error (PVE)

Speech rhythm was shown in our previous work to be helpful for estimating qualitative aspects of aphasic speech [81, 83]. 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. [154] 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:

$$PVE_M = \frac{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |(c_i - c_{i-m}) - (r_i - r_{i-m})|}{\sum_{i=2}^{N} \sum_{m=1}^{min(M,i-1)} |c_i - c_{i-m}| + |r_i - r_{i-m}|} \tag{6.2}$$

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 [83]. 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 [81, 83, 154].

6.5.6 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 [79]. The DTW-based features, inspired by Lee et al. [84–86], were shown to outperform Goodness of Pronunciation (GOP) [163] 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:

$$D(c_i, r_j) = -\log(c_i \cdot r_j) \tag{6.3}$$

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.

6.5.7 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

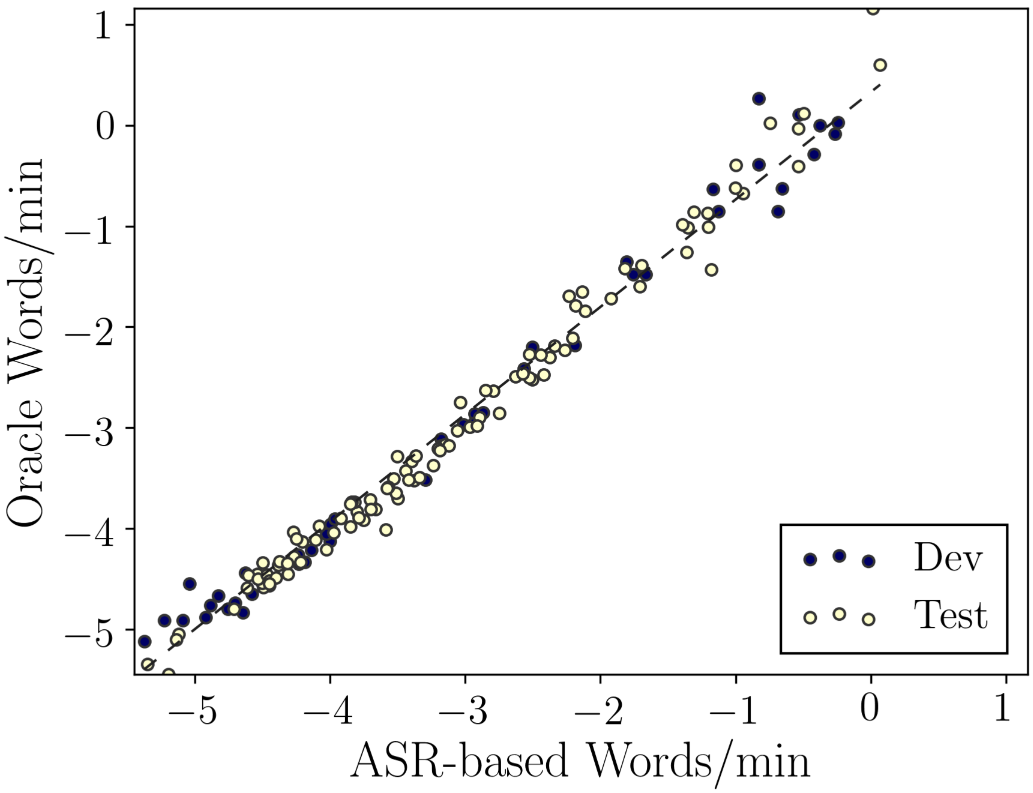


Figure 6.3: Example calibration of *words/min* feature. A linear transformation model is trained on development speakers ($y = 1.07x + .33$) and applied to test speakers. Feature values are z-normalized using statistics extracted from healthy controls.

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 (Figure 6.3). 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.

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., Figure 6.3). 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 6.7.1.

6.6 WAB-R AQ Prediction

The system's goal is to automatically predict WAB-R AQ, an assessment score closely tied to aphasia severity [72]. 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.

6.6.1 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 6.3.3, 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 ensures that test features 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 [121] 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 [41]. 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 6.7.2. Investigation of this modeling approach will be left for future work.

6.6.2 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 [68, 125]. 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.

6.7 Results and Discussion

6.7.1 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 6.2, 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.

Table 6.4: Comparison of oracle and speech recognition-based quantitative measures, using two-tailed paired t-test of equal means for regular features and two-way repeated measures Analysis of Variance (ANOVA) for statistics features, both with $p = .05$ (▲: not significantly different before calibration; ▼: not significantly different after calibration).

Group	ID	Feature	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}		▲▼	

It can be seen from Table 6.4 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***), {$\mathit{PVE}_{1,2,4}$*/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 [37,53,68]. 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.

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)

Table 6.5: Revised Western Aphasia Battery Aphasia Quotient (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 6.6).

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.

6.7.2 WAB-R AQ Prediction

The WAB-R AQ prediction results, measured in MAE and Pearson's correlation, are summarized in Table 6.5. The results are partitioned based on two factors. First, **transcript type** (*Oracle*, *Auto*, *Calibrated*) specifies the source from which features are computed (Section 6.6.1). Second, **feature extraction protocol** (*All*, *Free*, *Semi*, *Combined*) indicates the type of speech used for feature extraction (Section 6.6.2). 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,

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)

Table 6.6: Performance breakdown of individual feature groups (Section 6.5) under the *Combined* protocol, measured in Mean Absolute Error (MAE) and Pearson's correlation.

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 6.6 lists the prediction results of individual **feature groups** (*DEN*, *DYS*, *LEX*,

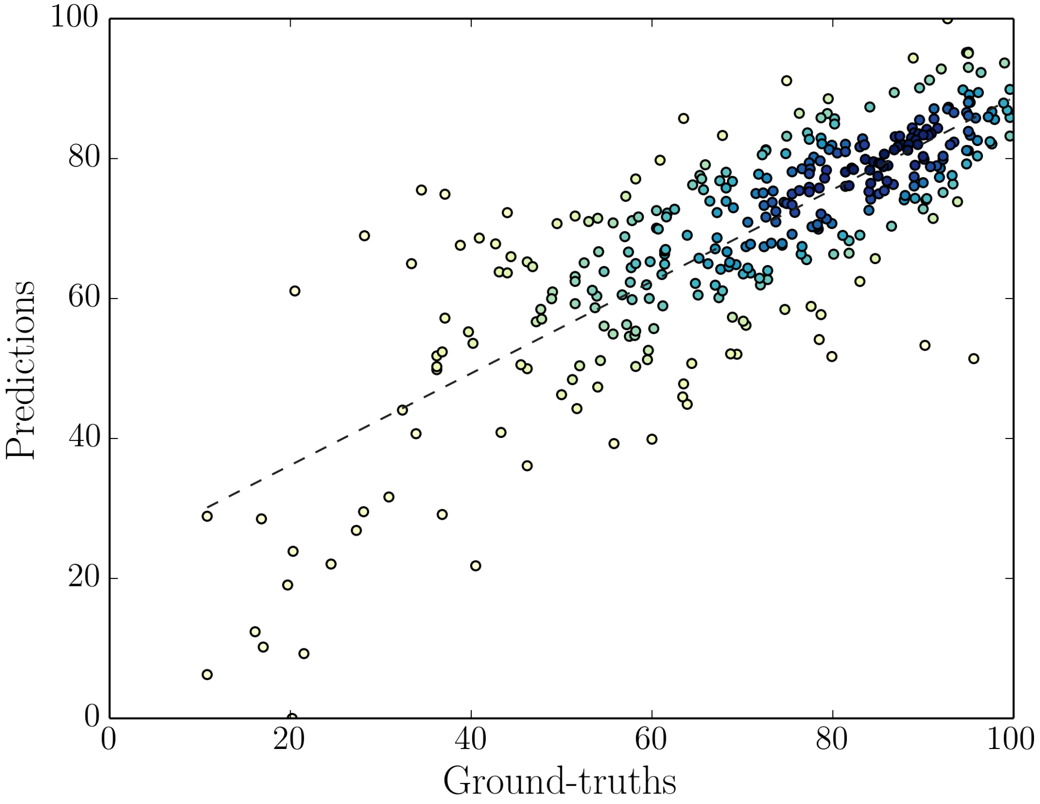


Figure 6.4: Aphasia Quotient (AQ) prediction plot. Darker color means higher density.

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.

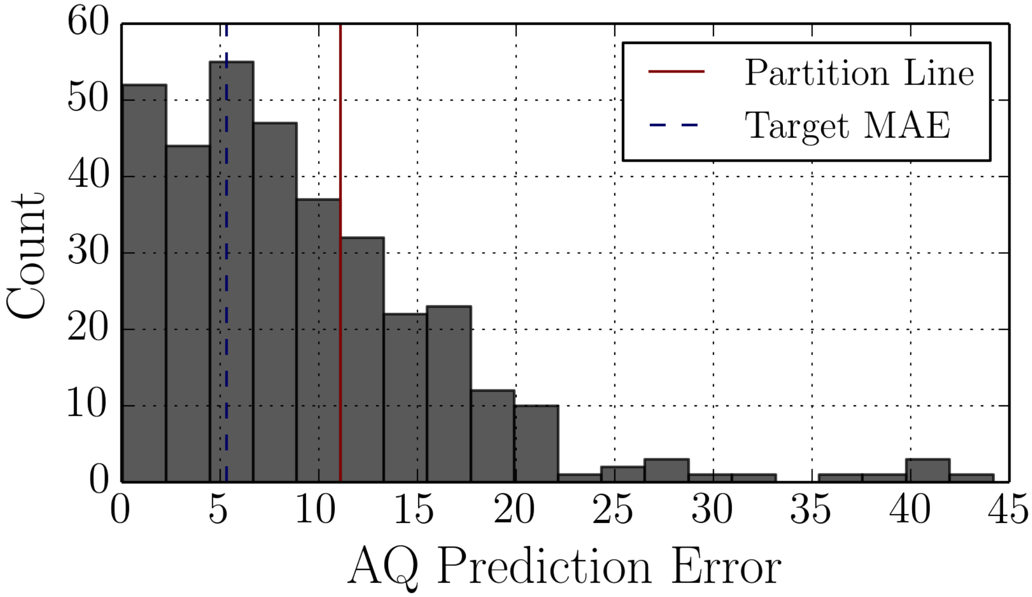


Figure 6.5: Histogram of Aphasia Quotient (AQ) prediction errors. The partition line divides subjects into two groups, *Low Errors* (left) and *High Errors* (right). The Mean Absolute Error (MAE) of the first group is approximately 5.316, the test-retest reliability of the AQ.

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. Figure 6.4 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.

Individual characteristics are not correlated with system performance. However, we

Property	**Low Errors**	**High Errors**	**Welch's t-test**	
			t	p
Word Error Rate	38.78 (14.09)	44.06 (17.09)	-2.824	.005
Aphasia Quotient	72.85 (17.03)	67.50 (23.27)	2.160	.032
Label Distance	20.16 (7.21)	24.64 (9.10)	-4.542	$< .001$

Table 6.7: Comparison of subjects with low and high Aphasia Quotient (AQ) prediction errors. Values shown are mean (standard deviation). *Label Distance* is the absolute difference between a subject's AQ and the average training AQ.

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 [148]. In other words, automatic AQ prediction can be considered satisfactory if the MAE does not exceed this value. As shown in Figure 6.5, 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 6.7). This suggests that we can roughly estimate the range of prediction errors given a PWA's WER level and/or current AQ score.

6.8 Conclusion

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 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 [26–28, 146, 147]. 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.

6.9 Work Published

The work presented in this chapter was published in the following article:

1. **Duc Le**, Keli Licata, and Emily Mower Provost. “Automatic Quantitative Analysis of Spontaneous Aphasic Speech.” *Speech Communication*. (in submission)

CHAPTER 7

Conclusions and Future Directions

In this dissertation, we have presented our work on how to computationally analyze and assess aphasic speech, with a long term goal of enabling automatic speech-based technology that can support aphasia rehabilitation. This chapter provides a high-level summary of our work and discusses potential avenues of research for future investigation.

7.1 Main Results and Contributions

The first part of the dissertation targeted the automatic intelligibility assessment of constrained speech data (i.e., speech with known target prompts), specifically the estimation of speech clarity, fluidity, and prosody (Chapter 3). We proposed a novel set of features that capture the pronunciation, rhythm, and intonation of PWAs by comparing with healthy speech patterns. We demonstrated that the system was able to reach human-level performance in estimating speech intelligibility, assuming that manually labeled transcripts are available. We subsequently lifted the dependence on manual transcripts by proposing a modified forced alignment method for transcript generation. Our fully automated system achieved competitive results compared to human evaluators.

A restriction of the above system is that it assumes the availability of known target prompts, an unrealistic assumption for unconstrained speech, the most common type of verbal communication in everyday interaction. Accurate automatic speech recognition (ASR) is required to reliably assess unconstrained speech. We addressed this need by performing one of the first large-scale studies on aphasic speech recognition (Chapter 4). We demonstrated the efficacy of utterance-level i-vectors and multi-task deep Bidirectional Long-Short Term Memory Recurrent Neural Network (BLSTM-RNN) for acoustic modeling. We showed that out-of-domain adaptation is a promising method for developing ASR systems on small datasets. Finally, our analysis indicated that there exists a moderate correlation between recognition errors and aphasia severity, and that ASR technology is more suited for non-conversational aphasic speech.

We applied the ASR techniques described in Chapter 4 to the problem of automatic paraphasia (naming error) detection (Chapter 5). We showed that for speech with restricted lexical content, task-specific language models can help improve recognition accuracy. We established the first framework for evaluating paraphasia detection at the word, utterance, and speaker level. Our results demonstrated that paraphasia detection at the word-level is feasible when target transcripts are available, whereas speaker-level analysis of phonemic paraphasia production rate can be estimated reasonably accurately based on ASR output.

Our final work targeted the automatic quantitative analysis of spontaneous aphasic speech (Chapter 6). The key questions we wanted answered in this work are, for features extracted from ASR-generated transcripts, are they (1) reflective of a PWA's true measures and (2) usable for estimating aphasia severity? We showed that the majority of the proposed features, many of which have important clinical implications, are highly robust against ASR errors after applying our calibration method. These ASR-based features yielded good performance in predicting the revised Western Aphasia Battery (WAB-R) Aphasia Quotient (AQ), a standard measure of aphasia severity, especially when free speech and semi-spontaneous speech are handled separately. However, results obtained from ASR-based features were not yet comparable to those achieved with oracle features. This calls for further advances in aphasic speech recognition and feature engineering.

7.2 Future Work

7.2.1 Specialized ASR Models for Aphasia

Despite the advances made in Chapter 4, ASR remains challenging for aphasic speech due to irregular speech-language patterns and high speaker variability. Acoustic and language model personalization is a promising method for improving ASR performance. High level of speaker variability hints at the potential of model personalization; however, it also makes identifying similar speakers for training difficult. Our preliminary experiments indicated that grouping PWAs by aphasia category (fluent vs. non-fluent), gender, or severity for acoustic/language model training has insignificant impact on ASR performance. To tackle this problem effectively, methods that explicitly take into account the prototypical error patterns associated with aphasia will likely yield better results.

Moreover, it was necessary to simplify CHAT transcripts to make them compatible with standard ASR systems. This reduces the informativeness of the transcripts, which originally contain important information for the analysis of aphasia, such as word-level errors, partial words, and different categories of sound fragments. While these special tokens can be recognized through post-hoc analysis of the simplified transcripts, such as what was done for paraphasia detection, a promising research direction is to model these tokens jointly as part of acoustic and language model training. This can potentially improve the transcript quality and widen the range of recognizable tokens.

7.2.2 Clinical Applications and Longitudinal Data Collection

The results and findings demonstrated in our work were obtained from post-hoc analysis of existing data. It is not yet clear what the advantages and disadvantages of our methods are when applied to real-world clinical applications. Since the long-term goal of this dissertation is to provide speech-based technology to support aphasia rehabilitation, it is important to perform human-centered studies to assess and quantify the practicality of our system. One possible way to perform such studies is to extend the mobile application described in Section 2.1.2.1 to incorporate the proposed automated speech-language assessment techniques and evaluate their effectiveness on realistic speech data collected as part of a therapeutic application.

The studies presented in this dissertation are cross-sectional by nature, which is necessitated by the way the data were collected. It is worthwhile to also carry out longitudinal studies to evaluate our system's ability to track clinically-relevant measures for the same subject over an extended period of time. Having access to longitudinal data will allow us to investigate adaptation methods to develop personalized prediction models, as opposed to the speaker-independent models targeted in this work. Collecting longitudinal data in traditional research settings is difficult given the time and resource requirements. Our mobile application, if deployed publicly, can make this task more feasible.

BIBLIOGRAPHY

[1] 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.

[2] A. Abad, A. Pompili, A. Costa, I. Trancoso, J. Fonseca, G. Leal, L. Farrajota, and I. P. Martins. Automatic word naming recognition for an on-line aphasia treatment system. *Computer Speech and Language*, 27(6):1235–1248, 2013.

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

[4] AphasiaBank. *Error Coding*, Accessed: 2015-11-13.

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

[6] H. Axer, J. Jantzen, G. Berks, D. Südfeld, and D. G. V. Keyserlingk. The aphasia database on the web: description of a model for problems of classification in medicine. In *European Symposium on Intelligent Technology*. Citeseer, 2000.

[7] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee. i-vector modeling of speech attributes for automatic foreign accent recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 24(1):29–41, 2016.

[8] P. Bell, M. Gales, P. Lanchantin, X. Liu, Y. Long, S. Renals, P. Swietojanski, and P. Woodland. Transcription of multi-genre media archives using out-of-domain data. In *IEEE Workshop on Spoken Language Technology (SLT)*, Miami, FL, USA, 2012.

[9] P. Bell and S. Renals. Complementary tasks for context-dependent deep neural network acoustic models. In *Interspeech*, Dresden, Germany, 2015.

[10] P. Bell and S. Renals. Regularization of context-dependent deep neural networks with context-independent multi-task training. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[11] V. Berisha, J. Liss, S. Sandoval, R. Utianski, and A. Spanias. Modeling pathological speech perception from data with similarity labels. In *IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 915–919, Florence, Italy, 2014.

[12] S. Bhogal, R. Teasell, M. Speechley, and M. L. Albert. Intensity of Aphasia Therapy, Impact on Recovery. *Stroke*, 34(4):987–993, Apr. 2003.

[13] S. K. Bhogal, R. W. Teasell, N. C. Foley, and M. R. Speechley. Rehabilitation of aphasia: more is better. *Topics in Stroke Rehabilitation*, 10(2):66–76, 2003.

[14] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[15] M. Black, J. Tepperman, S. Lee, and S. S. Narayanan. Predicting children's reading ability using evaluator-informed features. In *10th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1895–1898, Brighton, United Kingdom, 2009.

[16] M. Black, J. Tepperman, S. Lee, P. Price, and S. S. Narayanan. Automatic detection and classification of disfluent reading miscues in young children's speech for the purpose of assessment. In *Interspeech*, Antwerp, Belgium, Aug. 2007.

[17] M. P. Black and S. S. Narayanan. Improvements in predicting children's overall reading ability by modeling variability in evaluators' subjective judgments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5072, Kyoto, Japan, 2012.

[18] 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*, 19(4):1015–1028, May 2011.

[19] S. E. Blumstein, W. E. Cooper, H. Goodglass, S. Statlender, and J. Gottlieb. Production deficits in aphasia: a voice-onset time analysis. *Brain and Language*, 9(2):153–170, 1980.

[20] S. D. Breedin, E. M. Saffran, and M. F. Schwartz. Semantic Factors in Verb Retrieval: An Effect of Complexity. *Brain and Language*, 63(1):1 – 31, 1998.

[21] M. Breen, L. C. Dilley, J. Kraemer, and E. Gibson. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory*, (8):277–312, 2012.

[22] M. Brysbaert and B. New. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41(4):977–990, Nov 2009.

[23] L. R. Cherney, A. S. Halper, A. L. Holland, and R. Cole. Computerized Script Training for Aphasia: Preliminary Results. *American Journal of Speech-Language Pathology*, 17(1):19–34, Feb 2008.

[24] L. R. Cherney, A. S. Halper, and R. C. Kaye. Computer-based script training for aphasia: Emerging themes from post-treatment interviews. *Journal of Communication Disorders*, 44(4):493–501, 2011.

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

[26] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain. Automatic Selection of Speakers for Improved Acoustic Modelling: Recognition of Disordered Speech with Sparse Data. In *IEEE Workshop on Spoken Language Technology (SLT)*, South Lake Tahoe, NV, USA, 2014.

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

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

[29] N. Côté. *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer, 2011.

[30] G. Dahl, D. Yu, L. Deng, and A. Acero. Large Vocabulary Continuous Speech Recognition With Context-Dependent DBN-HMMS. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011.

[31] G. Dahl, D. Yu, L. Deng, and A. Acero. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech & Language Processing (TASLP)*, 20(1), 2012.

[32] M. Danly and B. Shapiro. Speech prosody in Broca's aphasia. *Brain and Language*, 16(2):171 – 190, 1982.

[33] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech & Language Processing*, 19(4), 2011.

[34] J. R. Duffy. *Motor speech disorders: Substrates, differential diagnosis, and management*. Mosby, 2 edition, 2005.

[35] L. A. Edmonds, S. E. Nadeau, and S. Kiran. Effect of Verb Network Strengthening Treatment (VNeST) on Lexical Retrieval of Content Words in Sentences in Persons with Aphasia. *Aphasiology*, 23(3):402–424, Mar 2009.

[36] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andr, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, pages 14–27, 2016.

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

[38] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication*, 11(3):165–175, 1995.

[39] M. M. Forbes, D. Fromm, and B. MacWhinney. Aphasiabank: A resource for clinicians. In *Seminars in Speech and Language*, volume 33, page 217. NIH Public Access, 2012.

[40] S. Fox, E. Armstrong, and L. Boles. Conversational treatment in mild aphasia: A case study. *Aphasiology*, 23(7-8):951–964, 2009.

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

[42] 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.

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

[44] N. Friedmann, M. Biran, and D. Dotan. *Lexical retrieval and its breakdown in aphasia and developmental language impairment*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2013.

[45] J. Gandour, S. H. Petty, and R. Dardarananda. Dysprosody in Broca's aphasia: A case study. *Brain and Language*, 37(2):232 – 257, 1989.

[46] S. Garimella, A. Mandal, N. Strom, B. Hoffmeister, S. Matsoukas, and S. H. K. Parthasarathi. Robust i-vector based adaptation of DNN acoustic model for speech recognition. In *Proc. of the 16th Annual Conference of the ISCA (INTERSPEECH)*, pages 2877–2881, Dresden, Germany, 2015.

[47] J. Gehring, Y. Miao, F. Metze, and A. Waibel. Extracting deep bottleneck features using stacked auto-encoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3377–3381, Vancouver, BC, Canada, 2013.

[48] O. Ghahabi, A. Bonafonte, J. Hernando, and A. Moreno. Deep neural networks for i-vector language identification of short utterances in cars. *Interspeech*, pages 367–371, 2016.

[49] K. J. Gilhooly and R. H. Logie. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12(4):395–427, Jul 1980.

[50] O. Glembek, L. Burget, P. Matejka, M. Karafiát, and P. Kenny. Simplification and optimization of i-vector extraction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4516–4519, Prague, Czech Republic, 2011.

[51] H. Goodglass, E. Kaplan, and B. Barresi. *Boston Diagnostic Aphasia Examination*. Philadelphia: Lippincott, Williams & Wilkins, 3 edition, 2000.

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

[53] M. Grande, K. Hussmann, E. Bay, S. Christoph, M. Piefke, K. Willmes, and W. Huber. Basic parameters of spontaneous speech as a sensitive method for measuring change during the course of aphasia. *International Journal of Language & Communication Disorders*, 43(4):408–426, 2008.

[54] A. Graves, A. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, Vancouver, BC, Canada, 2013.

[55] K. L. Haley. Temporal and spectral properties of voiceless fricatives in aphasia and apraxia of speech. *Aphasiology*, 16(4–6):595–607, 2002.

[56] K. L. Haley, A. Jacks, M. de Riesthal, R. Abou-Khalil, and H. L. Roth. Toward a quantitative basis for assessment and diagnosis of apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 55(5):S1502–S1517, 2012.

[57] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Exploration Newsletter*, 11(1):10–18, Nov. 2009.

[58] F. Hamidi, M. Baljko, N. Livingston, and L. Spalteholz. CanSpeak: A Customizable Speech Interface for People with Dysarthric Speech. In *Proc. of the 12th International Conference on Computers Helping People with Special Needs (ICCHP)*, pages 605–612, Vienna, Austria, 2010. Springer.

[59] M. Hawley, P. Enderby, P. Green, S. Brownsell, A. Hatzis, M. Parker, J. Carmichael, S. Cunningham, P. O'Neill, and R. Palmer. Stardust; speech training and recognition for dysarthric users of assistive technology. In *Advancement of Assistive Technology (AAATE)*, pages 959–963, Dublin, Ireland, 2003.

[60] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics*, 29(5):586–593, 2007.

[61] M. S. Hawley, P. Green, P. Enderby, S. Cunningham, and R. K. Moore. Speech technology for e-inclusion of people with physical disabilities and disordered speech. In *Interspeech*, Lisbon, Portugal, September 2005.

[62] N. Helm-Estabrooks, M. L. Albert, and M. Nicholas. *Manual of Aphasia and Aphasia Therapy*. Pro-Ed, 3 edition, 2013.

[63] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMM systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1630–1635, Istanbul, Turkey, 2000.

[64] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, G. Dahl, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[65] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(9):1735–1780, Nov. 1997.

[66] W. Hu, Y. Qian, F. K. Soong, and Y. Wang. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, 67:154 – 166, 2015.

[67] H. Huang, H. Xu, X. Wang, and W. Silamu. Maximum f1-score discriminative training criterion for automatic mispronunciation detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 23(4):787–797, 2015.

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

[69] M. Kalinyak-Fliszar, N. Martin, E. Keshner, A. Rudnicky, J. Shi, and G. Teodoro. Using virtual technology to promote functional communication in aphasia: Preliminary evidence from interactive dialogues with human and virtual clinicians. *American Journal of Speech-Language Pathology*, pages S974–S989, 2015.

[70] R. C. Katz, B. Hallowell, C. Code, E. Armstrong, P. Roberts, C. Pound, and L. Katz. A multinational comparison of aphasia management practices. *International Journal of Language & Communication Disorders*, 35(2):303–314, 2000.

[71] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3):345–354, 2005.

[72] A. Kertesz. *The Western Aphasia Battery - Revised.* Texas: Harcourt Assessments, 2006.

[73] A. Kertesz and E. Poole. The aphasia quotient: the taxonomic approach to measurement of aphasic disability. *Canadian Journal of Neurological Sciences*, 1(1):7–16, Feb 1974.

[74] J. Kim, N. Kumar, A. Tsiartas, M. Li, and S. S. Narayanan. Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech and Language*, 2014.

[75] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[76] P. Kitzing, A. Maier, and V. L. Åhlander. Automatic speech recognition (asr) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logopedics Phonatrics Vocology*, 34(2):91–96, 2009.

[77] S. Landa, L. Pennington, N. Miller, S. Robson, V. Thompson, and N. Steen. Association between Objective Measurement of the Speech Intelligibility of Young People with Dysarthria and Listener Ratings of Ease of Understanding. *International Journal of Speech-Language Pathology*, 16(4):408–416, 2014.

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

[79] D. Le, K. Licata, and E. Mower Provost. Automatic Paraphasia Detection from Aphasic Speech: A Preliminary Study. In *Interspeech*, Stockholm, Sweden, 2017.

[80] D. Le, K. Licata, and E. Mower Provost. Automatic Quantitative Analysis of Spontaneous Aphasic Speech. *Speech Communication*, in submission.

[81] D. Le, K. Licata, C. Persad, and E. Mower Provost. Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE Transactions on Audio, Speech, and Language (TASLP)*, 24:2187–2199, 2016.

[82] D. Le and E. Mower Provost. Improving Automatic Recognition of Aphasic Speech with AphasiaBank. In *Interspeech*, San Francisco, USA, 2016.

[83] D. Le and E. M. Provost. Modeling Pronunciation, Rhythm, and Intonation for Automatic Assessment of Speech Quality in Aphasia Rehabilitation. In *Proc. of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014.

[84] A. Lee and J. Glass. A comparison-based approach to mispronunciation detection. In *IEEE Spoken Language Technology Workshop (SLT)*, pages 382–387, Dec 2012.

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

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

[87] T. Lee, A. Kong, V. Chan, and H. Wang. Analysis of Auto-aligned and Auto-segmented Oral Discourse by Speakers with Aphasia: A Preliminary Study on the Acoustic Parameter of Duration. In *Procedia - Social and Behavioral Sciences*, volume 94, pages 71–72, 2013.

[88] T. Lee, A. Kong, and W. Lam. Measuring prosodic deficits in oral discourse by speakers with fluent aphasia. *Frontiers in Psychology*, (47), 2015.

[89] T. Lee, Y. Liu, P. Huang, J. Chien, W. Lam, Y. Yeung, T. Law, K. Lee, A. Kong, and S. Law. Automatic Speech Recognition for Acoustical Analysis and Assessment of Cantonese Pathological Voice and Speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[90] B. Li and K. C. Sim. Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. In *Proc. of the 11th Annual Conference of the ISCA (INTERSPEECH)*, Chiba, Japan, 2010.

[91] K. Li, X. Qian, and H. Meng. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 25(1):193–207, 2017.

[92] K. Livescu and J. Glass. Feature-based pronunciation modeling for speech recognition. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 81–84, Boston, Massachusetts, 2004.

[93] K. Livescu, P. Jyothi, and E. Fosler-Lussier. Articulatory feature-based pronunciation modeling. *Computer Speech & Language*, 36(C):212–232, 2016.

[94] L. Lu, A. Ghoshal, and S. Renals. Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 374–379, Olomouc, Czech Republic, 2013.

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

[96] B. Macwhinney, D. Fromm, M. Forbes, and A. Holland. AphasiaBank: Methods for Studying Discourse. *Aphasiology*, 25(11):1286–1307, 2011.

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

[98] K. Makin, B. McDonald, L. Nickels, C. Taylor, and M. Moses. The facilitation of word production in aphasia: What can it do for the clinician. *Acquiring Knowledge in Speech, Language and Hearing*, 6(90-92), 2004.

[99] L. M. Manheim, A. S. Halper, and L. Cherney. Patient-Reported Changes in Communication After Computer-Based Script Training for Aphasia. *Archives of Physical Medicine and Rehabilitation*, 90(4):623–627, Apr 2009.

[100] J. Mayer and L. Murray. Functional measures of naming in aphasia: Word retrieval in confrontation naming versus connected speech. *Aphasiology*, 17(5):481–497, 2003.

[101] T. McAllister Byun, P. F. Halpin, and D. Szeredi. Online crowdsourcing for efficient rating of speech: a validation study. *Journal of Communication Disorders*, 53:70–83, 2015.

[102] I. McGraw, I. Badr, and J. R. Glass. Learning lexicons from speech using a pronunciation mixture model. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, 21(2):357–366, 2013.

[103] M. Meinzer, D. Djundja, G. Barthel, T. Elbert, and B. Rockstroh. Long-term stability of improved language functions in chronic aphasia after constraint-induced aphasia therapy. *Stroke*, 36(7):1462–1466, 2005.

[104] M. Meinzer, S. Streiftau, and B. Rockstroh. Intensive language training in the rehabilitation of chronic aphasia: Efficient training by laypersons. *Journal of the International Neuropsychological Society*, 13(5):846–853, 2007.

[105] K. T. Mengistu and F. Rudzicz. Comparing Humans and Automatic Speech Recognition Systems in Recognizing Dysarthric Speech. In *Proceedings of the 24th Canadian Conference on Artificial Intelligence*, pages 291–300, St. John's, Canada, 2011.

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

[107] N. Miller. Measuring up to speech intelligibility. *International Journal of Language and Communication Disorders*, 48(6):601–612, 2013.

[108] N. Miller, L. Allcock, D. Jones, E. Noble, A. J. Hildreth, and D. J. Burn. Prevalence and pattern of perceived intelligibility changes in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 78(11):1188–1190, 2007.

[109] N. Miller, K. Willmes, and R. De Bleser. The psychometric properties of the English language version of the Aachen Aphasia Test (EAAT). *Aphasiology*, 14(7):683–722, 2000.

[110] A. Mohamed, G. E. Dahl, and G. E. Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech & Language Processing*, 20(1):14–22, 2012.

[111] S. O. C. Morales and S. J. Cox. Modelling Errors in Automatic Speech Recognition for Dysarthric Speakers. *EURASIP Journal on Advances in Signal Processing*, pages 2:1–2:14, Jan. 2009.

[112] M. B. Mustafa, F. Rosdi, S. S. Salim, and M. U. Mughal. Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Systems with Applications*, 42(8):3924 – 3932, 2015.

[113] A. K. Namasivayam, M. Pukonen, D. Goshulak, J. Hard, F. Rudzicz, T. Rietveld, B. Maassen, R. Kroll, and P. van Lieshout. Treatment intensity and childhood apraxia of speech. *International Journal of Language and Communication Disorders*, 50(4):529–546, Jul 2015.

[114] 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)*, pages 841–848. MIT Press, 2001.

[115] L. E. Nicholas and R. H. Brookshire. Presence, Completeness, and Accuracy of Main Concepts in the Connected Speech of Non-Brain-Damaged Adults and Adults With Aphasia. *Journal of Speech, Language, and Hearing Research*, 38(1):145–156, 1995.

[116] L. Nickels. Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, 16(10-11):935–979, 2002.

[117] M. Nicolao, A. V. Beeston, and T. Hain. Automatic assessment of english learner pronunciation using discriminative classifiers. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5351–5355, South Brisbane, Queensland, Australia, 2015.

[118] K. Odell, M. R. McNeil, J. C. Rosenbek, and L. Hunter. Perceptual characteristics of consonant production by apraxic speakers. *Journal of Speech and Hearing Disorders*, 55(2):345–359, May 1990.

[119] 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*, 23(3):165–177, Sep 2010.

[120] P. Pedersen, K. Vinter, and T. S. j. Olsen. Aphasia after stroke: type, severity and prognosis. *Cerebrovascular Diseases*, 17(1):35–43, 2003.

[121] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[122] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. G. 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.

[123] 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*, 27(8):1226–1238, 2005.

[124] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The Kaldi Speech Recognition Toolkit. In *Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, 2011.

[125] R. Prins and R. Bastiaanse. Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, 18(12):1075–1091, 2004.

[126] X. Qian, H. Meng, and F. Soong. A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 24(6):1020–1028, 2016.

[127] F. Ramus, M. Nespor, and J. Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition*, 75(1):AD3 – AD30, 2000.

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

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

[130] 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*, 19(7):2081–2090, 2011.

[131] R. R. Robey. A Meta-Analysis of Clinical Outcomes in the Treatment of Aphasia. *Journal of Speech, Language, and Hearing Research*, 41:172–187, 1998.

[132] F. Rudzicz. Phonological features in discriminative classification of dysarthric speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009.

[133] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos. The ibm 2016 speaker recognition system. *arXiv preprint arXiv:1602.07291*, 2016.

[134] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos. Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5040–5044, 2016.

[135] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015.

[136] H. Sak, A. W. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH*, Singapore, 2014.

[137] H. Sak, A. W. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Interspeech*, Singapore, 2014.

[138] H. Sak, A. W. Senior, K. Rao, and F. Beaufays. Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition. *CoRR*, abs/1507.06947, 2015.

[139] H. Sak, O. Vinyals, G. Heigold, A. W. Senior, E. McDermott, R. Monga, and M. Mao. Sequence Discriminative Distributed Training of Long Short-Term Memory Recurrent Neural Networks. In *Interspeech*, Singapore, 2014.

[140] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In *Automatic Speech Recognition and Understanding (ASRU)*, pages 55–59, Olomouc, Czech Republic, 2013.

[141] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez. Tools and technologies for computer-aided speech and language therapy. *Speech Communication*, 51(10):948–967, 2009.

[142] M. L. Seltzer and J. Droppo. Multi-task learning in deep neural networks for improved phoneme recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013.

[143] A. Senior and I. Lopez-Moreno. Improving DNN Speaker Independence with I-vector Inputs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[144] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel. Mixture of plda models in i-vector space for gender-independent speaker recognition. In *Interspeech*, pages 25–28, Florence, Italy, 2011.

[145] W. K. Seong, J. H. Park, and H. K. Kim. Dysarthric speech recognition error correction using weighted finite state transducers based on context–dependent pronunciation variation. In *Computers Helping People with Special Needs*, pages 475–482. Springer, 2012.

[146] H. V. Sharma and M. Hasegawa-Johnson. State-transition Interpolation and MAP Adaptation for HMM-based Dysarthric Speech Recognition. In *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 72–79, Los Angeles, CA, USA, 2010.

[147] H. V. Sharma and M. Hasegawa-Johnson. Acoustic model adaptation using in-domain background models for dysarthric speech recognition. *Computer Speech & Language*, 27(6):1147 – 1162, 2013.

[148] C. M. Shewan and A. Kertesz. Reliability and Validity Characteristics of the Western Aphasia Battery (WAB). *Journal of Speech and Hearing Disorders*, 45(3):308–324, 1980.

[149] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. TOBI: A Standard for Labeling English Prosody. In *Proc. of the 2nd International Conference on Spoken Language Processing (ICSLP)*, pages 867–879, Banff, Alberta, Canada, 1992.

[150] N. Simons-Mackie, A. Raymer, E. Armstrong, A. Holland, and L. Cherney. Communication partner training in aphasia: A systematic review. *Archives of Physical Medicine and Rehabilitation*, 91(12):1814–1837, December 2010.

[151] H. Stadthagen-Gonzalez and C. J. Davis. The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38(4):598–605, Nov 2006.

[152] A. Stolcke, J. Zheng, W. Wang, and V. Abrash. SRILM at sixteen: Update and outlook. In *Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[153] T. Tan, Y. Qian, D. Yu, S. Kundu, L. Lu, K. C. Sim, X. Xiao, and Y. Zhang. Speaker-aware training of LSTM-RNNS for acoustic modelling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5280–5284, 2016.

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

[155] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016.

[156] S. Thomas, M. L. Seltzer, K. Church, and H. Hermansky. Deep neural network features and semi-supervised training for low resource speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6704–6708, Vancouver, BC, Canada, 2013.

[157] G. Van Nuffelen, C. Middag, M. De Bodt, and J. P. Martens. Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language and Communication Disorders*, 44(5):716–730, 2009.

[158] O. Vinyals, L. Deng, D. Yu, and A. Acero. Discriminative pronounciation learning using phonetic decoder and minimum-classification-error criterion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4445–4448, 2009.

[159] J. Wambaugh, J. R. Duffy, M. McNeil, D. Robin, and M. Rogers. Treatment guidelines for acquired apraxia of speech: A synthesis and evaluation of the evidence. *Journal of Medical Speech-Language Pathology*, 14(2):xv–xxxiii, Jun 2006.

[160] J. L. Wambaugh, M. M. Kalinyak, and J. E. West. A Critical Review of Acoustic Analyses of Aphasic and/or Apraxic Speech. *Clinical Aphasiology*, 24:35–63, 1996.

[161] G. Weismer and J. S. Laures. Direct magnitude estimates of speech intelligibility in dysarthria: effects of a chosen standard. *Journal of Speech, Language, and Hearing Research*, 45(3):421–433, 2002.

[162] S. M. Witt. Automatic error detection in pronunciation training: Where we are and where we need to go. In *International Symposium on Automatic Detection of Errors in Pronunciation Training*, Stockholm, Sweden, 2012.

[163] S. M. Witt and S. J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(23):95 – 108, 2000.

[164] S.-C. Yin, R. C. Rose, O. Saz, and E. Lleida. Verifying pronunciation accuracy from speakers with neuromuscular disorders. In *Interspeech*, Brisbane, Australia, September 2008.

[165] C. Yu, C. Zhang, S. Ranjan, Q. Zhang, A. Misra, F. Kelly, and J. H. Hansen. UTD-CRSS system for the NIST 2015 language recognition i-vector machine learning challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5835–5839, 2016.

[166] D. Yu and L. Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2014.

[167] D. Yu and M. Seltzer. Improved Bottleneck Features Using Pretrained Deep Neural Networks. In *Proc. of the 12th Annual Conference of the ISCA (INTERSPEECH)*, Florence, Italy, 2011.

[168] D. Yu, K. Yao, H. Su, G. Li, and F. Seide. KL-Divergence Regularized Deep Neural Network Adaptation For Improved Large Vocabulary Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, 2013.

[169] D. Yu, K. Yao, H. Su, G. Li, and F. Seide. Kl-divergence Regularized Deep Neural Network Adaptation for Improved Large Vocabulary Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[170] W. Ziegler and A. Zierdt. Telediagnostic assessment of intelligibility in dysarthria: a pilot investigation of MVP-online. *Journal of Communication Disorders*, 41(6):553–577, 2008.