A Computational Model of the Relationship Between Speech Intelligibility and Speech Acoustics

by

Yishan Jiao

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2019 by the
Graduate Supervisory Committee:

Visar Berisha, Co-chair
Julie Liss, Co-chair
Yi Zhou

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

Speech intelligibility measures how much a speaker can be understood by a listener. Traditional measures of intelligibility, such as word accuracy, are not sufficient to reveal the reasons of intelligibility degradation. This dissertation investigates the underlying sources of intelligibility degradations from both perspectives of the speaker and the listener. Segmental phoneme errors and suprasegmental lexical boundary errors are developed to reveal the perceptual strategies of the listener. A comprehensive set of automated acoustic measures are developed to quantify variations in the acoustic signal from three perceptual aspects, including articulation, prosody, and vocal quality. The developed measures have been validated on a dysarthric speech dataset with various severity degrees. Multiple regression analysis is employed to show the developed measures could predict perceptual ratings reliably. The relationship between the acoustic measures and the listening errors is investigated to show the interaction between speech production and perception. The hypothesize is that the segmental phoneme errors are mainly caused by the imprecise articulation, while the sprasegmental lexical boundary errors are due to the unreliable phonemic information as well as the abnormal rhythm and prosody patterns. To test the hypothesis, within-speaker variations are simulated in different speaking modes. Significant changes have been detected in both the acoustic signals and the listening errors. Results of the regression analysis support the hypothesis by showing that changes in the articulation-related acoustic features are important in predicting changes in listening phoneme errors, while changes in both of the articulation- and prosody-related features are important in predicting changes in lexical boundary errors. Moreover, significant correlation has been achieved in the cross-validation experiment, which indicates that it is possible to predict intelligibility variations from acoustic signal.

To my beloved family

ACKNOWLEDGMENTS

This dissertation would not have been possible without the support of my advisors, committee members, colleagues, family and friends.

Formost, I would like to express my deepest application to my advisors, Dr. Visar Berisha and Dr. Julie Liss, for their guidance, encouragement and persistent support during my PhD study. Dr. Visar Berisha has advised me to propose meaningful research projects and develop effective solutions. I learned from him that if a problem could be solved in a simple way, there was no need to make it complicated. Dr. Julie Liss has advised me to answer scientific problems in scientific ways. Besides, her amazing personality made me believe that a person could be so blessed if she or he could stay positive and be nice to people around.

I would like to thank Dr. Yi Zhou for being my committee member and providing helpful advice to my study. Her passion and devotion to research set a wonderful example for me and teach me to always keep a positive attitude no matter how hard the problem we are facing.

I am highly indebted to the Speech and Hearing Science Department for providing the diverse multidisciplinary environment and caring the development of each individuals. Being connected with professors in various fields and learning their unique studies opened my mind and enriched my knowledge.

I would like to thank the professors that have taught me and help me during my PhD study. Thanks to Dr. Michael Dorman, Dr. Andrea Pittman, Dr. Corianne Rogalsky, Professor Myra Schatzki, and Professor Kelly Ingram. As an international student, I am grateful to them for not only teaching me knowledge but also being thoughtful and respecting the culture differences.

Great thanks to my lab mates and friends during my time at ASU for their companion and support. Thanks to Alan Wisler, Megan Willi, Beatriz Barragan, Lourdes

Nieto and many others who have helped me and cared me. Thanks to Aishwarya Pratap Singh, Anna Grabek, Marissa Miller, and Karly Dierckens for their help in the data collection part of this dissertation.

In addition, many thanks to my husband, Ming Tu. Thanks for walking with me during the past eight years of school life. Thanks for his support and encouragement in both study and life. Thanks to my parents for their selfless love and their indispensable support during my PhD study. Special thanks for my daughter, Yimi Tu, for coming to my life and bringing me unprecedented feeling of happiness.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

1.1 Problem Statement and Hypothesis

Speech enables human to communicate efficiently. Speech disorders affect the way a person produces speech sounds. Being unable to articulate their thoughts clearly and fluently impacts communicative ability and overall quality of life (Smith et al., 1996). Speech-language pathologists help people with speech disorders find and overcome the issues in their speech production and develop practical communication strategies. Since intelligibility is an indication of general communicative competence, improving intelligibility is the central goal of speech therapy.

In laymans terms, intelligibility is the degree to which a speaker can be understood by a listener, or how comprehensible the speech is. Although the word "intelligibility" appears frequently in the literature, there has not been a commonly accepted agreement on the definition and measurement of intelligibility. The reason is that intelligibility is an aggregative indicator of speech which can be affected by many factors, such as linguistic structure, articulatory precision, perceptual strategies, and sometimes physiological causes (Kent, 1992). Many studies have attempted to assess intelligibility using a variety of methods.

Early literature has largely focused on intelligibility tests that provide an index of severity (Enderby, 1980; Yorkston, Beukelman, & Traynor, 1984). Weismer, Martin, and Kent (1992) suggest that an analysis model based on acoustic-phonetics can be potentially more useful in guiding the decision-making process in clinical practice. Moreover, they suggest that intelligibility is not only a speaker characteristic, rather it is a two-way measure of the speaking-listening process. The degraded speech signal may present more difficulties to the listener than the expected mismatch between acoustic-phonetic events.

In the studies by Liss and colleagues (Liss, Spitzer, Caviness, Adler, & Edwards, 1998; Liss, Spitzer, Caviness, & Adler, 2002; Liss, Spitzer, Caviness, Adler, & Edwards, 2000), evidence has been offered for this bidirectional relationship between the degraded speech signal and the strategies listeners use to decipher it. They suggest that in the perception of disordered speech, more higher-level cognitive processing is involved when the lower-level acoustic information becomes degraded or unreliable. It has been found that listeners relied on temporal rhythmic cues, such as the contrasts between stress-unstressed syllables to make lexical segmentation decisions. This also suggests that when assessing intelligibility, both segmental and suprasegmental information should be considered.

From the speakers perspective, acoustic analysis is necessary to identify the cues in the speech signal that are related to intelligibility degradation. In the literature, a number of acoustic features have been developed and used for analyzing disordered speech, such as speaking rate, vowel spaces, formant slopes, jitter/shimmer, voice onset time (VOT), etc. (Kent & Kim, 2003; Weismer, Jeng, Laures, Kent, & Kent, 2001) It is known that to explore the sources of intelligibility degradation, acoustic analysis must be done on various aspects of speech, such as articulation, prosody, vocal quality, nasality, etc. However, traditional acoustic analysis largely relies on human labor, which results in the difficulties of studying acoustics of speech disorders in a broader range. Our lab has focused during the last several years on developing automated acoustic analysis tools for the study of speech disorders. We aim to explore clinically useful information from the speech signal and providing reliable, quantitative and comprehensive indexes of speech to facilitate clinical practice. Although the validity of each of the developed methods has been shown separately in our previous work, they have not been used together in the context of an explanatory model of intelligibility.

This dissertation describes the work of the following:

1. The development of a systematic listener transcript analysis method. A multidimensional intelligibility profile (MIP) will be estimated automatically from the listener transcripts. Different from traditional perceptual intelligibility measures, the MIP reveals how degraded speech challenges listeners attempt to decipher it by characterizing speech perception process across multiple levels of granularity, from segmental (phoneme errors) to suprasegmental (lexical boundary errors). Multiple regression analysis will show the relationship between the estimated MIP metrics with the multi-dimensional perceptual ratings of dysarthric speech.

2. The development of a suite of automatic acoustic measurements that quantify speech signal along multiple temporal and frequency scales. These measures provide analysis of speech from different aspects, including articulation, prosody, rhythm and phonation. Multi-task learning based multiple regression will help us study the relationship between the developed acoustic features with each dimension of the perceptual ratings of dysarthric speech.

3. The exploration of the relationship between variations in the acoustic measures and those in the MIP metrics. It will reveal the interaction between acoustic signal and listener perception strategies. Speakers will be instructed to make changes in their speaking manners so as to elicit changes in acoustic signals. Regression models are built to estimate the relationship between change in the acoustic measures and that in MIP metrics.

Acoustic signal will drive the percepts of the listeners. The level of speech intelligibility depends on the pronunciation as well as the rhythm controls of the speaker. Since English is a stress timed language, when the low-level phonemic information is degraded, listeners tend to rely on the contrastivity between stressed and unstressed syllables to segment a continuous acoustic stream into words so as to facilitate speech understanding. We hypothesize that the patterns of change in the MIP metrics can be explained by changes in the acoustic features, as a consequence of modifying speech based on intervention instructions. The specific hypothesis is as follows:

1. The MIP metrics measures phoneme recognition accuracy and lexical segmentation accuracy of the listeners. We hypothesize that the phoneme error metrics are closely related to the perceptual ratings on articulation, and the lexical segmental error metrics are closely related to the perceptual ratings on prosody.

2. The acoustic features measure speech production from the aspects of articulation, prosody, and voice quality. We hypothesize that the the articulation-related features are important in predicting perceptual ratings on articulation, the prosody-related features are important in predicting perceptual ratings on prosody, and the voice quality-related features are important in predicting perceptual ratings on vocal quality. All of the three categories of features are important in predicting the overall intelligibility of dysarthric speech.

3. We expect that speech produced in different speaking manners will be detected by the developed MIP metrics and acoustic features. We hypothesize that variations observed in the acoustic features can explain the variations in listener perceptual strategies measured by MIP metrics. Articulation-related acoustic features account for the most variations in phoneme error metrics. However, variations in lexical boundary errors should be explained by the degradations

in both articulation- and prosody-related features. The voice quality-related features measure the noisiness of the signal, which can have more impact on phoneme perception than prosody perception.

1.2 Significance of the Study

Speech disorders affect millions of people. Speech intelligibility, as a general indicator of communication ability, is central in the diagnosis and treatment of speech disorders. In the current clinical setting, the most commonly used intelligibility assessment is the clinicians informal perceptual estimation of the patients speech. However, ample evidence exists to suggest that auditory-perceptual judgements are inherently biased, especially those of the treating clinician whose perceptual system has been adapted to the patients speech patterns. There is an urgent need to develop a suite of reliable and comprehensive measurements to assess the intelligibility of disordered speech objectively and facilitate clinical practice. The current study investigates the relationship between the available acoustic information in the degraded speech and how listeners use that information. The mismatches between the availability of acoustic cues and the ones that listeners need to decipher speech reveal sources of intelligibility degradation. As a result, intelligibility can be modeled computationally by a number of automated acoustic features and measures automatically extracted from listener transcripts. The contribution of the current study is as follows:

1. We develop algorithms for holistic automated listener transcript analysis, including segmental and suprasegmental metrics based on phoneme and lexical segmentation errors. This promotes the development of objective perceptual intelligibility assessment.

2. We develop novel automated and clinically interpretable acoustic features, including spontaneous speech rate estimation (Jiao, Berisha, Tu, & Liss, 2015;

Jiao, Tu, Berisha, & Liss, 2016), entropy based phonemic inventory estimation (Jiao, Berisha, Liss, Hsu, et al., 2017), interpretable phonological features (Jiao, Berisha, & Liss, 2017), etc. We link these acoustic features to speech aspects that clinicians care, such as prosody and articulatory precision. Moreover, this study explores the different contribution of these features to intelligibility degradation.

3. This study provides a theoretical explanation of intelligibility degradation in adverse listening conditions, by considering the impact of available acoustic information and strategies employed by listeners to decode speech.

4. By estimating intelligibility in a computational and clinically meaningful way, the importance of the factors affecting intelligibility are identified. It will provide valid and reliable outcome measures for progress tracking which is important for decision-making and the evaluation of interventions.

1.3 The Outline of the Dissertation

The dissertation is divided into 7 chapters. Chapter 1 is the current chapter and it introduces the scientific problem, our hypotheses, and the significance of the s-tudy. Chapter 2 introduces the up-to-date research progress in speech intelligibility assessment and the existing methods of pathological speech acoustic analysis. Chapter 3 describes the experiment design and data collection. Chapter 4 introduces the automated MIP metrics developed for analyzing listener transcript errors, including phoneme errors and lexical boundary errors. Chapter 5 introduces the automated acoustic measures developed for analyzing disordered speech across several subsystems and temporal scales (articulation, prosody, rhythm, and phonation). Chapter 6 investigates the impact of changes in acoustic signal on the changes in speech percepts by exploring the relationship between variations in the acoustic features and in the MIP metrics. Chapter 7 discusses the findings from the experiments.

Chapter 2

LITERATURE REVIEW

This chapter reviews the literature related to the study of intelligibility in speech disorders. Section 2.1 introduces two traditional perceptual intelligibility assessment methods commonly used in the literature: assessment based on scaling procedures and assessment based on word identification/transcription tasks. In the same section, we provide an overview of problems associated with these methods and recently-proposed improvements. Section 2.2 introduces some emerging studies in automated intelligibility assessment. Section 2.3 reviews the existing studies investigating the relationship between acoustic features and intelligibility percepts.

2.1 Perceptual Intelligibility Assessment

Schiavetti et al. (1992) stated that any measure of speech intelligibility is a measurement of the interaction between a speaker, a transmission system, and a listener. Therefore, he suggested that speech intelligibility could be defined as the match between the intention of the speaker and the response of the listener to the speech passed through the transmission system. In the speech-language pathology field, the measure of intelligibility is typically treated as a criterion for assessing the severity of speech disorders (Metz, 1980). Traditionally, there are two methods to assess intelligibility:

1. **Word identification test in which the listener is required to transcribe what the speaker says:** The outcome variable in these tests is the percentage of words that the listener's responses match the speaker's indention. The test can be done on single-word stimuli by using a carefully designed word list

that considers phonemic contrasts (Kent, Weismer, Kent, & Rosenbek, 1989; Tikofsky, 1970; Yorkston & Beukelman, 1980), or at the utterance level by considering contextual impact on intelligibility (Dongilli, 1993; Hammen et al., 1991; K. K. Tjaden & Liss, 1995; Yorkston & Beukelman, 1981b) Figure 2.1 shows the procedures of the two tasks.

2. **The scaling procedures in which the listener makes judgements about the speaker's intelligibility:** Equal appearing interval (EAI) (Frearson, 1985) is one of the techniques where the listeners, typically a speech-language pathologist (SLP), rates the speech on a pre-defined scale (e.g., 7-point or 5-point interval scale). Different from the interval scaling procedure where a constrained scale is provided, another technique is the direct magnitude estimation (DME) (S. S. Stevens & Galanter, 1957; Weismer & Laures, 2002) in which listeners scale each speech sample with or without a given standard stimulus. A standard stimulus or modulus is chosen by the experimenter to represent low, middle, or high intelligibility and is assigned to a number. Listeners then rate other samples against the modulus. In a free modulus setting, listeners assign any number to the first speech sample and then scale subsequent samples as magnitude ratios relative to preceding stimuli (Schiavetti et al., 1992). Figure 2.2 shows the procedures of EAI, free-modulus DME, and with-modulus DME.

Due to their complementary benefits and drawbacks, both methods are widely used in the research and clinical studies of pathological speech. On one hand, the word identification test is interpretable to the patients and other professionals by quantifying intelligibility using an interpretable percentage value (e.g. percent words correct). It has also been shown to have a close relationship with the information transferred during communication (Beukelman & Yorkston, 1979). Moreover, listen-

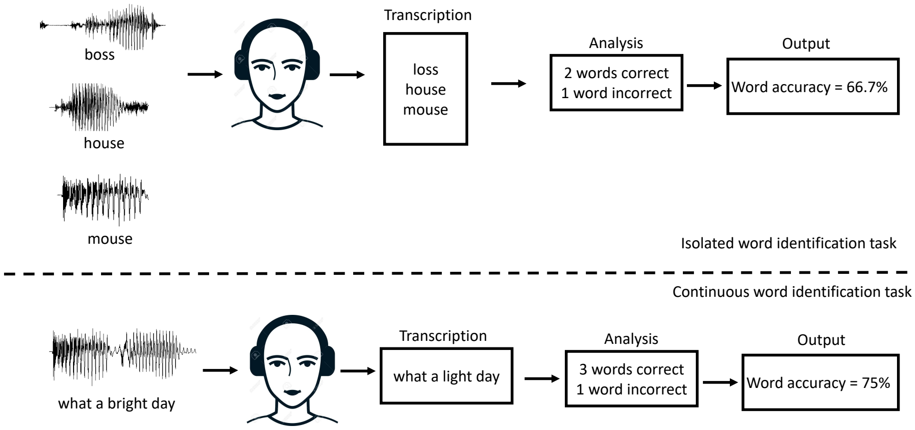


Figure 2.1: Illustration of the Procedures of the Isolated (Upper) and Continuous (Lower) Word Identification Tasks.

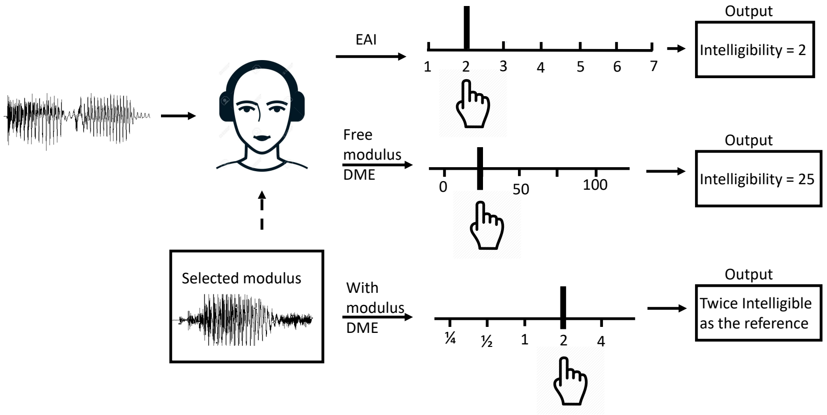


Figure 2.2: Illustration of the Scaling Procedures of Intelligibility Scoring Tasks.

ers can be recruited from the general population which makes the conduction of the test efficient and economical. However, the intelligibility scores are frequently derived from single word identification task, which may not be sensitive to non-segmental contributors of intelligibility deficits, such as prosody and voice quality. Compared to single word intelligibility test, evaluation on connected speech is closer to the functional level of communication, and evidence has shown that the results are different to that in isolated word intelligibility tests (Fogerty & Kewley-Port, 2009; Weismer et al., 1992).

On the other hand, the scaling procedure is a direct assessment of perceptual intelligibility. It can be applied to various dimensions of speech, such as articulatory precision, speech naturalness, voice quality, etc. EAI has been frequently used in the early relevant studies (Frearson, 1985). One of the most concerns regarding the use of EAI is its validity. S. Stevens (2012) suggests that the equal partition of the scale may not be consistent with the nonlinear perception of some dimensions. An alternative method, DME, has been widely accepted in the communication disorders field. DME with modulus is usually preferred because free modulus scaling may make listeners uncomfortable and the post-processing of the data complicated (Weismer & Laures, 2002). The advantage of DME over EAI is that it does not make linear assumptions and is not bound by fixed minimum/maximum values and thus no constraints on the scales (Zraick & Liss, 2000). Despite the popularity of the scaling methods, there are some crucial problems with it. One of the inherent issues is that the scaling method is a subjective procedure which is prone to human bias. The bias can derive from many factors, such as the different internal standards, the familiarization with the material or the speaker, and the varied experience (Hustad & Cahill, 2003; Liss et al., 2002; McHenry, 2011) Another problem comes from the design of the test. Studies have shown that the perceptual scaling of a fixed set of utterances depends on the identity of the standard (Poulton & Poulton, 1989; Weismer & Laures, 2002) For example, in the DME method, there is not a standard modulus can be broadly used across different studies so that the selection of the modulus relies heavily on the expertise of the experimenters. Last but not least, the participants of the scaling test are either experienced SLPs or people who have received a certain amount of training. Unlike the word identification test which can be done by any normal hearing listeners, the cost of the scaling procedures is much higher and more time consuming.

Although the drawbacks of the traditional methods are acknowledged, they are still widely used in the current studies (Lansford, Luhrsen, Ingvalson, & Borrie, 2018; McAuliffe, Fletcher, Kerr, O'Beirne, & Anderson, 2017) and there have been several efforts focused on improving the reliability of intelligibility assessment criteria. De Bodt, Huici, and Van De Heyning (2002) stated that intelligibility is the product of a series of interactive processes as phonation, articulation, resonance and prosody. Therefore, they proposed to estimate intelligibility using a linear combination of ratings on voice quality, articulation, nasality, and prosody (See Equation 2.1). The coefficients are shown as below. Their findings suggest that articulation and prosody have stronger impact on intelligibility than the other two dimensions.

$$\begin{aligned} \mathit{Intelligibility} =& 0.1626 \times (\mathit{voice\ quality}) + 0.66 \times (\mathit{articulation}) \\ &+ 0.0141 \times (\mathit{nasality}) + 0.3139 \times (\mathit{prosody}) \end{aligned} \tag{2.1}$$

However, this method is still based on auditory-perceptual judgements, which are inherently biased as mentioned in the previous subsection. To evaluate intelligibility objectively, Liss et al. (1998, 2002, 2000) attempt to derive more information from listener transcripts than the word identification test. They suggest that phonemic degradation forces listeners to use more robust acoustic cues, such as syllabic contrastivity, to segment speech into word-size frames, in which to resolve phoneme identity and comprehend speech. Thus, lexical segmentation is critical in speech intelligibility. Therefore, in their studies, a list of phrases was designed especially for studying lexical segmentation strategies used by the listeners so that the suprasegmental impact on intelligibility deficits can be estimated from listener transcripts by estimating lexical boundary errors.

2.2 Automated Intelligibility Assessment

As we can see, the current methods for intelligibility assessment involve significant human effort, which can make them impractical for use in-clinic. The development of automated methods which require little to no human involvement will not only make intelligibility assessment more reliable but also provide a handy research and clinical tool for the community. Automatic speech recognition (ASR) is a technique that automatically transforms an acoustic representation of human speech into a text of word sequence. The training of an ASR engine usually requires a large amount of speech samples. In the speech disorders field, ASR has been used to recognize pathological speech and intelligibility is estimated as the percentage of words correctly decoded by ASR (Maier et al., 2009; Middag, Martens, Van Nuffelen, & De Bodt, 2009; Middag, Van Nuffelen, Martens, & De Bodt, 2008; Tu, Wisler, Berisha, & Liss, 2016). Although there exists a correlation between the estimated values and the perceptual intelligibility scores, the underlying difference of speech recognition in human and ASR makes it unreliable to reflect real speech perception process.

Bocklet and colleagues (Bocklet, Haderlein, Hönig, Rosanowski, & Nöth, 2009) borrow the idea from speaker identification and propose to predict intelligibility using the super-vector extracted from the Gaussian mixture models (GMM). It also has been proven to be language-independent (Middag, Bocklet, Martens, & Nöth, 2011). Other automated methods include predicting intelligibility from the statistics of acoustic features. Falk, Chan, and Shein (2012) use 6 acoustic features to characterize atypical speech from multiple perceptual dimensions, such as voice quality, temporal dynamics, nasality, and prosody. These features are used to classify speech samples into four categories with low to high intelligibility scores. Similarly, Kim and colleagues (J. Kim, Kumar, Tsiartas, Li, & Narayanan, 2012) predict intelligibility using acoustic features from multiple aspects of speech. Instead of using a linear combination of all the features, multiple classification models based on each category of features are fused together to predict intelligibility.

2.3 Acoustic Measures Related to Intelligibility

Apart from the progress in the assessment of intelligibility, attentions have also been placed on investigating acoustic cues related to intelligibility. Speaking rate is one of the most prominent symptoms of speech disorders. Although no significant correlation is found between speaking rates and intelligibility scores (Weismer, Laures, Jeng, Kent, & Kent, 2000; Yorkston & Beukelman, 1981b), studies have shown that rate control strategies had positive impact on speech intelligibility (Yorkston, Hammen, Beukelman, & Traynor, 1990). Yorkston and Beukelman (1981a) suggest that there exists an optimal speaking rate for a speaker to achieve a relatively good speech intelligibility. The explanation could be that rate control intervention does not only modify the speaking rate of a speaker, it may also increase articulatory precision, as well as help to coordinate various speech processes. Moreover, listeners may also use different perception strategies when listening to different rates of speech (Blanchet & Snyder, 2010). This indicates that the intelligibility and the changes in acoustic signals have a complex relationship due to the interaction among different acoustic features and the hierarchical structure of speech perception.

This effect can be also reflected in the loudness treatment method. Treatment of loudness, usually referring to the Lee Silverman Voice Treatment (LSVT), focuses on increasing the speech loudness of dysarthric speakers. It has also been shown to have positive effects on improving speech intelligibility (Ramig, Sapir, Fox, & Countryman, 2001; Wenke, Theodoros, & Cornwell, 2008). Again, the effects are not only due to the increased sound pressure level (SPL), but also the subsequent changes in other acoustic features, such as formants (Sapir, Spielman, Ramig, Story, & Fox, 2007). K. Tjaden and Wilding (2004) investigated the effects of rate and loudness control on the acoustic signal and the intelligibility of dysarthric speech. The results showed that the vowel space area was maximized in slow speech, while the first-moment difference measures, indexing stop consonant acoustic distinctiveness, was maximized in loud speech, with intelligibility improved in both conditions. The authors also suggest in another study (K. Tjaden & Wilding, 2011) that dysarthric speech with slowed rate and increased vocal loudness has distinctive F0 variations than the habitual speech. All of these findings suggest that the improved intelligibility is related to the changes of a series of acoustic features even if the treatment focuses on a single aspect. However, it is unclear how the changes in acoustic features are related to intelligibility degradation or improvement, and if they are reliable enough to assess intelligibility.

Articulation has been shown as the strongest contributor to intelligibility among other perceptual dimensions, such as voice quality, nasality, and prosody (De Bodt et al., 2002). Therefore, measures related to articulation, such as the vowel formant frequencies and the vowel space measures have been widely used to assess speech intelligibility. It has been shown that the speakers with larger vowel spaces are more intelligible than those with reduced spaces, and vowel space measures are significantly correlated with speech intelligibility (Bradlow, Torretta, & Pisoni, 1996; H.-M. Liu, Tsao, & Kuhl, 2005; Turner, Tjaden, & Weismer, 1995). It suggests that the vowel space measure is a good predictor of intelligibility. Some latest studies indicate that there exist other articulatory measures that could better represent speech intelligibility, such as the overlap degree among vowels (H. Kim, Hasegawa-Johnson, & Perlman, 2011), the distinctiveness among neighboring vowels (Neel, 2008), etc. Other acoustic features have also been shown related to intelligibility, such as F0 variability, segment durations, format slopes, modulation energies, residual signal distributions, cepstral coefficients (Bunton, Kent, Kent, & Duffy, 2001; Falk et al., 2012; Weismer et al., 2001).

Although the above studies have shown a relationship between acoustic features to intelligibility, it has not been shown how they are related to intelligibility gains, which should be more meaningful in helping identify treatment targets and select intervention strategies. To simulate intelligibility variation, Fletcher and colleagues (Fletcher, McAuliffe, Lansford, Sinex, & Liss, 2017a) (Fletcher, Wisler, McAuliffe, Lansford, & Liss, 2017) recorded the same group of speakers reading the same material in different speaking modes (habitual, loud, and slow). A set of acoustic features, related to prosody and articulation were extracted manually or automatically to predict intelligibility gains obtained in a subjective listening experiment. Their results suggest that variance in intelligibility gains can be partially explained by their explored acoustic measures.

Chapter 3

METHOD OVERVIEW AND DATA COLLECTION

3.1 Method Overview

To investigate the research questions and test the hypotheses described in Chapter 1, the following steps will be employed and shown in Figure 3.1.

*Step*1 Audio data collection. Speech samples from people with and without motor speech disorders were collected. Intelligibility variations was simulated with different speaking modes, which are habitual, slow, loud, and clear. The procedures will be described in Section 3.2.2.

*Step*2 Listener transcript collection. Transcripts for each sample were collected from multiple non-expert listeners. We conducted the listening experiment through an online crowd-sourcing platform, MTurk. The procedures will be described in Section 3.2.3

*Step*3 Transcript scoring. MIP metrics were used to quantify perceptual intelligibility segmentally and suprasegmentally from listener transcripts. Algorithms were developed to estimate three phoneme errors and four lexical boundary errors. The validity of the algorithms was proved by comparing the estimated metrics with hand labels. The capability of the metrics to predict intelligibility was examined using linear regression. Details of the algorithms and the experiments will be described in Chapter 4.

*Step*4 Acoustic analysis. A variety of acoustic features were developed for analyzing the collected speech signal from different aspects, including articulation, prosody, and voice quality. Their relationship with perceptual ratings was examined using multi-task learning. The algorithms of each acoustic measure along with their perceptual interpretation, and the experiment settings and results will be described in Chapter 5.

*Step*5 Relationship investigation. We investigated the relationship between acoustic features and MIPs. Statistics (e.g. means) of the acoustic features and the MIP metrics were calculated from the collected audios and transcripts for each speaker. Changes were identified from habitual to the other speaking modes in every measure. Regression analysis was conducted to reveal how changes in acoustic features impact the strategies listeners used to understand speech (measured by MIP metrics). Details of the experiment and results will be described in Chapter 6.

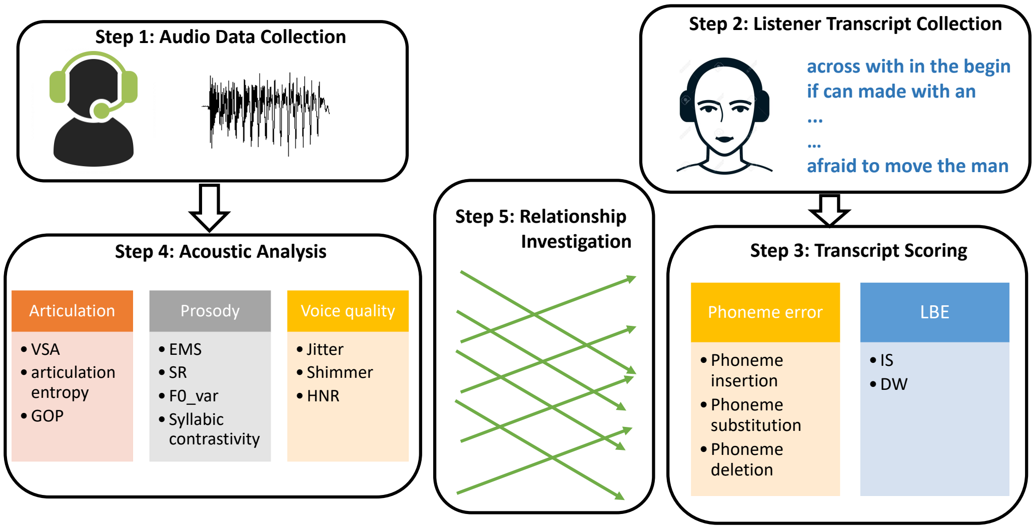


Figure 3.1: Experimental Procedures.

3.2 Data Collection

3.2.1 Stimuli

The speech stimuli consist of 80 phrases, which had a rich and balanced phoneme inventory. They were designed especially for studying the relationship of intelligibility and lexical boundary error (LBE) in dysarthric speech (Liss et al., 1998, 2002, 2000). Please see Appendix A for the whole list of the phrases. Briefly, they are comprised of 6 syllables that create 3 to 5 mono- and di-syllabic words, which form grammatically plausible phrases with low inter-word predictability. The phrases were designed to alternate strong (S) and weak (W) syllables in either trochaic or iambic stress patterns to induce LBEs. Some examples of the phrases and their stress patterns are shown in Table 3.1, where spaces in the stress pattern indicate the word boundaries.

Table 3.1: Examples of the Stimuli Phrases and the Stress Patterns.

Phrase	**Stress Pattern**
address her meeting time	WS W SW S
bolder ground from justice	SW S W SW
beside a sunken bat	WS W SW S
cool the jar in private	S W S W SW

3.2.2 Audio Data Collection

Two speech datasets were used in this study. The first dataset was used in our previous study and described in the paper by Liss et al. (2009). Briefly, it contains speech samples from 73 dysarthric speakers with 34 females and 39 males. The dysarthria subtypes included ataxic dysarthria secondary to cerebellar degradation

(Ataxic, N = 16), hyperkinetic dysarthria secondary to Huntington's disease (HD, N = 6), mixed spastic-flaccid dysarthria secondary to amyotrophic lateral sclerosis (ALS, N = 14), and hypokinetic dysarthria secondary to idiopathic Parkinsons disease (PD, N = 37). These speakers provided a variety of speech error patterns and represent mild to severe intelligibility decrements within each subtype (see Table 3.2). Each speaker in this dataset read the above mentioned 80 phrases in their normal voice.

Table 3.2: Descriptions of the Four Subtypes of Dysarthria in the Dataset.

Dysarthria Types	Etiology	Speech Characteristics
Ataxic	Cerebellar degeneration	Irregular articulatory breakdown, distorted vowels, prolonged phonemes, monopitch
Hyperkinetic	Huntington's disease	Irregular and intermittent consonant and vowel distortion, inappropriate,silences, bursts of loudness change
Mixed Spastic-Flaccid	ALS	Imprecise consonants, hypernasality, slow rate, distorted vowels, strained-strangled vocal quality
Hypokinetic	Parkinson's disease	Imprecise consonants, breathiness, monopitch, reduced stress, inappropriate silences, short rushes of speech

For this dataset, perceptual ratings were also collected. Fifteen second-year master students enrolled in the SLP program at ASU rated each speaker along five perceptual dimensions: severity, nasality, vocal quality, articulatory precision, and prosody on a scale from 1 to 7 (from normal to severely abnormal). Their ratings were integrated into a single set using the evaluator weighted estimator (EWE) method.

The other dataset was newly collected for this study. It contains 20 healthy participants, including 10 females and 10 males with average age as 69.2. To simulate acoustic changes within speakers, each speaker read 40 phrases randomly selected from the 80 ones in four different speaking modes: habitual, loud, clear, and slow. Following habitual, the other three modes were randomized for each speaker and each phrase. The instructions for the four speaking modes are as follows:

- *Please read the following phrase in your typical voice.* (Habitual)
- *Please read the following phrase loud enough for a person across the room to hear.* (Loud)
- *Please read the following phrase using a very clear voice.* (Clear)
- *Please read the following phrase about half as fast as you usually talk.* (Slow)

Note that these instructions were for eliciting variants of speech change within a given speaker. However, there was no experimental need to ensure speakers produce speech changes in any particular way. Instead, the potential changes would be defined acoustically and perceptually.

All speech samples were recorded either in the Motor Speech Disorders Laboratory at Arizona State University (ASU) or IRB-approved auxiliary research sites (Liss et al., 2009). An elicitation interface created on DMDX, an experimental interface freeware (Forster & Forster, 2003), was used for the audio collection. Participants were seated in a sound-attenuating booth or a quiet room and fitted with a head-mounted microphone. The reading instruction and the text of the phrase were presented to the speaker visually on a computer screen. Before each recording the speaker was prompted by a tone to start.

To induce perceptual intelligibility degradation on healthy speech, white noises were added to the speech signals. In a pilot study, 30 randomly selected habitual phrases were embedded in the white noise at different signal-to-noise ratios (SNR) measured by root-mean-squares (RMS) from -5dB to 0dB. Listener transcripts were collected (in the way described in the following subsection) to calculate an average word error rate. We intended to achieve 50% intelligibility degradation, which was a level that is known to lead listeners to syllabic contrastivity cues for lexical segmentation (Borrie, McAuliffe, & Liss, 2012; Liss et al., 2002). As a result, 0dB was selected with 57% WER. In the formal listener experiment, all healthy speech samples were embedded into white noises to reach 0dB SNR. As such, the sound pressure level that may affect intelligibility variations was excluded.

3.2.3 Listener Transcript Collection

Non-expert listeners were recruited to transcribe the audio samples in the two speech datasets. The consent form approved by IRB can be found in Appendix B. Different from experienced clinicians and listeners after training, these naïve listeners represented a realistic perceptual audience, in which a wide cross-section of typical listeners found themselves in the position of attempting to decipher pathological speech. For each speech sample, 10 transcripts were collected from different listeners because our previous study (Berisha, Liss, Sandoval, Utianski, & Spanias, 2014) showed that 10 transcripts per speech sample by unfamiliar non-expert listeners could achieve stable reliability. To avoid any familiarization with the speakers and the speech material, we set the rule as that each listener transcribed no more than two phrases from the same speaker and never transcribed the same phrase more than once.

To facilitate the recruitment of a large number of listeners, Amazon Mechanical Turk (MTurk) was used in this study. MTurk is an online crowdsourcing platform which enables individuals (Workers) and organizations (Requesters) to coordinate the use of human intelligence to perform tasks. In our study, a website was developed to which listeners are directed from MTurk. Participants were instructed to listen to each spoken phrase using headphones, and type what they think the speaker was saying in a given area. They were asked to make their best guesses if they were unsure about what was said. Each phrase could be played no more than twice.

The biggest advantage of using MTurk comparing to conventional subject recruitment methods is its efficiency. For example, in our study, the speed of data collection could be up to 50 participants per day. Moreover, the reward given to the participants on MTurk was less than that given to the on-site subjects. For example, in our study, each participant received $1-$2 for transcribing 40 to 80 phrases. However, the disadvantage we found in using MTurk was the lack of control and supervision. For example, we wanted to recruit listeners with English as their first languages. However, we were only able to specify their locations as US on MTurk. To obtain such demographic information we need, we designed a questionnaire and asked the participants to provide information about their native languages and mental/hearing conditions before leading them to the transcribing task. Moreover, since the workers completed the task without being monitored, it is questionable whether people pay enough attention to the task. Although the study by (Paolacci & Chandler, 2014) have shown that the rate of failing attention on MTurk was no higher than other formats, it is better to identify those listeners. Therefore, four easy-to-understand phrases read by two healthy speakers were randomly embedded to the task for identifying any inattentive participants.

Chapter 4

THE DEVELOPMENT OF AUTOMATED MIP ANALYSIS

4.1 Introduction

Although perceptual evaluation is central to the differential diagnosis of motor speech disorders (see, The Preferred Practice Patterns for the Profession of Speech-Language Pathology; ASHA), abundant evidence suggests that perceptual estimates of speech intelligibility are inherently biased, unreliable and are particularly unsuitable for the tracking of speech change secondary to intervention or disease progression by the treating clinicians (Liss et al., 2002; McHenry, 2011; Sheard, Adams, & Davis, 1991). Despite that, the most commonly used clinical method for characterizing changes in speech intelligibility is the treating clinicians informal perceptual estimation of their patients speech (Duffy, 2013; King, Watson, & Lof, 2012; Miller, 2013). Survey results suggest that SLPs highly value subjective perceptual assessment and feel comfortable with the common practice of informal estimation of intelligibility; and they regard objective metrics estimated from transcribed speech as a nice to have rather than a must have for clinical practice (Alice & O., 2008; Miller, 2013).

There are at least two factors that contribute to the current clinical practice of preferring subjective over objective assessments for speech characterization. First, the quality of speech is ultimately judged by a human listener, as it impacts the ability to communicate. In the absence of such human factors, objective measures lack clinical interpretability. Second, more objective approaches are resource-heavy and may involve manual coding and scoring of speech transcripts. In this Research Note, we present an automated approach for scoring transcripts that provides a holistic and objective representation of intelligibility that derives from its underlying perceptual tasks of phoneme identification and lexical segmentation of the speech stream (Hustad, 2006).

Traditional objective measures derived from transcripts usually capture word accuracy which provides a percentage of words that the listener's responses match the speaker's intention (Kent et al., 1989; Yorkston & Beukelman, 1981b). Manual scoring is also required if we want to extract additional information from the transcripts, such as lexical segmentation errors. Since this is time consuming and requires specialized training, it has only been done in a few research studies (Liss et al., 1998, 2002, 2000). More recently, there have been several approaches aimed at automating transcript scoring (Borrie, Barrett, & Yoho, 2019; Le, Licata, Persad, & Provost, 2016). These approaches result in objective measures of word or phoneme errors and show a good correlation with perceptual ratings. However, to the best of our knowledge, no existing literature addresses the problem of automated estimation of lexical segmentation errors.

In this chapter, we present a family of algorithms for automating transcript s-coring (relative to a target transcript). The approach described herein automatically extracts information related to phoneme and lexical segmentation errors directly from the transcripts. We call it the multidimensional intelligibility profile (MIP). Specifically, the scoring scheme yields information related to the perceptual task of phoneme identification (phoneme substitution, insertion, and deletion errors) by aligning the target transcript with the transcript the listener produces. In addition, by comparing the word boundaries and the stress patterns of the target and listener transcript-s, we automated the extraction of lexical segmentation metrics previously used in the literature to capture the perceptual task of lexical segmentation of the speech stream (Liss et al., 1998). We demonstrate the validity of the automated metrics by comparing them against manual labels from trained coders. In addition, linear regression analysis provides evidence that these metrics are significant predictors of clinical auditory-perceptual ratings provided by trained listeners and word accuracy from transcription.

4.2 Method

4.2.1 Data Collection

The audio stimuli used in this study was the 73 dysarthric speaker dataset. To collect transcripts, 819 listener participants were recruited via MTurk. Our target was to collect approximately 10 different transcripts per phrase. As a result, a total number of 63,840 phrase transcriptions were collected.

Using a survey, non-native English speakers, listeners with hearing loss, head injury, psychiatric disorders, and attention deficit disorders (ADD) were identified, and their data was excluded from analysis. Additionally, only the data from listeners who transcribed all four easy-to-underastand phrases correctly were included for analysis. After filtering, 498 listeners (60.8%) and 33,969 phrase transcriptions (53.2%) were left. In other words, 39.2% speakers and 46.8% transcripts were discarded from analysis.

4.2.2 Pre-processing of the Transcripts

Because listener responses were not constrained in any way, the collected data contains misspellings, non-English words, acronyms, and other unanalyzable responses. Some examples of these errors are shown in Table 4.1 Although a protocol for manually assessing these entries could be undertaken, it defeats the purpose of an

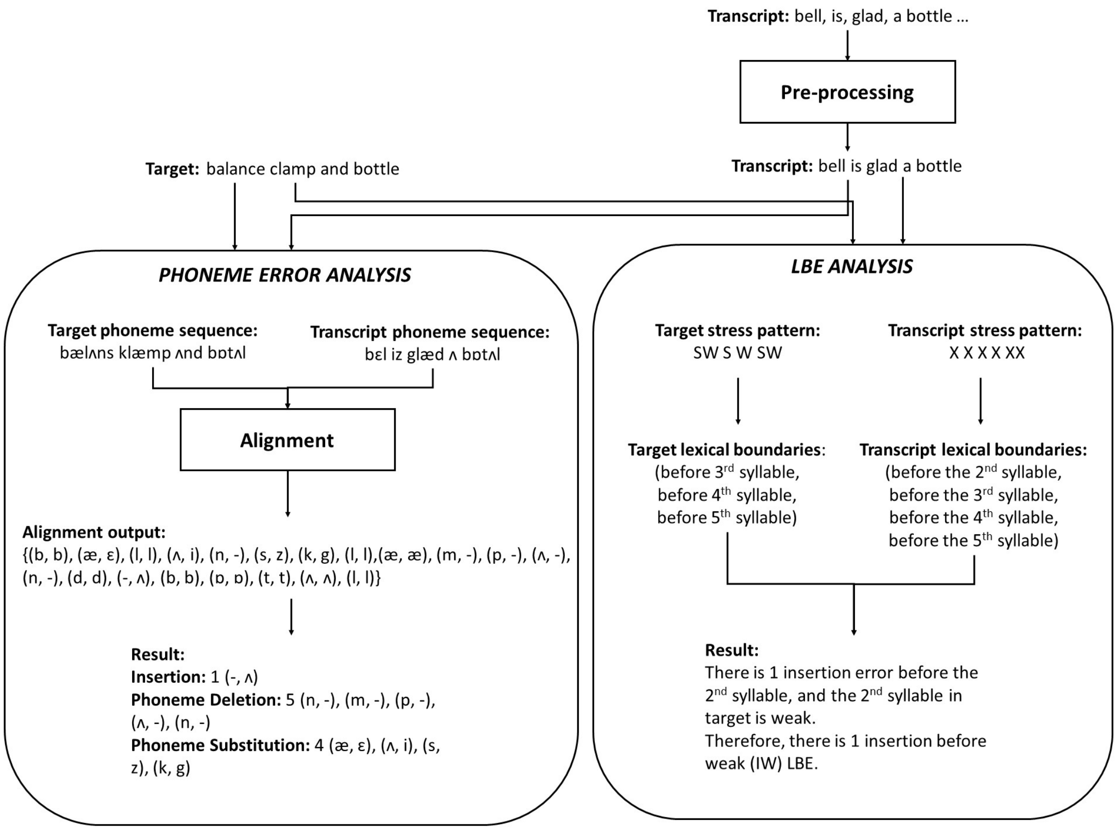
Table 4.1: Errors Appeared in the Collected Transcripts and the Corresponding Examples.

Error Type	**Example**
Contractions	dont, hed
Loan words	Lakh
Hyphen Space Holders	jack___
Hyphens Instead of Spaces	man-to-state
No Spaces	fellingrecklessviolet
Proper Names	spock, amanda
Truncations	fam; vid;
Acronyms	pe
Ambiguous Pronunciation	herby [eɹbi] or [heɹbi]
Morphologized Real Word Results in Nonword	unfortune; precoat
Nonwords	awa
Misspellings	aprthied, cancle

automated assessment measure. Therefore, these transcripts were subject to a series of automated corrections as follows: 1) all special characters were removed except for single quotes and hyphens since they are allowed in the dictionary; 2) an automated spell checking and correction tool was used to correct any misspelling and typos; 3) the whole transcript was discarded if there was any out-of-dictionary word in it. A total of 2,645 transcripts (7.8% of the 33,969 transcripts) were discarded. For the above procedure, we used the Carnegie Mellon University (CMU) English pronouncing dictionary (CMUdict, http://www.speech.cs.cmu.edu/cgi-bin/cmudict)

4.2.3 Automated Transcript Analysis

The automated transcript analysis was then conducted on the qualified transcripts to calculate measures related to phonemic identification (phoneme insertion, deletion, and substitution errors), and to lexical segmentation (4 LBEs). Figure 4.1 provides a schematic example of the scoring of a transcribed phrase relative to the target.



Transcript: bell, is, glad, a bottle ...
Pre-processing
Transcript: bell is glad a bottle
Target: balance clamp and bottle
PHONEME ERROR ANALYSIS
Target phoneme sequence: bælʌns klæmp ʌnd bɒtʌl
Transcript phoneme sequence: bɛl iz glæd ʌ bɒtʌl
Alignment
Alignment output: {(b, b), (æ, ɛ), (l, l), (ʌ, i), (n, -), (s, z), (k, g), (l, l),(æ, æ), (m, -), (p, -), (ʌ, -), (n, -), (d, d), (-, ʌ), (b, b), (ɒ, ɒ), (t, t), (ʌ, ʌ), (l, l)}
Result:
Insertion: 1 (-, ʌ)
Phoneme Deletion: 5 (n, -), (m, -), (p, -), (ʌ, -), (n, -)
Phoneme Substitution: 4 (æ, ɛ), (ʌ, i), (s, z), (k, g)
LBE ANALYSIS
Target stress pattern: SW S W SW
Transcript stress pattern: X X X X XX
Target lexical boundaries: (before 3rd syllable, before 4th syllable, before 5th syllable)
Transcript lexical boundaries: (before the 2nd syllable, before the 3rd syllable, before the 4th syllable, before the 5th syllable)
Result:
There is 1 insertion error before the 2nd syllable, and the 2nd syllable in target is weak.
Therefore, there is 1 insertion before weak (IW) LBE.

Figure 4.1: An Example of Automated Transcript Analysis.

Phoneme errors include phoneme insertion, deletion, and substitution errors. An example of the phoneme error analysis is shown in the left box of Figure 4.1. To calculate phoneme errors, we aligned the target and transcript using algorithms for phonetic sequence alignment that considers the articulation similarity between phonemes (Kondrak, 2003). We used the CMUdict to generate phonetic sequences from each word. The output of the alignment algorithm is a list of aligned phoneme pairs. There are four types of pairs from the output, and we define them as follows: 1) two identical phonemes are aligned to each other, e.g., (b,b), which is a correct transcription; 2) two distinct phonemes are aligned to each other, e.g., (æ, ɛ), which is counted as a substitution error; 3) the target phoneme cannot be aligned to any transcribed phoneme, e.g., (n,-), which is counted as a deletion error; 4) the transcribed phoneme cannot be aligned to any target phoneme, e.g., (-,ʌ), which is counted as an insertion error. The number of errors were then normalized by the total number of phonemes in the target phrase.

LBEs include 4 subtypes: 1) lexical boundary insertion error before a strong syllable (IS); 2) lexical boundary insertion error before a weak syllable (IW); 3) lexical boundary deletion error before a strong syllable (DS); 4) lexical boundary deletion error before a weak syllable (DW). In the example shown in Figure 4.1, instead of perceiving the first word in the target phrase as “balance”, the participant perceived it as “bell is”. Thus, a lexical boundary was wrongly inserted before the unstressed second syllable of balance, resulting in an IW error.

To automate the calculation of LBEs, we developed the algorithms as follows. For LBE analysis we only need to know the stress pattern of the target phrase. We first used a placeholder X to represent a syllable, and a space to represent the lexical boundary. (See the right box of Figure 4.1) By comparing the location of the lexical boundaries with the target stress pattern, we counted the number of the four possible LBEs automatically. Taking the transcript in Figure 4.1 as an example, the lexical boundary pattern of the target phrase is [before the 3rd syllable, before the 4th syllable], while the pattern of the transcript is [before the 2nd syllable, before the 3rd syllable, before the 4th syllable]. Therefore, there is an insertion error before the 3rd syllable, and the 3rd syllable in the target phrase is an unstressed syllable. Thus, the

LBE is IW. There were no other types of LBEs for this example. It should also be noted that at this stage of development, the algorithm performs best on transcripts which match the target in terms of syllable numbers, therefore all analyses included are comprised of transcript and targets comprised of 6 syllables (exclusions of more or fewer syllables account for approximately 15.4% of the data).

For each speaker, we calculated the statistics of the metrics as follows: 1) phoneme errors were calculated as the total number of errors normalized by the total number of phonemes in the target phrases; 3) LBEs were calculated as the number of each error type normalized by the number of the corresponding error opportunities in the target phrases. Taking the sample in Figure 4.1 as an example, the error opportunity of IS, IW, DS and DW was 0, 2, 2, 1, respectively. Across each entire corpus of phrases transcribed by each listener, the opportunities to produce the four categories of errors were roughly equivalent.

To formulate the MIP metrics, suppose a transcript t_s comes from a speaker s, where $t = 1, 2, ..., T, s = 1, 2, ..., S$, and T is the total number of the collected transcripts for speaker s, and S is the number of speakers in the dataset, which is 73 here. We count the number of the insertion, deletion, substitution phoneme errors and denote them as $ins(t_s)$, $del(t_s)$, $sub(t_s)$. The total number of phonemes in t_s is denoted as $N(t_s)$. Therefore, for speaker s, the phoneme errors can be calculated as Equation 4.1, 4.2, and 4.3.

$$INS(s) = \frac{\sum_{t=1}^{T} ins(t_s))}{\sum_{t=1}^{T} N(t_s))} \tag{4.1}$$

$$DEL(s) = \frac{\sum_{t=1}^{T} del(t_s))}{\sum_{t=1}^{T} N(t_s))} \tag{4.2}$$

$$SUB(s) = \frac{\sum_{t=1}^{T} sub(t_s))}{\sum_{t=1}^{T} N(t_s))} \tag{4.3}$$

For LBEs, we count the number of IS and DW errors and denote them as $is(t_s)$ and $dw(t_s)$. The opportunities of IS and DW errors in a target phrase of t_s is denoted as $Ois(t_s)$ and $Odw(t_s)$. Therefore, the LBEs can be calculated as Equation 4.4 and 4.5 The complete MIP metrics for speaker s is represented as $MIP(s) = [INS(s), DEL(s), SUB(s), IS(s), DW(s)]$.

$$IS(s) = \frac{\sum_{t=1}^{T} is(t_s))}{\sum_{t=1}^{T} Ois(t_s))} \tag{4.4}$$

$$DW(s) = \frac{\sum_{t=1}^{T} dw(t_s))}{\sum_{t=1}^{T} Odw(t_s))} \tag{4.5}$$

4.2.4 Validity of the Automated Transcript Analysis

To assess the validity of the algorithms for transcript analysis, the estimated metrics were compared to the results of the gold-standard manual scoring. Two research assistants enrolled in the speech language pathologist master program of ASU were trained by an LBE analysis expert (AL) to independently analyze phoneme errors and LBEs of a randomly selected subset of 40 samples from the collected data. They were given 40 pairs of target phrase and the corresponding listener transcription. The target-transcript phoneme alignment generated by the alignment tool were also provided for their reference (the estimated LBEs were not provided). They were trained to transform the text to phoneme sequences using CMUdict and align the target and transcript phonemes based on their knowledge. The research assistants were also instructed to manually code each lexical boundary error as occurring before strong or weak syllables (IS, IW, DS, DW).

The results of comparing the calculations between the two research assistants and between each research assistant and the algorithm are shown in Table 4.2. We use the Pearson correlation and mean absolute error (MAE) to evaluate the reliability of the algorithm relative to the reliability of the two raters.

Table 4.2: Validity Evaluation Results of the Automated Transcript Analysis.

		Correlation	**MAE**
Inter-rater	Phoneme errors	0.97	0.15
	LBEs	0.59	0.12
Algorithm-Rater1	Phoneme errors	0.89	0.38
	LBEs	0.90	0.03
Algorithm-Rater2	Phoneme errors	0.90	0.34
	LBEs	0.69	0.09

For phoneme error analysis, the algorithm is strongly correlated with the results provided by the individual research assistants, but the MAE is much larger. This is because the raters tend to use alignment strategies that differ slightly from the alignment algorithm. One of the common disagreements between manual coding and the algorithm was the alignment of diphthongs. In some cases, the alignment algorithm treats a diphthong as one vowel and sometimes as two. For example, for "beside" in the target phrase and "they say" in the transcript, the output of the alignment algorithm is (b, ð), (i, ei), (s,s), (a, e), (i, i), (d, -), where the first 'ei' was not separated, but the second one was separated into two monophthongs. Another type of disagreement came from the different alignment decisions made by the manual coders when compared to the algorithm. For example, when the target was "used" and the transcript was "good", the research assistants aligned the first phoneme 'j' with ‘g’ (1 substitution error), while the algorithm did not align them together based on their phonological distance (1 deletion and 1 insertion error).

In contrast to the high correspondence between the coding of the two research assistants for phoneme errors, the LBE analysis results yielded low inter-rater reliability. This is not unexpected because coding consistency is a function of experience, which is why it is common protocol in LBE studies to include multiple coders, including at least one with coding expertise to resolve discrepancies among coders (Liss et al., 1998). Also, as expected, the algorithm achieved higher correlation coefficients and lower MAEs with each rater. The algorithm was more stable because decision rules are clearly defined instead of relying on unstable internal rules that different coders likely have. As per manual coding protocol, the teams expert coder (AL) coded the LBEs on the same set as those coded by the research assistants and calculated the correlation and MAE between the algorithm and the experts codings. This achieved 1.0 correlation coefficient and 0.0 MAE on 34 of the selected phrases. For the other 6 phrase transcriptions, the algorithm could not analyze due to the different number of syllables in the transcript and target. However, the expert could code LBEs for them because she was not constrained by the requirement of a 6-syllable transcription to make coding decisions.

4.2.5 Regression Analysis with Perceptual Ratings Related to Intelligibility

To examine the ability of the estimated metrics to predict perceptual intelligibility, we performed a linear regression analysis using the statistics of the estimated metrics as independent variables. For the dependent variable, we used three perceptual ratings and word accuracy, which is a traditional intelligibility measure derived from listener transcripts. We obtained perceptual scores of the 73 dysarthric speakers from 15 master students enrolled in the SLP program of ASU in a previous study (Tu, Berisha, & Liss, 2017). They were instructed to listen to five sentences from each speaker and provide ratings for severity, articulatory precision, and prosody on a 1-7 scale (typical to severely atypical). Their ratings were integrated using the evaluator weighted estimator (EWE). The word accuracy was calculated as the number of correctly transcribed words over the total number of words in the target phrase. Mean value was calculated for each speaker. In this study, we examine the relationship between the proposed metrics and the three perceptual dimensions along with word accuracy. For the metrics, since the four LBEs were strongly correlated, we only used IS and DW in the regression analysis, which are the theoretically most commonly produced insertion and deletion error types in English (Cutler & Carter, 1987). Because different speakers have different number of transcripts, we normalized IS and DW errors by their corresponding opportunities in the target phrases. Statistical Package for the Social Sciences (SPSS) was used for this analysis.

Table 4.3: The Estimated Linear Regression Model for Predicting Severity.

Coefficients					
Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std. Error**	**Beta**		
(Constant)	2.367	.194		12.177	.000
Phoneme insertion	-11.852	11.939	-.440	-.993	.324
Phoneme deletion	-.975	4.684	-.059	-.208	.836
Phoneme substitution	12.734	7.305	.679	1.743	.086
IS	5.957	2.557	.714	2.330	.023
DW	-4.970	3.935	-.139	-1.263	.211

$R^2 = 0.667, p < 0.001$

Table 4.4: The Estimated Linear Regression Model for Predicting Articulatory Precision.

Coefficients					
Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std. Error**	**Beta**		
(Constant)	1.741	.175		9.950	.000
Phoneme insertion	-16.044	10.747	-.566	-1.493	.140
Phoneme deletion	1.253	4.216	.072	.297	.767
Phoneme substitution	17.343	6.576	.879	2.637	.010
IS	4.847	2.302	.552	2.106	.039
DW	-4.211	3.542	-.112	-1.189	.239

$R^2 = 0.757, p < 0.001$

4.3 Results

Table 4.3 4.4 4.5 4.6 show the coefficients of the metrics when predicting the four dependent variables respectively. From the results, we can see that all the models fit the data well with significance level $p < 0.001$. The R^2 for predicting severity, articulatory precision, prosody, and word accuracy are 0.667, 0.757, 0.620, and 0.973, respectively. It indicates that the metrics are reliable features to predict perceptual ratings and the traditional intelligibility measure. From the standardized coefficients, we can see how changes in the proposed metrics accounted for changes in each response variable. Taking severity as an example, one deviation in phoneme substitution errors accounted for a change of 0.679 in severity (on a 7-point scale), while IS errors accounted for 0.714.

The importance of the metrics was different when predicting different perceptual ratings and word accuracy. The most significant predictor of articulatory precision was phoneme substitution errors, while IS errors emerged as the most significant pre-

Table 4.5: The Estimated Linear Regression Model for Predicting Prosody.

Coefficients					
Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std. Error**	**Beta**		
(Constant)	2.379	.193		12.296	.000
Phoneme insertion	-18.414	11.884	-.735	-1.549	.126
Phoneme deletion	.229	4.662	.015	.049	.961
Phoneme substitution	13.238	7.272	.758	1.820	.073
IS	6.290	2.545	.809	2.471	.016
DW	-4.419	3.917	-.132	-1.128	.263

$R^2 = 0.620, p < 0.001$

dictor of prosody and severity. Apart from phoneme substitution errors and IS errors, the other metrics are less prominent ($p > 0.05$) in predicting the three perceptual ratings. When predicting word accuracy, all metrics except for phoneme insertion errors emerged as significant predictors ($p < 0.05$). If we consider word accuracy as an overall intelligibility score, and integrate the three phoneme errors as a single articulation measure and two LBEs as a single prosodic measure, we are able to tell the relative importance of articulation and prosody to intelligibility. To do that, in Table 4.6, we averaged the absolute values of the standardized coefficients of three phoneme errors and two LBEs respectively and got 0.346 for articulation and 0.174 for prosody. This result coincides with a previous study by De Bodt et al. (2002) where intelligibility was represented as a linear combination of multiple dimensions, and the relative importance of articulation to prosody is also nearly 2:1 (0.66 articulation and 0.3139 prosody).

Table 4.6: The Estimated Linear Regression Model for Predicting Word Accuracy.

Coefficients					
Model	**Unstandardized Coefficients**		**Standardized Coefficients**	**t**	**Sig.**
	B	**Std. Error**	**Beta**		
(Constant)	.898	.008		113.935	.000
Phoneme insertion	.681	.484	.177	1.406	.164
Phoneme deletion	-.403	.190	-.171	-2.120	.038
Phoneme substitution	-1.853	.296	-.690	-6.254	.000
IS	-.254	.104	-.212	-2.449	.017
DW	-.699	.160	-.136	-4.379	.000

$R^2 = 0.973, p < 0.001$

4.4 Discussion

The automation of scoring to capture the various contributors to speech intelligibility should have dramatic implications for both clinical practice in speech-language pathology and for moving forward communication sciences. However, there exists a natural tension in this goal because human perception is the final arbiter of speech goodness. Automated algorithms are only useful to the extent they can be designed to extract perceptually-meaningful aspects from the speech signal. In this report, we presented a method for achieving this goal by focusing on measures that tap two fundamental components of speech intelligibility–the listeners ability to identify constituent phonemes and to segment the acoustic stream at its word boundaries. While the automated phoneme measures were highly correlated with the manually coded measures, human coders were more flexibly able to deal with coding of diphthongs and unexpected transcription errors. The automated lexical boundary error coding was superior to trained human coders in its consistent application of operational definitions. This level of stability matched the coding of an expert LBE coder, but unlike the expert, the algorithm could not navigate coding of transcribed phrases that exceeded or did not include the target six syllables. Within these parameters, the automated coding enjoyed a high level of success.

The algorithms were also largely successful in generating perceptually- and clinically-meaningful data in their correspondence with perceptual ratings for severity, articulation and prosody. The automated measures that accounted for the most variability in articulatory precision were phoneme substitution errors. The automated measures that accounted for most of the variability in prosody were IS errors. These patterns make intuitive sense because phoneme errors are mainly due to the imprecise articulation, while LBEs reveal the lack of prosodic control. For severity, IS accounted for most of the variability among the automated measures. It can be interpreted as that variations in phoneme errors did not necessarily result in severity variation. However, the judgement of how severe a speaker was depended on whether this speaker was able to use syllabic contrasts to help listeners make right decisions on lexical boundaries. It coincides with the underlying theory that humans reply on lexical segmentation cues to understand speech when low-level phonemic information is degraded (Liss et al., 1998). When lexical segmentation cues were also degraded, it would be an extreme challenge for the listener to understand the speech.

The overarching goal of this research is to deliver objective intelligibility assessments that are reliable and clinically meaningful. By scoring listener transcripts from segmental (phoneme errors) and suprasegmental (LBEs) levels, clinicians could have a view of their patients speech from the listeners perspectives and conduct clinical interventions by considering listener processing strategies to improve intelligibility.

4.5 Limitations

The limitations of the current study and the projected improvements for future research is as follows. Due to the large range of severity levels of the speakers and the unsupervised transcript collection through MTurk, 48.3% listener transcripts were discarded during the analysis. In the future, listener selection will be refined to include a pre-screening survey to ensure listeners which meet our criteria. Additionally, the LBE analysis algorithm could only process transcripts with six syllables, while humans are capable of coding transcripts with fewer or more syllables than the target phrase at most times. In follow up analyses, we found that a large portion (approximately 39.4%) of the excluded transcripts only missed the first unstressed syllable. For those cases, we plan to add an extract placeholder ('X') to the beginning of the transcript and analyze them in the same way as we did for the six-syllable transcripts. Finally, the proposed metrics are not able to predict other variables which underly intelligibility degradations, such as vocal quality and nasality. For that, we will conduct acoustic analysis on the speech signals to extract relevant features and combine them with the current metrics to form a more comprehensive objective assessment of speech intelligibility.

Chapter 5

INTERPRETABLE AUTOMATED ACOUSTIC ANALYSIS

5.1 Introduction

In the study of motor speech disorders, acoustic analysis is a useful tool to quantify changes in speech signal. As we presented in Chapter 2, commonly used acoustic measures include VSA, segment durations, formants transitions, etc. We noticed that acoustic analysis was usually conducted by researchers for testing a specific hypothesis, but not often in clinical applications. One of the substantial reasons is that traditional acoustic measures usually require human labors, such as segmenting and labeling, which make it time-consuming and therefore, not flexible to be used frequently or on a large scale. However, we believe that if they are easy to access, acoustic measures can be appreciated in clinical practice because they directly measure changes in speech, which reveals the actual speech characteristics objectively without involving the auditory-perception processes. With the aim of making acoustic analysis accessible and meaningful in the clinical practice, we developed a group of acoustic measures which possess the following three characteristics: comprehensive, interpretable, and automated.

First, the acoustic measures analyze speech from different perceptual aspects and at different spectral-temporal scales, including vowel/consonant/syllable articulation, speaking rate and rhythms, prosody variation, syllable contrasts, and voicing qualities. Second, the acoustic features we employed were all clinically interpretable. Although high dimensional speech engineering features, such as the mel-frequency cepstral coefficients (MFCC) were widely used in speech signal processing, they were not included in our study due to their lack of interpretability. Besides their calculation process, in this chapter, we will provide interpretations for each acoustic feature with examples and illustrations. Third, all of the features can be extracted from the raw speech signal automatically using the computer-based algorithms.

To be specific, we developed 36-dimension acoustic measures as shown in Table 5.1. In the articulation category, we measure vowel distinctiveness by using an automated VSA analysis algorithm. In order to also cover consonant pronunciations, we developed a novel measure, articulation entropy, by measuring the overall phonemic inventory of a speaker. Besides these two long-term features, at segmental level, we employed the acoustic model of a pre-trained ASR system and calculated the goodness of pronunciation (GOP) of phonemes. For measuring prosody and rhythm, speaking rate and pitch variations were calculated first. Low-rate amplitude fluctuations were measured using six-dimension envelope modulation spectrum (EMS) features, which quantified the fluctuations of speech temporal envelopes. Moreover, the degree of syllable contrasts was measured by the ratio of the duration and intensity of the stressed and unstressed syllables. For characterizing voice quality, we measured the period and amplitude disturbances using jitter and shimmer features. Moreover, the general quality of the speech was measured with the harmonic-to-noise ratio (HNR) and the fraction of unvoiced frames. Section 5.2 introduces the calculation process of the above acoustic features and provides examples to interpret them.

To examine the validity of the developed acoustic features, we investigated their relationship with perceptual ratings. Instead of using a traditional multiple regression model, we employed the multi-task learning technique. It helped us answer what features were important to all perceptual dimensions, and what were important to a specific dimension. Section 5.3 introduces the multi-task learning technique and the model we selected for our study. Section 5.4 describes the experiments and the results.

Table 5.1: The Comprehensive and Interpretable Automated Acoustic Measures.

Perceptual category	**Acoustic feature**	**Dimension**	**Interpretation**
Articulation	VSA	1	Vowel distinctiveness
	Articulation entropy	1	Phonemic inventory
	GOP	9	Goodness of pronunciation
Prosody	EMS	6	Fluctuations in envelop modulation
	Speaking rate	1	Speed of speech
	F0 variation	1	Pitch variation
	Syllable contrast	2	Stress-unstressed syllable contrast
Vocal quality	Voice breaks	3	Unvoiced fractions
	Jitter	4	Periodicity Stability
	Shimmer	5	Amplitude Stability
	HNR	3	Harmonicity

5.2 Acoustic Measures

5.2.1 Articulation-Related Acoustic Features

Automated VSA. In pathological speech analysis, VSA is often used as a measure of articulatory precision (Roy, Nissen, Dromey, & Sapir, 2009; Turner, Tjaden, & Weismer, 1995). It measures vowel distinctiveness by calculating the area of the quadrilateral in a 2D space formed by the first and second formants of the corner vowels. It is a proxy of articulatory working space and perceptual separability between vowels. The traditional measure of VSA requires manual segmentation of individual vowels. Sandoval and colleagues (Sandoval, Berisha, Utianski, Liss, & Spanias, 2013) developed an automated algorithm to estimate VSA on a continuous speech by including all vowels instead of corner vowels only. The diagram of the VSA calculation is shown in Figure 5.1. The speech signal, which contained a variety of vowels, was first processed into consecutive frames with 20ms length. A voiced/unvoiced detection module identified the voiced segments. The first and second formants (F1/F2) were extracted automatically from each voiced frame. After removing the outliers, the remaining points were clustered into 12 groups (corresponding to the 12 English vowels) using the k-means algorithm. The convex hull spanned by the cluster centers was then determined. The area of the resulting convex polygon was calculated as the VSA. Sandoval et al. (2013) has shown that the measures estimated in this approach are strongly correlated with the traditional VSAs. Figure 5.2 shows an example of VSAs of two speakers. The left speaker (A) has a mild hyperkinetic dysarthria and his perceptual score in articulatory precision is 1 (least severe). On contrary, the right speaker (B) has a severe mixed spastic-flaccid dysarthria with a perceptual score as 7 (most severe). It is clear from the figure that the F1/F2 points of speaker A are distributed in a larger space than speaker B. It indicates that in speaker A's speech, different vowels are pronounced differently due to the effective movement of articulators. However, the vowels produced by speaker B are gathered together, which indicates that he was not able to make his articulators, such as tongues and lips, to reach the target positions for correctly and clearly producing a specific vowel.

Articulation entropy. VSA has been a prevalent metric in evaluating the articulation of the disordered speech, but it has some limitations. First of all, VSA is designed to only measure vowel pronunciation but ignoring consonants. Second, the

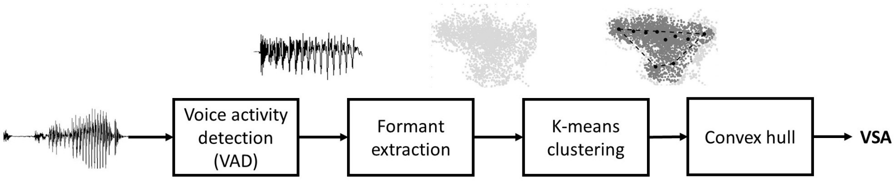


Figure 5.1: The Diagram of Automated VSA Estimation.

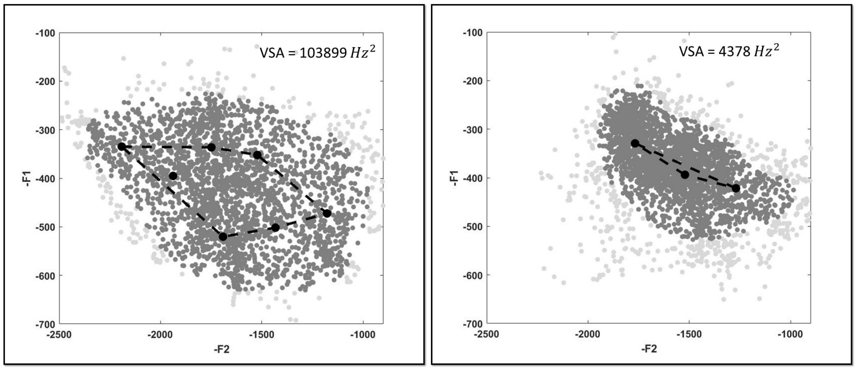


Figure 5.2: The Comparison of VSAs Between a Mild Dysarthric Speaker (left) and a Severe Dysarthric Speaker (right).

VSA calculation relies on precise formant estimation, which implies that it could be unreliable when the formant estimation is less accurate. To avoid those issues, we proposed an unsupervised metric, called articulation entropy (Jiao, Berisha, Liss, H-su, et al., 2017), that considered both vowel and consonant production and did not require formant estimation. We extented the idea of entropy in information theory to the acoustic representation of speech. We estimated the entropy of the distribution of someone's sounds and used it to characterize his/her working phonemic inventory. The framework of articulation entropy calculation is shown in Figure 5.3. A continuous speech signal with various phonemes was first pre-processed by removing the silent periods and normalized into a uniform intensity level. From each frame (20ms) of the speech, we extracted mel-filterbank features with cubic root compression (MelRoot3) (Tu, Xie, & Jiao, 2014). Features from consecutive frames within a phoneme-length window (100-200ms) were stacked into a long feature vector. The entropy of the distribution of these feature vectors was calculated using a nonparametric estimation method (Berisha, Wisler, Hero, & Spanias, 2016). The hypothesis is that when two speakers read the same content, we expect to see that the distribution of acoustic features from the speaker who had more precise articulation should have larger variation, and a larger entropy, than that of the speaker who has imprecise articulation. We can also interpret the articulation entropy using a similar concept as VSA, which is that the larger the features span in the space, the better the articulation is. For VSA, the features are F1/F2 points and the space is 2D, while for articulation entropy, the features are the stacked MelRoot3 features and the space is high-dimensional. For visualization, we reduced the high-dimensional features into 2D space using principal component analysis (PCA). Figure 5.4 shows an example of articulation entropy by comparing two speakers. The dot in the plots can be treated as a sound segmented from the speaker's speech. (The edges are used to calculate entropy.) The closer the dots are in the space, the similar those sounds are. Therefore, it is clear that sounds produced by the right speaker are more distinct to each other than those produced by the left speaker. We denote articulation entropy as artEnt in the experiment.

GOP GOP is a measure based on the log-posterior probabilities calculated from a pre-trained ASR system. It was originally developed for evaluating the degree of mispronunciation in non-native speech (Witt & Young, 2000). Here we employ it to measure the articulation precision of disordered speech. Figure 5.5 shows the diagram of the GOP calculation. First, an ASR system was trained using a large spoken speech dataset from normal speakers. The pronunciation of each phoneme was represented with computational models. After that, on the collected audio samples,

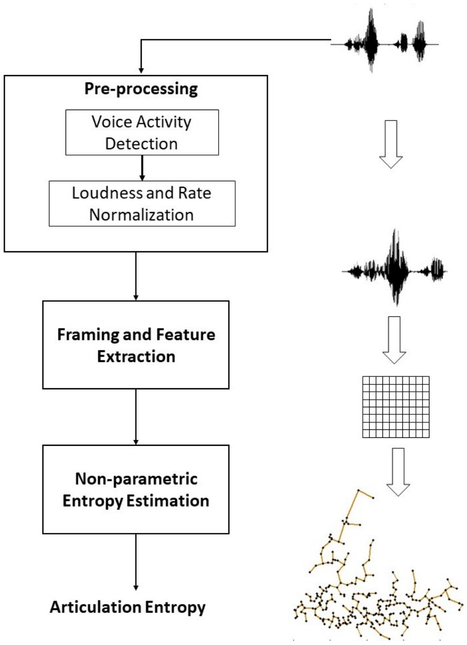

Pre-processing
Voice Activity Detection
Loudness and Rate Normalization
Framing and Feature Extraction
Non-parametric Entropy Estimation
Articulation Entropy

Figure 5.3: Procedures of the Articulation Entropy Estimation.

the acoustic models from the trained ASR system were used to align the speech signal with the phonemes in the target phrase (forced-alignment). For a target phoneme p, the GOP was calculated using Equation 5.1, where the numerator is the probability of the acoustic features belonging to the target phoneme, and the denominator is the probability of the acoustic features belonging to the other phonemes in the dictionary, and $|O^p|$ is the duration of the acoustic segment. We can interpret it as how much this segment of sound looks like the target phoneme compared to how much it looks like the other phonemes. When the speaker produces the target phoneme correctly, we would expect to see that the posterior of the target phoneme (numerator) is high, and the posterior of the other phonemes (denominator) is low, therefore, the GOP score is high. In the experiment, for each speaker, we calculated the minimum, mean and the standard deviations of GOP scores for the vowels, consonants, and syllables. We denote them as GOP_minV, GOP_minC, GOP_minS, GOP_meanV, GOP_meanC, GOP_meanS, GOP_stdV, GOP_stdC, and GOP_stdS.

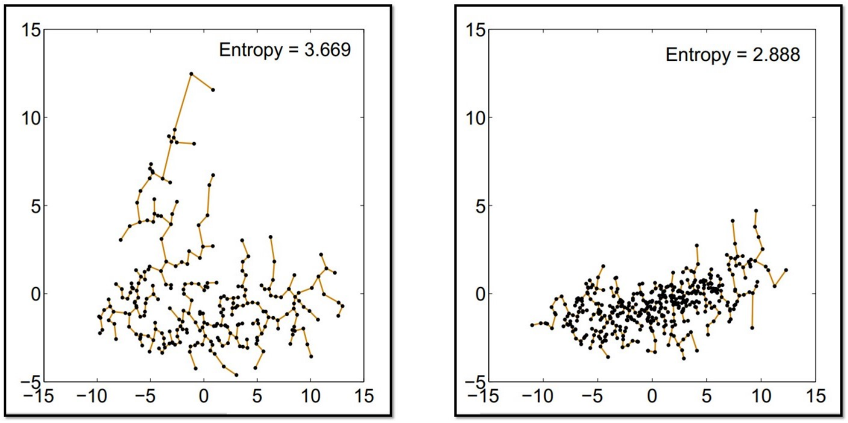


Figure 5.4: The Comparison of Articulation Entropy Between a Mild Dysarthric Speaker (left) and a Severe Dysarthric Speaker (right).

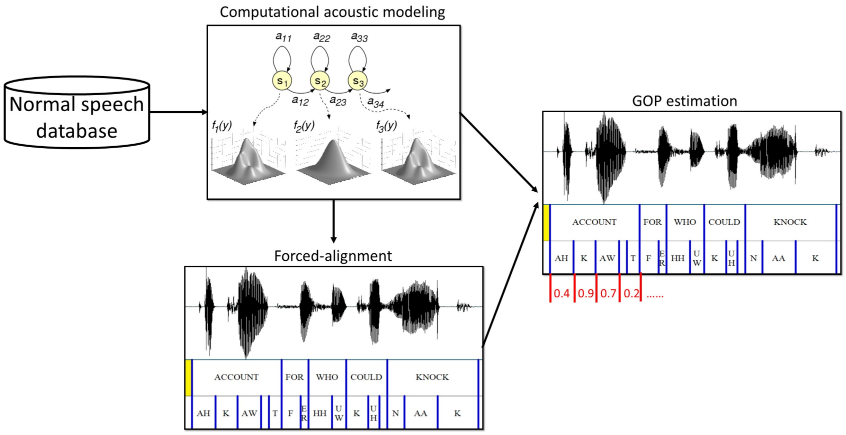

Normal speech database
Computational acoustic modeling
a_{11} a_{22} a_{33}
s_1 s_2 s_3
a_{12} a_{23} a_{34}
$f_1(y)$ $f_2(y)$ $f_3(y)$
Forced-alignment
GOP estimation
ACCOUNT FOR WHO COULD KNOCK
AH K AW T F ER HH UW K UH N AA K
0.4 0.9 0.7 0.2 ......

Figure 5.5: The Procedures of GOP Estimation.

$$GOP(p) = log\left[\frac{P(O^p|p)P(p))}{\sum_{q\epsilon Q} P(O^q|q))P(q))}\right] / |O^p| \tag{5.1}$$

5.2.2 *Prosody-Related Acoustic Features*

EMS The envelope modulation spectrum (EMS) is a spectral analysis of the slow amplitude modulations of the speech envelope. EMS has been shown to be a useful indicator of atypical rhythm patterns in pathological speech analysis (Liss, LeGendre, & Lotto, 2010). The calculation of EMS features is shown in Figure 5.6. The envelope of a continuous speech signal was first obtained by passing the half rectification of the signal through a low-pass filter with 30Hz cutoff frequency. The resulting envelope contained temporal variations in amplitude such as those that corresponded to syllables and the stressed-unstressed rhythmic patterns. EMS features were extracted from the power spectrum of the envelope signal. We extracted six variables: (1) peak frequency (EMS_pFreq) and (2) peak amplitude (EMS_pAmp) normalized by the total energy of the signal were related to the dominant modulation rates, by indexing the dominant fluctuation rate and the degree of the dominance. The third variable was (3) the normalized energy between 3Hz-6Hz (EMS_E3-6), corresponding to periods from 167ms to 333ms, which covered the majority of syllable durations in normal English (Arai & Greenberg, 1997). This frequency band was also across the 4Hz rate, which had been shown as the dominant rate in normal speech (Divenyi, Greenberg, & Meyer, 2006). The last three variables were (4) the normalized energy below 4Hz (EMS_E0-4), (5) the normalized energy above 4Hz and up to 10Hz (EMS_E4-10), and (6) the ratio of the energy below 4Hz and that within 4Hz-10Hz (EMS_ratio4) (Liss et al., 2010). They measured rhythm variations at syllable levels. Figure 5.7 shows an example of the temporal envelopes and the logarithmic power spectrum from a healthy speaker and an Ataxic speaker who has equal stressed rhythm patterns. From the temporal envelopes (middle), we can see that the left normal speaker showed prominent variations in duration and intensity, while the right Ataxic speaker showed less variations. From the power spectrum (bottom), we can see that the left normal speaker showed a peak around 4Hz and the energy distribution below and above 4Hz was significantly different, while for the right Ataxic speaker, the peak was deviated from 4Hz and the energy was uniformly distributed.

Speaking rate. Changes in speaking rate is a critical index in the evaluation of

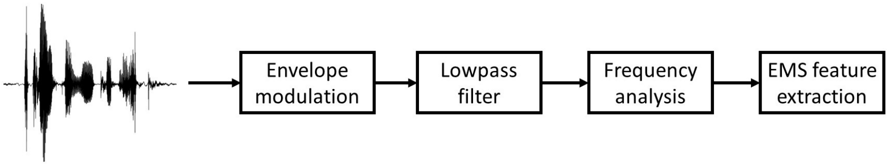


Figure 5.6: The Procedures of EMS Feature Extraction.

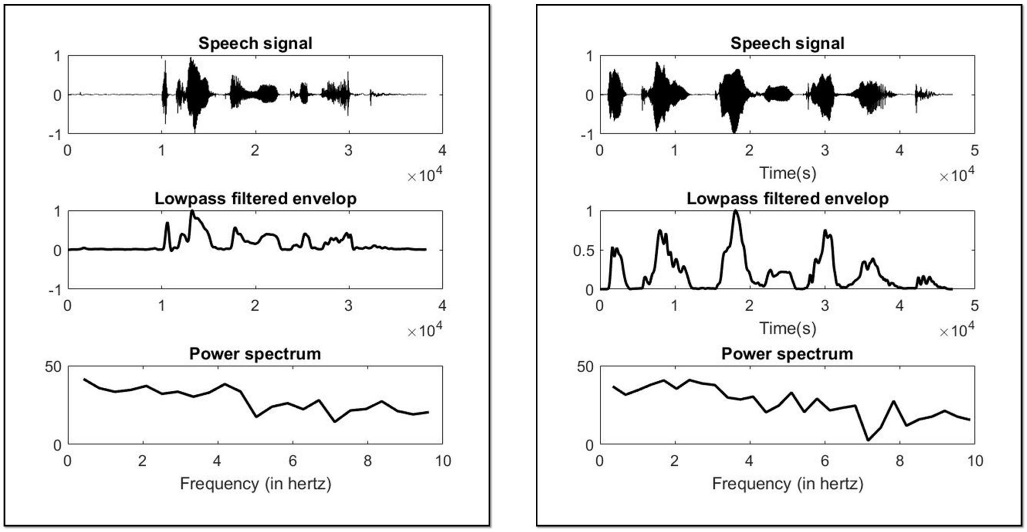


Figure 5.7: A Comparison of the Envelope Modulations Between a Normal Speaker (left) and an Ataxic Dysarthric Speaker (right).

speech disorders. In clinical practice, speaking rate is usually measured on reading speech which has a provided transcripts because the number of syllables are fixed. To measure speaking rate in spontaneous speech or estimate speaking rate variations during speech, we need to rely on computational algorithms. Traditional speaking rate estimation algorithms were usually based on peak detection from the amplitude modulation of the speech signal, which was heuristic and had issues when applied to disordered speech. We developed a data-driven speaking rate estimation method using machine learning (Jiao et al., 2015, 2016). A diagram of speaking rate calculation is shown in Figure 5.8. For a given speech segment with any length, it extracted acoustic features such as EMS and the statistics of MFCC. A recurrent neural network (RNN), which had been trained on a large dataset with variable speaking rates, was applied to estimate speaking rates second by second. Besides outputting an overall speaking rate of the entire speech sample, it also provided speaking rate variations over time which allowed clinicians and patients to monitor the change of speaking rate while talking. In our experiment, since our data was from reading task with a fixed number of syllables per utterance, the speaking rate was estimated by measuring the duration from speech onsets to offsets. We denote speaking rate as SR in the experiment.

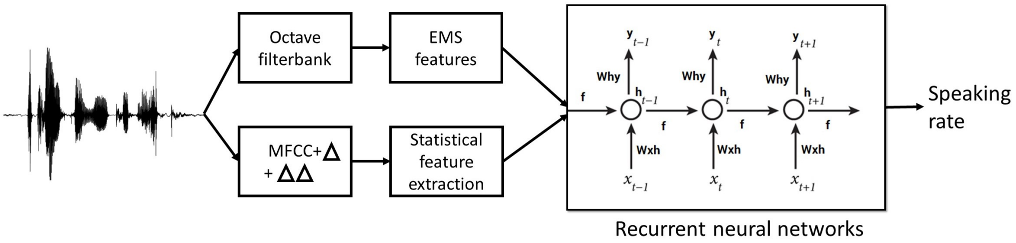


Figure 5.8: Data-Driven Based Speaking Rate Estimation Method.

F0 variations. Pitch variation is related to speech prosody and can be estimated by the standard deviation of F0 contours. Due to the unstable vocal fold vibrations of dysarthric speakers, traditional pitch estimation methods are usually unreliable. In this study, we used an ensemble method by combining three state-of-the-art pitch estimation methods (Camacho & Harris, 2008; Kasi & Zahorian, 2002; Tan & Alwan, 2013). F0 contours were extracted using each individual method, and we only kept and averaged the values where there were agreements (within 10 Hz differences) among the three methods (Hsu et al., 2017). Speech from people with monopitch would have a smaller F0 variation than normal speech. We denote F0 variations as F0_var in the experiment.

Syllable contrast. Studies have shown that listeners rely on syllable contrast to identify word boundaries when phonemic information is degraded in dysarthric speech. In this study, we measure the degree of syllable contrast by the ratio of duration (syllCont_dur) and average intensity (syllCont_int) of stressed and unstressed syllables. To obtain the time boundaries of stressed and unstressed syllable nuclei (vowels), we made use of the forced-alignment output from the GOP extraction steps.

5.2.3 Voice Quality-Related Acoustic Features

In the study of speech disorders, voice quality is usually estimated on speech with a sustained vowel. In our study, we used the forced-alignment method introduced in the GOP features to find the time boundaries for all phonemes. We removed the consonants and concatenated all vowels for each speaker. The following measures were then calculated on the audio signals with the concatenated vowels using the voice report function in Praat (Boersma, 2006).

Voice breaks. People with normal voices are able to maintain phonation while pronouncing vowels. Due to the defected motor control of vocal folds, pathological voices tend to have more unvoiced frames and voice breaks. Therefore, we measured it by calculating three variables: the fraction of locally unvoiced frames (voicing_uv),the number of voice breaks (voicing_brks), and the degree of voice breaks (voicing_dbrks). We expect that people with dysphonia have more voice breaks than those who do not.

Jitter and shimmer. Jitter and shimmer are two common acoustic features measuring the instability of laryngeal controls. Jitter is defined as the frequency variation from cycle to cycle in the sound wave (Zwetsch, Fagundes, Russomano, & Scolari, 2006), which is mainly due to the lack of vocal cord control. Shimmer is the variation in amplitude, which is caused by the reduction of glottal resistance and is correlated with noise emission and breathiness (Teixeira, Oliveira, & Lopes, 2013). The features related to jitter include local jitter (jitter_abs), local jitter normalized by the average period (jitter_norm), the relative average perturbation of jitters (jitter_rap), the five-point period perturbation quotient (jitter_ppq5), local shimmer (shimmer_local), local shimmer in dB (shimmer_localdB), the three-point, five-point and 11-point amplitude perturbation quotient (shimmer_apq3, shimmer_apq5, and shimmer_apq11).

HNR. The harmonic-to-noise ratio (HNR) is the ratio between periodic and non-periodic components presented in the speech signal (Murphy & Akande, 2005). It is related to the ability of the speaker to coordinate source and filter acoustics. The periodic components arise from the vibration of the vocal cords, and the non-periodic components come from the glottal noise. A high HNR value is associated with sonorant and harmonic voice, while a low HNR indicates an asthenic voice and dysphonia (Teixeira et al., 2013). HNR features include the HNR calculated by autocorrelation (HNR_auto), the HNR calculate by cross-correlation or the absolute HNR (HNR_abs), and the HNR in dB (HNR_dB).

5.3 Multi-Task Learning

Regression analysis is widely used in studying how the variations of a set of predictors impact the changes in the response variables. Linear regression is a simple, powerful and interpretable model to help understand this question. However, a traditional linear regression model assumes the response variables are independent to each other and estimates coefficients without considering the possible relationship between the tasks. Multi-task learning (MTL) is a method that assumes the learning of a desired target may benefit from the learning of several relevant targets so that they can be jointly trained. When we can identify multiple relevant targets in a study which may or may not have a common set of features, training the models simultaneously may help us understand the relationship of the predictors with the target variables as a whole. Moreover, it has been shown that MTL is potential to improve the learning performance than the learning on individual tasks separately. MTL has been applied in many different fields, such as natural language processing (Collobert & Weston, 2008), speech recognition (Deng, Hinton, & Kingsbury, 2013), computer visions (Girshick, 2015), etc.

In our study, the response variables are five perceptual ratings, which are severity, nasality, vocal quality, articulatory precision, and prosody. It is obvious that these dimensions are related to each other. For example, when someone is severe, it is expected that his or her articulation and prosody are both affected. Depending on the type of neurological disorders the speaker has, degradations in nasality and vocal quality are also likely to appear. Therefore, it is appropriate to use MTL in the current study.

MTL methods are different based on the assumptions of the relatedness of the tasks. For example, when we assume all tasks are related, we can consider using regularized MTL (Evgeniou & Pontil, 2004), joint feature learning (J. Liu, Ji, & Ye, 2009; Obozinski, Taskar, & Jordan, 2006), low rank MTL (Ji & Ye, 2009), alternating structure optimization (ASO) (Ando & Zhang, 2005) and so on. When the tasks are assumed to distribute in a graph or tree structure, clustered MTL (Jacob, Vert, & Bach, 2009), network MTL (Yan, Ricci, Subramanian, Lanz, & Sebe, 2013) and tree MTL (S. Kim & Xing, 2010) can be considered. Some other methods, such as deep neural networks with shared hidden layers (Ruder, 2017), make no assumption of the tasks, but lack interpretability. In our study, we assume all tasks are latently related and we want to seek an interpretable model. Thus, we restricted our search to the methods based on multiple linear regression.

A standard multiple linear regression model is shown in Equation 5.2, where $X^k \in \mathbb{R}^{n \times p}$ and $\vec{y^k} \in \mathbb{R}^n$ are the feature matrix and the response variable for the k-th task. The learning of the model is to estimate an optimal coefficient vector $\vec{\theta^k} \in \mathbb{R}^p$ to fit the given data as well as possible under a certain optimization criterion, such as the least squares.

$$\vec{y^k} = X^k\vec{\theta^k}, k = 1, ..., r \tag{5.2}$$

Multi-task learning is when we have $r > 1$ response variables. In the setting of multiple regressions, the r tasks are usually assumed to be "simultaneously sparse" (Tropp, Gilbert, & Strauss, 2005), where the number of relevant features for each task is small, and there is a large overlap of these relevant features across different tasks. Applying it to our study, we developed multiple measures in each perceptual categories, such as the VSA and articulatory entropy for measuring articulation, and the EMS and speaking rate for measuring prosody. However, we are uncertain their importance in predicting perceptual ratings. We expect to see that some features are identified as more important than others ("sparsity"). Moreover, we also expected to see the important features across different tasks are largely overlapped because for motor speech disorders the symptoms in speech usually appear from different aspects as shown in Table 3.2 ("simultaneous"). Therefore, the "simultaneously sparse" assumption meets our demand.

Under this assumption, block sparsity or group sparsity was proposed in the early literatures (J. Liu et al., 2009; Obozinski et al., 2006). Here we represent multiple regression in a format of matrix in Equation 5.3 where $Y \in \mathbb{R}^{n\times r}$, $X \in \mathbb{R}^{n\times p}$, and $\theta \in \mathbb{R}^{p\times r}$. Thus, in the coefficient matrix θ , each column corresponds to a task, and each row to a feature dimension. Block sparsity is the structure of the coefficient matrix where each row is either all zero or mostly non-zero, and the number of non-zero rows is small (See Figure 5.1). To encourage such structure during training, l_1/l_q ($q > 1$) norm regularization is usually used, such as the l_1/l_∞ norm (Negahban & Wainwright, 2008; Turlach, Venables, & Wright, 2005) and the l_1/l_2 norm (Lounici, Pontil, Tsybakov, & van de Geer, 2009; Obozinski, Wainwright, & Jordan, 2011).

$$Y = X\theta \tag{5.3}$$

The problem of the block sparsity is that the sparsity is strictly shared, which means that a feature is either important to all tasks or not important to any. It is not realistic because we expect that each task depends on features specific to itself although a common set of features can be shared. Moreover, studies have shown that the l_1/l_q norms can easily result in the non-sparse rows in the coefficient matrix taking nearly identical values (Negahban & Wainwright, 2008; Obozinski et al., 2011). This is an even more strict assumption that not only do the features have to be exactly the same, but also their importance in prediction.

To solve this problem and make the model more realistic, Jaladi and colleagues proposed a dirty model (Jalali, Sanghavi, Ruan, & Ravikumar, 2010), where it decomposed the coefficient matrix into a group sparse matrix (Q in Figure 5.9) corresponding to the shared features, and an element sparse matrix (P in Figure 5.9) corresponding to the task-specific features. It integrated both block sparsity and l1 regularization, and had been shown to outperform each of them. The algorithm of the dirty model optimization is shown in Equation 5.4.

$$min_{P,Q} \left\| Y - X(P+Q) \right\|_F^2 + \lambda_1 \left\| P \right\|_{1,q} + \lambda_2 \left\| Q \right\|_1 \tag{5.4}$$

It estimates a sum of two coefficient matrices P and Q with different regularization for each: encouraging block-structured row-sparsity in P using l_1/l_q norm and element sparsity in Q using l_1 norm.

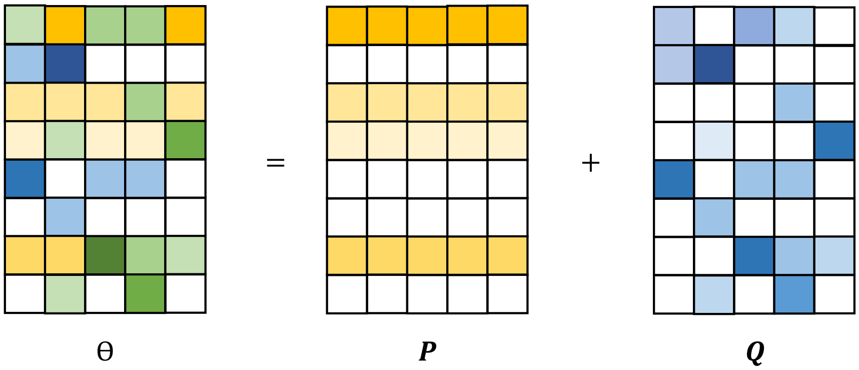


Figure 5.9: An Illustration of the Dirty Model.

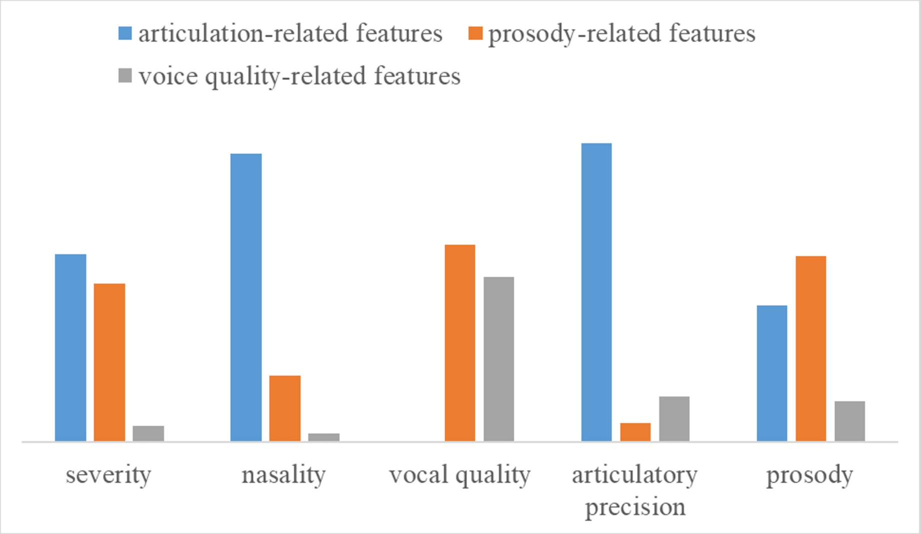


Figure 5.10: The Importance of the Acoustic Features in Predicting Different Perceptual Dimensions.

5.4 Experiments and Results

On the 73-speaker dysarthric speech dataset, we extracted the acoustic features in Table 5.1, and used them as independent variables. The 5-dimension perceptual ratings were used as dependent variables. Referring to Equation 5.4, X corresponds to the acoustic features, and Y corresponds to the perceptual ratings. The acoustic

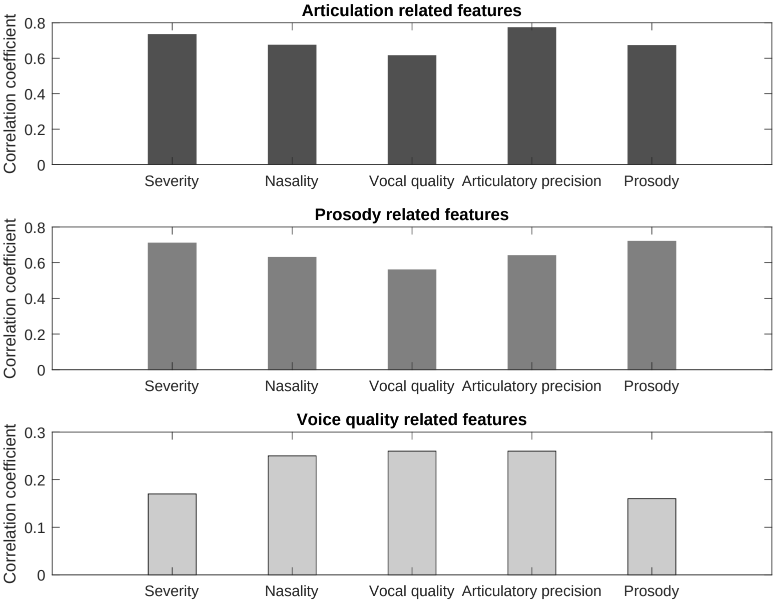


Figure 5.11: The Correlation Coefficients of the Regression Model Using a Single Set of Acoustic Features.

features are normalized by Z-score so that the estimated coefficients can be treated as importance degrees. A MATLAB based multi-task learning toolbox, MALSAR (Zhou, Chen, & Ye, 2011), was used. We applied the dirty model on the data and tuned the parameters λ_1 and λ_2 using cross validation. The group sparsity matrix P and the element sparsity matrix Q were then estimated. Table 5.2 shows the selected acoustic features with non-zero values in P and Q. We can see that in all the 15 dimensional acoustic features, 13 of them were selected as useful shared predictors across tasks. Among them, there were 6 features related to articulation, 4 features related to prosody, and 3 features related to voice quality. It is consistent with our hypothesis that all the three categories of acoustic features are important in predicating intelligibility, which can be measured as a linear combination of the multi-dimensional perceptual ratings (De Bodt, Huici, & Van De Heyning, 2002). For each individual task, different features were selected based on their relevance to each perceptual dimension.

Suppose $\vec{\theta^k} = [\theta_1, \theta_2, ..., \theta_p]$ is the estimated coefficients for the k-th task in Q. The importance of the 3 sets of acoustic features (articulation, prosody, voice quality) can be estimated using Equation 5.5, where $m = 1, 2, 3$ is the index of the acoustic feature set, N_m is the number of dimensions in the m-th feature set. Table 5.3 shows the importance values of the 3 acoustic feature sets in predicting the 5 perceptual dimensions. Based on this, we calculated their relative (normalized by their sum) importance within each task and plot it in Figure 5.10. We can see that when predicting different perceptual dimensions, the importance of the acoustic features varies. For example, articulation related features are the most important factor in predicting articulatory precision, while it is less important in predicting vocal quality and prosody. Similarly, prosody related features are more relevant to prosody ratings than the other dimensions. As for severity, which is a more general dimension, articulation and prosody related features are equally important along with a slightly less contribution from features related to voice quality. This result support our hypothesis that the acoustic features measured from different aspects of speech are most related to their corresponding perceptual dimensions.

To further check how these 3 sets of acoustic features are related to the 5 perceptual dimensions, we used each of them to predict the perceptual ratings and calculated the correlation coefficients between the predicted values and the real values. Figure 5.11 shows the result. From the figure, we can see that each individual feature set is the most correlated with its corresponding perceptual dimension: articulation features to articulatory precision, prosody features to prosody, and voice quality features to vocal quality. However, we also notice that by only using the voice quality features, the correlation coefficients (0.2) are much lower than only using the other two set of features (0.7), even when predicting vocal quality ratings. It indicates that the features we are using to measure voice quality, such as jitter, shimmer, and HNR, are not as reliable as the other features like GOP and EMS. The reason is that these features were usually measured on sustained phonation instead of continuous spoken speech as we did in this experiment. Therefore, it implies that a more robust voice quality feature needs to be developed in the future. However, the correlation coefficient when predicting perceptual voice quality was still significant ($p < 0.05$) although the number was low, which means that we could still claim that those measures are related to vocal quality.

To examine how well the acoustic features predict perceptual ratings, we used the leave-one-speaker-out cross validation. To be specific, each time during training, we left one speaker out, and used the other 72 speakers to train the model and obtained a coefficient matrix θ. At test phase, we applied each trained model on the left-out speaker. Table 5.4 shows the correlation coefficient and the mean absolute errors (MAE) between the predicted and the real perceptual ratings. As a comparison, we also trained 5 linear regression models with LASSO regularization in the same cross validation fashion, and presented the result in Table 5.4. From the result, we can see that the predicted values are strongly correlated to the perceptual ratings with a low MAE. It indicates that the developed acoustic features are capable of making reliable predictions of perceptual ratings. Compared to LASSO based single task learning, the multi-task learning models achieved a better result, which indicates that it is beneficial to jointly train these tasks than training them separately.

In addition, we used the acoustic features to predict word accuracy, which was a general measure of intelligibility, with l1 regularization in a cross-validation fashion. The predicted results achieved $R^2 = 0.72$ with a significance $p < 0.01$. This indicates that by using our development acoustic features, we are able to make reliable predictions of someone's intelligibility degree only from his/her speech signals.

$$Importance(m) = \frac{\sum_{i=1}^{N_m} |\theta_i|}{N_m} \tag{5.5}$$

Table 5.2: The Selected Features ("yes") in the Multi-Task Learning Model.

		P	**Q**				
			Severity	**Nasality**	**Vocal quality**	**Articulatory Precision**	**Prosody**
Articulation-related features	VSA	yes					
	artEnt	yes				yes	
	GOP_minV		yes				yes
	GOP_minC	yes	yes			yes	
	GOP_minS		yes				yes
	GOP_meanV	yes	yes			yes	
	GOP_meanC		yes			yes	
	GOP_meanS		yes				
	GOP_stdV		yes				
	GOP_stdC	yes	yes	yes			
	GOP_stdS	yes	yes				yes
Prosody-related features	EMS_pFreq	yes		yes		yes	
	EMS_pAmp						yes
	EMS_E3-6				yes		yes
	EMS_E0-4				yes		
	EMS_E4-10				yes		
	EMS_ratio4			yes	yes		
	SR	yes	yes		yes		yes
	F0_var	yes	yes		yes		yes
	syllCont_dur						yes
	syllCont_int	yes			yes		yes
Voice quality-related features	voicing_uv	yes					
	voicing_brks						
	voicing_dbrks						yes
	jitter_abs				yes	yes	
	jitter_norm	yes	yes		yes	yes	
	jitter_rap				yes	yes	
	jitter_ppq5				yes	yes	
	shimmer_local		yes		yes		
	shimmer_localdB			yes			
	shimmer_apq3	yes		yes	yes	yes	yes
	shimmer_apq5				yes		
	shimmer_apq11		yes				
	HNR_auto						
	HNR_abs				yes		
	HNR_dB						

Table 5.3: The Importance of the Three Acoustic Feature Sets to the Five Perceptual Dimensions.

	Severity	**Nasality**	**Vocal quality**	**Articulatory precision**	**prosody**
Articulation-related features	0.0160	0.0104	0	0.0426	0.0122
Prosody-related features	0.0134	0.0024	0.0286	0.0025	0.0167
Voice quality-related features	0.0014	0.0003	0.0240	0.0065	0.0037

Table 5.4: The Correlation Coefficients and the MAEs Between the Predicted and the Actual Perceptual Ratings for Dirty Model based MTL and Single Task Learning uisng LASSO.

Multi-task learning						
	Severity	Nasality	Vocal quality	Articulatory precision	Prosody	**Average**
Correlation Coefficient	0.807	0.749	0.669	0.820	0.724	**0.754**
MAE	0.724	0.721	0.822	0.730	0.849	**0.769**
Single task learning						
Correlation Coefficient	0.776	0.769	0.497	0.829	0.792	**0.733**
MAE	0.732	0.689	0.974	0.720	0.765	**0.776**

Chapter 6

THE RELATIONSHIP BETWEEN ACOUSTIC SIGNALS AND LISTENER ERROR PATTERNS.

6.1 Introduction

In the previous two chapters, we assessed intelligibility from the listener's perspective using the MIP metrics, and quantified information in the acoustic signals related to intelligibility from the speaker's aspect. In this chapter, we investigate the interaction between acoustics and listener error patterns. First, the correlation between acoustic features and transcript errors measured by MIP and word accuracy were studied across speakers on the 73-speaker dysarthric speech dataset. Second, we simulated speech changes within speakers by using different cues, which were slow, clear, and loud. It resulted in variations in both acoustic features and transcript errors. By investigating the relationship between them, we aim to show how changes in acoustic signals affect the ways listeners perceive and understand speech, and whether we can predict intelligibility gains and listener error variations from the changes we observed in the acoustic signal.

6.2 Data Collection

In this chapter, we used both of the datasets described in Chapter 3: the 73-speaker dysarthric speech dataset and the healthy speech dataset. Twenty healthy speakers produced each of the 40 speech samples (phrases) four times under different conditions, including habitual, slow, clear, and loud. Studies have shown that in the treatment of dysarthria, patients potentially improved their intelligibility when being cued with these prompts (K. Tjaden, Sussman, & Wilding, 2014). Although we did not aim to investigate which intervention strategy is more effective than others, these cues were able to trigger changes in speaking styles so that variations can be detected from the acoustic signals, and therefore, resulting variations in intelligibility.

To induce changes in intelligibility, we embedded the speech signal into background noise. With the same level of signal to noise ratio (SNR), we were able to compare the impact of spectral temporal variations in different conditions on intelligibility. The total number of speech samples was 3,200. We separated them into 80 batches with each containing 40 samples. In each batch, there were two samples from each of the 20 speakers, and no duplicate phrases existed in any batch.

We collected the transcripts from MTurk. For each batch, we aimed to recruit 10 listeners, but some listeners were not able to complete the tasks for unknown reasons. As a result, 675 listeners located in the US participated in the experiment. We filtered the listeners based on their answers to the questionnaire. Details about the questionnaire have been described in Chapter 3. After filtering, 24,862 transcripts qualified for our study, with 311 transcripts on average collected for each speaker in each condition.

6.3 Correlation Between Acoustic Measures and Transcript Errors Across Speakers

On the 73-speaker dysarthric speech dataset, we extracted the MIP metrics, word accuracy, and 36-dimensional acoustic features in the study of Chapter 4 and Chapter 5. Here we calculated the bivariate Pearson correlations between each of the acoustic features and the six transcript errors (five MIP metric and word accuracy). Table 6.1 shows the result with $*$ indicating a significance at $p < 0.05$, and $**$ a significance at $p < 0.01$ (highlighted in the table). From the result, we can see that the absolute values of some acoustic features did not show correlations with intelligibility level directly, such as the VSA and F0 variation. It was reasonable because these two measures could be affected by many speaker-related factors, such as age and gender (Pettinato, Tuomainen, Granlund, & Hazan, 2016). However, most of the articulation- and prosody-related features showed significant correlations with word accuracy and the MIP metrics. The GOP features appeared to be strongly correlated with all the transcript error metrics. The EMS features and the stressed-unstressed syllable duration contrast (SyllCont_dur) also moderately correlated with them. It is not surprising that the voice quality metrics did not yield a strong correlation with the MIP metrics. Several studies have shown that vocal quality has the lowest correlation with intelligibility among the various perceptual dimensions (De Bodt, Huici, & Van De Heyning, 2002). For the two types of transcript errors, the GOP features showed stronger correlations with phoneme errors and IS errors than DW errors. The EM-S features showed higher correlations with IS errors than phoneme errors, while the correlation with DW errors was not significant. This is consistent with the hypothesis that segmental phoneme errors are primarily related to articulation-related features, while suprasegmental lexical boundary errors can be affected by both articulation- and prosody-related features. From the table, we also found that the correlation of the acoustic features with the DW errors was lower than the other transcript errors. This suggests that the DW errors could be affected by a combination of multiple acoustic features instead of a single one.

6.4 The Relationship Between Variations in Acoustic Measures and Transcript Errors Within Speakers

As we discussed in the previous section, variations in the acoustic features across speakers may not necessarily correlate with the absolute intelligibility scores (word accuracy). In this section, we study whether changes in the acoustic measures within speakers could predict intelligibility gains or degradations. We triggered variations in speech signals by asking the same speaker to read the materials in different conditions (habitual, slow, clear, and loud). In this way, we could learn the particular changes in which acoustic features may have positive effects on intelligibility gains, and how the intervention strategies affect variations in individual acoustic features.

6.4.1 Variations in Transcript Errors in Different Speaking Modes

We calculated word accuracy and the MIP metrics for each speaker in the four conditions from the collected transcripts using the methods described in Chapter 4. Table 6.2 shows the average word accuracy and five MIP metrics over the 20 speakers in each speaking mode. From the table, we can see that on average, the cued speech show improvement in all metrics. The improvement of slow and clear conditions compared to habitual is nearly the same, but more improvement is observed in loud condition. One exception can be found in DW errors where slow cued speech shows more improvement. We can interpret it as follows. DW errors appear when the listeners did not detect the word boundary before an unstressed syllable. For example, it happens when a listener perceives “beside a” as “decided”. It makes sense that in slow condition, there are fewer liaisons so that listeners could detect that the unstressed syllable does not belong to the previous word. However, in the previous subsection, the correlation between speaking rate and DW errors is not significant. It indicates that although DW errors are related to speaking rate, their relationship may not be in a linear fashion.

Figure 6.1 shows the word accuracy for each speaker in different speaking conditions. We can see that the three intervention instructions are able to trigger variations in intelligibility. However, the gains vary greatly from speaker to speaker. For each speaker, we calculated the gains from habitual to the cued conditions as a percentage of increment (positive) or decrement (negative). Table 6.3 shows the mean, standard deviation, minimum and maximum values in the word accuracy gains. From the result, we can see that on average there are improvements in the cued conditions, but the variation is large. For example, in the slow condition, one speaker showed 117.54% increment compared to his/her habitual speech, while another speaker showed 22.51% decrement. Previous studies noticed that although slow and loud were two commonly used treatment strategies to help improve intelligibility, they might not work for all patients (Fletcher, McAuliffe, Lansford, Sinex, & Liss, 2017b). Here we also notice that speech produced with intervention cues has the potential for intelligibility improvement, but there are exceptions.

For the MIP, we also calculated the percentage of decrement in each error metric. Figure 6.2 shows the result. We can see that all the MIP errors decrease in the cued conditions, but the decrements are different between segmental measures and suprasegmental measures. For instance, the improvements in phoneme errors are smaller in the slow condition than the other two conditions. However, the slow cued speech showed greater improvements in lexical segmentation errors. This is because "slow" is a direct instruction on rhythm control. Comparing "clear" and "loud", the results indicate that "loud" is more effective in improving phoneme recognition accuracy than "clear" even though "clear" is more related to articulation control.

6.4.2 Variations in Acoustic Measures in Different Speaking Modes

The acoustic measures described in Chapter 5 were calculated from the acoustic signals for all speakers in the four conditions. Changes from the habitual speech to the cued speech were calculated as a percentage value similar to what we did in Section 6.4.1. To have a clearer view, we selected 15 features from the 36 dimensions as representatives. These features covered all acoustic measures and are more interpretable than the other dimensions so that they can help us explain the observations with emphasis. Variations of the 20 speakers in these dimensions are shown in Table 6.4.

For articulation-related measures, we notice that the VSA, and the GOP show significant improvements in the three cued conditions, but variations in articulation entropy are less prominent. Comparing between the three intervention strategies, loud cued speech shows the greatest changes in VSA and articulation entropy measures, while slow cued speech shows the greatest changes in GOP measures. We can interpret this finding as that clear and loud cues make the speakers exaggerate articulations which can be detected by VSA and articulation entropy. However, some of the exaggerated phonemes may sound unnatural so as to increase the ASR likelihood measured by GOP. However, the slowed speech makes the speakers have more time to focus on phoneme pronunciations, which increases the GOP values.

For prosody-related measures, we first notice that the speaking rate decreases in all three cued conditions with more decrements in the slow cued speech. The variation of pitch is also larger in the cued conditions than in habitual with more impact from loud cues. Changes in the EMS features indicate that the three cues are able to increase fluctuations in envelope modulation, but the slow cue is more efficient. From the two syllable contrast measures, we notice that only loud cue can increase the contrast ratio between stressed and unstressed syllables. It indicates that when speaking in a loud voice, speakers tend to put more energies on stressed syllables than unstressed ones. Although the slow and clear cues are also able to increase envelope fluctuations, it does not result in the speakers paying more attention on syllable contrasts.

For voice quality-related measures, we notice that in the three cued conditions, the speakers show more stable vocal control by having fewer voice breaks, less jitter and shimmer and more harmonicity. Voice breaks and jitter show the highest decrements in the loud condition. We can interpret it as that the air pressure increases greatly so as to make the vocal fold vibrate in a more stable way. Shimmer and HNR show more improvements in the slow condition. We suspect that when speaking slowly, the speakers are prolonging the syllable nuclei more than the transitions.

From Table 6.4, we can see that the standard deviations are large regarding the corresponding mean values, which indicates that different speakers may display significant difference in the changes of acoustic signals even when being cued with the same instructions. This is similar to what we noticed in the MIP metrics.

6.4.3 The Relationship Between Acoustic Variations and Transcript Error Variations

Here we investigate the relationship between acoustic variations and the variations in MIP and word accuracy. We want to answer two questions: (1) changes in which acoustic features are important in predicting intelligibility gains, and what is their relative importance; (2) can we use acoustic features to make reliable predictions of MIP and word accuracy so that we can learn how much the intelligibility (word accuracy) is expected to change given the observed acoustic variations and in what way it affects listener strategies (MIP metrics).

To answer these questions, we used multiple linear regression models to fit the data. The dependent variables were the variations of the acoustic measures in percentage from habitual to the three cued conditions. Correspondingly, the independent variables were the increment in word accuracy and the decrement in the five MIP metrics.

The group importance of the three categories of acoustic features, which were articulation-related, prosody-related, and voice quality-related, were calculated using Equation 5.5 and plotted in Figure 6.3.

Table 6.5 shows the five most important predictors and their standardized coefficients for the six response variables respectively. From the figure, we can see that articulation-related features were the most important predictors of phoneme errors. From the table, we can see that changes in GOP features account for the most variations. This is consistent with our hypothesis that articulation-related acoustic features are linked closely with segmental phoneme metrics. Among the nine dimensional GOP features, the standard deviation of consonant GOP values is the most important for predicting phoneme errors. Voice quality-related features also emerge as important predictors of phoneme error variations. It implies that when the speaker shows better voice control, the speech signal becomes more harmonic and have fewer voice breaks which helps listeners perceive phonemes more precisely and continuously.

For predicting variations in the lexical segmentation errors, prosody-related features especially the EMS features became more important compared to them in predicting phoneme errors. This is also consistent with our hypothesis that prosody-related acoustic features should have a closer relationship with lexical segmentation strategies than phoneme perception. In addition, articulation-related features, such as GOPs, also played a critical role in predicting lexical segmentation errors. It indicates that listeners rely on both phonemic and rhythmic cues for lexical segmentation. The changes in voice quality features were less important in predicting lexical segmentation errors than in predicting phoneme errors with only one shimmer feature appeared in the five most important predictors.

For predicting word accuracy, the five most important predictors include two GOP features, two EMS features, and one jitter feature, which indicates that the overall intelligibility gains is predictable by considering all articulation, prosody, and vocal quality variations in the acoustic signal, but the former two are more important than the latter one.

To examine how well the acoustic changes can predict transcript error changes, we trained a linear regression model using the dirty model based MTL in a leave-one-speaker-out fashion. Table 6.6 shows the correlation coefficients and the MAEs between the predicted and actual changes in word accuracy and MIPs. Figure 6.4 shows the scatter plots and the fitted lines. From the result, we can see that the predicted values are significantly correlated with the real values, but with a relatively large MAE. From the scatter plots, we can find some reasons for this finding. First, we only had 60 training samples (20 speakers in three cued condition), which was small for model learning. Second, the data was not uniformly distributed. The number of speakers who showed a moderate changes was larger than those who showed great or slight changes. Therefore, we noticed a larger prediction error in the region with fewer training samples. However, for the majority of the speakers, we could make reasonable predictions with a significant correlation. By having a model like this, we expect to predict intelligibility gains (e.g., word accuracy improvement) from acoustic variations. Moreover, by predicting the MIP metrics, we could also interpret such gains from the perspective of a listener when the transcripts are not available.

6.4.4 Discussion of the Number of Phrases

In this study, there are 80 phrases in the stimuli. For the dysarthric speech dataset, we collected all 80 phrases from the speakers. However, for the healthy speech dataset that was used in this chapter, only 40 phrases were collected from each speaker in each speaking mode. To examine if it is sufficient enough as 80 phrases, and if we can further reduce the number of phrases, we extracted the MIP metrics and the acoustic features by using 5, 10, 20, 40, 60 number of phrases and calculated their correlation coefficients with the measures we extracted from all 80 phrases on the dysarthric speech dataset. Figure 6.5 shows the result. From the figure, we can see that the measures extracted from reduced phrases have strong correlation with those extracted from all 80 phrases. Especially, when the number increased to 40, the benefit of including more phrases became limited. Another thing we noticed from this experiment was the difference between VSA and articulation entropy. Although they both measured the general articulation, the articulation entropy was stable (high correlation) when the number of phrases reduced. However, we were not able to calculate VSA when the number of phrases was smaller than 20, and the correlation was lower than the articulation entropy measure when using 20 to 60 phrases. This indicates that we may replace VSA with articulation entropy especially when the audio samples are limited.

Table 6.1: Bivariate Pearson Correlation Between Acoustic Features and Transcript Error Metrics.

		Word accuracy	**phoneme insertion**	**Phoneme deletion**	**Phoneme substitution**	**IS**	**DW**
Articulation-related features	**VSA**	0.174	-0.136	-0.154	-0.152	-0.089	-0.217
	artEnt	**.465****	**-.540****	**-.559****	**-.465****	**-.497****	-0.147
	GOP_minV	**.726****	**-.695****	**-.707****	**-.697****	**-.691****	**-.526****
	GOP_minC	**.869****	**-.871****	**-.902****	**-.848****	**-.880****	**-.589****
	GOP_minS	**.867****	**-.861****	**-.886****	**-.842****	**-.864****	**-.603****
	GOP_meanV	**.757****	**-.752****	**-.757****	**-.736****	**-.742****	**-.506****
	GOP_meanC	**.855****	**-.866****	**-.928****	**-.827****	**-.884****	**-.568****
	GOP_meanS	**.859****	**-.864****	**-.907****	**-.832****	**-.872****	**-.579****
	GOP_stdV	**-.695****	**.649****	**.660****	**.663****	**.645****	**.529****
	GOP_stdC	**-.870****	**.868****	**.893****	**.851****	**.879****	**.601****
	GOP_stdS	**-.860****	**.842****	**.856****	**.835****	**.842****	**.617****
Prosody-related features	**EMS_pFreq**	0.093	-0.202	-0.208	-0.114	-0.159	0.189
	EMS_pAmp	**-.480****	**.581****	**.578****	**.492****	**.600****	0.080
	EMS_E3-6	.270*	**-.368****	**-.388****	-.276*	**-.412****	0.045
	EMS_E0-4	**-.384****	**.463****	**.434****	**.415****	**.474****	0.045
	EMS_E4-10	**.384****	**-.463****	**-.434****	**-.415****	**-.475****	-0.045
	EMS_ratio4	**-.422****	**.503****	**.477****	**.451****	**.519****	0.068
	SR	.273*	**-.361****	**-.320****	**-.302****	**-.385****	0.077
	F0_var	0.057	-0.073	-0.086	-0.082	-0.071	-0.056
	syllCont_dur	**.424****	**-.430****	**-.406****	**-.443****	**-.469****	-.249*
	syllCont_int	0.008	-0.045	-0.025	-0.006	-0.090	0.037
Voice quality-related features	**voicing_uv**	-0.155	0.191	0.179	0.140	0.197	-0.100
	voicing_brks	0.069	-0.111	-0.111	-0.101	-0.088	-0.003
	voicing_dbrks	0.079	-0.102	-0.110	-0.122	-0.120	-0.024
	jitter_abs	0.187	-.233*	-0.224	-0.226	-.264*	-0.011
	jitter_norm	0.078	-0.123	-0.140	-0.112	-0.182	0.070
	jitter_rap	0.181	-0.222	-0.223	-0.221	-.272*	0.007
	jitter_ppq5	0.162	-0.215	-0.210	-0.203	-.251*	0.044
	shimmer_local	0.095	-0.112	-0.135	-0.127	-0.145	-0.016
	shimmer_localdB	0.128	-0.154	-0.168	-0.159	-0.180	-0.040
	shimmer_apq3	0.100	-0.089	-0.122	-0.128	-0.147	-0.061
	shimmer_apq5	0.111	-0.130	-0.145	-0.146	-0.164	-0.016
	shimmer_apq11	0.114	-0.163	-0.164	-0.152	-0.167	0.046
	HNR_auto	-0.148	0.164	0.166	0.178	0.170	0.037
	HNR_abs	0.144	-0.166	-0.163	-0.171	-0.162	-0.028
	HNR_dB	-0.147	0.160	0.153	0.181	0.175	0.043

Table 6.2: Word Accuracy and MIP Metrics in Different Speaking Modes.

	Habitual	**Slow**	**Clear**	**Loud**
Word accuracy	0.41	0.48	0.49	0.53
Phoneme insertion	0.11	0.09	0.09	0.07
Phoneme deletion	0.13	0.11	0.10	0.08
Phoneme substitution	0.17	0.14	0.14	0.12
IS	0.37	0.26	0.26	0.24
DW	0.05	0.02	0.03	0.03

Table 6.3: Gains in Word Accuracy from Habitual to the Cued Conditions.

	Mean (%)	**Std (%)**	**Min (%)**	**Max (%)**
From habitual to slow	22.58	34.64	-22.51	117.54
From habitual to clear	24.5	24.6	-19.69	71.39
From habitual to loud	38.05	35.45	-2.7	119.82

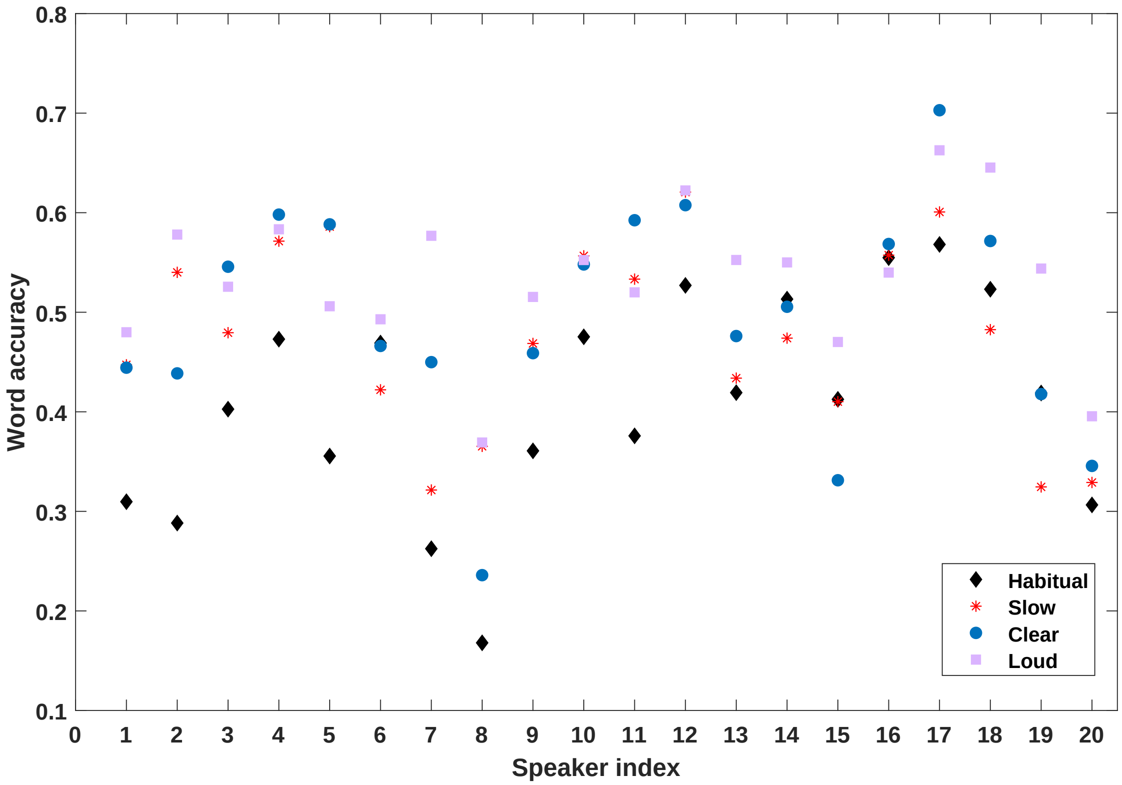


Figure 6.1: Word Accuracy of Individual Speakers in Different Speaking Modes.

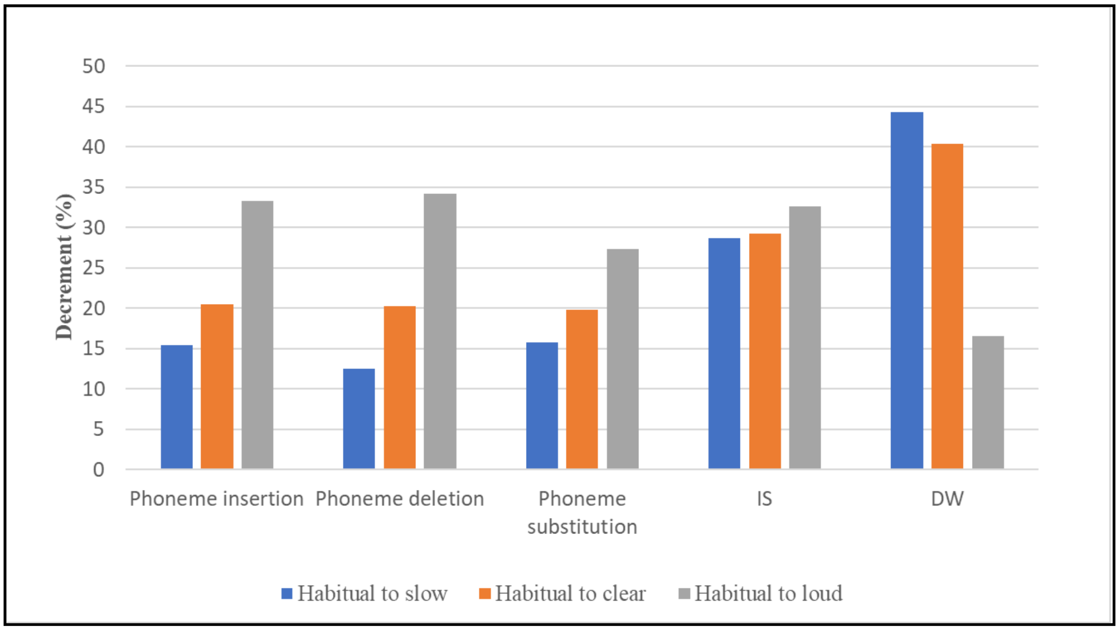


Figure 6.2: Decrements in the MIP Error Metrics from Habitual to the Cued Speech.

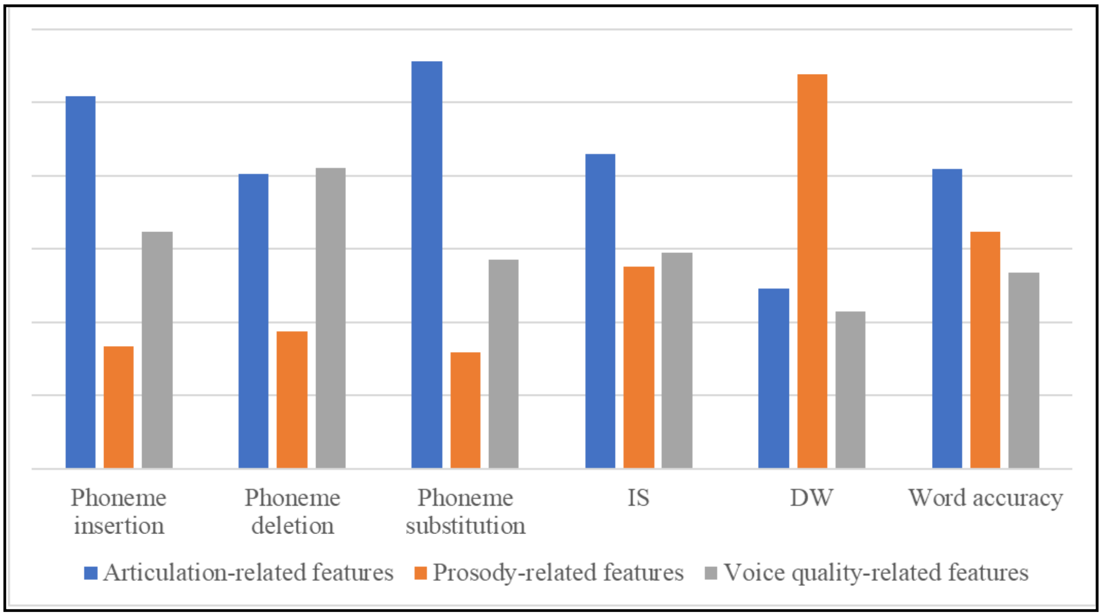


Figure 6.3: Group Importance of the Changes in Articulation-, Prosody- and Voice Quality-Related Acoustic Features for Predicting Changes in MIPs and Word Accuracy.

Table 6.4: Changes in Acoustic Measures from Habitual to the Cued Conditions.

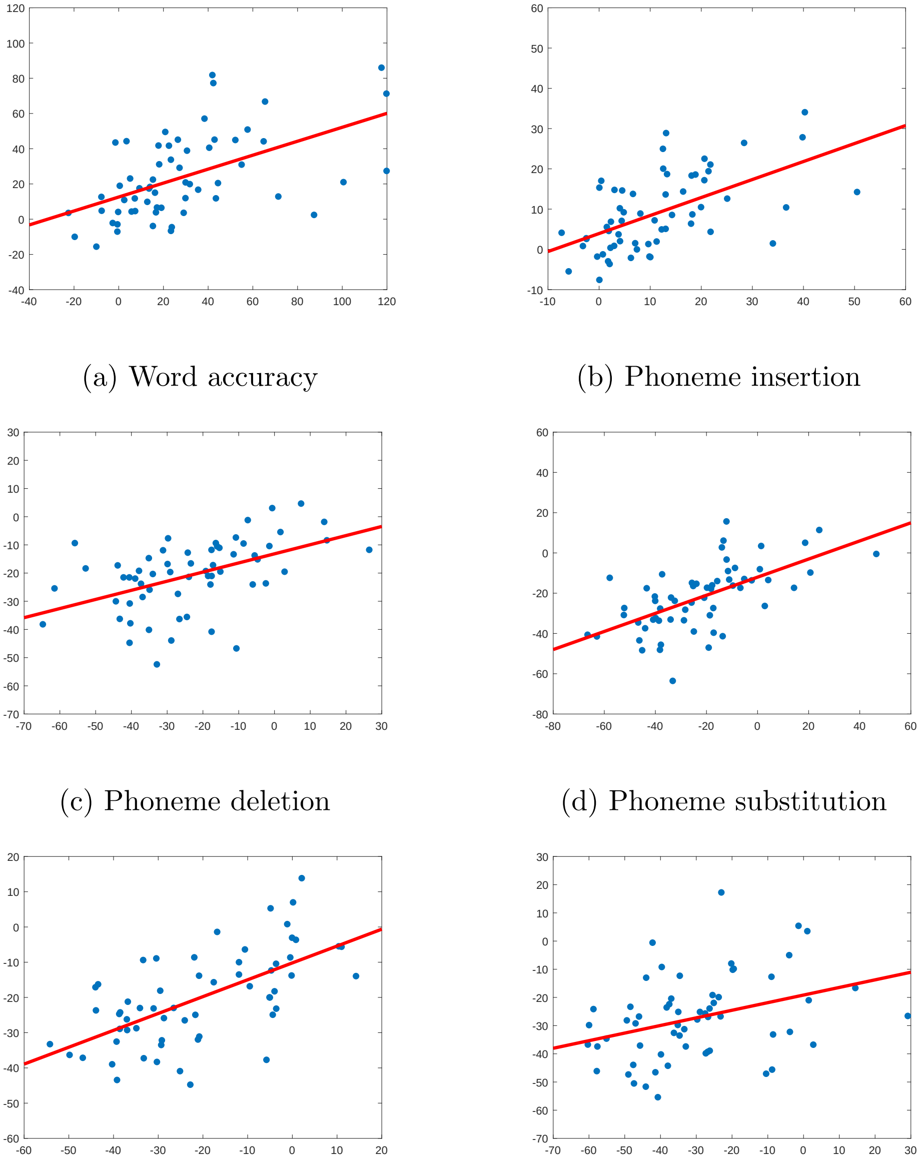
	From habitual to slow		**From habitual to clear**		**From habitual to loud**	
	Mean(%)	**Std(%)**	**Mean(%)**	**Std(%)**	**Mean(%)**	**Std(%)**
VSA	110.26	304.05	138.81	300.45	**289.47**	459.75
artEnt	-8.49	6.29	1.49	10.37	**2.42**	7.30
GOP_meanV	**66.09**	65.05	38.54	63.15	42.01	49.94
GOP_meanC	**56.90**	68.71	21.66	42.00	8.00	33.53
GOP_meanS	**62.99**	58.95	30.76	51.13	27.33	40.17
EMS_pAmp	**12.40**	8.01	4.78	5.68	1.28	4.88
EMS_ratio4	**15.47**	9.34	6.83	7.23	2.13	7.28
SR	**-29.95**	17.70	-13.83	14.69	-4.51	15.09
F0_var	28.00	61.58	48.57	94.08	**54.08**	70.61
syllCont_dur	-10.13	14.78	-6.48	9.49	**1.88**	9.37
syllCont_int	-6.81	20.95	-2.52	18.55	**9.73**	22.13
voicing_dbrks	-4.37	26.45	-7.60	23.76	**-33.60**	15.07
jitter_norm	-21.81	14.99	-19.48	13.38	**-32.08**	14.49
shimmer_localdB	**-14.17**	17.97	-5.56	18.10	-2.10	18.48
HNR_dB	**8.94**	9.05	3.00	8.50	-0.37	10.22

Table 6.5: The Five Most Important Features and Their Standardized Coefficients in Predicting Changes in MIPs and Word Accuracy.

Phoneme insertion				
GOP_stdC	GOP_minC	jitter_rap	HNR_db	SR
3.921	-3.427	-1.703	1.48	1.03
Phoneme deletion				
GOP_stdC	GOP_minC	shimmer_local	shimmer_localdB	shimmer_apq3
2.543	-2.277	-1.865	1.765	1.485
Phoneme substitution				
GOP_stdC	GOP_stdS	GOP_minV	GOP_meanS	jitter_norm
2.04	-2.026	1.224	1.102	1.085
IS				
GOP_stdC	EMS_ratio4	GOP_minC	shimmer_local	EMS_E0-4
4.278	-2.879	-2.748	-2.735	2.683
DW				
EMS_E0-4	EMS_ratio4	shimmer_apq5	GOP_minV	GOP_stdV
-8.967	6.736	-3.331	-2.221	2.105
Word accuracy				
EMS_E0-4	GOP_minV	GOP_stdV	EMS_ratio4	jitter_rap
-2.938	-2.92	2.784	1.551	1.54

Table 6.6: The Correlation Coefficient and MAEs of between the Predicted Changes and the Actual Changes in Word Accuracy and MIPs.

	Correlation coefficient	**MAE**
Word accuracy	0.58	19.71
Phoneme insertion	0.67	7.58
Phoneme deletion	0.65	13.08
Phoneme substitution	0.74	9.62
IS	0.62	11.32
DW	0.43	14.98



(a) Word accuracy

(b) Phoneme insertion

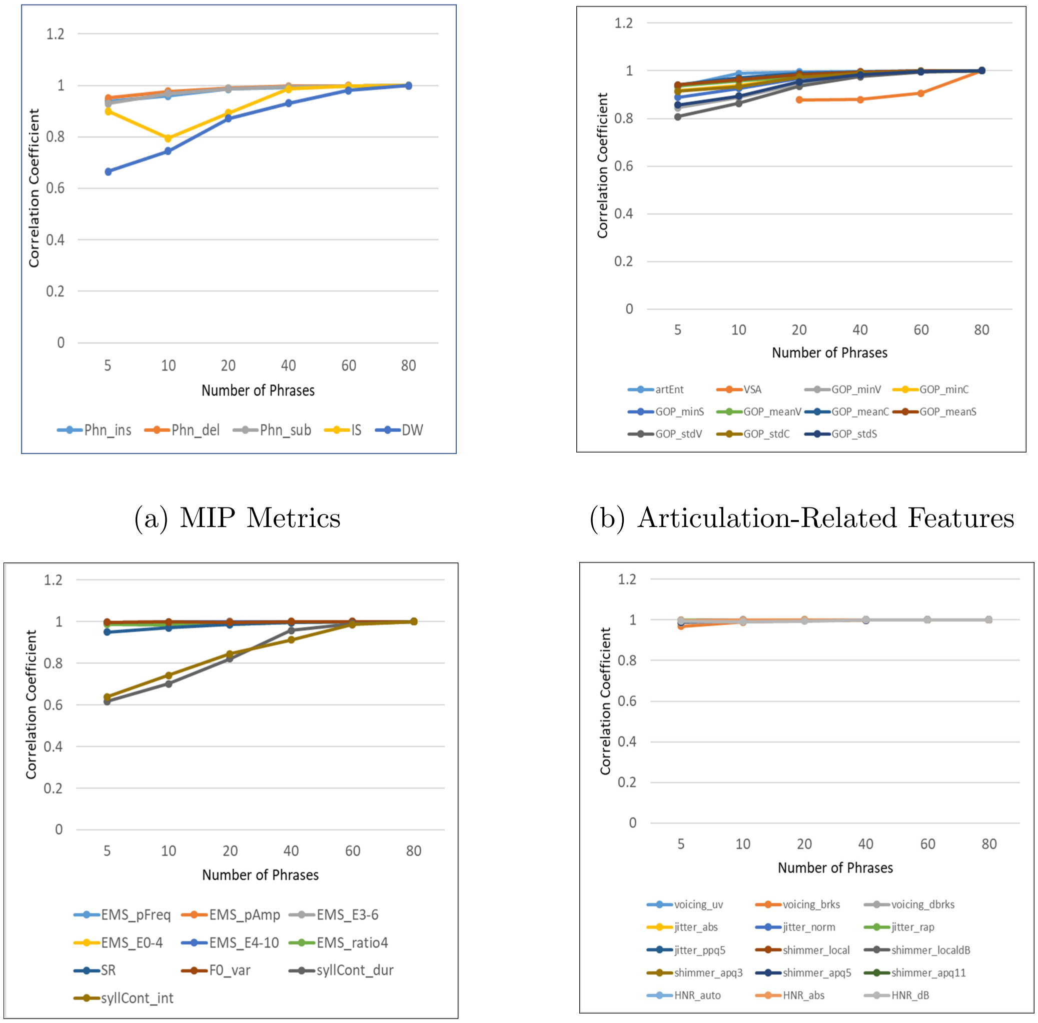
(c) Phoneme deletion

(d) Phoneme substitution

(e) IS

(f) DW

Figure 6.4: Scatter Plots Showing the Predicted Changes (y-axis) and the Actual Changes (x-axis) in Word Accuracy and MIPs.



(a) MIP Metrics

(b) Articulation-Related Features

(c) Prosody-Related Features

(d) Voice Quality-Related Features

Figure 6.5: Correlation Coefficients of the Measures Extracted from Reduced Number of Phrases with Those Extracted from 80 Phrases.

Chapter 7

CONCLUSION AND FUTURE WORK

Speech intelligibility is the amount of information that has been successfully transmitted from the speaker to the listener during the speech communication process. The quality of the information in the acoustic signal is determined by the control of the speaker's production system. Different quality of the acoustic information results in different strategies that listeners use to comprehend speech. Improving speech intelligibility is the central goal of speech pathology. Perceptual measures of intelligibility suffer from subjective bias. Word accuracy is a traditional objective measure of intelligibility calculated from listener transcripts. However, it is not able to reveal the underlying sources of intelligibility degradation. This study attempts to understand speech intelligibly on a deeper level. We want to answer not only how much the intelligibility decreases, but also how the speech challenges listeners' attempt to understand it, what characteristics of the speaker make the speech degrade, and what is the relationship between acoustic information and listening strategies.

To answer those questions, we have proposed to quantify intelligibility from both listener transcripts and the acoustic signals, and studied their relationship using a computational model. We assumes that intelligibility degradation is due to phoneme misperception and incorrect word segmentation, which was caused by the imprecise articulation and abnormal rhythm patterns of the speaker. When the phonemic cues in the acoustic signal are less reliable, listeners make use of the more robust rhythmic cues, such as the stress-unstressed syllable contrasts, to segment acoustic streams into words to facilitate speech comprehension. Based on this, we hypothesize that the articulatory precision affects the phoneme recognition accuracy, while both the articulation and rhythm controls of the speaker affect the listeners' lexical segmentation strategies.

To test the hypothesis, we have developed a multidimensional intelligibility profile (MIP) to measure intelligibility degradations from listener transcripts by using both segmental phoneme errors and suprasegmental word segmentation errors. From the acoustic signal, we have developed a comprehensive set of automated measures to evaluate the speaker's articulation, prosody/rhythm, and voice quality. We have triggered acoustic variations within speakers and studied the relationship between changes observed in acoustic features and those in listening error patterns. The main findings can be summarized as follows:

1. The five dimensional MIP metrics are a complete representation of speech intelligibility. Evidence has been shown in Table 4.6 that when predicting word accuracy using the three phoneme errors and two LBEs, the model achieved $R^2 = 0.97$, which means the MIP metrics can explain 97% variations in word accuracy.

2. The developed acoustic features are significantly related to the perceptual dimensions they aim to measure. Evidence has been shown in Figure 5.11 that when predicting perceptual ratings with their corresponding acoustic features, the predicted values achieved significant correlation with the actual values. However, the voice quality-related features showed less reliability than the articulation- and prosody-related features.

3. It is beneficial to predict multiple perceptual ratings (which are highly correlated) simultaneously. Evidence has been shown in Table 5.4 that models that are trained with MTL could make better predictions than those trained separately.

4. The articulation-related acoustic features are important in predicting both phoneme

errors and lexical segmentation errors, while the prosody-related features are more important in predicting lexical segmentation errors. This is consistent with the hypothesis and evidence has been shown in Figure 6.3.

5. We can make reliable predictions of listening errors from changes observed in acoustic signals. From changes calculated in the acoustic features, we can predict how each dimension of MIP metrics and word accuracy change within a speaker. Evidence has been shown in Table 6.6 and Figure 6.4.

In addition to the above findings, we also noticed some results from the experiment that beyond our hypothesis. First, Figure 6.3 shows that changes in voice quality-related features are significantly important in predicting changes in phoneme errors and segmentation errors. This is surprising because previous studies have shown that voice quality only accounts for a small amount of variations in intelligibility (See Equation 2.1), although this is not in a within speaker setting. For within speakers, in a relevant study (Fletcher, McAuliffe, et al., 2017b), one of the experiment results showed that a speaker's baseline voice quality was the best determinate of whether one treatment strategy could be more successful than another. Even though they used a different set of voice quality measures and did not measure variations from habitual to the intervention strategies, their findings may suggest the same thing as ours, which was improving someone's voice quality may have positive effects on improving speech intelligibility. Second, from Table 6.5 we found that the GOP of consonants was one of the most important predictors of phoneme recognition errors. This suggests that it might be more efficient to focus more on consonant pronunciation than vowels in clinical practice.

Since this is in the early stage of the study, there are some limitations in the methods and algorithms and need to be improved in the future work. For example, the procedures of recruiting listeners need to be more efficient without or with few collected data being discarded. The algorithms for LBE calculation need to be enhanced to handle more variant transcripts, such as those with different number of syllables as the targets. More reliable voice quality-related features need to be developed to align perceptual ratings better.

Besides the improvement on the current measures, it is also important to refine and expand the model because intelligibility is a complicated concept and has not been fully understood. More informal and finer measures need to be extracted to better quantify intelligibility and assess it from more aspects. For example, we aligned the target-listener transcripts based on the phonological distance. However, only the number of errors were counted to measure listeners' phoneme recognition accuracy. It would be useful if we could make use of the distance information and develop measure with higher precision.

In this study, we have developed a computational model to investigate the relationship between speech intelligibility and speech acoustics. By measuring listening error patterns, we have understood that intelligibility reduced because the listener perceived phonemes wrongly and segmented words incorrectly. By measuring acoustic signals, we have understood that the speech degraded because the speaker had difficulties in making clear pronunciations, maintaining normal prosody and rhythm patterns, and keeping a stable vocal fold vibrations. We expect that a model like this would provide more useful information to clinicians. It tells them how one patient is different from another even if they may have the same level of severity. Being facilitated with this model, the clinicians could develop more personalized treatment plans for different patients. Moreover, it could also help predict treatment outcomes given a specific intervention strategy. The patients would be clearer about how it may help listeners perceive their speech if they could modify it in a certain way.

REFERENCES

Alice, L., & O., F. (2008). Speech-language pathologists' experience on perceptual evaluation of speech disorders: A survey. *ASHA Convention.*

Ando, R. K., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*(Nov), 1817–1853.

Arai, T., & Greenberg, S. (1997). The temporal properties of spoken japanese are similar to those of english. In *Fifth european conference on speech communication and technology.*

Berisha, V., Liss, J., Sandoval, S., Utianski, R., & Spanias, A. (2014). Modeling pathological speech perception from data with similarity labels. In *Proceedings of the ... ieee international conference on acoustics, speech, and signal processing / sponsored by the institute of electrical and electronics engineers signal processing society. icassp (conference)* (Vol. 2014, pp. 915–919).

Berisha, V., Wisler, A., Hero, A. O., & Spanias, A. (2016). Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing*, *64*(3), 580–591.

Beukelman, D. R., & Yorkston, K. M. (1979). The relationship between information transfer and speech intelligibility of dysarthric speakers. *Journal of communication disorders*, *12*(3), 189–196.

Blanchet, P. G., & Snyder, G. J. (2010). Speech rate treatments for individuals with dysarthria: A tutorial. *Perceptual and motor skills*, *110*(3), 965–982.

Bocklet, T., Haderlein, T., Hönig, F., Rosanowski, F., & Nöth, E. (2009). Evaluation and assessment of speech intelligibility on pathologic voices based upon acoustic speaker models. In *Proceedings of the 3rd advanced voice function assessment international workshop* (pp. 89–92).

Boersma, P. (2006). Praat: doing phonetics by computer. *http://www. praat. org/.*

Borrie, S. A., Barrett, T. S., & Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America*, *145*(1), 392–399.

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual learning of dysarthric speech: A review of experimental studies. *Journal of Speech Language and Hearing Research*, *55*(1), 290–305.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech* *communication*, *20*(3-4), 255–272.

Bunton, K., Kent, R. D., Kent, J. F., & Duffy, J. R. (2001). The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics & Phonetics*, *15*(3), 181–193.

Camacho, A., & Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, *124*(3), 1638–1652.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning* (pp. 160–167).

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the english vocabulary. *Computer Speech and Language*, *2*, 133–142.

De Bodt, M. S., Huici, M. E. H.-D., & Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of communication disorders*, *35*(3), 283–292.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8599–8603).

Divenyi, P., Greenberg, S., & Meyer, G. (2006). *Dynamics of speech production and perception* (Vol. 374). Ios Press.

Dongilli, P. A. (1993). Semantic context and speech intelligibility.

Duffy, J. R. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management.* Elsevier Health Sciences.

Enderby, P. (1980). Frenchay dysarthria assessment. *British Journal of Disorders of Communication*, *15*(3), 165–173.

Evgeniou, T., & Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 109–117).

Falk, T. H., Chan, W.-Y., & Shein, F. (2012). Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, *54*(5), 622–631.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., Sinex, D. G., & Liss, J. M. (2017a). Predicting intelligibility gains in individuals with dysarthria from baseline speech features. *Journal of Speech, Language, and Hearing Research*, *60*(11), 3043–3057.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., Sinex, D. G., & Liss, J. M. (2017b). Predicting intelligibility gains in individuals with dysarthria from baseline speech features. *Journal of Speech, Language, and Hearing Research*, *60*(11), 3043–3057.

Fletcher, A. R., Wisler, A. A., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2017). Predicting intelligibility gains in dysarthria through automated speech feature analysis. *Journal of Speech, Language, and Hearing Research*, *60*(11), 3058–3068.

Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *The Journal of the Acoustical Society of America*, *126*(2), 847–857.

Forster, K. I., & Forster, J. C. (2003). Dmdx: A windows display program with millisecond accuracy. *Behavior Research Methods Instruments and Computers*, *35*(1), 116–124.

Frearson, B. (1985). A comparison of the aids sentence list and spontaneous speech intelligibility scores for dysarthric speech. *Australian Journal of Human Communication Disorders*, *13*(1), 5–21.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the ieee international conference on computer vision* (pp. 1440–1448).

Hammen, V., Yorkston, K., Dowden, P., Moore, C., Yorkston, K., & Beukelman, D. (1991). Index of contextual intelligibility: Impact of semantic context in dysarthria. *Dysarthria and apraxia of speech: Perspectives on management*, 43–53.

Hsu, S.-C., Jiao, Y., McAuliffe, M. J., Berisha, V., Wu, R.-M., & Levy, E. S. (2017). Acoustic and perceptual speech characteristics of native mandarin speakers with parkinson's disease. *The Journal of the Acoustical Society of America*, *141*(3), EL293–EL299.

Hustad, K. C. (2006). A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners. *American Journal of Speech-Language Pathology*.

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*.

Jacob, L., Vert, J.-p., & Bach, F. R. (2009). Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems* (pp. 745–752).

Jalali, A., Sanghavi, S., Ruan, C., & Ravikumar, P. K. (2010). A dirty model for multi-task learning. In *Advances in neural information processing systems 23* (pp. 964–972).

Ji, S., & Ye, J. (2009). An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning* (pp. 457–464).

Jiao, Y., Berisha, V., & Liss, J. (2017). Interpretable phonological features for clinical applications. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5045–5049).

Jiao, Y., Berisha, V., Liss, J., Hsu, S.-C., Levy, E., & McAuliffe, M. (2017). Articulation entropy: An unsupervised measure of articulatory precision. *IEEE Signal Processing Letters*, *24*(4), 485–489.

Jiao, Y., Berisha, V., Tu, M., & Liss, J. (2015). Convex weighting criteria for speaking rate estimation. *IEEE/ACM transactions on audio, speech, and language processing*, *23*(9), 1421–1430.

Jiao, Y., Tu, M., Berisha, V., & Liss, J. (2016). Online speaking rate estimation using recurrent neural networks. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5245–5249).

Kasi, K., & Zahorian, S. A. (2002). Yet another algorithm for pitch tracking. In *2002 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, pp. I–361).

Kent, R. D. (1992). *Intelligibility in speech disorders: Theory, measurement and management* (Vol. 1). John Benjamins Publishing.

Kent, R. D., & Kim, Y.-J. (2003). Toward an acoustic typology of motor speech disorders. *Clinical linguistics & phonetics*, *17*(6), 427–445.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482–499.

Kim, H., Hasegawa-Johnson, M., & Perlman, A. (2011). Vowel contrast and speech intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica*, *63*(4), 187–194.

Kim, J., Kumar, N., Tsiartas, A., Li, M., & Narayanan, S. (2012). Intelligibility classification of pathological speech using fusion of multiple high level descriptors. In *Thirteenth annual conference of the international speech communication* *association.*

Kim, S., & Xing, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Icml* (Vol. 2, p. 1).

King, J. M., Watson, M., & Lof, G. L. (2012). Practice patterns of speech-language pathologists assessing intelligibility of dysarthric speech. *Journal of Medical Speech-language Pathology*, *20*(1), 1–17.

Kondrak, G. (2003). Phonetic alignment and similarity. *Computers and The Humanities*, *37*(3), 273–291.

Lansford, K. L., Luhrsen, S., Ingvalson, E. M., & Borrie, S. A. (2018). Effects of familiarization on intelligibility of dysarthric speech in older adults with and without hearing loss. *American journal of speech-language pathology*, *27*(1), 91–98.

Le, D., Licata, K., Persad, C., & Provost, E. M. (2016). Automatic assessment of speech intelligibility for individuals with aphasia. *IEEE Transactions on Audio, Speech, and Language Processing*, *24*(11), 2187–2199.

Liss, J. M., LeGendre, S., & Lotto, A. J. (2010). Discriminating dysarthria type from envelope modulation spectra. *Journal of Speech, Language, and Hearing Research*.

Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *The Journal of the Acoustical Society of America*, *104*(4), 2457–2466.

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America*, *112*(6), 3022–3030.

Liss, J. M., Spitzer, S. M., Caviness, J. N., Adler, C., & Edwards, B. W. (2000). Lexical boundary error analysis in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America*, *107*(6), 3415–3424.

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., & Caviness, J. N. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of speech, language, and hearing research*.

Liu, H.-M., Tsao, F.-M., & Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America*, *117*(6), 3879–3889.

Liu, J., Ji, S., & Ye, J. (2009). Multi-task feature learning via efficient l 2, 1-norm minimization. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence* (pp. 339–348).

Lounici, K., Pontil, M., Tsybakov, A. B., & van de Geer, S. A. (2009). Taking advantage of sparsity in multi-task learning. *conference on learning theory*.

Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., & Nöth, E. (2009). Peaks–a system for the automatic evaluation of voice and speech disorders. *Speech Communication*, *51*(5), 425–437.

McAuliffe, M. J., Fletcher, A. R., Kerr, S. E., O'Beirne, G. A., & Anderson, T. (2017). Effect of dysarthria type, speaking condition, and listener age on speech intelligibility. *American Journal of Speech-Language Pathology*, *26*(1), 113–123.

McHenry, M. (2011). An exploration of listener variability in intelligibility judgments. *American Journal of Speech-Language Pathology*.

Metz, D. (1980). Toward an objective description of the dependent and independent variables associated with intelligibility assesments of hearing-impaired adults. *Speech Assessment and Speech Im-provement for the Hearing Imparied*, 72–81.

Middag, C., Bocklet, T., Martens, J.-P., & Nöth, E. (2011). Combining phonological and acoustic asr-free features for pathological speech intelligibility assessment. In *Twelfth annual conference of the international speech communication association.*

Middag, C., Martens, J.-P., Van Nuffelen, G., & De Bodt, M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing*, *2009*(1), 629030.

Middag, C., Van Nuffelen, G., Martens, J.-P., & De Bodt, M. (2008). Objective intelligibility assessment of pathological speakers. In *9th annual conference of the international speech communication association (interspeech 2008)* (pp. 1745–1748).

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

Murphy, P. J., & Akande, O. O. (2005). Cepstrum-based estimation of the harmonics-to-noise ratio for synthesized and human voice signals. In *International conference on nonlinear analyses and algorithms for speech processing* (pp. 150–160).

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of Speech, Language, and Hearing Research*.

Negahban, S., & Wainwright, M. J. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of l 1,8 -regularization. In *Nips'08 proceedings of the 21st international conference on neural information processing systems* (pp. 1161–1168).

Obozinski, B. Y. G., Wainwright, M. J., & Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, *39*(1), 1–47.

Obozinski, G., Taskar, B., & Jordan, M. (2006). Multi-task feature selection. *Statistics Department, UC Berkeley, Tech. Rep*, *2*.

Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188.

Pettinato, M., Tuomainen, O., Granlund, S., & Hazan, V. (2016). Vowel space area in later childhood and adolescence: Effects of age, sex and ease of communication. *Journal of Phonetics*, *54*, 1–14.

Poulton, E. C., & Poulton, S. (1989). *Bias in quantifying judgements.* Taylor & Francis.

Ramig, L. O., Sapir, S., Fox, C., & Countryman, S. (2001). Changes in vocal loudness following intensive voice treatment (lsvt®) in individuals with parkinson's disease: A comparison with untreated patients and normal age-matched controls. *Movement disorders: official journal of the Movement Disorder Society*, *16*(1), 79–83.

Roy, N., Nissen, S. L., Dromey, C., & Sapir, S. (2009). Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy. *Journal of communication disorders*, *42*(2), 124–135.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sandoval, S., Berisha, V., Utianski, R. L., Liss, J. M., & Spanias, A. (2013). Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America*, *134*(5), EL477–EL483.

Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H., & Fox, C. (2007). Effects of intensive voice treatment (the lee silverman voice treatment [lsvt]) on vowel articulation in dysarthric individuals with idiopathic parkinson disease: Acoustic and perceptual findings. *Journal of Speech, Language, and Hearing Research*.

Schiavetti, N., et al. (1992). Scaling procedures for the measurement of speech intelligibility. *Intelligibility in speech disorders*, 11–34.

Sheard, C., Adams, R. D., & Davis, P. J. (1991). Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. *Journal of Speech Language and Hearing Research*, *34*(2), 285–293.

Smith, E., Verdolini, K., Gray, S., Nichols, S., Lemke, J., Barkmeier, J., ... Hoffman, H. (1996). Effect of voice disorders on quality of life. *Journal of Medical Speech-Language Pathology*, *4*(4), 223–244.

Stevens, S. (2012). Perceptual magnitude and its measurement. *Handbook of perception*, *2*, 361–389.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of experimental psychology*, *54*(6), 377.

Tan, L. N., & Alwan, A. (2013). Multi-band summary correlogram-based pitch detection for noisy speech. *Speech communication*, *55*(7-8), 841–856.

Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, *9*, 1112–1122.

Tikofsky, R. S. (1970). A revised list for the estimation of dysarthric single word intelligibility. *Journal of Speech and Hearing Research*, *13*(1), 59–64.

Tjaden, K., Sussman, J. E., & Wilding, G. E. (2014). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in parkinson's disease and multiple sclerosis. *Journal of Speech, Language, and Hearing Research*, *57*(3), 779–792.

Tjaden, K., & Wilding, G. (2011). The impact of rate reduction and increased loudness on fundamental frequency characteristics in dysarthria. *Folia Phoniatrica et Logopaedica*, *63*(4), 178–186.

Tjaden, K., & Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria. *Journal of Speech, Language, and Hearing Research.*

Tjaden, K. K., & Liss, J. M. (1995). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics*, *9*(2), 139–154.

Tropp, J. A., Gilbert, A. C., & Strauss, M. J. (2005). Simultaneous sparse approximation via greedy pursuit. In *Proceedings. (icassp '05). ieee international conference on acoustics, speech, and signal processing, 2005.* (Vol. 5, pp. 721–724).

Tu, M., Berisha, V., & Liss, J. (2017). Interpretable objective assessment of dysarthric speech based on deep neural networks. In *Proceedings of the annual conference of the international speech communication association, interspeech* (pp. 1849–1853).

Tu, M., Wisler, A., Berisha, V., & Liss, J. M. (2016). The relationship between perceptual disturbances in dysarthric speech and automatic speech recognition performance. *The Journal of the Acoustical Society of America*, *140*(5), EL416–EL422.

Tu, M., Xie, X., & Jiao, Y. (2014). Towards improving statistical model based voice activity detection. In *Fifteenth annual conference of the international speech communication association.*

Turlach, B. A., Venables, W. N., & Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, *47*(3), 349–363.

Turner, G. S., Tjaden, K., & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research*, *38*(5), 1001–1013.

Weismer, G., Jeng, J.-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica*, *53*(1), 1–18.

Weismer, G., & Laures, J. S. (2002). Direct magnitude estimates of speech intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research.*

Weismer, G., Laures, J. S., Jeng, J.-Y., Kent, R. D., & Kent, J. F. (2000). Effect of speaking rate manipulations on acoustic and perceptual aspects of the dysarthria in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica*, *52*(5), 201–219.

Weismer, G., Martin, R., & Kent, R. (1992). Acoustic and perceptual approaches to the study of intelligibility. *Intelligibility in speech disorders*, 67–118.

Wenke, R. J., Theodoros, D., & Cornwell, P. (2008). The short-and long-term effectiveness of the lsvt® for dysarthria following tbi and stroke. *Brain Injury*, *22*(4), 339–352.

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2-3), 95–108.

Yan, Y., Ricci, E., Subramanian, R., Lanz, O., & Sebe, N. (2013). No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *Proceedings of the ieee international conference on computer vision* (pp. 1177–1184).

Yorkston, K. M., & Beukelman, D. R. (1980). A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility. *Journal of Communication Disorders*, *13*(1), 15–31.

Yorkston, K. M., & Beukelman, D. R. (1981a). Ataxic dysarthria: Treatment sequences based on intelligibility and prosodic considerations. *Journal of Speech and Hearing Disorders*, *46*(4), 398–404.

Yorkston, K. M., & Beukelman, D. R. (1981b). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders*, *46*(3), 296–301.

Yorkston, K. M., Beukelman, D. R., & Traynor, C. (1984). *Assessment of intelligibility of dysarthric speech*. Pro-ed Austin, TX.

Yorkston, K. M., Hammen, V. L., Beukelman, D. R., & Traynor, C. D. (1990). The effect of rate control on the intelligibility and naturalness of dysarthric speech. *Journal of Speech and Hearing Disorders*, *55*(3), 550–560.

Zhou, J., Chen, J., & Ye, J. (2011). Malsar: Multi-task learning via structural regularization. *Arizona State University*.

Zraick, R. I., & Liss, J. M. (2000). A comparison of equal-appearing interval scaling and direct magnitude estimation of nasal voice quality. *Journal of Speech, Language, and Hearing Research*, *43*(4), 979–988.

Zwetsch, I. C., Fagundes, R. D. R., Russomano, T., & Scolari, D. (2006). Digital signal processing in the differential diagnosis of benign larynx diseases [abstract in english]. *Scientia Medica*, *16*(3), 109–114.

APPENDIX A

PHRASE LIST OF THE STIMULI

1. account for who could knock
2. address her meeting time
3. admit the gear beyond
4. advance but sat appeal
5. afraid beneath demand
6. amend estate approach
7. and spoke behind her sin
8. attack became concerned
9. avoid or beat command
10. appear to wait then run
11. assume to catch control
12. attend the trend success
13. award his drain away
14. balance clamp and bottle
15. beside a sunken bat
16. bolder ground from justice
17. bush is chosen after
18. butcher in the middle
19. career despite research
20. cheap control in paper
21. commit such used advice
22. confused but roared again
23. connect the beer device
24. constant willing walker
25. cool the jar in private
26. darker painted baskets
27. define respect instead

28. distant leaking basement

29. divide across retreat

30. done with finest handle

31. had eaten junk and train

32. embark or take her sheet

33. for coke a great defeat

34. forget the joke below

35. frame her seed to answer

36. functions aim his acid

37. its harmful note abounds

38. hold a page of fortune

39. increase a grade sedate

40. indeed a tax ascent

41. kick a tad above them

42. listen final station

43. mark a single ladder

44. mate denotes a judgment

45. may the same pursed it

46. measure fame with legal

47. mistake delight for heat

48. mode campaign for budget

49. model sad and local

50. narrow seated member

51. her owners arm the phone

52. pain can follow agents

53. perceive sustained supplies

54. pooling pill or cattle

55. push her equal culture

56. rampant boasting captain

57. remove and name for stake

58. resting older earring

59. rocking modern poster

60. rode the lamp for testing

61. round and bad for carpet

62. rowing farther matters

63. seat for locking runners

64. secure but lease apart

65. signal breakfast pilot

66. sinking rather tundra

67. spackle enter broken

68. or spent sincere aside

69. stable wrist and load it

70. submit his cash report

71. support with dock and cheer

72. target keeping season

73. technique but sent result

74. thinking for the hearing

75. to sort but fear inside

76. transcend almost betrayed

77. unless escape can learn

78. unseen machines agree

79. vital seats with wonder

80. pick a chain for action

APPENDIX B

CONSENT FORM FOR TRANSCRIPT COLLECTION ON MTURK

Introduction

The purposes of this form are to provide you (as a prospective research study participant) information that may affect your decision as to whether or not to participate in this research and to record the consent of those who agree to be involved in the study.

Researchers

Dr. Julie Liss, a Professor in the Department of Speech and Hearing Sciences (College of Health Solutions) at ASU, and Dr. Visar Berisha, an Assistant Professor in the Department of Speech & Hearing Sciences and the School of Electrical, Computer, and Energy Engineering, have invited your participation in a research study.

Study purpose

We are collecting speech transcriptions from people aged 18 and older who have normal hearing. We will use these transcriptions to help us find out more about the perception of disordered speech. This step will allow us to devise better strategies for the treatment of these speech problems.

Description of research study

If you decide to participate, then you will join a study funded by NIH/NIDCD involving research of the perception of disordered speech. Your participation will be completely online and will last no longer than 1 hour. If you agree to participate, we ask that you be seated in a quiet room in front of a computer. You will listen to a series of words and phrases spoken by either individuals with or without speech disorders in quiet or in background noise and asked to transcribe what you hear. Research completed based on these transcriptions will provide an understanding of the impact of dysarthria on communicative function.

Risks

There are no known risks from taking part in this study.

Benefits

Although there may be no direct benefits to you, these transcriptions may improve our understanding of how dysarthria affects speech. This may, in turn, allow for the development of better speech therapy treatments.

Confidentiality

All information obtained in this study is strictly confidential unless disclosure is required by law. The results of this research study may be used in reports, presentations, and publications, but the researchers will not identify you. Your responses will be anonymous.

Withdraw privilege

Your participation in this project is completely voluntary. There is no penalty for not participating, or for choosing to withdraw from participation at any time. Your decision will in no way affect your relationship with ASU or your grade in any course. Should you choose to withdraw from the study, your digital audio-video files will not be saved and will be discarded electronically.

Costs and payments

The researchers want your decision about participating in the study to be absolutely voluntary. Yet they recognize that your participation may pose some inconvenience. You will receive $1-$2 for your participation, paid via Amazon Mechanical Turk.

Voluntary consent

Any questions you have concerning the research study or your participation in the study, before or after your consent, will be answered by Dr. Julie Liss at (480) 965-9136. If you have questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk; you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at 480-965 6788. This form explains the nature, demands, benefits and any risk of the project. By signing this form you agree knowingly to assume any risks involved. Remember, your participation is voluntary. You may choose not to participate or to withdraw your consent and discontinue participation at any time without penalty or loss of benefit. In signing this consent form, you are not waiving any legal claims, rights, or remedies. A copy of this consent form will be offered to you.

By clicking "Agree", you consent to participate in the above study and indicated that:

1. you have read the above information
2. you voluntarily agree to participate
3. you are at least 18 years of age

Agree