Objective Estimation of Dysarthric Speech Intelligibility

by

Richard Hummel

A thesis submitted to the
Department of Electrical and Computer Engineering
in conformity with the requirements for
the degree of Master of Applied Science

Queen's University
Kingston, Ontario, Canada
September 2011

Abstract

The de-facto standard for dysarthric intelligibility assessment is a subjective intelligibility test, performed by an expert. Subjective tests are often costly, biased and inconsistent because of their perceptual nature. Automatic objective assessment methods, in contrast, are repeatable and relatively cheap. Objective methods can be broken down into two subcategories: reference-free, and reference based. Reference-free methods employ estimation procedures that do not require information about the target speech material. This potentially makes the problem more difficult, and consequently, there is a deficit of research into reference-free dysarthric intelligibility estimation.

In this thesis, we focus on the reference-free intelligibility estimation approach. To make the problem more tractable, we focus on the dysarthrias of cerebral palsy (CP) . First, a popular standard for blind speech quality estimation, the ITU-T P.563 standard, is examined for possible application to dysarthric intelligibility estimation. The internal structure of the standard is discussed, along with the relevance of its internal features to intelligibility estimation. Afterwards, several novel features expected to relate to some of the acoustic properties of dysarthric speech are proposed. Proposed features are based on the high-order statistics of parameters derived from linear prediction (LP) analysis, and a mel-frequency filterbank.

In order to gauge the complimentariness of P.563 and proposed features, a linear intelligibility model is proposed and tested. Intelligibility is expressed as a linear combination of acoustic features, which are selected from a feature pool using speaker-dependent and speaker-independent validation methods. An intelligibility estimator constructed with only P.563 features serves as the 'baseline'. When proposed features are added to the feature pool, performance is shown to improve substantially for both speaker-dependent and speaker-independent methods when compared to the baseline. Results are also shown to compare favourably with those reported in the literature.

Acknowledgments

My time with Dr. Wai-Yip Geoffrey Chan has been more than memorable. I have learnt many valuable things under his direction as my supervisor that will benefit me for decades to come, whether or not I continue to work in signal processing. I would also like to thank Dr. Tiago Falk, a recently appointed professor at INRS-EMT and graduate of Mc^{2}L, for his valuable contributions to my learning and to my appreciation for doing research. Finally, I want to thank Christina, my sister, and my parents for their love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Dysarthria is a motor speech disorder, affecting the ability of its sufferers to articulate properly, and consequently, they may be poorly understood. As a result, the 'intelligibility' of a dysarthric subject is often assessed by a physician. Assessment of dysarthric speech intelligibility involves an expert judge and a subjective rating system (such as the Frenchay Dysarthria assessment [1]). Intelligibility assessment is usually done as part of a larger battery of assessments to gauge the severity of the disease or progress of treatment. Subjective techniques are often costly, however, since they usually require the time of an expert in the field of speech language pathology. Furthermore, perceptual intelligibility scales do not usually allow for any detailed assessment of the *nature* of any intelligibility deficit. Despite these setbacks, subjective, perceptual techniques remain the de-facto standard in assessing intelligibility and other related quantities of dysarthric speech. To combat some of the above issues, computer-based speech analysis has begun to play an important role

in dysarthric speech intelligibility assessment. In particular, using computer based methods reduces some of the costs associated with subjective tests. Computer based methods also offer a repeatable technique for assessment.

Existing approaches for estimating subjective scores employ one of two different approaches. The first approach is to use automatic speech recognition (ASR) systems, trained on healthy speech. The recognition rate is then the measure of intelligibility (e.g., [2, 3, 4]). This approach assumes that ASR systems will behave similarly to human perception when presented with dysarthric speech. ASR systems can be difficult to train, and usually require a template 'reference' signal or knowledge about the phonetic content of the target word.

The second approach is to estimate intelligibility without a reference signal. For this task, features which do not require a reference signal, otherwise known as 'blind' or 'single-ended' features, are required. This approach has several benefits over reference-based approaches, including language independence, and robustness to speech errors unrelated to dysarthria (e.g., accidental misreading of text). There have been only a few published attempts at the blind estimation of pathological speech intelligibility. Of these works, even fewer attempt to estimate intelligibility directly using acoustic properties related to the perceptual characteristics of dysarthric speech [5, 6, 7]). These works only explore the usefulness of their proposed parameters in a narrow context (i.e., [5, 6] present speaker-dependent results for one subtype of dysarthria only, and [7] explores the use of only one blind parameter as an intelligibility correlate). Therefore, much work is needed to create blind features which are applicable to multiple forms of dyasarthria, and are robust to individual speaker characteristics.

1.2 Contributions

This thesis makes the following contributions:

1. The use of blind speech quality features is proposed for dysarthric speech intelligibility estimation. More specifically, the ITU-T P.563 standard for the blind estimation of speech quality is discussed in detail. Experimental results using cerebral palsied speech show that speech quality estimates computed using the P.563 standard perform poorly as intelligibility correlates. In contrast, many of P.563's internal features are shown to correlate well with intelligibility. Additionally, when several internal P.563 features are re-combined using linear regression, performance improves over the use of individual features. Estimation performance obtained by linearly combining P.563 features serves as a baseline for later experiments.

2. Features related to some of the acoustic properties of dysarthric speech are proposed for dysarthric speech intelligibility estimation. The proposed features exploit the co-occurrence of several deviant speech dimensions in select dysarthrias to compactly represent dysarthric speech intelligibility. This is accomplished by proposing novel features that are sensitive to several deviant dysarthric speech characteristics (e.g., slow speaking rate and distorted vowel sounds). Proposed features include statistics based on the linear predictive (LP) model of speech, and the popular mel-frequency filterbank. Experiments with a dysarthric speech database using P.563 and proposed features give results similar to those reported in the literature, in addition to consistently outperforming a baseline (P.563 features only).

1.3 Organization of Thesis

This thesis is organized the following manner. Chapter 2 gives background information on the perceptual and acoustic properties of dysarthric speech. It also elaborates on similar works discussed in the literature. Chapter 3 begins by discussing the relationship between speech quality and intelligibility. A popular standard for blind speech quality estimation, ITU-T P.563, is then discussed in some detail. Chapter 4 proposes several features expected to reflect the acoustical properties discussed in Chapter 2. Finally, experiments using blind features from P.563 as well as proposed features from Chapter 4 are discussed in Chapter 5.

Chapter 2

Assessing Dysarthric Speech Intelligibility

2.1 Introduction

As discussed in the previous chapter, a major part of the treatment of dysarthria includes assessing how intelligible that person's speech is to others. Perceptual analysis of dysarthric speech has long been the de-facto standard for intelligibility assessment; many different techniques exist for performing this task [8]. However, there is currently a sparsity of automatic systems for blindly estimating dysarthric speech intelligibility in the literature.

To be able to develop an objective estimation system for assessing dysarthric speech, we take a 'ground up' approach by first examining the perceptual characteristics of dysarthric speech. To limit the scope of the study, and because of limitations of available data, the dysarthrias associated with cerebral palsy are studied exclusively herein. The perceptual characteristics examined can then be related to studies of the

acoustic properties of pathological speech. This gives a foundation for the further development of blind dysarthric speech intelligibility measurements that can be tested experimentally.

The organization of this chapter is as follows. First, various definitions of 'intelligibility' are discussed. Afterwards, an overview of the common perceptual characteristics of the various dysarthria subtypes is given. Finally, systems discussed in the literature for estimating intelligibility, blindly or otherwise, are discussed.

2.2 Defining Intelligibility

In the most general sense, intelligibility is the ability of an individual to be understood verbally by another. For deeper study, however, this definition is too vague and not easily quantifiable. In many perceptual studies, subjective intelligibility is assessed using one of a few different scale-based systems, (e.g., where overall perception of intelligibility is graded from 1 to 5); alternately, listeners may be asked to transcribe what they have heard, with the transcription accuracy (in %) serving as the intelligibility score [8]. In either case, many different 'scopes' of intelligibility are possible, and each one offers a distinct view of the effects of dysarthria. Intelligibility can be assessed from a 'high-level' (e.g., conversational intelligibility) to a 'low-level' (e.g., phoneme intelligibility). Examining intelligibility from a higher level gives a better perspective of the effects of dysarthria in everyday life. In contrast, studying intelligibility at a lower level allows researchers to better understand the precise articulatory difficulties of dysarthric speakers.

2.3 Perceptual Dimensions of Dysarthria

Dysarthria is a blanket term for a set of related disorders that affect the movement of speech articulators, which in turn affect speech intelligibility. There are roughly six major subtypes of dysarthria: spastic, flaccid, ataxic, hyperkinetic, hypokinetic, and 'mixed' (a combination of two or more of the first five types). The subtypes were defined by Darley et al. in [9] based on commonalities of the neurological/motor conditions causing the dysarthria.

While this thesis explores the objective estimation of dysarthric speech intelligibility, the subjective perceptual characteristics of dysarthria serve as a useful starting point for the design of an objective intelligibility estimation system. The perceptual characteristics of each subtype can be then related to the acoustic properties of the speech signal. In the next few subsections, some of the common perceptual characteristics between the subtypes are discussed, as well as the specific characteristics of the dysarthrias related to cerebral palsy.

2.3.1 Common Perceptual Characteristics

In addition to defining the dysarthric subtypes, Darley [9] examined the perceptual characteristics of each subtype. Three judges rated 38 perceptual dimensions of speech on a 7 point severity scale (1=normal,7=severe deviation from normal). Inter-judge reliability was 85% (within 1 scale degree). Similar reliability is reported in more recent work [10].

The following seven of the 38 perceptual characteristics were useful in describing the speech in *five or more* of the seven neurological groups studied in [9]: monopitch, monoloudness, harsh voice, imprecise consonants, short phrases, reduced stress, and

distorted vowels. This finding corroborates the more recent work in [11], where subjective intelligibility was expressed as a linear combination of four coarse-grained speech characteristics (most to least correlated with intelligibility): articulation, prosody, voice quality and nasality. Table 2.1 groups the seven perceptual characteristics from [9] listed above in terms of three of the coarse-grained characteristics from [11].

Category (from [11])	Description of Category	Perceptual Attributes (from [9])
Articulation	Relating to production of correct sounds for speech	Imprecise consonants, distorted vowels
Prosody	Relating to suprasegmental factors of speech production, including pitch (vocal fold vibration frequency), pauses, rate, loudness [12]	Monopitch, monoloudness, short phrases, reduced stress
Voice Quality	Relating to the timbre or pleasantness of the voice	Harsh voice

Table 2.1: Significant Perceptual Attributes of Dysarthric Speech

Ideally, for the purpose of estimating intelligibility, the deviant aspects of dysarthric speech would not vary across subgroups. In this unrealistic scenario, an objective intelligibility estimation system would be easier to develop, as it could exploit the common deviant attributes among subtypes. While the subtypes share some common characteristics (Table 2.1), several perceptual characteristics are only applicable to one or two subtypes, and deviations along the listed dimensions vary in extent between subtypes. Therefore, any system not tested for use on a particular subtype may not perform well on that subtype.

2.3.2 Dysarthrias of Cerebral Palsy

Optimally, dysarthrias from every subtype would be used in the design and testing of an intelligibility estimation system. Unfortunately, data from a small subset of subtypes is usually all that is available. In this thesis, we focus on the dysarthric subtypes present in cerebral palsy. Cerebral Palsy (CP) is caused by a brain lesion, which affects the neurological functioning of the patient. Cerebral Palsy consists of several subgroups; the main subgroups of CP are spastic, athetoid (dyskinetic), mixed (athetoid/spastic), and ataxic [13]. The three subtypes studied here are spastic, athetoid, and mixed. Spastic cerebral palsy often results in spastic dysarthria, and athetosis causes a type of slow/fast hyperkinesia. The term 'hyperkinesia' refers to excess movement, and the slow/fast descriptor refers to the type of excess movement present. The corresponding dysarthria is classified as a 'hyperkinetic' dysarthria [14]. Cerebral palsy patients with the 'mixed' subtype have a combination of spastic and athetoid symptoms.

The symptoms of spastic dysarthria include distorted vowels and consonants, reduced speaking rate, harsh voice quality, hypernasality, and abnormal prosody (e.g., monoloudness and monopitch) [15]. Hyperkinetic dysarthria is more difficult to qualify perceptually, since each etiology appears cause different set of associated deviant perceptual characteristics. To our knowledge, there are no studies in the literature that directly qualify the dysarthria associated with athetosis in terms of perceptual characteristics. However, two other etiologies causing hyperkinesia (chorea and dystonia) were studied in [9]. Both subtypes deviated significantly from normal speech along several dimensions, including imprecise consonants, distorted vowels, harsh voice, monopitch, slow rate and monoloudness. Some hyperkinetic subjects present

with excess loudness variations in place of monoloudness.

Besides overlapping perceptual characteristics, there are other commonalities between the dysarthric subtypes of cerebral palsy. The study in [16] found that the types of phonetic errors made by patients were nearly identical in nature to those of athetoid patients. Furthermore, the types of phonetic errors made by speakers with mixed CP were found to be similar to those of spastic and athetoid speech [17]. In addition, naive listeners were asked to identify the subtypes of dysarthria in [18] and frequently misidentified hyperkinetic dysarthria as spastic dysarthria.

2.4 Acoustic Properties of Dysarthric Speech

Section 2.3 discussed various perceptual aspects of dysarthric speech. In this section, the relationship between some of the perceptual properties of dysarthric speech are related to its acoustic properties. In particular, speech properties relating to spastic and hyperkinetic dysarthrias are explored. First, some background on linear predictive coding (LPC) and formant analysis is given. Linear prediction analysis is a common tool for representing the vocal tract of a speaker. Formant analysis is centered around determining the resonances of the vocal tract through the use of LPC analysis. Afterwards, selected acoustic characteristics of dysarthric speech relating to the dysarthrias of cerebral palsy are discussed. These characteristics will be exploited later in this thesis for the blind estimation of dysarthric speech intelligibility. They include vowel articulation, prosody, hypernasality and hoarseness.

2.4.1 LPC Analysis, Formants, and the Residual

Linear predictive coding (or LPC) analysis is a tool often used in speech processing for estimating an all-pole model of the speaker's vocal tract. This allows the separation of the effects of the vocal tract from the 'excitation' signal from the vocal folds. This relationship is commonly expressed in the z-domain:

$$X(z) = E(z)H(z), \tag{2.1}$$

where $H(z)$ is the z-domain representation of the vocal tract filter, $E(z)$ is the z-transform of the excitation signal, and $X(z)$ is the z-transform of the speech signal.

For voiced speech, $E(z)$ is often modeled as the z-transform of an impulse train, which represents periodic input from the vocal folds.

In LPC analysis, and for voiced speech in particular, the filter $H(z)$ is commonly expressed in the form:

$$H(z) = \frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}}, \tag{2.2}$$

where p is the number of filter coefficients. Several techniques are available for computing the filter coefficients a_k, with one of the more common techniques being the autocorrelation method [19]. Before the filter coefficients are computed, a pre-emphasis filter of the form

$$G(z) = 1 - az^{-1} \tag{2.3}$$

is applied to the speech signal, with the coefficient a ranging from approximately 0.9 to 1. The pre-emphasis filter emphasizes the higher frequency components and is often applied before coefficient estimation to prevent the poles from clustering near the low frequencies. Since speech is a rapidly changing signal, a 30-50 msec window is usually applied to the speech section of interest before LPC analysis is performed.

2.4.1.1 LPC Poles and Formants

After the LP coefficients are computed, several other related parameters can be obtained. Of interest are the roots of the polynomial $1 + \sum_{k=1}^{p} a_k z^{-k}$. The i^{th} root ($i = 1..p$) can be expressed as $b_i e^{\frac{f_i 2\pi i}{F_s}}$, where F_s is the sampling frequency, b_i is related to the 'bandwidth' of the pole (also the distance of the pole from the origin), and f_i is the pole's frequency location. In voiced speech, these roots represent the effect of resonant frequencies of the vocal tract. Note that we are only interested in *half* of the poles (i.e., in the first and second quadrants of the complex plane), because the roots will have conjugate symmetry.

Not all the resonant frequencies are of equal importance; in vowel perception, the most important resonances are the two with the lowest f_i values. The smallest f_i is typically labeled F1 (the first formant), the second smallest f_i is labeled F2, etc., choosing only f_i in the range $\frac{F_s}{2} > f_i > 0$ for the reasons stated above. More sophisticated approaches involve techniques such as smoothing to remove discontinuities [20].

2.4.1.2 LPC Residual Signal

Referring back to equation (2.1), if we let $G(z) = \frac{1}{H(z)}$, we can obtain the excitation signal, $E(z)$:

$$E(z) = X(z)G(z) = X(z)\frac{1}{H(z)}. \tag{2.4}$$

The filter $G(z)$ is called the *inverse filter*, and can be obtained in a manner similar to $H(z)$. By applying $G(z)$ to the speech signal, the signal spectrum is flattened (since $G(z)$ is an all-zero filter) and an estimate of the excitation signal $E(z)$ can be obtained.

2.4.2 Vowel Articulation

Many of the dysarthrias affect the ability of the sufferer to position their tongue properly, and as a result, vowel distortions are a common characteristic of dysarthria. This is because tongue positioning is essential to vowel production. The F1 frequency is proportional to how 'low' the tongue is, and the F2 frequency is proportional to how 'front' the tongue is [21]. The 'vowel space' consists of all possible F1/F2 combinations, and the vowels that lie on the extrema of the F1/F2 space are known as 'corner vowels'. In dysarthric speech, the corner vowels are shifted towards the center of the F1/F2 space. Consequently, the area enclosed by the triangle formed by connecting three of the corner vowels decreases in proportion to intelligibility [22]. This effect is known as 'vowel centralization'.

Figure 2.1 compares the corner vowels of two male speakers: one severely dysarthric speaker and one mildly dysarthric speaker. The figure was generated from single vowel tokens of dysarthric speakers from the UA database [23]. Plotted results correspond to similar figures presented in [22], in that the vowel triangle area of the severely dysarthric speaker is much less than that of the speaker with mild dysarthria.

2.4.3 Prosody

Pitch, loudness and speaking rate are commonly affected in dysarthric speech. Fundamental frequency, which is closely related to pitch, is often computed using autocorrelation or peak-picking [24] and corresponds to the cycling time of the vocal folds. In dysarthric speakers, only limited control of fundamental frequency is possible [25]. Loudness and the related concept of speaking rate are also perceptual

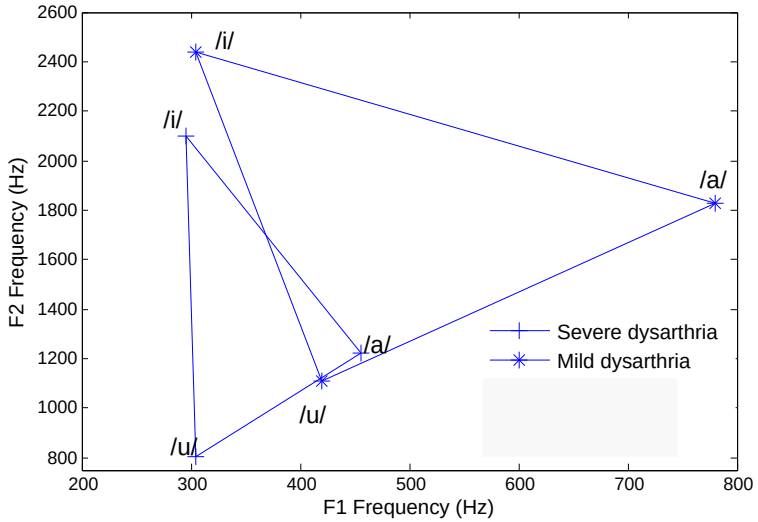

Figure 2.1: Comparison of vowel triangle areas: Severe vs. mild dysarthria

quantities. The perceived loudness of a signal is related to the distribution of energy in the speech spectrum, whereas speaking rate is more related to the *modulation* of energy over time. The works in [26, 5] explored the use of amplitude envelopes in assessing dysarthric speech. Both works found that properties of the amplitude envelope reflected speaking rate and intelligibility in dysarthric speech.

2.4.4 Hoarseness

A commonly reported deviant perceptual characteristic of dysarthric speech is 'hoarse voice'. This characteristic usually arises from impaired laryngeal functioning in dysarthric subjects. Vocal hoarseness has been studied extensively in the literature, but very little has been done in the way of quantifying the relationship between hoarseness (and voice quality in general) in dysarthric speech to intelligibility. Many

acoustic studies of hoarseness have found the proportion of harmonic to non-harmonic energy in voiced speech to be a useful indicator of hoarseness, as well as amplitude and fundamental frequency variations (a good summary is given in [27]. Furthermore, hoarseness is often correlated with other voice characteristics, such as roughness and breathiness [28].

2.4.5 Hypernasality

Hypernasality refers to the excess contribution of the nasal cavity to the acoustics of speech. As discussed in [29, 30], hypernasality causes a decline in the energy contained in existing high frequency resonances, and introduces new 'nasal' resonances (including a strong low-frequency resonance), as well as several anti-resonances which cause noticeable dips in the spectrum. A severe instance of hypernasality is presented in Figure 2.2. The top plot shows the vowel spectrum of a speaker with a severely nasal sounding voice, and the bottom plot shows a voice with normal nasality. The target vowels are the same. As can be seen from the figure, the nasal voice has a prominent low-frequency resonance, as well as several anti-resonances (the large dips in the spectrum).

2.5 Objective Intelligibility Estimation

In the pathological speech intelligibility literature, there are two main approaches to the objective estimation of dysarthric speech intelligibility: reference based and reference-free. The reference based approach involves the use of a 'reference' signal for the computation of the intelligibility score. The reference-free approach does not

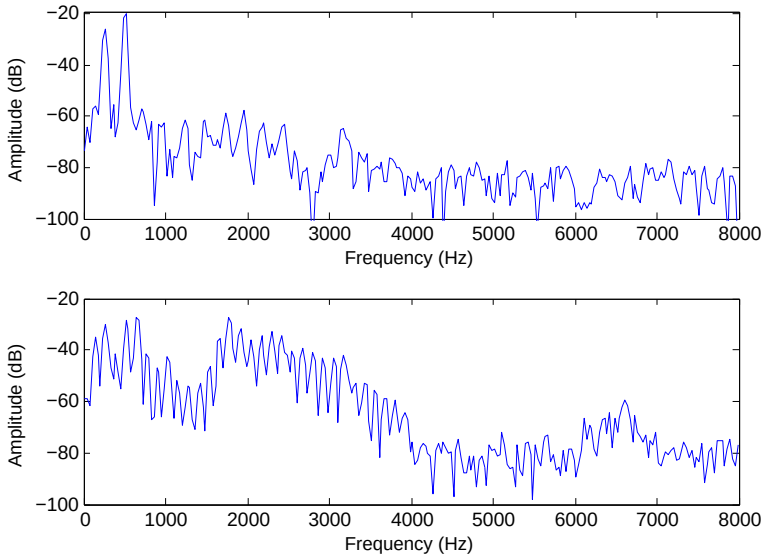

Figure 2.2: Vowel spectrum of /a/: Nasal (top) vs. normal (bottom) speech

require a reference signal; rather, one or more 'features' are computed and combined to form the intelligibility score. In this section, both approaches are discussed, along with relevant works in the literature.

2.5.1 Reference Based Intelligibility Estimation

The basic idea behind any reference based approach is to first compute a quantity of interest (e.g., a phonetic transcription of an uttered word) and then compare the result to a model or reference. In pathological speech intelligibility estimation, the reference-based approach is often explored through the application of automatic speech recognition (ASR) tools [2, 3, 4]. The ASR system is usually first trained on normal speech with the target vocabulary of interest, using the popular mel-frequency cepstral coefficients (MFCCs). The pathological speech can then be fed through the

ASR system; the transcription provided by the ASR can then be compared to the 'reference' signal (often a list of target words or phonemes). The recognition rate then becomes the subjective intelligibility assessment metric. This approach gives a fairly direct method of estimating intelligibility scores; the ASR system serves as a 'listener' and the intelligibility score is computed by comparing the ASR transcription to the reference signal.

The works in [31, 32] also use ASR reference-based systems, but use a slightly different approach. In [31], a forced alignment procedure is used to fit the incoming speech signal into a 'reference' model of the uttered word. The reference model is a hidden Markov model (HMM) trained on healthy speech, which is a standard statistical model for ASR systems. The log-likelihood that the reference model could have generated the input dysarthric speech sample serves as a measure of dysarthric speech severity. Similarly, the work described in [32] used a forced alignment procedure, but for the purpose of estimating phoneme intelligibility. Features consisted of probabilities generated from an ASR model, which were combined through a linear regressor to estimate phoneme intelligibility. This approach can be considered an objective approach to the studies in [16, 33], where intelligibility is related to a small number of phonetic error types.

2.5.2 Reference Free Intelligibility Estimation

In reference-free systems, the goal is to estimate intelligibility without prior knowledge of speech material. Generally, this means that intelligibility deficits are not measured directly, i.e. through comparing a list of transcribed words/phonemes to

a list of 'reference' words. Reference-free systems have the advantage of being potentially applicable to new types of speech data, e.g., spontaneous speech, or speech of a different language. Unlike most ASR-based approaches, this type of assessment usually requires the use of pathology specific characteristics, particularly when employing prosodic parameters. The advantage gained is that a reference signal is no longer required, which potentially makes the development of language and vocabulary independent systems easier.

Reference-free intelligibility assessment systems for pathological speech take varied approaches. One approach to intelligibility estimation is to combine acoustic parameters known to relate to intelligibility or dysarthric speech severity. Features related to the three subgroups presented in Table 2.1 (articulation, prosody, voice quality) were combined using a linear regression function in [5, 6], and included parameters related to pitch, articulation difficulties, reduction in speaking rate, and voice quality. Most of the features quantified time-based deviations in various energy and amplitude characteristics of dysarthric speech. Second formant (F2) slopes were investigated in [7]. The slope of the second formant was found to increase in proportion to intelligibility. The second formant slope is related to both slow rate of speech (prosody) and the articulatory ability of the speaker.

The second approach to blind intelligibility estimation of pathological voice involves computing common speech features (e.g., MFCCs), and then employing a statistical model (such as a gaussian mixture model or a neural network) to compute some features, followed by a regressor to estimate intelligibility. In [34], a system was developed similar to the reference based system in [32], but omits the forced-alignment system described previously. Instead, several quantities are derived

directly from an artificial neural network (ANN) based phonological feature detector, covering 12 different statistical properties of 11 different phonological feature types (voiced/unvoiced, nasal/non-nasal, etc.). Lastly, the work presented in [35] fitted a Gaussian mixture model (GMM) to computed MFCC and perceptual linear prediction (PLP) features, and then mapped the estimated mixture parameters (covariance and means) to a final intelligibility score.

2.6 Summary

In this chapter, the fundamentals regarding the characteristics of dysarthric speech intelligibility estimation were presented. Intelligibility was first defined in the context of dysarthria. Then, the general perceptual characteristics of dysarthria were discussed, followed by a focused discussion on the dysarthrias of cerebral palsy. Several acoustic properties of the dysarthrias of cerebral palsy were discussed, followed by an overview of the different approaches to intelligibility estimation present in the literature.

Chapter 3

Blind Speech Quality Estimation

As a starting point for further investigation into dysarthric speech intelligibility, we propose the use of blind speech quality estimation algorithms. In this chapter, we briefly discuss the relationship between speech quality and intelligibility. Afterwards, the ITU-T P.563 standard for blind speech quality estimation is discussed. Finally, we describe the feature re-mapping strategy we employ to make better use of P.563's internal features.

3.1 Motivation

While speech quality and intelligibility are related, their relationship is not trivial [21]. For example, broad bandwidth speech may be intelligible and pleasant, whereas synthesized speech may be intelligible but artificial sounding and therefore deemed poor in quality. As another example, if a speech signal is masked by an external noise source (e.g., by a plane passing overhead), the relationship between speech quality and intelligibility may be more direct.

Despite the somewhat complex relationship between the two concepts, several parameters used in the estimation of speech quality may also be sensitive to speech characteristics distorted by dysarthria. Therefore, as a starting point, we seek a blind speech quality estimation system that we can modify to instead estimate dysarthric speech intelligibility. The speech quality estimates may also be useful as correlates of dysarthric speech intelligibility. For this purpose, we use the ITU-T P.563 standard for the blind estimation of narrow-band speech quality [36]. The P.563 standard estimates the mean opinion score (MOS), which is a standard subjective quality rating scale for speech. The mean opinion score is derived by averaging many individual rankings. Each ranking is based on a standardized scale: "bad", "poor", "good", "fair", "excellent", represented numerically by integers from 1 to 5 for a score computation. The P.563 standard estimates the MOS score by selectively combining various features related to various aspects of speech quality. The source code of the P.563 standard is freely available, and the feature mappings can easily be modified to suit our needs. Re-mapping P.563's features for a particular purpose has precedence; for example, the work in [37] proposed a new feature mapping to improve MOS estimates of noise suppressed speech. Similarly, the work in [38] re-mapped P.563's internal features to instead estimate speech 'naturalness' ratings rather than speech quality (MOS) ratings.

3.2 The ITU-T P.563 Standard

The ITU-T P.563 standard was designed for the estimation of narrowband speech quality [39]. 'Narrowband' speech means a sampling frequency of 8000 Hz. The standard is a popular benchmark for the blind, objective estimation of the mean

opinion score (MOS). In the next few sections, we discuss the high-level view of P.563, and then examine the three categories of features computed by P.563.

3.2.1 High-Level Overview

The P.563 standard approaches the speech quality estimation problem by first classifying the signal into one of six distortion classes. This is done by first computing a set of 'key parameters' which determine which of the six types of distortion are present. The six classes cover distortions related to noise, clipping, muting, and 'unnatural' voice. Class assignment is performed using a hierarchical approach that chooses the most relevant distortion class when multiple distortion types are detected. The six classes are ranked according to their 'annoyance'. If P.563 detects that an input signal meets the entrance criteria of two or more distortion classes, the perceptual model employed is the model of the most 'annoying' class. The reasoning behind this is psycho-acoustic; some types of distortion affect subjective quality scores more than others. Based on the selected class, P.563 computes an intermediate MOS estimate. The intermediate quantity is then combined with a class-independent MOS estimate to produce the final quality score. A high-level system overview is given in Figure 3.1.

3.2.2 Feature Overview

In order to compute an estimate of speech quality over a wide range of degradation conditions, P.563 computes a wide variety of features. In total, P.563 computes 43 different features, which can be grouped into three categories: linear predictive coding (LPC) and vocal tract features, a 'pseudo-reference' basic quality model, and other

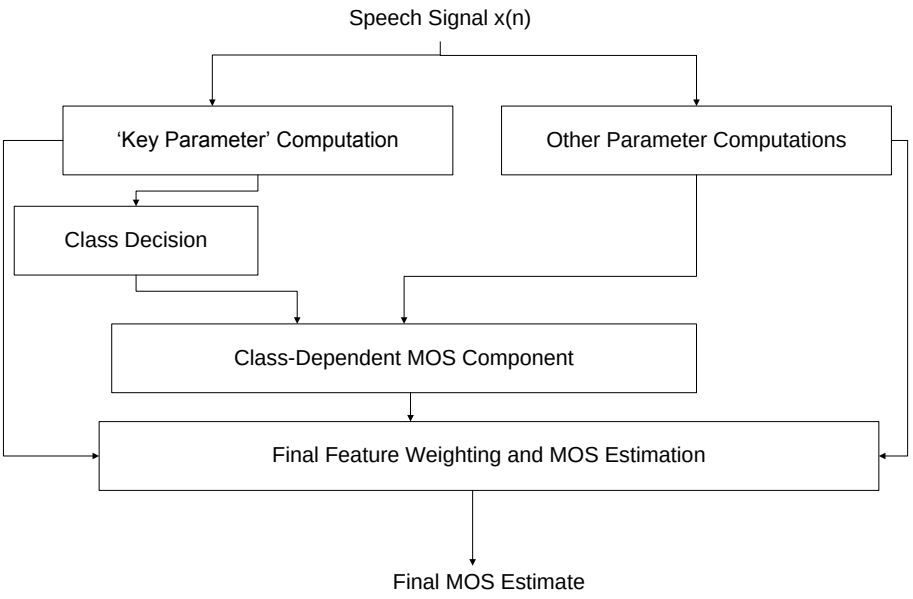

Figure 3.1: High-level system overview of P.563

distortion parameters (clipping, mutes, etc.) [39]. We examine each category in turn, and discuss the applicability of each category's features to the assessment of dysarthric speech intelligibility. Since there are many features (43) and only a select few are relevant to our problem, their computational details are left out. Instead, details of features found to be experimentally useful are left to Chapter 5.

3.2.2.1 Speech Production Model Features

Several statistics of the linear predictive coding coefficients and cepstrum are analyzed over speech segments for their 'naturalness'. The statistics of cepstral and LPC coefficients lie within particular ranges for natural speech. Cepstral processing is a

standard technique for speech signal processing, designed to separate glottal excitation from vocal tract shape. In addition to LPC/cepstral statistics, an acoustic tube model is employed to detect unnatural/physically impossible vocal tract shapes. Since these features describe statistics about the *speech* signal, and not background noise or other types of distortions, it is expected that they will be useful in later experiments.

3.2.2.2 Basic Quality Model

In the basic quality model feature category, P.563 computes a set of speech quality parameters (not the final MOS estimate) computed using a 'pseudo-reference' model. This is accomplished by first 'cleaning up' the signal, and then using a reference-based model to compare the clean version to the input signal. The first stage (the 'cleaning') is done with the use of a vocal tract model and re-synthesis of the speech signal. Then, a double-ended (reference-based) perceptual model computes a speech quality rating by comparing the 'clean' version to the input signal. The potential applicability of this approach depends on the manner by which the signal is 'cleaned'. Estimation of the vocal tract parameters was not developed with dysarthric speech in mind. Consequently, the vocal tract model may not provide a useful norm against which a reference based perceptual compares the input signal.

3.2.2.3 Distortion Parameters

The remaining P.563 features target specific types of distortions, such as clipping, muting, etc. Many of the features in this category describe signal attributes that are irrelevant to dysarthric speech intelligibility. However, a few features in this category

could be useful, particularly those features that quantify the temporal dynamics and the spectral characteristics of the dysarthric speech segments.

3.3 Re-mapping Blind Speech Quality Features

The feature mappings in P.563 were designed to map feature values to a speech quality estimate, and therefore are potentially not ideal for estimating dysarthric speech intelligibility. Perceptually, there should be some correlation between speech quality and intelligibility. However, P.563 uses features which were selected based on their sensitivity to degradations encountered in telephony transmissions, and feature mappings were not optimized for intelligibility. Therefore, we consider a new feature mapping explicitly for intelligibility estimation. To re-map the features, we use a linear model to combine selected features into an intelligibility estimate. Let ϕ_{est} be the intelligibility estimate, N the number of features, $\mathcal{F}$ the set of selected features, and a_i the model coefficient for feature f_i. Our proposed model can then be written as

$$\phi_{est} = a_0 + \sum_{i=1}^{N} a_i f_i, \qquad f_i \in \mathcal{F}. \tag{3.1}$$

To obtain the weights a_i, we use a least-squares estimation approach [40]:

$$a = (M^T M)^{-1} M^T \Phi, \tag{3.2}$$

where $a = [a_0 a_1 .. a_N]^T$, M is our matrix of feature values (with the leftmost column of M being all ones, and each row comprising the feature values for one subjective intelligibility score), and Φ is the vector of subjective intelligibility scores.

3.4 Summary

In this chapter, we have discussed the use of the ITU-T P.563 blind speech quality assessment standard in estimating dysarthric speech intelligibility. A system-level overview of the P.563 standard was then given, outlining the internal classification system and MOS computation. Afterwards, we introduced the possibility of combining some of P.563's internal features into a new intelligibility estimator using equation 3.1. In summary, the P.563 standard gives two different avenues of exploration into dysarthric speech intelligibility: the quality estimates from the six distortion classes, and the 43 feature values. In the next chapter, we introduce several new candidate features, and discuss their significance with respect to the acoustic and perceptual properties of dysarthric speech.

Chapter 4

Linear Prediction and Delta Energy Features

To accurately estimate the intelligibility of dysarthric subjects, we need a compact set of features to describe the deviant characteristics of each individual. The feature set computed by the P.563 standard serves as a starting point, but were not designed for dysarthria assessment. Consequently, it may not target the deviant aspects specifically enough to be useful in dysarthric intelligibility estimation. For that reason, we define several low-level acoustic features that are expected to reflect many of the high-level deviant *perceptual* dimensions of dysarthric speech.

4.1 Motivation

The purpose of many features discussed in the literature relate to describing properties of the dysarthrias being studied that are most often deviant or abnormal when compared to normal, intelligible speech. To recapitulate, these properties include

articulation problems, disturbed prosody, poor voice quality and nasality [11]. While many of these quantities have a perceptual basis, there has been much work in objectively quantifying these quantities in terms of acoustic features (e.g. [41, 42, 43, 22]). As discussed in Chapter 2, the majority of speech intelligibility prediction systems use automatic speech recognition (ASR) and/or phonetic features, typically performing word or phoneme identification, followed by comparison to a 'reference' signal. Consequently, there are very few studies and corresponding features in the literature suitable for the 'blind' estimation of dysarthric speech intelligibility.

In an attempt to fill in this gap in the literature, this thesis proposes the use of two particular types of sliding-window spectral quantities: parameters reflecting the change in energy content of a mel-frequency filterbank, and statistics of various quantities derived from the linear prediction (LP) filter model of speech. In particular, we define quantities related to the slow rate of dysarthric speech and to the limited articulatory capabilities of dysarthric speakers.

The rest of this chapter is structured as follows. First, we discuss the construction of our mel-frequency filterbank. We then show how the features are computed, and how the filterbank relates to the properties of dysarthric speech, with emphasis on the imprecise vocal tract configurations typical of athetoid/spastic dysarthrias. Next, we define LP based features expected to be relevant to the estimation of speech intelligibility.

4.2 Delta subband energies

Several spectral subbands are critical for speech intelligibility; prior work has shown that intelligibility is mostly preserved when only sparse spectral information is kept

[44]. It is also known that certain subbands in the 1000 - 3000 Hz range are integral to speech intelligibility, in and around 1500 Hz in particular [45]. Formant measures offer a more targeted approach to the spectral behaviour of speech, and may offer clues to variations in intelligibility, but formant detectors are often error prone. This is especially true when dealing with breathy/hoarse speech, as is typical of dysarthria. Considering the above, features consisting of the delta-energies of a mel-spaced triangular filterbank are proposed. The term 'delta' refers to using the absolute value of energy differences between consecutive frames as a feature, instead of just the energy values.

One way to describe the energy distribution over a range of frequencies is to use a filterbank; typically, filters are employed whose center frequencies are spaced along the spectrum according to some criterion. For this purpose, the filterbank used here is a mel-spaced triangular filterbank, often used for computing the common mel-frequency cepstral coefficients (MFCCs). The articulation index (AI), discussed in [21], uses signal-to-noise ratio measurements for each subband to compute a measure related to intelligibility. To our knowledge, however, the use of mel-filterbank derived delta energy parameters to estimate dysarthric speech intelligibility is novel. It is hoped that by quantifying the spectrum of dysarthric speech, using mel-frequency spaced triangular filters, the subtle differences between various levels of intelligibility can be captured. Since one of the symptoms of athetoid/spastic dysarthric speech is reduced speaking rate, it is hoped that the *difference* in energy values between consecutive frames, or *delta* energies, will be more highly correlated with intelligibility than the absolute energy values. In this section, we define the mel-frequency filterbank used in our experiments; then, the details of the actual computation are discussed. Finally,

we discuss which filters are expected to be the most useful in assessing intelligibility.

4.2.1 Defining the Filterbank

For our filterbank, we use 20 triangular filters, whose center frequencies are linearly spaced on the mel scale, and range in frequency from 89 to 7016 Hz. Conversion from frequency f to mel m is done with the formula described in [46], viz.

$$m = 2595 \log_{10}(1 + \frac{f}{700}). \tag{4.1}$$

A plot showing the response of the filters is given in Figure 4.1. As can be seen from the figure, the entire frequency range from 0 to 8000 Hz is covered by one or more filters. The choice of 20 filters provides good coverage of the speech spectrum, with many of the filters lying within regions of the spectrum that are important to speech intelligibility (more on this in Section 4.2.3.)

4.2.2 Computing Delta Energy Parameters

A voice activity detection (VAD) algorithm [47] is employed to keep only the active speech sections. Then, a discrete Fourier transform and a shifting 50 ms Hamming window with 50% overlap are employed to get the magnitude spectrum. Choosing a longer window length means less contrast between heavily impaired and normal speech; a smaller window length means less frequency specificity in the filterbank. The chosen window length of 50 ms provides a compromise between the two.

The filterbank energy values are:

$$M_{fc}(i) = \sum_{k=0}^{L/2} |X(i,k)|^2 Y_{fc}(k), \tag{4.2}$$

where $X(i,k)$ denotes the k^{th} Fourier transform coefficient of the i^{th} windowed frame,

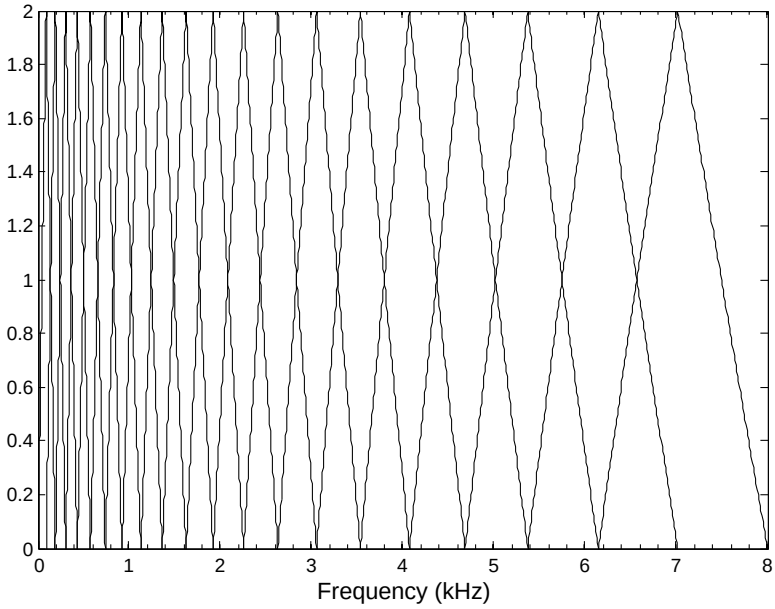

Figure 4.1: Mel-filterbank response

L is the length of the Fourier transform, fc is the center frequency of a triangular filter, and $Y_{fc}(k)$ the triangular filter coefficients. The per-frame energy values are then averaged:

$$E_{fc} = \frac{1}{N_w} \sum_{i=1}^{N_w} M_{fc}(i) \tag{4.3}$$

The delta mel-band energies can finally be defined as:

$$\delta_{fc} = \frac{1}{N_w - 1} \sum_{i=1}^{N_w - 1} |M_{fc}(i+1) - M_{fc}(i)|. \tag{4.4}$$

4.2.3 Relating the Filters to Physical Reality

As discussed, one of the symptoms of dysarthria is reduced articulation proficiency. Formant measures may offer clues to variations in intelligibility, but formant detectors are often error prone. Therefore, features consisting of average subband energies

and delta-energies of a mel-spaced triangular filterbank are proposed. In spastic and athetoid dysarthrias, vowels lying on the extremities of the F1-F2 space shift closer towards the center (see Section 2.4). The shift in formant frequencies should correspond to a decrease in energy of spectral components around 800-1000 Hz, which is the upper range of F1 for /a/, and 2000-2500 Hz, which is the upper range of the F2 value of /i/ [48].

The higher frequency subbands (> 3000 Hz) may also capture spectral changes due to hypernasality. Poor voice quality may also play a role in high frequency spectral content, since breathy voice can manifest as masking of higher formants [43]. As discussed in Chapter 2, the perceptual dimension of breathiness may be correlated with that of hoarseness [28], which was shown to be a prominent dysarthric speech characteristic.

4.3 LP Based Features

Linear prediction analysis is a tool with a long history of use in speech analysis, and is often the first step in performing many different types of analysis (see Chapter 2). In this thesis, we describe and examine three different quantities related to the LP filter shape and residuals of *voiced* dysarthric speech: the mean spectral flatness of the LP filter, the kurtosis of the spectral flatness of the LP filter, and what we term 'forward Itakura distance', which is the Itakura distance [49] between consecutive frames.

Traditionally, LP analysis was solely used to perform formant analysis on pathological voice signals. In [50], however, Davis described several quantities derived from a linear prediction (LP) filter. The derived quantities offered insights into the nature of the pathological voice signal. Since the work in [50], there has been some more

recent success in the use of LP based features in the classification of dysarthric speech samples (such as the work in [51]), but relatively little work has been done in understanding how such features can be used to estimate intelligibility. In this section, we discuss the details of the LPC filter computation, and a high-level explanation of the physical significance of each feature is given.

4.3.1 Computing the LP filter models

The chosen model order is $p = 20$, and the coefficients are computed using a sliding window of 0.05s and a window shift of 0.025s over voiced sections only. For our sampling rate of 16 kHz, the choice of pre-emphasis and p are a little larger than usual, but we are more concerned about the *shape* of the LP filter than we are about the precise placement of poles. To determine which sections are voiced, we employ the robust pitch tracking algorithm described in [52]. In general, all-pole filter models with a relatively low order may only be used to model *voiced* sections of speech; a more complex model, such as a pole-zero filter or auto-regressive moving average (ARMA) model, is needed for unvoiced speech [20].

4.3.2 Spectral Flatness of the LP Filter

To compute the spectral filter flatness (SFF), the magnitude frequency responses of the computed LP filters are sampled at regular intervals, and a measure of spectral flatness is computed from the samples. Spectral flatness is computed for each frame from the samples using the following formula:

$$SFF = \log\left(\frac{\frac{1}{N}\sum_{i=0}^{N-1}|H(\frac{i\pi}{N-1})|}{\sqrt[N]{\prod_{i=0}^{N-1}|H(\frac{i\pi}{N-1})|}}\right), \tag{4.5}$$

where N is the number of frequency samples. The SFF feature measures the log of the ratio of the arithmetic mean to geometric mean of the spectral envelope. Thus, SFF is non-negative, and its value increases with spectral non-flatness. Also, any multiplicative term applied to $H(z)$ has no effect on the spectral flatness measure. Note that SFF is a measure of the spectral flatness of the computed all-pole filter model, and not the inverse filter, as was originally defined in [50].

The value $N = 101$ was chosen because it gave a good representation of the filter shape without heavy computational costs. After the spectral flatness feature is calculated, we compute the kurtosis of the spectral filter flatness (SFF) values using the following formula:

$$kurt\_SFF = \frac{\frac{1}{M}\sum_{i=1}^{M}(SFF_i - \mu_{SFF})^4}{(\frac{1}{M}\sum_{i=1}^{M}(SFF_i - \mu_{SFF})^2)^2} - 3, \tag{4.6}$$

where μ_{SFF} is the mean spectral flatness, M is the number of spectral flatness values, and SFF_i is the spectral flatness of the i^{th} frame.

The kurtosis of a distribution quantifies its 'peakedness' or the size of the distribution's tails. A low kurtosis means the distribution has very few 'peaks' (e.g., a uniform distribution), whereas a high kurtosis value means the distribution is very 'peaky' (e.g., a Laplace distribution). The subtraction of three in the kurtosis computation is to force the kurtosis of a Gaussian distribution to zero.

Figure 4.2 shows the difference in the distribution of spectral flatness values of high and low intelligibility speakers. The top plot is a histogram of a male speaker with low intelligibility; the bottom plot is the spectral flatness data from a different male speaker with high intelligibility. In particular, we can note that the difference in the means is not obviously very large between the two distributions; rather, the difference in *peakedness* is the distinguishing quantity. Furthermore, we see that the

distinction holds across gender.

The kurtosis of the spectral flatness of the LP filter is a good measure of intelligibility because it represents the range of the possible filter shapes. Since the articulatory range of severely dysarthric speakers is typically limited, the range of filter flatness values should be correspondingly low.

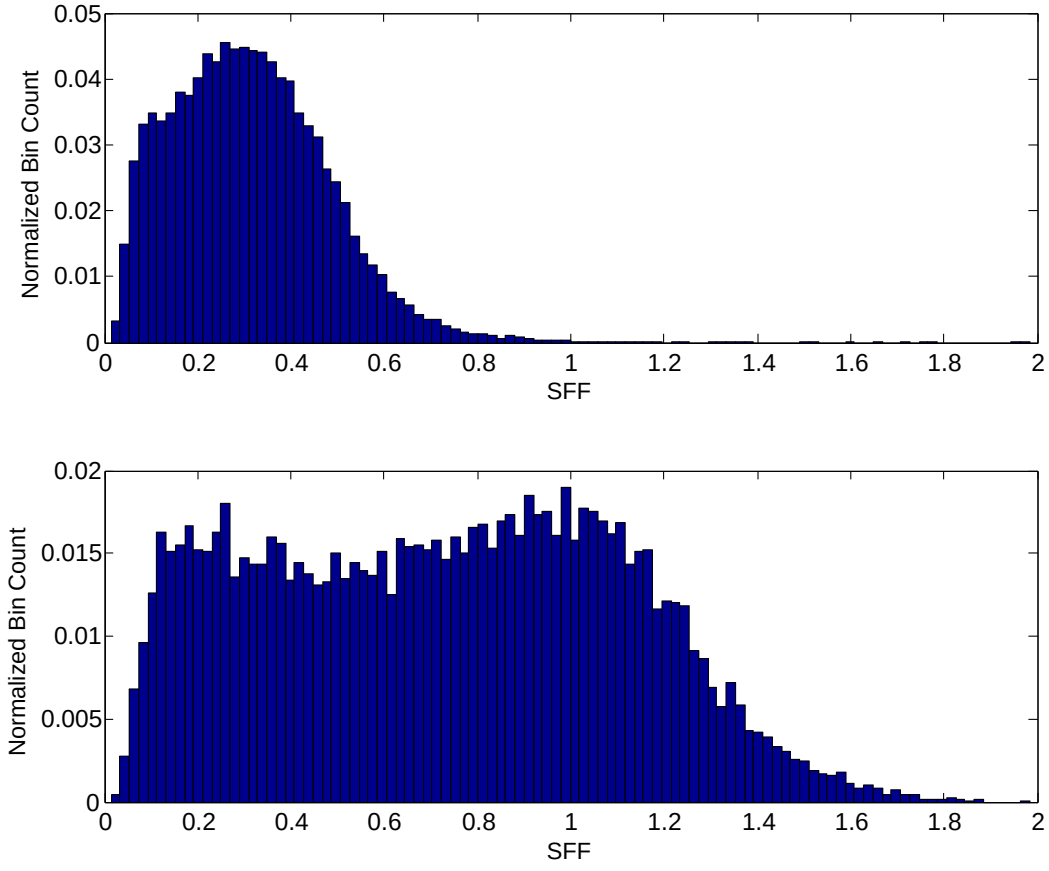

Figure 4.2: Comparison of SFF distributions for low (top) vs. high (bottom) intelligibility speakers

4.3.3 Forward Itakura Distance

As discussed, one of the common characteristics of the dysarthrias of cerebral palsy is a slow rate of speech and poor articulation. Consequently, it should be expected that for any given instant in time, the spectrum of a dysarthric speaker's speech should change more slowly than a healthy speaker's relative to the next frame of the person's speech. To quantify this idea, we employ a spectral distortion measure, Itakura distance (ID), to measure the ability of an LP model to predict the spectral shape of the *next frame*.

Itakura distance (ID) is a gain-independent spectral distortion measure [49]. The Itakura distance can be computed as a ratio of residual prediction energy between two sets of LP coefficients:

$$ID = \log\left(\frac{yRy^T}{xRx^T}\right) \tag{4.7}$$

where x and y are LP coefficients, R_i is the autocorrelation matrix from frame i, and equation 4.7 is equation (8) from [53]. A related measure, the Itakura-Saito distance (ISD), has been used in a reference-based pathological voice assessment method in [54]. The typical use for these types of distortion measures is to compare two signals to one another: the input signal, and a 'reference' signal. The distortion measure then serves as a degree of (dis)similarity between the two signals. In this instance, however, we desire a *blind* measure that can be used to estimate speech intelligibility. To accomplish this, we use the ID measure to quantify the similarity between two adjacent frames in our speech signal. The idea was further inspired by the predictive power measurement described in [55], which used a measure of the 'predictive power' of a wavelet based representation of a pathological voice signal to perform classification.

We first define the vector x as [1 $a_i(1)$ $a_i(2)$ .. $a_i(p)$], and the vector y as [1 $a_{i+1}(1)$ $a_{i+1}(2)$ .. $a_{i+1}(p)$], where $a_i(n)$ is the n^{th} LP coefficient from the i^{th} LP analysis frame. The matrix R_i is the autocorrelation matrix of data from frame i. It consists of values of the autocorrelation function computed at various time lags (see [19]). As noted in [53], the quantities xRx^T and yRy^T are prediction residuals when LP coefficients x and y are used to predict the data whose autocorrelation matrix is R. More specifically, xRx^T is the *energy* of the prediction residual. The Itakura distance as defined here is always greater than or equal to zero. This is because the LP coefficients x are chosen to minimize the error residual xRx^T. Therefore, $yRy^T \geq xRx^T$, which implies $\log\left(\frac{yRy^T}{xRx^T}\right) \geq 0$. If the computed LP model does not change very much between consecutive frames, the forward Itakura distance is expected to be close to zero. We can interpret the Itakura distortion measure used in this way as a quanity related to *how well the current LP model predicts the LP model of the next frame.*

There are two main perceptual characteristics of dysarthric speech that are related to the forward Itakura distance: distorted articulation, and slowed rate of speech. The reasoning is similar to that used to understand the usefulness of the second formant slope [7] as an intelligibility correlate. Since the rate of speech is often slowed in dysarthria, the rate at which the spectral content changes should be correspondingly slow as well. As discussed in Chapter 2, the shape of the LP filter corresponds to the configuration of the vocal tract. If the vocal tract is not able to take on a large extent of configurations due to articulation difficulties, the computed LP model will not likely change significantly between frames. It is important to note that since the formula for Itakura distance uses the LP model coefficients, it does not encode significant information about difference in the amount of energy between frames.

Rather, it only encodes differences in spectral shapes.

After the ID values are calculated, we then calculate the sample skewness of the distribution with the following formula:

$$skew\_ID = \frac{\frac{1}{M}\sum_{i=1}^{M}(ID_i - \mu_{ID})^3}{(\frac{1}{M}\sum_{i=1}^{M}(ID_i - \mu_{ID})^2)^{3/2}}, \tag{4.8}$$

where μ_{ID} is the mean forward Itakura distance, M is the number of forward Itakura distance values, and ID_i is the i^{th} forward Itakura distance. With our 'forward Itakura distance' measure, the skewness of the feature values was more useful than the sample mean as a correlate of speech intelligibility. The 'skewness' of a distribution measures its asymmetry. As an example, the skewness of a Gaussian distribution is 0, but the skewness of an exponential distribution is 2. The usefulness of the skewness statistic is in part to the definition of Itakura distortion. The use of the skewness statistic emphasizes the difference in the number of fast changing segments between frames of highly intelligible vs. low intelligibility speech. Kurtosis was not useful in combination with Itakura distortion, as the peakedness of the distributions did not vary for subjects with different intelligibility scores.

4.4 Summary

In this chapter, we have presented several features related to the properties of dysarthric speech. In particular, we presented several blind measures related to some of the deviant characteristics of dysarthric speech: breathiness/hoarseness, slow speaking rate, and distorted vowel space. These aspects were captured using parameters related to a mel-frequency based representation of the speech spectrum, as well as the linear predictive (LP) model of speech. These features represent a novel addition to the

growing list of blind dysarthric speech intelligibility measurements, in that their usefulness does not directly depend on any form of comparison or reference signal. In the next chapter, several experiments using P.563 and proposed features demonstrate their effectiveness in speaker dependent and independent intelligibility estimation models.

Chapter 5

Experimental Results

In previous chapters, we have outlined the P.563 standard for blind speech quality assessment, and have defined several features expected to be relevant to dysarthric speech intelligibility. In this chapter, we begin by first describing the Universal Access (UA) dysarthric speech database which is used in all subsequent experiments. Next, the MOS estimates produced by the six distortion classes are explored as potential intelligibility estimates. The P.563 and proposed features are then examined *individually* for their correlations with intelligibility. The features significantly correlated with subjective intelligibility are discussed, and put in physical context with dysarthric speech. Finally, two validation procedures are defined, and are used to test the candidate features and linear intelligibility model defined by (3.1).

5.1 The Universal Access Database

The Universal Access database [23] is a publicly available database containing single word recordings of dysarthric speakers with cerebral palsy (CP). There is a total of 15

speakers in the database. Eleven of the 15 subjects have spastic CP, two have athetoid CP, one has mixed CP, and one subject has an unknown subtype. Demographics and dysarthria subtypes for each speaker are given in Appendix A. The speech samples were recorded at a sampling rate of 16 kHz. Recording was done using an eight channel microphone array. Microphone six was used for all speakers (except speaker F04, where the data from microphone seven was found to contain less background noise). The vocabulary represented in the database consists of 300 uncommon English words, and 155 other words. The 300 uncommon words were selected from children's novels, and the 155 other words consist of spoken single digits (10 words), words from the radio alphabet (26 words), computer commands (19 words), and the 100 most common English words from the Brown Corpus of Written English. Each speaker has 765 files in the database, consisting of one utterance each of the 300 'uncommon' words, and three utterances of each of the 155 'other' words.

The UA database contains intelligibility scores for each speaker, in addition to speech data. The intelligibility scores were calculated as follows. First, a subset of 200 words in total was taken from all word categories; 25 of the 200 words were repeated to assess intra-speaker reliability. Five naive judges, all native speakers of American English between 18 and 40 years of age, were employed to transcribe what they heard for each word. Listening was done through headphones in a quiet environment. For each word, every listener was asked to rate how 'sure' they were of their transcription: 0 = not sure, 1 = somewhat sure, 2 = very sure. For all words labeled as very sure, average agreement rate (intra-speaker reliability) was found to be 91.64% over the five listeners. The five transcription accuracy percentages were averaged to give one intelligibility score per speaker. We label these average scores as $\{\phi_1, \phi_2, .., \phi_{15}\} = \Phi$,

where ϕ_i is the average intelligibility score for speaker i. They are used as our 'ground truth' intelligibility ratings, with which we measure the performance of our proposed estimators.

5.2 Feature Computation and Pre-Processing

The proposed features were computed by first normalizing the average power of the active speech sections to -26 dBov using the P.56 voltmeter [56]. The unit 'dbOv' denotes a power level measured relative to the power level where clipping starts to occur. The relevant equations from Chapter 4 were then applied to the normalized data. To obtain the internal P.563 features and speech quality estimates, the 765 utterances per speaker were first downsampled to 8 kHz, and then processed by P.563. The P.563 standard uses its own implementation of the P.56 standard where needed, so P.56 normalization did not need to be applied a priori.

5.3 P.563 MOS estimates

For each of the six distortion classes, an average MOS estimate is calculated over each speaker's 765 utterances; Pearson's correlation coefficient (r) is then computed between the averages and Φ. These MOS estimates are computed by P.563 and are distinct from the intelligibility scores, Φ. Pearson's correlation coefficient is a measure of the 'linear' relationship between two variables. It is defined for any two equal length, indexed data sets $\{x(n)\}$ and $\{y(n)\}$ as:

$$r = \frac{\sum_{k=1}^{K}(x(k)-\bar{x})(y(k)-\bar{y})}{\sqrt{\sum_{k=1}^{K}(x(n)-\bar{x})^2 \sum_{j=1}^{K}(y(j)-\bar{y})^2}}, \tag{5.1}$$

where $\bar{x}$ and $\bar{y}$ are the mean values for $x(n)$ and $y(n)$, respectively, and K is the length of both data sets. A two-tailed Student's t-test is also used to determine the significance of the obtained correlation result, and gives the likelihood of the null hypothesis ($r = 0$).

The result from applying (5.1) to the average MOS estimates and Φ is that none of the six classes of MOS estimates are significantly correlated ($p < 0.05$) with Φ. These results are corroborated by the results reported by Maier [30]. The work by Maier, however, used the 3-SQM algorithm [57] and only investigated the final MOS estimate, and not the MOS estimate from each distortion class.

5.4 P.563 and Proposed Feature Correlations

Before combining features in our linear intelligibility model (5.1), we first survey each feature's correlation with respect to intelligibility. Average feature values are computed over each speaker's 765 utterances; then, Pearson's and Spearman's correlation coefficients, denoted r and ρ, are calculated between average feature variables and Φ (the subjective intelligibility scores). Spearman's correlation ρ is also used and is defined in a way similar to equation (5.1), but uses the ranks of the values instead of the values. Given the limited amount of available data, Spearman's correlation values may be more informative overall than Pearson's correlation values.

By computing the correlations between the features and Φ values, we can obtain an idea as to which features are likely to be selected in any feature selection process. It should be stressed that examining any feature *by itself* is not enough to understand its potential usefulness. When combined, some features may offer complementary

information to other features. Thus, combinations of features may yield better results than any feature individually. Table 5.1 lists the features that have significant Pearson's correlation ($p < 0.05$) with Φ. The P.563 feature names (FrameRepeats, LPCcurt, LPCskew, LPCskewAbs, SpecLevelRange, SpecLevelDev) are taken from the P.563 documentation.

Table 5.1: Correlation Values of P.563 and Proposed Features

Feature Name	Pearson Correlation r	Spearman Correlation ρ
kurt_SFF	-0.89	-0.94
skew_ID	0.82	0.83
LPCskew	-0.78	-0.77
LPCskewAbs	0.76	0.71
LPCcurt	0.75	0.70
mean_SFF	0.66	0.60
δ_{3535}	0.66	0.73
δ_{3056}	0.62	0.69
SpecLevelRange	-0.61	-0.70
δ_{738}	0.60	0.68
SpecLevelDev	-0.60	-0.55
δ_{4075}	0.59	0.60
δ_{2631}	0.57	0.65
FrameRepeats	-0.56	-0.52
δ_{1624}	0.54	0.54
δ_{1361}	0.52	0.56
δ_{1920}	0.52	0.63

It is important to note that the P.563 standard includes features that measure various aspects of signal noise. Several of these features correlated significantly with subjective intelligibility scores, but were excluded from consideration because of their lack of relevance to the problem at hand. Many P.563 features had more to do with the signal recording conditions than the intelligibility of dysarthric speech. Excluded features include parameters relating to signal to noise ratios and background noise.

The high correlation values in Table 5.1 (up to 0.94) show that many internal P.563 and proposed features have a strong relationship with intelligibility. However, it is not immediately clear why the listed features are relevant to the estimation of dysarthric speech intelligibility. A more thorough discussion is needed to understand why, in particular, the δ_{fc} and P.563 features listed are significant. In the following sections, the δ_{fc} and the six listed P.563 features are put into context with the acoustic and perceptual properties of dysarthric speech.

5.4.1 Delta Energy Features

The δ_{fc} features listed in Table 5.1 corroborate with the discussion in Section 4.2.3. The listed δ_{fc} features can then be grouped in accordance with the acoustic characteristics of dysarthric speech discussed previously. Features δ_{1361}, δ_{1624} and δ_{1920} capture differences around the critically important 1500 Hz region of speech. The delta features in the 3000-4000 Hz range (δ_{3535}, δ_{3056} and δ_{4075}) capture reductions in energy due to a reduction in upper formant energy related to hoarseness and hypernasality. Finally, the δ_{738} and δ_{2631} features lie on the extremities of the F1-F2 space.

5.4.2 LPC Coefficient Statistics

The LPCcurt, LPCskew and LPCskewAbs features quantify changes in the statistics of the LPC coefficient distribution. Their values are expected to lie within a certain range for normal speech. The P.563 standard maps deviations from these ranges to low quality scores. More specifically, LPCcurt and LPCskew are the kurtosis and skewness of the LPC coefficient distributions, respectively. They are computed in a manner similar to equations 4.6 and 4.8. Unlike the LPC features described in Chapter 4,

LPCcurt, LPCskew and LPCskewAbs are computed over 0.032 msec frames, including voiced and unvoiced segments. The per-frame values are then averaged to give a final value. The LPCskewAbs feature is the absolute value of LPCskew.

5.4.3 Spectral Amplitude Features

The SpecLevelDev and SpecLevelRange reflect aspects of the spectral *amplitude* characteristics of speech segments (estimated from a short term Fourier transform). Both parameters are computed over every active speech frame and then averaged to give a final value. These features differ from the spectral flatness measures introduced in Chapter 4 as they are computed over both voiced *and* unvoiced speech. Furthermore, since P.563 is designed for narrowband speech, only frequencies in the 0-4000 Hz range are used. The proposed spectral flatness features use the entire available (0-8000 Hz) spectrum. More details on both features are given below.

5.4.3.1 SpecLevelDev

The SpecLevelDev feature is the average standard deviation of the spectral amplitudes in the 1000-2000 Hz range, and so can be regarded as a type of spectral flatness measure. The larger SpecLevelDev is, the larger the average variance in amplitudes in that range. Note that this feature is different from the spectral flatness measures described previously because SpecLevelDev only uses spectral amplitudes over the specified 1 kHz range, whereas the proposed features use the full 8 kHz bandwidth. In severely dysarthric speech, many speech segments that should be unvoiced are unnaturally voiced [5]. This means that there should be a greater proportion of sections with strong periodicity, and as a result, frequent harmonics in the 1000-2000

Hz range. These harmonics are the likely cause for the larger standard deviation in the 1000-2000 Hz range for low intelligibility speakers.

5.4.3.2 SpecLevelRange

SpecLevelRange is calculated by computing the average difference between the 85^{th} and 20^{th} percentiles (P_{85} and P_{20}) of the spectral amplitude distribution. The relevance of SpecLevelRange is in part due to the hypernasal quality of dysarthric speech. Many observed dysarthric speech samples had one prominent low frequency formant (see Figure 2.2 in Chapter 2). The increased ratio of voiced to unvoiced speech segments in low intelligibility speech may also impact this feature in a manner similar to SpecLevelDev.

5.4.4 FrameRepeats

The FrameRepeats feature is computed using a local search algorithm for every frame, and reflects the number of adjacent frames that are strongly correlated. There are two likely perceptual reasons for the usefulness of this feature: reduced speaking rate, and monopitch (reduced variability of fundamental frequency, $f0$). Since speaking rate is slowed, adjacent frames are more likely to be correlated. Furthermore, pitch remains relatively consistent in nearby frames, further increasing cross-correlation values. These two characteristics explain why the correlation between the FrameRepeats feature and Φ is negative (i.e., as intelligibility decreases, the number of 'repeated' adjacent frames decreases).

5.5 Model Validation

In this section, we explore the usefulness of weighted *combinations* of P.563 and proposed features. Three different experiments are undertaken; two experiments employ a speaker dependent setup with changing vocabulary, and the third experiment examining results from a leave-one-speaker-out cross-validation (LOSO).

5.5.1 Experimental Setup

For every iteration of feature selection, the data is first split into training and testing data. Using the training data, we wish to choose N features from our feature pool. The first step in our feature selection procedure is to perform an initial 'pre-screening' using (5.1). Any feature not significantly correlated with Φ ($p > 0.05$) is removed. With the remaining features, we employ the forward variant of the sequential feature selection (SFS) algorithm [58] to select up to N of the remaining features. The order in which the features are selected is also recorded. Let N_f be the number of features after pre-screening, J be our objective function, $R(f)$ be the feature selection order (or rank) for feature f, $\mathcal{F} = \{f_i | i = 1..N_f\}$ be the set of candidate features after pre-screening, and let $\mathcal{G}_n = \{g_j | j = 1..n\}$ be the set of selected features at iteration n. The algorithm works as follows:

1. Initialize the variables: $n \leftarrow 0$, $\mathcal{G}_n \leftarrow \{\emptyset\}$
2. Pick a new feature that maximizes J: $g_{new} \leftarrow \arg\max_{f \in \mathcal{F} - \mathcal{G}_n} J(f + \mathcal{G}_n)$
3. Update the feature set: $\mathcal{G}_{n+1} \leftarrow \mathcal{G}_n + g_{new}$
4. Increment n: $n \leftarrow n + 1$

5. Record the rank of the selected feature: $R(g_{new}) \leftarrow n$

6. Termination check: If $n < N_f$ and $n < N$, goto (2); otherwise, goto (7)

7. Assign rank of $n + 1$ to unselected features: $R(\mathcal{F} - \mathcal{G}_n) \leftarrow n + 1$

The algorithm is slightly different than the standard sequential forward feature selection algorithm because of the added termination condition $n \geq N_f$. This extra condition is used because of the feature screening; when N is large, there *may* not be enough features left in the feature pool to continue for n sufficiently large.

After completing M cross-validation trials, where each trial selects N features, a method is needed for determining the best performing features from the results. This is because each trial may select a different sequence of features. One method of doing this is by computing the average feature rank (AFR) . AFR can be expressed as

$$AFR_f = \frac{1}{M} \sum_{j=1}^{M} R_f(j), \tag{5.2}$$

where $R_f(j)$ is the rank of feature f for trial j. The features with the N smallest AFR values then become our final choice for $\mathcal{F}$.

5.5.2 Speaker Dependent Validation

Ideally, any features employed in the blind estimation of speech intelligibility should be insensitive to the choice of vocabulary. One way of assessing this sensitivity is by training the intelligibility model using vocabulary that is disjoint from the vocabulary used to test the model. The first step is to partition the available vocabulary into two. Each partition would consist of utterances from half of the 455 word vocabulary available in the UA database. Utterances from one vocabulary partition become the

'training' data, and the 'test' data would consist of the remaining utterances. Repeating this process over M random partitionings allows examination of the sensitivity of the model to the choice of vocabulary.

Each of the M feature selection trials proceeds as follows. First, 233 of the 455 word vocabulary are randomly selected (including 150 uncommon words). Utterances corresponding to the selected vocabulary serve as data for feature selection and model training. The feature selection process, described in 5.5.1, can then be used to select N features. Pearson's correlation r serves as the objective function, J. The selected features constitute a temporary feature set $\mathcal{F}_i$ used to form the linear model for trial i, with model coefficients obtained using least squares regression. For model validation, the remaining unseen data (i.e., not used for feature selection and model training) is applied to the trained model to compute one ϕ_{est} value per speaker using (3.1). Model performance is quantified using r, Spearman's rank correlation (ρ), and root mean square error (ϵ) between ϕ_{est} values and Φ. Root mean square error (RMSE) is given by

$$\epsilon = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\phi_{est}(i) - \phi_i)^2}, \tag{5.3}$$

where $\phi_{est}(i)$ is the intelligibility estimate for speaker i, and ϕ_i is the 'true' subjective intelligibility score. The RMSE value can be interpreted as the square root of the variance of the error residuals.

Since feature selection is repeated M times for a given value of N, we average the M values of r, ρ and ϵ and denote the results $\bar{r}$, $\bar{\rho}$ and $\bar{\epsilon}$ to give measures of average model performance.

5.5.3 Spastic Dysarthric Speakers Only

In our work published in [59], we test the capability of a speaker dependent intelligibility model using only the spastic dysarthric speakers listed in Table A.1 in Appendix A (omitting speaker M09). Using this subset of the 15 UA speakers enables comparisons to the works in [5, 6]. For use as a baseline, we use the feature selection process described above using only P.563. The baseline can then be compared with model performance when features are selected from P.563 and δ_{fc} (proposed) features.

The first task is finding a suitable value for N, the number of included features. To do this, we vary N from one to 20 and examine $\bar{r}$ as N increases; for each value of N, we compute $\bar{r}$ using $M = 200$ feature selection trials. Two feature pools are tested: the 43 P.563 features (our baseline), and the 43 P.563 features with the 20 δ_{fc} features (the 'composite' estimator). Performance begins to degrade severely for $N > 6$; therefore, performance results versus number of features, up to $N = 6$, are plotted in Figure 5.1. We choose $N = 3$ as it offers near optimal performance with a small number of features. Average performance values $\bar{r}$, $\bar{\rho}$ and $\bar{\epsilon}$ for $N = 3$ are listed in Table 5.3, alongside similar results from [5, 6]. The resultant intelligibility estimates far outperform the MOS estimates as subjective intelligibility correlates (see Section 5.3).

The experimental procedures in [5, 6] are similar to the procedure used here, with the principal difference being how the train/test partitoning was done. In both [5] and [6], the training/testing data partitioning was only done once, and involved a 60-40 train/test split instead of the 50-50 train/test split used in this work. The f_{raw} intelligibility scores listed were computed using feature values and a least-squares fit (as in this work). The $f_{class,map}$ scores improved upon the f_{raw} scores through the

use of two additional steps. The first involved using two estimator functions: one for mid-low intelligibility speakers ($< 50\%$), and the other for mid-high intelligibility speakers ($> 50\%$). The second step used a 3^{rd} order monotonic polynomial mapping, as per the P.563 recommendation [36], for better mapping of objective scores onto subjective ones.

Figure 5.1 shows that combining proposed δ_{fc} features with P.563 features provides better results than using P.563 features alone for any $N > 1$. Table 5.3 shows that when proposed δ_{fc} features are combined with P.563 features in a 'composite' feature pool, performance is improved substantially over the baseline. The statistical significance of the improvements over the baseline were investigated using a two-tailed t-test. All three measures (r, ρ, ϵ) were found to have significantly increased $(p < 0.01)$. Furthermore, the composite estimator yields results comparable to the results in [5, 6], but using a slightly smaller percentage of available data for training (50% vs. 60%). Additionally, the proposed composite estimator uses only 3 features, whereas the works in [5, 6] use 5 features to obtain the performance values listed in Table 5.3.

To summarize feature selection results, AFR is applied to the results from the $M = 200$ feature selection trials ($N = 3$). The top three selected features, ordered by AFR, are listed in Table 5.2. The set of listed features can be compared with the features listed in 5.1.

As expected, a few of the features listed in Table 5.1 appear in Table 5.2. The features common to both tables are LPCcurt, SpecLvlRange and δ_{1361}. The two features not listed in Table 5.1, however, can be understood along the same lines as the features discussed in Section 5.4. The CepCurt feature is the average kurtosis of

Feature Pool	AFR	Feature Name
P.563	1.00	LPCcurt
	2.70	CepCurt
	3.34	SpecLvlRange
P.563 + δ_{fc}	1.00	LPCcurt
	2.04	δ_{992}
	3.03	δ_{1361}

Table 5.2: Top ranked features by AFR: Spastic dysarthria only, speaker dependent validation

the cepstrum, computed over speech segments. Its usefulness is along the same lines as the LPCcurt and LPCskew features. The δ_{992} feature is on the extreme end of the first formant (F1) range.

The curves in Figure 5.1 show that while one feature alone can be used to predict intelligibility, when *several features are combined*, the intelligibility estimate improves in a statistically significant way (see discussion above). The baseline estimator combines two features related to deviant speech statistics (CepCurt and LPCcurt), with another (SpecLevelRange) related to the hypernasal/unnaturally voiced aspect of dysarthric speech. The composite estimator combines one speech statistic (LPCcurt) with two features related to the slow speaking rate and disordered articulation of dysarthria (δ_{992} and δ_{1361}).

The discrepancies between the features listed in Table 5.2 and 5.1 suggest that the results obtained using only spastic dysarthric speakers may not be applicable to the rest of the dataset. Features are desired that can estimate intelligibility across subtypes. To address this issue, all 15 speakers are used to validate the model; the results are described in the next section. Additionally, the LP features proposed in Chapter 4 are explored.

Estimator	Performance Measure		
	$\bar{r}$	$\bar{\rho}$	$\bar{\epsilon}$ (%)
P.563 re-mapping (this thesis)	0.95	0.93	11.0
Composite (this thesis)	0.98	0.95	7.1
f_{raw} [5]	0.93	0.86	18.6
$f_{class,map}$ [5]	0.97	0.96	8.6
$f_{proposed}$ [6]	0.85	0.87	N/A

Table 5.3: Mean performance measures: Spastic dysarthria only, speaker dependent validation

5.5.4 All Speakers

In a separate experiment, all 15 speakers from the UA database with the speaker dependent validation method are used. Again, a baseline is created by using only P.563 features in the feature pool. The baseline results can then be compared to results obtained when all available features are employed (the 'composite' estimator). The experimental procedure from the previous section is re-used, with N varying from one to 20, and $M = 200$ feature selection trials for each N. The feature selection results are plotted in Figure 5.2 up to $N = 10$ (performance continues to degrade for $N > 10$).

Examining the plot in Figure 5.2, we notice immediately that the composite estimator outperforms the baseline for all N. The best feature selection performance for the baseline is achieved at $N = 4$, and at $N = 5$ for the composite estimator. The top four and top five features from the baseline and composite estimator experiments, respectively, are given in Table 5.4. Performance results are given in Table 5.5. As before, the statistical significance of the improvements over the baseline were investigated using a two-tailed t-test. All three measures (r, ρ, ϵ) were found to have significantly improved in the composite estimator ($p < 0.01$).

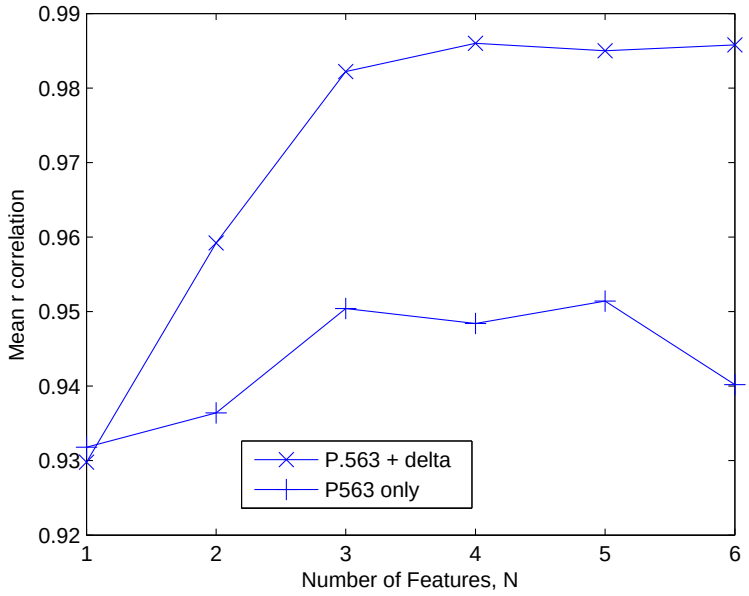

Figure 5.1: Average SFS results as a function of N; spastic dysarthric speakers only, speaker dependent validation

Comparing Tables 5.3 and 5.5, it is obvious that including more speakers with new dysarthric subtypes has caused a performance drop. Because of the limited amount of available data, it is difficult to ascertain the cause of the decrease in performance precisely. However, the performance drop is likely due to the inclusion of new dysarthric subtypes that vary along acoustic dimensions not accounted for by the regressor.

5.6 Speaker Independent Validation

In a clinical setting, a system such as the one proposed in this thesis would be used to predict the intelligibility of a speaker that had not been used to train the system.

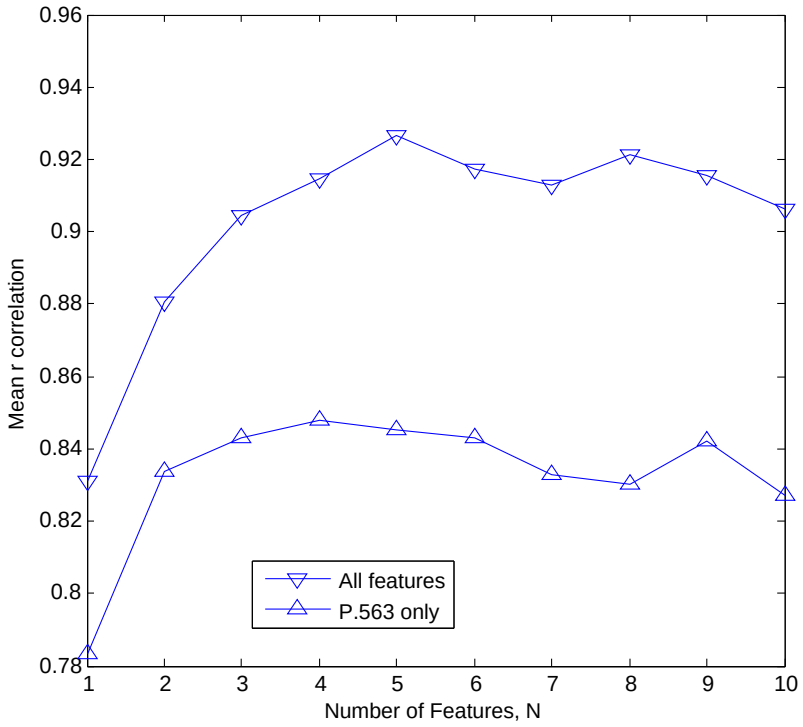

Figure 5.2: Average SFS results as a function of N; all speakers, speaker dependent validation

Therefore, an important test is to see how the system behaves when it is used to estimate the intelligibility of an 'unseen' speaker. To accomplish this, we use a leave-one-speaker-out cross validation technique, and again employ the sequential feature selection algorithm with ρ serving as our objective function J instead of r. For each trial, one speaker is left out as the 'test' speaker, and the other 14 speakers are used to train the model (the 'training' speakers). All available utterances for the 14 'training' speakers are used in training. As before, r, ρ and ϵ give measures of performance.

Feature Pool	Feature Name	AFR
P.563	LPCskew	1.0
	SpecLevelDev	2.1
	FrameRepeats	4.1
	LPCcurt	4.2
All Features	kurt_SFF	1.5
	skew_ID	3.8
	δ_{738}	4.3
	LPCskew	4.4
	FrameRepeats	4.8

Table 5.4: Top ranked features by AFR; all speakers, speaker dependent validation

Estimator	Performance Measure		
	$\bar{r}$	$\bar{\rho}$	$\bar{\epsilon}$ (%)
P.563 baseline	0.85	0.87	18.3
All features (composite)	0.92	0.92	14.0

Table 5.5: Mean performance measures; all speakers, speaker dependent validation

The choice of Spearman's vs. Pearson's correlation coefficient as objective criteria is important; using Spearman's correlation chooses features that consistently rank subjects according to intelligibility. If the number of speakers is small, as in our case, outliers may have a large effect on correlation values. Using Spearman's correlation avoids this issue.

The number of features N is varied from one to 20 and the results are plotted in Figure 5.3 up to $N = 10$. From the Figure, we can see that the best Spearman's correlation is obtained at both $N = 1$ and $N = 6$. To decide which value of N to choose, both Pearson's and Spearman's correlation results are plotted. The value $N = 5$ provides a good compromise between Pearson's and Spearman's correlation results.

Average feature rankings given in Table 5.6 show that the proposed feature,

kurt_SFF, is the top-ranked feature. The next three top ranked features are δ_{fc} features, with a large gap noticeable for the AFR value of kurt_SFF and δ_{1624}. This seems likely to be because the δ_{fc} features encode a large amount of information regarding individual speaker characteristics (such as fundamental frequency, formant locations, etc.). Despite the large AFR gap, inclusion of several δ_{fc} improves performance.

Performance measures $\bar{r}$, $\bar{\rho}$ and $\bar{\epsilon}$ for $N = 5$ are 0.91, 0.94 and 14.3%. These performance values are comparable with the speaker dependent results presented in Table 5.5. This suggests that the error induced by changing the train and test vocabulary is not trivial, and is about the same magnitude as the error obtained from testing with an unseen speaker. Thus, when considering the performance of a blind assessment system, the impact on the system to the choice of speech material should be considered.

The baseline (P.563 features only) experiment is omitted from Figure 5.3 because results were extremely poor for any choice of objective function and worsened for increasing N. The best performance obtainable with only P.563 features was at $N = 1$; $\bar{r}$, $\bar{\rho}$ and $\bar{\epsilon}$ values were 0.57, 0.45 and 29.1%, respectively. The top selected feature was LPCskew.

Ideally, features used in the assessment of dysarthric speech intelligibility should be able to estimate the intelligibility of a new, unseen speaker, as well as be robust to the choice of vocabulary. Features appearing in tables 5.4 and 5.6 should be robust to both types of variability. The only feature that meets this criteria is kurt_SFF. This feature is therefore a strong potential candidate for further investigation into dysarthric speech intelligibility.

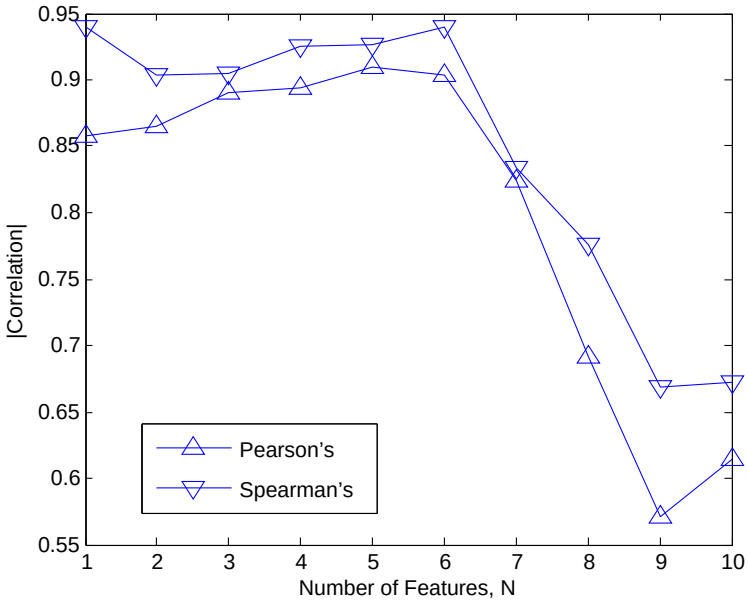

Figure 5.3: SFS results as a function of N; all speakers, speaker independent

5.7 Summary

The class-specific MOS estimates, internal P.563 and proposed features were first examined as intelligibility correlates. The MOS estimates were found to be insignificantly correlated with Φ, but many of the P.563 and proposed features were strongly correlated with subjective intelligibility scores (up to 0.94). Afterwards, P.563 and proposed features were combined in both speaker dependent and independent validation methods. Two different speaker dependent experiments were conducted and analyzed: one using only spastic dysarthric speakers, and the other using all 15 available speakers. The speaker dependent experiments using only spastic dysarthric speakers was shown to produce better results than those reported in the literature

Feature Name	AFR
kurt_SFF	1.0
δ_{1624}	4.5
δ_{3535}	5.1
δ_{4075}	5.3
SpecLevelRange	5.3

Table 5.6: Top ranked features by AFR: All speakers, speaker independent validation

using the same dataset. Lastly, a speaker independent validation procedure showed that the proposed model and features could be used to estimate the intelligibility of an unseen speaker. In all of the aforementioned experiments, intelligibility estimators consisting of both proposed *and* P.563 features were shown to outperform estimators consisting of only P.563 features.

Chapter 6

Summary and Conclusions

6.1 Conclusion

Traditionally, subjective methods were used in the assessment of dysarthric speech intelligibility. In order to cut costs and improve repeatability, objective intelligibility methods have begun to complement existing assessment methods. Currently, there is a sparsity of blind dysarthric assessment tools in the literature. This thesis has provided a number of novel features in an attempt to fill this gap.

Chapter 3 discussed the use of blind speech quality features for potential use in the assessment of dysarthric speech intelligibility. In particular, the P.563 standard for blind speech quality assessment was discussed in detail. In Chapter 4, several features related to the perceptual qualities of dysarthric speech (particularly spastic and hyperkinetic) are proposed. The results from several experiments were explored in Chapter 5; experiments included speaker-dependent and speaker-independent validation methods. One of the proposed features, kurt_SFF (kurtosis of the spectral flatness of the LP filter), was shown to perform well in both speaker-dependent *and*

speaker-independent validation methods. Examined blind speech quality features were shown to contain information complementary to proposed features. Reported results were shown to be comparable to those in the literature on the *same data set*.

6.2 Future Work

There is a continuing need for dysarthric speech assessment features and systems. In order to provide reliable assistance in a clinical setting, more work needs to be done. Therefore, there are many available avenues for future research. Among them are:

1. Analysis of spontaneous speech: Currently, the system studied herein is designed and tested to work on word intelligibility scores. Since the approach taken in this thesis is the 'blind' approach, the system could potentially be extended to work on spontaneously generated speech. This is a clinically useful task, because 'everyday' intelligibility is important to the wellbeing of dysarthric speakers.

2. Inclusion of other dysarthric subtypes: Only a few dysarthric subtypes were studied here (mixed, spastic, hyperkinetc). Including speech data from other dysarthric subtypes would likely require more features, but would simultaneously increase the usefulness of the designed system.

3. Objective quality measurement of dysarthric speech: Results suggest that some important components of objective speech quality measurements are affected when the speaker is dysarthric. An interesting question to ask is whether an objective speech quality algorithm (such as P.563) would still be useful in modeling subjective quality scores when the input speech is dysarthric.

Bibliography

[1] P. Enderby and R. Palmer, *Frenchay Dysarthria Assessment*. Austin, TX: Pro-Ed, 1983.

[2] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Noth, "PEAKS-A system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[3] L. Ferrier, H. Shane, H. Ballard, T. Carpenter, and A. Benoit, "Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition," *Augmentative and Alternative Communication*, vol. 11, no. 3, pp. 165–175, 1995.

[4] P. Doyle, H. Leeper, A. Kotler, N. Thomas-Stonell, C. O'Neill, M. Dylke, and K. Rolls, "Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility," *Journal of Rehabilitation Research and Development*, vol. 34, no. 3, pp. 309–316, 1997.

[5] T. H. Falk, W.-Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, 2011.

[6] T. H. Falk, R. Hummel, and W.-Y. Chan, "Quantifying perturbations in temporal dynamics for automated assessment of spastic dysarthric speech intelligibility," in *ICASSP*, 2011, pp. 4480–4483.

[7] R. Kent, J. Kent, G. Weismer, R. Martin, R. Sufit, B. Brooks, and J. Rosenbek, "Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects," *Clinical Linguistics & Phonetics*, vol. 3, no. 4, pp. 347–358, 1989.

[8] K. Hustad, "Estimating the intelligibility of speakers with dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 3, p. 217, 2006.

[9] F. Darley, A. Aronson, and J. Brown, "Differential diagnostic patterns of dysarthria," *J. Speech Lang. Hear. Res.*, vol. 12, no. 2, p. 246, 1969.

[10] K. Bunton, R. Kent, J. Duffy, J. Rosenbek, and J. Kent, "Listener agreement for auditory-perceptual ratings of dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 6, p. 1481, 2007.

[11] M. De Bodt, M. Hernández-Díaz Huici, and P. Van De Heyning, "Intelligibility as a linear combination of dimensions in dysarthric speech," *Journal of Communication Disorders*, vol. 35, no. 3, pp. 283–292, 2002.

[12] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* New Jersey: Prentice-Hall, 2001.

[13] K. Krigger, "Cerebral palsy: an overview," *American Family Physician*, vol. 73, no. 1, pp. 91–100, 2006.

[14] W. Hardcastle and J. Laver, *The handbook of phonetic sciences*. Wiley-Blackwell, 1999, vol. 5.

[15] D. Freed, *Motor speech disorders: Diagnosis and treatment*. Singular Pub Group, 2000.

[16] L. Platt, G. Andrews, and P. Howie, "Dysarthria of adult cerebral palsy: Ii. phonemic analysis of articulation errors," *Journal of Speech and Hearing Research*, vol. 23, no. 1, p. 41, 1980.

[17] B. Ansel and R. Kent, "Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy," *Journal of Speech and Hearing Research*, vol. 35, no. 2, p. 296, 1992.

[18] B. Zyski and B. Weisiger, "Identification of dysarthria types based on perceptual analysis," *Journal of Communication Disorders*, vol. 20, no. 5, pp. 367–378, 1987.

[19] J. Markel and J. Gray, *Linear Prediction of Speech Signals*. Springer, Berlin, 1976.

[20] X. Huang, A. Acero, H. Hon *et al.*, *Spoken language processing*. Prentice Hall PTR New Jersey, 2001.

[21] J. Deller Jr, J. Proakis, and J. Hansen, *Discrete Time Processing of Speech Signals*. Prentice Hall PTR Upper Saddle River, NJ, USA, 1993.

[22] H. Liu, F. Tsao, and P. Kuhl, "The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy," *J. Acoust. Soc. Am.*, vol. 117, pp. 3879–3889, 2005.

[23] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proceedings of the International Conference on Spoken Language Processing*, 2008, pp. 1741–1744.

[24] M. Schroeder, *Computer speech*. Springer, 1999.

[25] K. Schlenck, R. Bettrich, and K. Willmes, "Aspects of disturbed prosody in dysarthria," *Clinical Linguistics & Phonetics*, vol. 7, no. 2, pp. 119–128, 1993.

[26] S. LeGendre, J. Liss, and A. Lotto, "Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra." *The Journal of the Acoustical Society of America*, vol. 125, p. 2530, 2009.

[27] V. Parsa and D. Jamieson, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech," *Journal of Speech, Language, and hearing research*, vol. 44, no. 2, p. 327, 2001.

[28] R. Prosek, A. Montgomery, B. Walden, and D. Hawkins, "An evaluation of residue features as correlates of voice disorders* 1," *Journal of communication disorders*, vol. 20, no. 2, pp. 105–117, 1987.

[29] M. McNeil, *Clinical management of sensorimotor speech disorders*. Thieme Medical Pub, 2008.

[30] A. Maier, "Speech of children with cleft lip and palate: Automatic assessment," Ph.D. dissertation, Technische Fakultät der Universität Erlangen-Nürnberg, January 2009.

[31] J. Carmichael and P. Green, "Revisiting dysarthria assessment intelligibility metrics," in *Proceedings of the International Conference on Spoken Language Processing*, 2004, pp. 742–745.

[32] C. Middag, J. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[33] R. Kent, G. Weismer, J. Kent, and J. Rosenbek, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, p. 482, 1989.

[34] C. Middag, Y. Saeys, and J. Martens, "Towards an asr-free objective analysis of pathological speech," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[35] T. Bocklet, T. Haderlein, F. Hönig, F. Rosanowski, and E. Nöth, "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.

[36] ITU-T P.563, "Single-ended method for objective speech quality assessment in narrowband telephony applications," Intl. Telecom. Union, 2004, geneva, Switzerland.

[37] A. Ekman and W. Kleijn, "Improving quality prediction accuracy of P. 563 for noise suppression," in *Proceedings of the Internationalal Workshop on Acoustic Echo and Noise Control*, 2008.

[38] T. Falk, S. Möller, V. Karaiskos, and S. King, "Improving instrumental quality prediction performance for the blizzard challenge," in *Proc. Blizzard Challenge Text-to-Speech Workshop*, vol. 5, 2008.

[39] L. Malfait, J. Berger, and M. Kastner, "P. 563 - The ITU-T Standard for Single-Ended Speech Quality Assessment," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 1924–1934, 2006.

[40] S. Kay, "Fundamentals of statistical signal processing, volume i: Estimation theory," 1993.

[41] R. Kent, G. Weismer, J. Kent, and J. Rosenbek, "Toward phonetic intelligibility testing in dysarthria." *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–99, 1989.

[42] G. Turner, K. Tjaden, and G. Weismer, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," *Journal of Speech and Hearing Research*, vol. 38, no. 5, p. 1001, 1995.

[43] E. Castillo-Guerra and A. Ruiz, "Automatic modeling of acoustic perception of breathiness in pathological voices," *Biomedical Engineering, IEEE Transactions on*, vol. 56, no. 4, pp. 932 –940, april 2009.

[44] S. Greenberg, T. Arai, and R. Silipo, "Speech intelligibility derived from exceedingly sparse spectral information," in *Proc. Int. Conf. Speech and Lang. Proc.*, 1998, pp. 2803–2806.

[45] R. Warren, K. Riener, J. Bashford, and B. Brubaker, "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Attention, Perception, & Psychophysics*, vol. 57, no. 2, pp. 175–182, 1995.

[46] D. O'Shaughnessy, *Speech communications: human and machine*. Universities Press, 1987.

[47] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[48] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," *Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[49] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice hall Englewood Cliffs, New Jersey, 1993, vol. 103.

[50] S. Davis, "Acoustic characteristics of normal and pathological voices," *Speech and Language: Advances in Basic Research and Practice*, vol. 1, pp. 271–335, 1979.

[51] V. Parsa and D. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 2, p. 469, 2000.

[52] D. Talkin, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[53] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 1, pp. 67 – 72, Feb. 1975.

[54] L. Gu, J. Harris, R. Shrivastav, and C. Sapienza, "Disordered speech assessment using automatic methods based on quantitative measures," *EURASIP Journal on Applied Signal Processing*, vol. 9, pp. 1400–1409, 2005.

[55] P. Scalassara, C. Maciel, and J. Pereira, "Predictability analysis of voice signals," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 28, no. 5, pp. 30–34, 2009.

[56] ITU-T P.563, "Objective measurement of active speech level," Intl. Telecom. Union, 1993, geneva, Switzerland.

[57] "3SQM™ ADVANCED NON-INTRUSIVE VOICE QUALITY TESTING," OPTICOM GmbH, Erlangen, Germany, 2004., Tech. Rep.

[58] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[59] R. Hummel, T. H. Falk, and W.-Y. Chan, "Spectral Features for the Automatic Assessment of Dysarthric Speech Intelligibility," in *Annual Conference of the International Speech Communication Association*, 2011, to appear.

Appendix A

UA Subject Data

Table A.1: UA Subject Data

Speaker ID	Age	Intelligibility (%)	Dysarthria subtype
M01	>18	15.0	Spastic
M04	>18	2.0	Spastic
M05	21	58.0	Spastic
M07	58	28.0	Spastic
M08	28	93.0	Spastic
M09	18	86.0	Spastic
M10	21	93.0	Not sure
M11	48	62.0	Athetoid (or mixed)
M12	19	7.4	Mixed
M14	40	90.4	Spastic
M16		43.0	Spastic
F02	30	29.0	Spastic
F03	51	6.0	Spastic
F04	18	62.0	Athetoid
F05	22	95.0	Spastic