언어학석사 학위논문

Automatic Detection and Assessment of Dysarthric Speech using Prosody-Based Measures

운율 정보를 이용한 마비말장애 음성 자동 검출 및 평가

2020년 8월

서울대학교 대학원

언어학과 언어학 전공

Abner Hernandez

Abstract

Automatic Detection and Assessment of Dysarthric Speech using Prosody-Based Measures

Abner Hernandez

Department of Linguistics

Graduate School

Seoul National University

One of the earliest cues for neurological or degenerative disorders are speech impairments. Individuals with Parkinson's Disease, Cerebral Palsy, Amyotrophic lateral Sclerosis, Multiple Sclerosis among others are often diagnosed with dysarthria. Dysarthria is a group of speech disorders mainly affecting the articulatory muscles which eventually leads to severe misarticulation. However, impairments in the suprasegmental domain are also present and previous studies have shown that the prosodic patterns of speakers with dysarthria differ from the prosody of healthy speakers. In a clinical setting, a prosodic-based analysis of dysarthric speech can be helpful for diagnosing the presence of dysarthria. Therefore, there is a need to not only determine how the prosody of speech is affected by dysarthria, but also what aspects of prosody are more affected and how prosodic impairments change by the severity of dysarthria.

In the current study, several prosodic features related to pitch, voice quality, rhythm and speech rate are used as features for detecting dysarthria in a given speech signal. A variety of feature selection methods are utilized to determine which set of features are optimal for accurate detection. After

selecting an optimal set of prosodic features we use them as input to machine learning-based classifiers and assess the performance using the evaluation metrics: accuracy, precision, recall and F1-score. Furthermore, we examine the usefulness of prosodic measures for assessing different levels of severity (e.g. mild, moderate, severe). Finally, as collecting impaired speech data can be difficult, we also implement cross-language classifiers where both Korean and English data are used for training but only one language used for testing.

Results suggest that in comparison to solely using Mel-frequency cepstral coefficients, including prosodic measurements can improve the accuracy of classifiers for both Korean and English datasets. In particular, large improvements were seen when assessing different severity levels. For English a relative accuracy improvement of 1.82% for detection and 20.6% for assessment was seen. The Korean dataset saw no improvements for detection but a relative improvement of 13.6% for assessment. The results from cross-language experiments showed a relative improvement of up to 4.12% in comparison to only using a single language during training. It was found that certain prosodic impairments such as pitch and duration may be language independent. Therefore, when training sets of individual languages are limited, they may be supplemented by including data from other languages.

Keyword: dysarthric speech, prosody, machine learning, classification, cross-linguistics, feature selection, acoustics

Student Number: 2018-23331

Table of Contents

List of Figures

List of Tables

List of Equations

Chapter 1. Introduction

1.1. Dysarthria

Neurological disorders often come with a range of cognitive and physical issues that can make life difficult. Speech is one aspect of neurological disorders that can be severely damaged and lead to issues in both articulation and communication. A common speech disorder known as dysarthria often occurs in individuals with a variety of neurological damage. Dysarthria occurs up to 90% of the time in patients with Parkinson's Disease (Muller et al., 2001), 50% of the time for individuals with multiple sclerosis (Sandyk, 1995), one of the first symptoms of Amyotrophic Lateral Sclerosis (ALS) in 25% of patients was dysarthria (Traynor et al., 2000). Given the prevalence of dysarthria in neurological disorders, more research into dysarthria could help individuals live a more comfortable life. The purpose of the current study is to use prosodic measurements to automatically detect dysarthria in continuous speech.

An important aspect of dysarthria is the spectrum of issues that may or may not occur depending on severity, disorder type, dysarthria type or individual differences. In general, the most common speech related issues in dysarthria are respiration (i.e. frequent or forcible inspiration, long respiration resting level), speech tempo (i.e. slow or variable speech rate, many pauses), pitch (i.e. too high or too low pitch, variable pitch), articulation and nasality (i.e. hypernasality). While individual differences

exist, the specific issue and degree of issue often depends on the specific type of dysarthria. The most common classification system for dysarthria was developed by Darley, Aronson, and Brown (1975) known as The Mayo Classification System for Differential Diagnosis of Dysarthria. Table 1. displays some of the most common types of dysarthria along with their associated brain damage and major speech impairments. A more detailed overview of studies related to prosodic deficits in dysarthric speech will be explored in Chapter 2.

Table 1. *Common types of dysarthria and related speech issues.*

Type of Dysarthria (Disease)	Location of Damage	Distinct Speech Issues
Flaccid (Bulbar Palsy)	Lower Motor Neuron	Hypernasality, breathiness, audible inspiration
Spastic (Cerebral Palsy, MS)	Upper Motor Neuron	Misarticulation, slow speech rate, low pitch, harsh/strained voice.
Ataxic (Cerebellar ataxia)	Cerebellum	Monostress, phoneme and interval prolongation, dysrhythmia, syllable repetition, slow speech rate
Hyperkinetic (Parkinson's)	Basal Ganglia	Monopitch, monoloudness, variable speech rate, short rushes of speech
Mixed (ALS)	Multiple Motor Systems	Misarticulation, slowed speech rate, hypernasality, disrupted prosody

As seen from Table 1, misarticulations are not the only factor involved in dysarthria. While the articulatory muscles in the vocal tract are

essential in correct articulation, they are also important for natural prosody. For example, individuals with dysarthria tend to have little control over the contractions of the vocal tract which reduces the range and speed of laryngeal movement. The lack of control of one's vocal folds can result in a more monopitch voice, or an absence of stress within stress syllables. Therefore, there is a growing research interest in not only focusing on the articulatory difficulties involved in dysarthria but also the prosodic irregularities.

1.2. Impaired Speech Detection

Typically, dysarthria is diagnosed by a trained speech pathologist who administers several tasks to the patient in order to perceptually evaluate their speech (Duffy, 2013; Kent et al., 1987). These assessments tend to involve a speech pathologist eliciting speech from the patient and determining whether any irregularities are present. For example, one can measure the voice quality and the ability for the patient to change loudness and pitch to assess the laryngeal or phonation damage. We can also determine prosodic damage by having patients read sentences and observe any irregular variations in pitch, duration or stress. Several, standardized assessments based on perceptual evaluation have been proposed, with the Mayo Clinic Rating System (Darley, Aronson & Brown, 1969) and

Frenchay Dysarthria Assessment (FDA) being the most detailed and commonly utilized test for English speakers (Enderby, 1980).

Despite the wide use of perceptual evaluation, the subjective nature of the task and overly long duration of administering these types of tests are common criticisms. Low identification accuracy was found in Zyski and Weisiger (1987), while low intra- and inter-rater reliability was found in Kearns and Simmons (1988) and Zeplin and Kent (1996) for the Mayo Clinic Rating System. Other more general methods have been proposed (Wannberg, Schalling & Hartelius, 2016; Hong et al., 2018) with higher intra- and inter-rater reliability but still contain a subjectivity problem.

Another solution to the subjectivity and long duration issue is to conduct an acoustic analysis. This approach involves measuring certain acoustic properties of speech such as formant frequency, fundamental frequency (F0), jitter, shimmer, segment duration and comparing those values to a standard healthy speaker. Kent et al. (1999) provides a detailed description of useful measures when examining dysarthric speech from a specific viewpoint such as vowels, fricatives, voice quality, and so on. In general, if enough deviancy from the norm is present, it is possible that the individual has some form of dysarthria. Usually, acoustic analyses are not the sole determiner of dysarthria and a speech pathologist would still administer a perceptual evaluation. However, this approach comes closer to an object assessment of dysarthria.

Lastly, the rise of machine and deep learning methods have introduced a variety of methods for automatically detecting and even assessing the severity level of dysarthric speech. The main approach to using machine learning for detecting dysarthria is extracting acoustic features and using the features as input to a classifier. The goal of this approach is to allow the machine learning algorithm to automatically detect dysarthria based on manually crafted features (López, Orozco-Arroyave, Gosztolya, 2019; Kodrasi & Bourlard, 2019; Tripathi, Bhosale & Kopparapu, 2020). A second approach is to simply use the raw speech signal as features and feed them into complex neural architectures then allow the network to automatically determine the important information that distinguishes between healthy and dysarthric speech (Kim, Cao & Wang, 2018; Millet & Zeghidour, 2019; Mayle et al., 2019).

The first approach requires more data pre-processing as we need to systematically choose appropriate features for our machine learning model, but allows for more interpretability as we can more easily examine the specific acoustic impairments that are most useful in distinguishing dysarthric speech from healthy. The second approach requires less data preparation as we only need the raw speech signal but may suffer from a lack of interpretability since the network inherently determines what features of the speech signal are important. Recent studies have attempted to reduce this interpretability issue with some success but tend to require

sophisticated post processing techniques to extract interpretable information (Tu, Berisha & Liss, 2017; Korzekwa et al., 2019).

1.3. Research Goals & Outline

The main research question our study asks is 'which set of prosodic features are most useful for automatically detecting dysarthria in continuous speech?'. However, we also explore other related problems such as: which specific prosodic measurements contribute more to classification accuracy? What aspects of prosody are more important for distinguishing different severity levels (mild, moderate, severe)? Are there language specific differences? Are there language independent features that can be trained jointly? These questions are examined via machine learning-based experiments.

The following thesis is organized as follows: Chapter 2 will briefly go over previous literature in prosodic impairments in dysarthric speech and machine learning-based approaches for automatic detection and severity assessment. Issues regarding previous related studies and how this study differs will also be mentioned. Chapter 3 will describe the English and Korean dysarthric speech datasets in detail. In Chapter 4 we go over the prosodic features used in our study and several feature selection methods for selecting the optimal set of prosodic features are also proposed. We also describe the classifiers (random forest, support vector machine, neural

network) in detail. Since our baseline models use Mel-Frequency Cepstral Coefficients (MFCC), we will go over the extraction process and parameters regarding MFCC's. Starting from Chapter 5 we go over all the experiments. Two experiments per language group, detection and assessment, and one experiment we refer to as a cross-language experiment where we train our models using data from both languages but only test with one language. Results in Chapter 6 are evaluated by using accuracy, precision, recall and F1-scores. Chapter 7 and 8 will conclude the paper with a discussion of the results and future directions for dysarthric speech research.

Chapter 2. Background Research

2.1. Prosodic Impairments

2.1.1 English

The most salient prosodic impairments in dysarthric speech are related to pitch and speech rate. One of the earliest studies of dysprosody in dysarthric speakers was by Schlenck, Bettrich and Willmes (1993). In their study, length of tone units, fundamental frequency, and standard deviation of fundamental frequency from spontaneous speech was collected from 84 dysarthric speakers with ALS and 154 healthy controls. Results revealed significant differences from both speaker groups and by severity level. Severe dysarthric speech had shorter tone units and a higher mean fundamental frequencies than mild dysarthria and normal controls. Patients with mild dysarthria had lower standard deviations of fundamental frequency (more monotonous speech) than normal controls and severe dysarthric speakers.

The findings of Schlenck et al. (1993) are further supported by later studies in speakers with multiple sclerosis, cerebral disease and motor neuron disease (Bunton, Kent, Kent & Rosenbek, 2000; Lowit-Leuschel & Docherty, 2001). In Bunton et al.'s (2000) study, mean F0, F0 standard deviation, F0 variation, and duration of tone units which was defined as word or syllable per second for the minimal unit which can carry intonation were collected from speakers with ALS, cerebral disorders (CD) and healthy

controls. Results showed that speakers with ALS (49 Hz) and CD (46 Hz) tended to have lower F0 variation compared to healthy controls (143 Hz). Similarly, control speakers had a longer tone unit duration, a larger number of words in a tone unit, a smaller average duration of words in a tone unit compared to dysarthric speakers.

Lowit-Leuschel and Docherty (2001) found similar results by taking the following measurements from read and spontaneous speech: articulation rate (syll/min), mean unstressed vowel duration (UVD), number of unstressed vowels (UV), percentage of unstressed vowels, range of intensity variation (dB), F0 range, mean F0 (male and female). A summary of their results can be seen in Table 2. In general, dysarthric speakers had a slower articulation rate, less intensity and F0 variation, longer vowel duration, a smaller percentage of unstressed vowels, and a higher mean F0 for males. However, no test of significance was conducted between speaker groups only within groups. Therefore, we are unable to make conclusions regarding significant differences.

Table 2. *Prosodic measurement from dysarthric and healthy speakers.*

	Dysarthric Group	Control Group
Prosodic Measure	Reading / Spontaneous	Reading / Spontaneous
artic. rate	249 / 255	279 / 284
Mean UVD (ms)	80 / 68	50 / 47
No. of UV	45 / 43	49 / 58
% of UV	26 / 27	29 / 33
dB range	5.5 / 6.25	6.85 / 7.75
Mean F0 (male) Mean F0 (female)	158 / 156 196 / 206	119 / 101 209 / 199
F0 range (Hz)	140 / 123	191 / 129

2.1.1 Korean

Research with Korean speakers also found similar prosodic impairments in dysarthric speakers. Nam and Kwan (2005) took several prosodic measurements for six interrogative and declarative sentences for patients with spastic and athetoid cerebral palsy (SCP, ACP respectively) associated dysarthria. Unlike the studies with English speakers, healthy controls had the narrowest F0 range while the group with ACP had the widest F0 range for full sentences. The range of the pitch in sentence endings was wider in the SCP and ACP groups than in the healthy group. The range of the loudness in sentence endings was also wider in the SCP and ACP group than in the healthy group. Lastly, the duration of utterances

and the duration of pauses were much longer and the frequency of pause was higher for dysarthric speakers than for healthy speakers.

Kang, Seong and Yoon (2011) found differences by gender. For males, mean F0 slope and semitone slope were the most important factors to distinguish healthy and dysarthric speech, while for females mean energy slope and max energy slope were the most important. In another study, Kang, Yoon, Seong and Park (2012), found that patients with Parkinson's had lower pitch values in interrogative sentences, and lower loudness values than the control group. The prosody of dysarthric speakers with a wide range of disorders (Cerebral Palsy, Motor Neuron disease, traumatic brain injury, Parkinson's, cerebral disease) were examined in Seo and Seong (2012). Researchers found reduced speaking and articulation rates, reduced F0 slope and question-tone slope for sentences, and all of intonation slope in the final word for sentential questions.

In general, results follow closely to English speakers who also display reduced speech rates, and longer durations of utterances. The only language difference seen was in F0 range. English speakers with dysarthria tend to have a reduced range, while the speakers in Nam and Kwon (2005) had a wider range than healthy controls.

2.2. Machine Learning Approaches

The literature on machine learning-based approaches to dysarthric speech detection and assessment is wide and contains many different approaches to the difficult issue. We will first go over classical machine learning approaches, particularly those which utilize prosodic measurements, and then go over to more recent deep learning approaches.

Early approaches using prosody for automatic detection of dysarthric speech have been argued based on findings that prosodic impairments tend to be one of the notable cues for early stage dysarthria (Darkins, Fromkin & Benson, 1988). Therefore, including prosodic measurement can be essential for accurately detecting dysarthria in its early stages. Bocklet et al. (2011) extracted features from a variety of read sentences based on phonation (glottis features), articulation (MFCCs), and prosody (F0, energy, duration, pauses, jitter and shimmer) from both healthy and dysarthric speakers. These acoustic features were then used as input to a SVM classifier. Results show that glottal features can achieve an accuracy of 83.3%, MFCCs features reached an accuracy of 100%, and the prosodic features obtained up to a 90.5% accuracy. While results are promising in showing that prosodic information can be helpful for detection, one issue with this study was a lack of explanations regarding the exact prosodic measures. The total set includes 292-dimensional features where 73 are related to F0, duration, shimmer, jitter, pauses, and energy, along with their mean, minimum,

maximum and standard deviation (73 *4=292). After a correlation-based feature selection, only 12-17 of these prosodic measures are determined to be the most useful for distinction, but those selected measures are never explained. In a clinical setting, knowing these prosodic features would be essential in determining what aspects of a patient's prosody should be attended to when developing proper speech therapy.

The issue of selecting relevant and explainable features is addressed in Kadi et al. (2013), where the most relevant of exactly 11 prosodic features are used to automatically assess the severity level of dysarthric speakers from the publicly available Neymours database (Menendez-Pidal et at., 1996). A Linear Discriminant Analysis (LDA) based feature selection methods was used to determining the most discriminative prosodic features as follows (from most to least discriminative): articulation rate, # of period, mean pitch, voice breaks, %V, HNR, jitter, shimmer, std pitch, std period, NHR. These features were shown to assess four levels of dysarthric speech with an accuracy of 88.89% when using a gaussian mixture model classifier, and an accuracy of 93% when using an SVM classifier.

Kadi et al,'s (2013) study shows how a small set of prosodic features can be sufficient in detecting sentence-level dysarthria, however, one serious limitation to this study relates to the database. First, the speakers in the Neymours database are composed of 12 males, 11 with dysarthria and only one healthy control. The lack of both healthy speakers and female

speakers may limit the generalizability or the model's capability of accurate classification with other speakers. Another issue relates to the limited sentences structure. The database is mostly composed of simple carrier sentences where the format is always: 'the X is Y-ing the Z". X and Z coming from a set of 74 monosyllabic nouns, while Y was selected from a set of 37 disyllabic verbs. Using carrier sentences can alter the natural prosody of language leading to an inaccurate representation of prosody.

A slightly more recent study by Kim et al, (2015) attempts to alleviate the issue with the Kadi et al.'s (2013) work by evaluating the performance of classifiers trained on two different datasets. The first being the TORGO database, which was developed by Rudzicz, Namasivayam and Wolff (2012) at the University of Toronto. More details regarding this database will be addressed in Chapter 3, but in general there is a more diverse set of speakers, which help increase the generalizability, and a diverse set of recorded utterances that contain sentences with more natural prosody. The second database Kim et al., (2015) used is the NKI CCRT Speech Corpus developed for the 2012 Interspeech speaker trait sub-challenge for pathological speech (Schuller et al., 2012). This database contains recordings from 55 speakers (10 females, 45 males). The prosodic features are separated into two categories voice quality and pitch-duration. The voice quality feature set contains 3 measures, HNR, shimmer and jitter, along with statistical estimates such as quantiles, mean, median and

standard deviation. The pitch-duration set includes F0 measures, utterance and phone duration, along with normalized values and several statistical measures. An LDA-based classifier was used to achieve an accuracy of 71.9% and 82.1% for voice quality and pitch-duration feature sets respectively. While this study shows promising results by both reducing the feature set to more explainable features and utilizing a more complex and realistic database, there is still an issue with the representation of prosody. Prosody is a multidimensional aspect of speech that should not be limited to just F0, duration and voice quality. As mentioned in section 2.1 speech rate and rhythm are also important prosodic elements affected in dysarthric speech and should be included for a more complete holistic representation of prosody.

Deep learning approaches are another group of machine learning methods that incorporate more sophisticated learning algorithms and architectures. The training procedure tends to be the same where acoustic features are extracted and used as input to a classifier. Although the use of deep learning is the standard approach in many audio and speech classification problems, several issues arise that prevent it from being the standard in impaired speech detection. First, the success of deep learning has largely been the result of big data and the ability to train on large datasets. Unfortunately, the collection of impaired speech data is difficult and available datasets are often very limited. Secondly, most deep learning

approaches use features that can be either difficult to interpret in a clinical setting where dysarthric speech detection is most likely to be conducted or minimally helpful for further analysis.

Mayle et al. (2019) used long short-term memory (LSTM) recurrent neural networks (RNN) to detect dysarthria from MFCCs. While the results were promising, no comparison was made against classical machine learning algorithms. Furthermore, MFCCs have already been shown to be accurate in detecting dysarthria even in classical machine learning classifier algorithms such as SVMs, LDA, GMM, HMM, KNN (Bocklet et al. 2011; Selouani et al. 2012; Kim et al. 2015).

Convolutional neural networks (CNN) were used for dysarthric speech detection in An et al. (2018). CNN's can naturally extract local features from a speech signal, in this case from filterbank energies, and later fed to a feed-forward neural network for classification. Results show that using filterbanks in a CNN-based classifier produce a specificity rate of 80.9% while using other acoustic features (MFCC, prosody, statistical variations) in a standard feed-forward network reached a specificity of 80.4%.

Lastly, filterbanks were fed to attention-based LSTMs in Millet et al. (2019). Results show that time-domain filterbanks outperform low-level descriptors (65.5 % vs 82.4% UAR). However, results are either comparable

or inferior to other studies using the same dataset but with fewer features and less complex models (Kim et al., 2015).

The previously mentioned studies are not an exhaustive representation of all deep learning-based studies on dysarthric speech detection but provides some examples of drawbacks or issues with deep learning. The deep learning approach should not only provide good results but also help speech pathologists interpret the results to aid patients who are diagnosed with dysarthria. A growing trend has gone towards explainable deep learning, and current/future studies are attempting to apply deep learning techniques for dysarthric speech detection in an interpretable manner.

Chapter 3. Database

Early studies on dysarthric speech used personal datasets collected within the university or in collaboration with a speech pathology clinic. Recently, publicly available datasets are being used more often in order to allow other researchers to validate or replicate studies. Few of these sets are available but the most commonly used datasets for English are the Neymours dataset (Menendez-Pidal et at., 1996), the UA-Speech database (Kim et at., 2008), and the TORGO database (Rudzicz et al., 2012).

The issues of the Neymours database was described in the previous section, mainly regarding the lack of diversity in both speakers and stimuli. The UA-Speech database is a larger database of 15 speakers with dysarthria ranging from very low intelligibility to highly intelligible. Each speaker recorded 765 isolated words; 300 distinct uncommon words and 3 repetitions of digits, computer commands, radio alphabet and common words. The only concern with the UA-Speech database is that lack of full sentences. Speakers with dysarthria not only vary in severity between speakers but also within speakers. Some words may show signs of dysarthria more than others even within the same speaker, so it would be more helpful to analyze a full sentence rather than a single word. Furthermore, while severe speakers may be easily identified just by a single word, this is not necessarily the case for speakers with mild dysarthria. Early detection of dysarthria is a case of mild dysarthria and is an important

factor since early diagnosis can lead to early therapy. Early diagnosis would require an evaluation of continuous speech to accurately diagnose the presence of dysarthria in speech. Lastly, given that we are using prosodic features for classification, compared to isolated words, prosodic tendencies are better represented in continuous speech. Therefore, we choose to use the most recently built database TORGO, as this database contains a diverse set of speakers, stimuli and continuous speech.

Few databases of dysarthric speech in other languages exist, and even fewer are publicly available. Some commonly used databases are the CUHK for Cantonese (Wong et al., 2015), for Spanish the Orozco-Arroyave et al. (2014) dataset has often been studied. However, for our cross-language experiments we chose to use the Quality of Life Technology (QoLT) dataset, which is a Korean database of dysarthric speakers with cerebral palsy (Choi et al., 2012). We choose this database as it has a large number of speakers, contains continuous speech data, and comes from a non-indo European language. The few cross-language dysarthric speech studies that have been conducted have always been between European languages (Orozco-Arroyave et al., 2016). Therefore, including Korean allows the evaluation of training datasets between two very different languages.

3.1. English-TORGO

The TORGO dataset was originally created to provide resources for developing personalized ASR systems for speakers with dysarthria but has been widely used in dysarthric speech detection and assessment. The publicly available dataset contains 8 dysarthric speakers, 5 males and 3 females, from speakers with cerebral palsy and ALS. Speakers with dysarthria were assessed by a trained speech pathologist using the Frenchay Dysarthria Assessment. Four speakers were categorized as having severe dysarthria, one speaker with moderate/severe, one moderate, and two mild. Recording from 7 healthy controls, 4 males, 3 females, were also collected. A mixture of short words, non-words, restricted sentences (read speech), and unrestricted sentences (spontaneous speech) was recorded from all speakers. Some examples of the speech stimuli can be seen in table 3.

Table 3. *Speech stimuli examples from the TORGO database.*

Short Words	Digits, computer commands International radio alphabet Phonetically contrasting pairs of words
Restricted Sentences	Preselected phoneme-rich sentences such as: o ‘‘The quick brown fox jumps over the lazy dog’’ The Grandfather Passage The 460 TIMIT-derived sentences used as prompts in the MOCHA database
Unrestricted Sentences	Spontaneous speech elicited from an image description task of 30 images.

3.2. Korean-QoLT

The QoLT database was created to improve the quality of life for individuals with disabilities by improving technology commonly used by healthy speakers. In particular, for improving ASR technologies in PC's or smart phones. The database contains recordings from 100 dysarthric speakers and 30 healthy controls. A speech therapist assessed the severity of speakers via Percentage of Consonant Correct (PCC) using the Assessment of Phonology and Articulation for Children (APAC) words, and divided speakers into four groups; mild (PCC: 85~100%), mild to moderate (PCC: 65~84.9%), moderate to severe; (PCC: 50~64.9%), and severe (PCC: less than 50%). A subset of assessments was re-evaluated and it was found that the intra-rater reliability was .957 and the inter-rater reliability was .901 using Pearson's product moment correlation.

Four main sets of speech stimuli were recorded. First, 37 words from APAC which include 19 Korean consonants with 70 speech sounds – word-initial, word-final, word-medial onset and word-medial coda consonants. Second, 100 Machine Control Commands and 36 Korean Phonetic Alphabets. Machine control commands are commands which are commonly used for PC, cell phone, TV, radio, and other electronic appliances. Third, 452 Phonetically Balanced Words (PBW) where $1/9^{th}$ are recorded by dysarthric speakers and $1/3^{rd}$ by healthy speakers. Lastly, 100 words and 5 sentences for investigating Korean consonants and vowels reflecting various

phonetic environments. The five sentences along with their translations are displayed in table 4. As we are interested in continuous speech, only the five recorded sentences are considered in our experiments.

Table 4. *Full sentence stimuli along with phonetic and English translation from QoLT.*

Korean Hangul	Yale Romanization	English Translation
추석에는 온 가족이 함께 송편을 만든다.	chwusekeynun on kacoki hamkkey songphyenul mantunta	In Chuseok, the whole family makes songpyeon together.
갑자기 미국에 있는 오빠 얼굴이 보고 싶다.	kapcaki mikwukey issnun oppa elkwuli poko siphta.	Suddenly, I want to see my brother's face who is in America.
어제 하늘이 컴컴해지더니 비가 쏟아졌다.	ecey hanuli khemkhemhayciteni pika ssotacyessta.	The sky turned dark yesterday and it rained.
동생이랑 싸워서 엄마한테 혼났다.	tongsayngilang ssawese emmahanthey honnassta.	My mom scolded me for fighting with my younger sibling.
시원한 물 한 잔 주세요.	siwenhan mwul han can cwuseyyo.	I would like a glass of cold water, please.

Chapter 4. Methods

4.1. Prosodic Features

4.1.1 Pitch

Pitch is a commonly studied cue of dysarthria, showing differences not only with healthy speakers but also between speakers of different severity levels. Mild dysarthric speakers tend to be more monotonic while severe speakers often have significantly higher pitch than both mild and healthy speakers (Schlenck et al., 1993). Therefore, we believe pitch measurements to not only be helpful in detecting dysarthria but also useful for distinguishing different severity levels. However, we also expect some language differences to arise given the opposite results found in Korean (Nam and Kwon, 2005).

The acoustic representation of pitch is known as fundamental frequency (F0) which is the lowest frequency of a periodic waveform. F0 is measured for all voiced segments of an utterance. We include standard pitch measurements such as mean, median, minimum and maximum F0 along with standard deviation, 25% and 75% quantiles. Figure 1 and 2 also display the mean values for English and Korean speakers respectively. From both figures we see generally higher F0 values for speakers with dysarthria. The only language difference appears to be with the max F0 values which is similar in English speakers but much higher in Korean dysarthric speakers.

Severity based measures can be seen in figures 3 and 4 for English and Korean respectively. Speakers with severe dysarthria tend to have a higher max and mean F0. Interestingly, Korean speakers with moderate dysarthria tended to have higher F0 values for all measures excluding max F0, even compared to the severe group. Another important finding was that with English speakers the mild dysarthric group had a lower standard deviation (25.35 Hz) compared to healthy speakers (35.5 Hz) as expected given the studies showing this group to be more monopitch. However, the opposite was found in Korean speakers where healthy speakers had a slightly lower standard deviation (30.2 Hz) compared to the mild group (35.2 Hz).

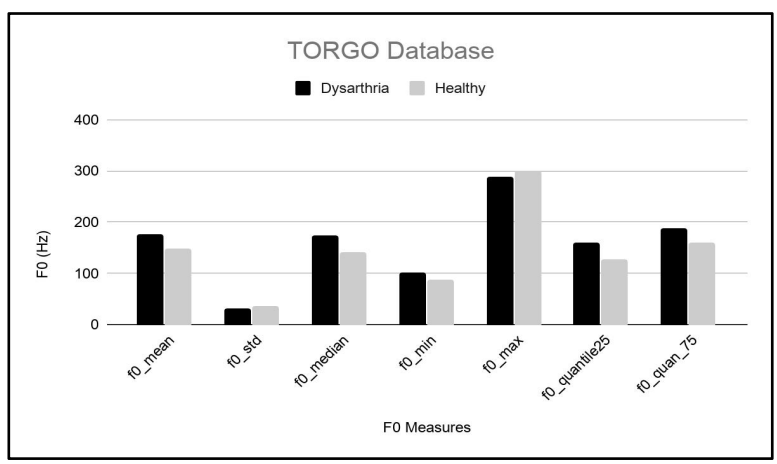

Figure 1. *Mean values for all pitch measures in healthy and dysarthric speakers*.

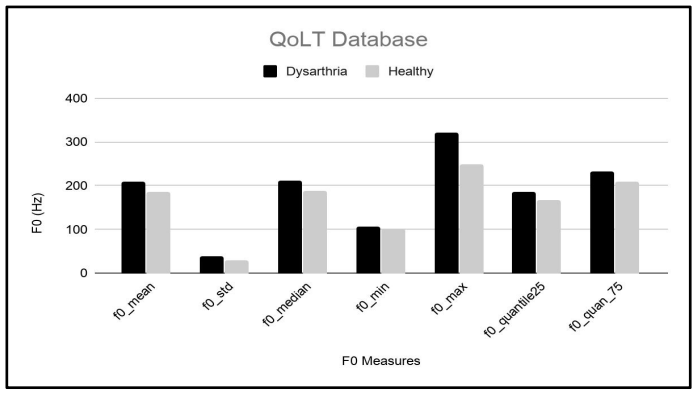

Figure 2. *Mean values for all pitch measures in healthy and dysarthric speakers.*

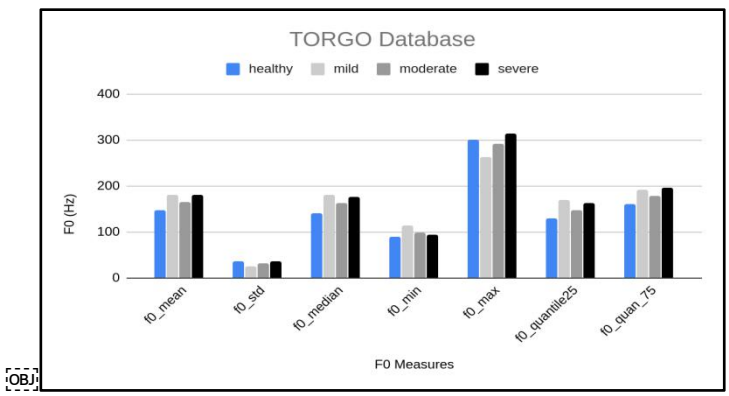

Figure 3. *Mean values for all pitch measures based on severity.*

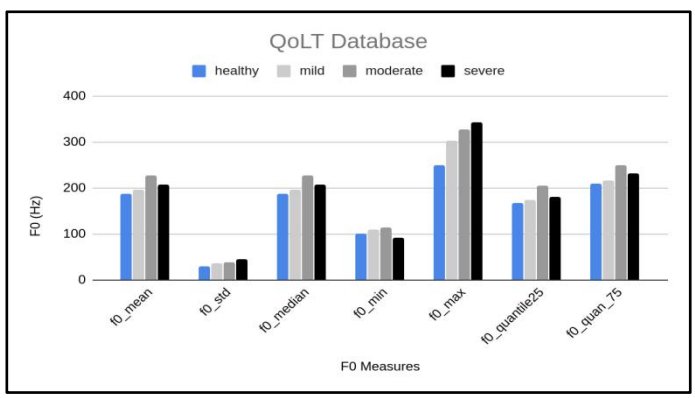

Figure 4. *Mean values for all pitch measures based on severity.*

4.1.2 Voice Quality

Voice quality refers to the properties of speech related to the vocal folds within the larynx. Individuals with dysarthria tend to have less control over their vocal folds leading to irregular measurements (Dogan et al., 2007). Speakers with multiple sclerosis caused dysarthria (spastic and ataxic dysarthria) had several voice quality based measurements taken, such as: jitter percent (jitt %), shimmer percent (shim %), soft phonation index (SPI), and noise to harmonics ratio (NHR). Results show that the mean jitter, shimmer, and SPI of MS patients were significantly increased compared to the control group ((Jitt, $p < 0.001$; Shim, $p <5$ 0.033; SPI, $p < 0.0001$). Voice quality features have also been shown to be useful in machine learning classification of impaired speech (Bocklet et al., 2011; Kadi et al., 2013; Kim et al., 2015). Our study extracts 5 voice quality measures: jitter, shimmer, Harmonics to noise ratio (HNR), # of voice breaks, and degree of voice breaks. These measures are extracted as they are the most commonly used measures for voice quality in clinical studies of dysarthric speech[1].

Jitter represents the variations of F0 within a time period. More specifically we can calculate relative local jitter by the average absolute difference between consecutive periods, divided by the average period. The

[1] Voice quality measures are not all directly related to prosody. For example, jitter and shimmer are related to perturbations of pitch, but voice break and HNR measure are more related to phonation. For completeness and fair comparisons with previous studies, we include voice breaks and HNR measures for our voice quality feature set.

calculation for jitter can be examined in equations 1-3, where T_i is the duration of the *i*th interval and *N* is the number of intervals.

$$\text{Absolute jitter (sec)} = \sum_{i=1}^{N} |T_i - T_{i+1}| / (N - 1) \quad (1)$$

$$\text{Mean Period (sec)} = \sum_{i=1}^{N} T_i / N \quad (2)$$

$$\text{Relative Jitter} = \text{Absolute Jitter} / \text{Mean Period} \quad (3)$$

Shimmer is similar to jitter except that perturbation of F0 falls in the amplitude domain, so we take the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

$$\text{Relative Shimmer} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \quad (4)$$

Next, HNR refers to the periodicity of a speech signal over noise. Harmonicity is measured in decibels (dB) by the ratio of the energy of the periodic part (Ep) related to the noise energy (En) as seen in equation 5.

$$\text{HNR (dB)} = 10*\log\left(\frac{Ep}{En}\right) \quad (5)$$

Furthermore, we take two measures related to breaks in voicing. In healthy speech, speakers can maintain the phonation of voiced segments such as a vowel for quite some time. However, speakers with dysarthria

have trouble with this task. We included two voice break related measures. The first being the number of voice breaks which is the number of distances between consecutive pulses that are longer than 1.25 divided by the pitch floor (in our case we set the pitch floor to 50 Hz). Secondly, we measure the degree of voice breaks, which is the total duration of the breaks over the signal, divided by the total duration, excluding silence at the beginning and the end of the sentence. Speakers from both our datasets were observed to generally have higher values for both voice break measurements. Mean values for all measures in Korean and English can be seen in Table 6. The only consistent trends we see are with voice breaks. In general, the more severe the dysarthria the higher number of voice breaks and larger degree of voice breaks. In English speakers, jitter is higher, but shimmer is lower than healthy controls.

Table 6. *Mean Voice Quality measure for all speaker groups.*

Corpus	Speaker Group	Jitter	Shimmer	HNR	# of VB	% of VB
TORGO	Healthy	1.85	11.46	9.59	6.00	17.13
	Dysarthric (All)	2.03	8.78	12.13	7.91	21.29
	Mild	2.02	9.76	10.23	6.7	16.71
	Moderate	1.80	7.98	13.75	7.80	27.10
	Severe	2.24	8.46	12.67	9.29	20.86
QoLT	Healthy	1.68	7.54	15.12	5.71	13.15
	Dysarthric (All)	1.61	7.11	15.83	9.50	29.21
	Mild	1.53	7.08	15.78	7.89	20.90
	Moderate	1.62	6.94	16.14	9.3	33.39
	Severe	1.69	7.37	15.50	11.84	34.58

4.1.3 Speech Rate

Several studies have found impairments in speech rate based measurements such as speaking rate (syll/per sec), articulation rate (syll/per sec without pause), # of pauses, segment duration (Ackermann & Hertrich, 1994; Le Dorze, Ouellet & Ryalls, 1994). Speakers with dysarthria tend to have both a lower speaking rate and articulation rate, more pauses, and longer syllable duration. The current study takes 7 relevant measures: full utterance duration, speaking duration, balance, speaking rate, articulation rate, number of syllables and number of pauses. To extract speech rate features, the approach taken by De Jong and Wempe (2009) is used where the syllable nuclei is automatically detected and no transcriptions are necessary. First, we use intensity to find peaks in the energy contour, since a vowel within a syllable (the syllable nucleus) has higher energy than surrounding sounds. Intensity contour is then used to make sure that the intensity between the current peak and the preceding peak is sufficiently low. With this procedure, multiple peaks within one syllable are deleted. Finally, we use voicedness to exclude peaks that are unvoiced, which is required to delete surrounding voiceless consonants that have high intensity. As expected, our data followed the trends of previous studies, dysarthric speakers tend to have longer durations, slower speaking and articulation rate, more pauses, and more syllables given the habit of repetition. The full range of mean values can be seen in table 7.

Table 7. *Mean Speech Rate measures for all speaker groups.*

Corpus	Speaker Group	# of syllables	# of pauses	Speaking Rate	Artic. Rate	Speaking Duration (sec)	Total Duration (sec)
TORGO	Healthy	9.12	0.17	2.00	4.135	2.22	4.50
	Mild	10.67	1.4	1.75	3.53	3.1	6.22
	Moderate	10.78	2.09	1.78	3.33	3.38	6.60
	Severe	12.21	1.82	1.69	3.17	3.77	7.13
QoLT	Healthy	11.75	0.09	3	4.58	2.52	3.83
	Mild	13.14	1.45	2.29	3.82	3.58	6.00
	Moderate	13.91	3.29	1.69	3.42	4.19	8.96
	Severe	17.53	4.74	1.64	3.47	5.29	11.49

4.1.4 Rhythm

The last group of prosodic measurements we extract are known as rhythm metrics. Unlike pitch or voice quality measures, rhythm does not have a specific acoustic cue. Instead, linguists have proposed several durational measures of vocalic and intervocalic segments. These measures have been shown to be correlates of rhythm (Ramus, Nespor & Mehler, 1999; Grabe & Low, 2002; Dellwo & Wagner, 2003). Traditionally, rhythm metrics have been used to classify between languages with different rhythm patterns. Such as comparing stress-timed, syllable-timed or mora-timed languages. The focus of the current study is not to compare the rhythm of Korean and English with rhythm metric, but instead use the metrics to distinguish between healthy and dysarthric speakers. Liss et al. (2009) were one of the first researchers to use rhythm metrics to classify healthy and dysarthric speakers, and showed an accuracy of 80% when classifying

different types of dysarthrias (ALS, PD, HD, Ataxic). Further studies have supported these results by using rhythm metrics in machine learning classifiers (Selouani et al., 2012; Dahmani et al., 2013).

One of the first group of rhythm metrics, formally known as 'the deltas', was proposed by Ramus et al. in 1999. They proposed 3 metrics; the average proportion of vocalic intervals (%V), and the average standard deviations of consonantal (ΔC) and vocalic (ΔV) intervals. For example, "next Tuesday on" (phonetically transcribed as /nɛkstjuzdeɪɒn/) would contain 3 vocalic and 4 consonantal intervals (/n/ /ɛ/ /kstj/ /u/ /zd/ /eɪɒ/ /n/). They found that the proportion of time of vocalic intervals in the sentence (%V) and the standard deviation of intervocalic intervals (ΔC) was the best correlate for distinguishing different rhythm classes. In general, stress-timed languages have high ΔC and low %V, in contrast syllable-timed languages have low ΔC but high %V. Figure 5 shows that our healthy speakers follow this trend as English speakers have a higher ΔC but lower %V compared to Korean speakers. On the other hand, regardless of the language, speakers with dysarthria have an overall high ΔC.

Researchers have tried to normalize delta values in order to reduce the interaction between speech rate and deltas. Dellwo and Wagner (2003) proposed a method where the values of deltas are divided by the mean duration of vocalic or consonantal intervals, then multiplied by 100. These normalized measures are known as the 'Varcos' and can be measured for

both vowel and consonant intervals. For example, Varco C can be calculated as such:

$$\text{VarcosC} = \frac{\Delta C * 100}{mean(c)} \tag{6}$$

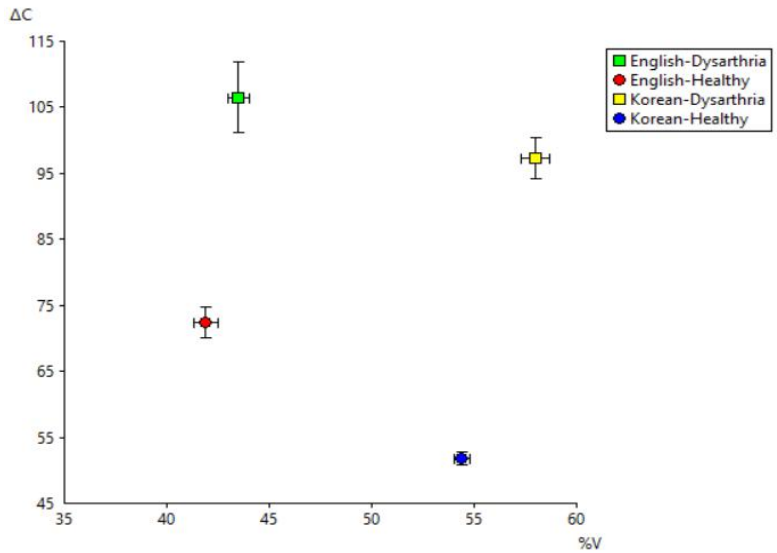

Figure 5. *Mean values of ΔC and %V for dysarthric and healthy groups in both Korean and English.*

The last group of speech metrics were proposed by Grabe and Low (2002). They take another approach to rhythm, where the temporal succession of the vocalic and consonantal intervals is taken into consideration instead of joining all the values and calculating the standard deviation. The influence of speech rate variation can be controlled by calculating the normalized PVI, which calculates the mean absolute normalized difference between durations of neighboring interval pairs. In

general, the raw PVI is used for consonantal intervals and normalized PVI for vocalic intervals. rPVI and nPVI can be defined as in eq. 7-8, where d_k is the length of the k^{th} vocalic or intervocalic segment and m is the number of segments.

$$\text{rPVI} = \sum_{k-1}^{m-1} |d_k - d_{k+1}|/(m-1) \quad (7)$$

$$\text{nPVI} = 100 * \sum_{k-1}^{m-1} \left|\frac{d_k - d_{k+1}}{\frac{d_k + d_{k+1}}{2}}\right| / \ (\text{m} - 1) \quad (8)$$

A full table of mean scores for all rhythm metrics for all speaker groups can be seen in Table 7. As seen from the table 7, healthy English speakers have a higher ΔC but lower %V compared to Korean speakers. English speakers also have lower varco and nPVI means compared to Korean speakers. Speakers with dysarthria from both language groups have overall higher means for deltas and rPVI metrics. This is likely due to difficulty in articulating, leading to highly variable durations of both consonantal and vocalic intervals. Both Varcos and nPVI measures show minimal difference between healthy and dysarthric speakers.

Table 7. *Mean values of rhythm metrics for all speaker groups.*

Speaker Group	%V	ΔV	ΔC	varco -V	varco-C	Vrpvi	Crpvi	Vnpvi	Cnpvi
English Healthy	41.72	60.70	73.28	53.18	50.89	66.20	81.85	55.85	56.89
English Dys.	43.54	93.30	107.56	50.66	55.07	102.86	116.34	54.03	58.58
Korean Healthy	54.37	65.69	51.79	57.59	55.23	67.52	65.78	61.51	70.36
Korean Dys.	57.83	139.57	96.65	58.43	60.05	148.56	110.56	60.90	69.18

4.2. Feature Selection

Choosing the right set of features is an important aspect when training machine learning models as not all features may be necessary. In order to select the optimal set of prosodic features, we conduct several feature selection methods and compare the performance for each method. For our study, we specifically implemented three major feature selection methods: the filter method, embedded method, and wrapper methods.

The filter method works by selecting the best features based on univariate statistical tests. The selection of features is independent of any machine learning algorithm. Features are ranked on the basis of statistical scores which tend to determine the features' correlation with the outcome variable. In our case we use ANOVA F-values since our groups are categorical. F-values in this case are the ratios of two Chi-distributions

divided by its degrees of Freedom (as in eq. 9) and is used since we are comparing the variance between the groups and variance within the groups.

$$F = (\chi_1^2/n1\text{-}1) / (\chi_2^2/ n2\text{-}1) \quad (9)$$

To calculate F-values for feature selection we first need to calculate the between sum of squares (SSB) and within sum of squares (SSW). The distance between each group average value g from grand means xbar is g-xbar to get eq. 10 where g_i is the i^{th} item in the set and $\overline{X}$ is the mean of all items in the set. The distance between each observed value within the group x from the group-mean g is given as x-g in equation 11. Lastly, our F-value is calculated as in equation 12. For each feature, if the null hypothesis is rejected that means variance exists between the groups and we will include this feature for model training.

$$SSB = (g_i\text{-}x)^2 \quad (10)$$

$$SSW = (x_i\text{-}g)^2 \quad (11)$$

$$F = (SSb/df_b) / (SSW/df_w) \quad (12)$$

Next, we tested two embedded feature selection methods, an L1-based (lasso) feature selection and tree-based feature importance method. The lasso method is a regularisation approach where a penalty is applied

over the coefficient of a linear model (see eq. 13). We then select the features with non-zero coefficients.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)}) + \frac{\lambda}{m}\sum_{j=1}^{n} |\theta_j| \quad (13)$$

Tree-based estimators such as random forest or extra forest can be used to compute impurity-based feature importances, which in turn can be used to discard irrelevant features. Running this technique allows us to test the top features used in an iterative manner. Lastly, the method that worked best for our models in all experiments was recursive feature elimination. Recursive feature elimination (RFE) performs a greedy search to find the best performing feature subset. It iteratively creates models and determines the best or the worst performing feature at each iteration. It constructs the subsequent models with the leftover features until all the features are explored. It then ranks the features based on the order of their elimination. A sample of the features selected can be visualised from table 8, which shows the features selected for binary detection in the TORGO dataset[2].

For our experiments we use the RFE feature set as it was the feature selection method which provided the best results and was consistent with all scenarios (detection, assessment, cross-language) for both languages. From table 8 we see that each feature selection method selects different features.

[2] Tables for other experiments are in the appendix.

For example, all methods with the exception of the filter method selected some pitch features. Furthermore, some methods select more features than others, as seen when comparing the RFE and lasso methods which have 8 and 21 selected features respectively.

Table 8. *Selected features for detection using various feature selection methods for the TORGO dataset.*

Features	Filter Method	Lasso Method	Tree-based	RFE
Pitch	None	f0_std, f0_quan_75, f0_min, f0_max, f0_quantile25	f0_quantile25, f0_quan_75, f0_mean, f0_median,	f0_std, f0_quantile25, f0_quan_75
Voice Quality	Shimmer, Mean HNR	Shimmer, Jitter, % of voice breaks, # of voice breaks	Mean HNR, Jitter, Shimmer	Jitter
Speech Rate	# of pauses, full duration, speaking duration, articulation rate	speaking rate, articulation rate, speaking duration, number of pauses	# of pauses, full duration, speaking duration	# of pauses, full duration
Rhythm	delta-V, Vrpvi, delta-C, Crpvi	Vnpvi, Cnpvi, varco-V, %V, delta-V, Crpvi, delta-C, Vrpvi	None	delta-V, varco-V
# of features	10	21	10	8

4.3. Classification Models

In our experiments we evaluate the performance of selected features by using them as input to three different machine-learning based classifiers: random forest, support vector machine, and a feed-forward neural network. Including multiple classifiers allows us to generalize the performance and reduces the chance of our data overfitting to one classifier. All classifiers were used for each experiment, detection and assessment for both Korean and English. As well as for the cross-language experiments.

4.3.1 Random Forest

A random forest (RF) classifier is an estimator that fits multiple decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. We use the Gini impurity function to measure the quality of a split. Gini Impurity is a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance was randomly classified according to the distribution of class labels from the data set. The Gini impurity can be computed by summing the probability p_i of an item with label i being chosen, times the probability $\sum_{k\neq 1} p_k = 1 - p_i$ of a mistake in categorizing that item. Figure 6 displays a simple example of how a random forest is structured. In this basic case our random forest produces 3 decision trees, where 2 trees have predicted an utterance to be dysarthric, while 1 tree made a healthy prediction. Given that the majority

of trees have predicted dysarthria, the final decision for our random forest classifier will be dysarthric. In actual practice we will have to tune several hyper parameters such as the number of decision trees and depth of trees (how many nodes). We could also train using information gain (entropy) instead of the Gini impurity, but we found the latter to produce better results.

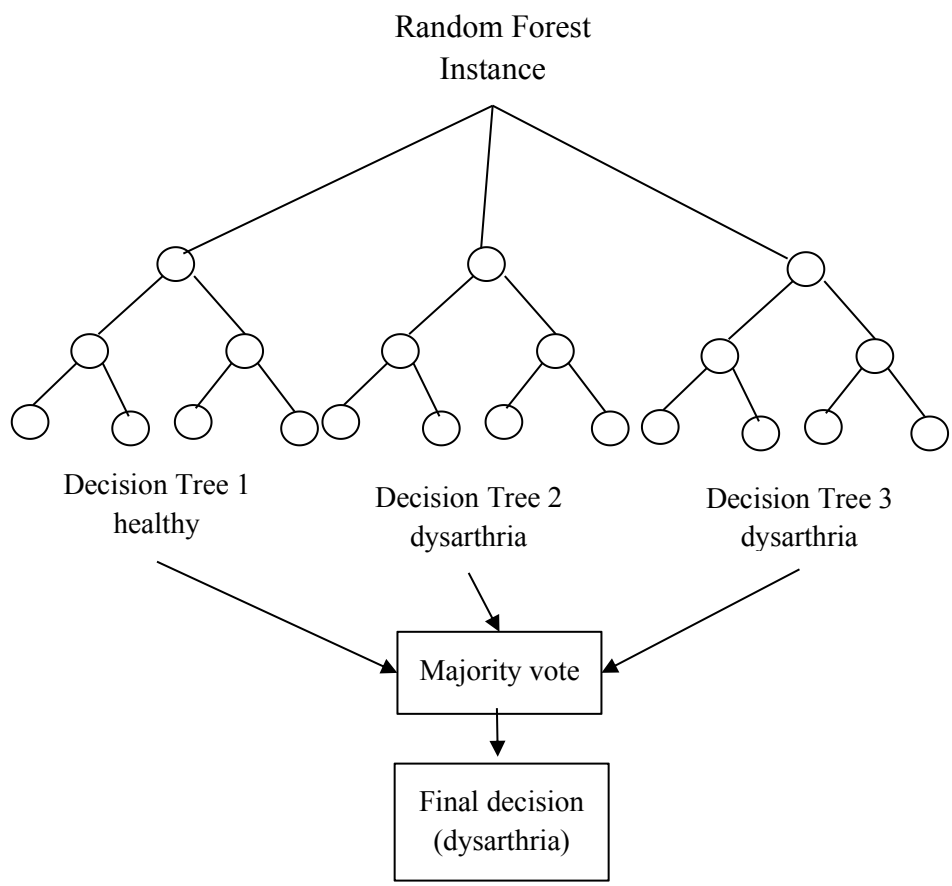

Figure 6. *Simplified example of a random forest classifier.*

4.3.2 Support Vector Machine

The next classifier is a support vector machine (SVM), which is the most commonly used classifier in machine learning, and in particular for impaired speech detection (Selouani et al., 2012; Dahmani et al., 2013; Kim et al., 2015; Orozco-Arroyave et al., 2016). The success of SVM's has not been limited to early studies, but continues to show good performance even in recent studies as they consistently perform well even with small datasets (López et al., 2019; Kodrasi & Bourlard, 2019; Tripathi et al., 2020). SVM is another supervised learning model which aims to find the maximum-margin hyperplane and margins for a given set of data points. Figure 7 shows an ideal case where the data points represent utterances from either healthy of dysarthric speakers. In order to maximize the margin we use the hinge loss function.

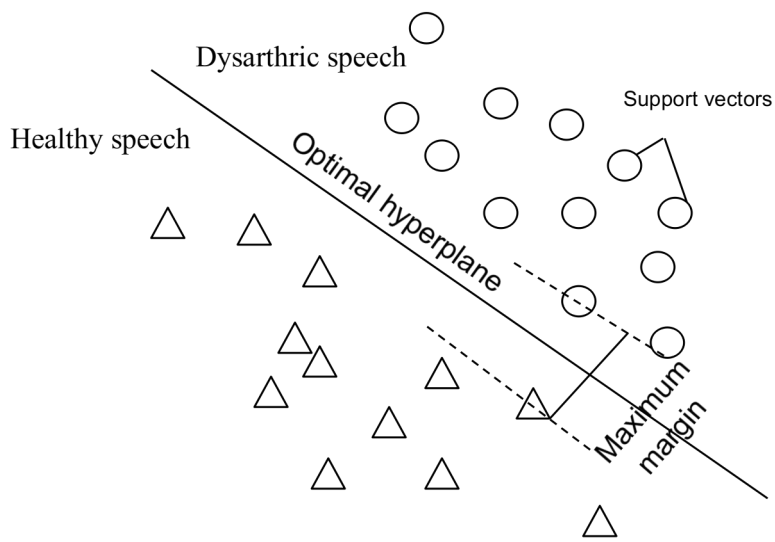

Figure 7. *Simplified example of a linear SVM.*

In many cases including the current study, the fact that we have several features means our data points are represented in a high-dimensional feature space. Therefore, we must account for the non-linear dimensionality by implementing a 'kernel trick' which will map our data points into the appropriate dimension space. For our SVM model we use a Gaussian radial basis function.

Another important aspect of SVM's are the C and gamma parameters, which must be optimized. C is the parameter for the margin cost function, which controls the influence of each individual support vector; this process involves trading error penalty for stability. A small C makes the cost of misclassification low (soft margin), allowing more of them for the sake of wider margin. A large C makes the cost of misclassification high (hard margin), forcing the algorithm to explain the input data stricter and potentially overfit. The goal is to find the balance between a too soft margin or a too hard margin.

The gamma parameter relates to the kernel method and defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close'. If the gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting. When gamma is very small, the model is too constrained

and cannot capture the complexity of the data. Again, we must find a good balance between a gamma with a too high value or a too small value.

4.3.3 Feed-Forward Neural Network

The last and most complex classifier is the feed-forward neural network (FFNN) which is a type of artificial network that sends information between nodes in a single direction. The literature on neural networks is vast and beyond the scope of this paper, but the most basic FFNN is a multilayer perceptron (MLP) that learns a function f(): $R^n \rightarrow R^0$ by training on a dataset, where n is the number of dimensions for the input and 0 is the number of dimensions for the output. Given a vector of acoustic features $X = x_1, x_2,.....,x_n$ and some targets y (labels regarding diagnosis) an MLP can learn a non-linear function approximator for classification.

Figure 8 shows a simplified MLP where we have an input layer $X = X_1\ldots$ of acoustic features with values that gets combined with some weights $a = a_1\ldots$, and eventually a prediction gets made on whether the acoustic vector was a representation of healthy speech (H) or dysarthric speech (D). Hyperparameter tuning of layers, nodes, learning rate, optimizer, epochs, is very important and we apply a grid search to find these optimal parameters.[3]

[3] The exact parameters are discussed in Ch. 5.

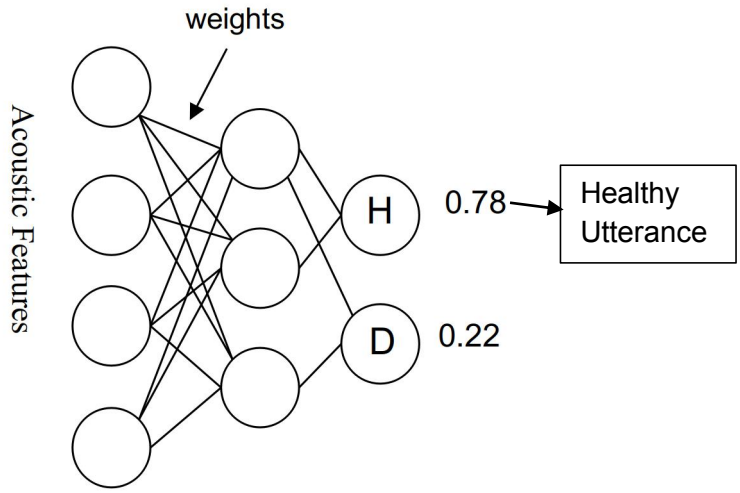

Figure 8. *Simplified example of MLP classifier.*

4.4. Mel-Frequency Cepstral Coefficients

As a baseline, we compared the performance of classifiers when solely trained on Mel-frequency cepstral coefficients compared to different sets of prosodic features. The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC and are commonly used in ASR systems as they are a good approximation to the vocal tract and thus model pronunciation well.

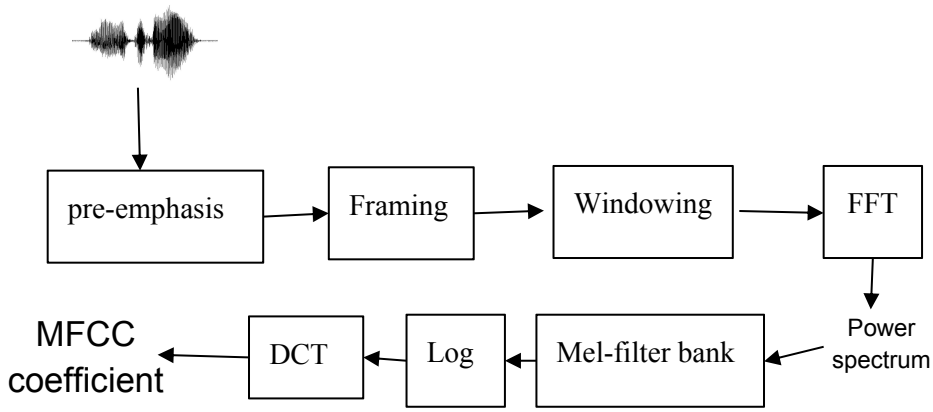

Figure 9. *Main steps for computing MFCCs.*

There are several steps to calculate MFCCs which can also be seen from figure 9. First, apply a pre-emphasis filter on the speech signal to amplify the high frequencies and then take the Fourier transform of the signal within a defined window. Second, we map the powers of the spectrum obtained above onto the mel scale with overlapping windows. The mel scale is used since it is a better representation of the human auditory system which is not linear. Third, we take the log of the powers at each of the mel frequencies. Fourth, we take the discrete cosine transform of the log filterbank energies. The MFCCs are the amplitudes of the resulting spectrum.

One issue with using MFCC as input for machine learning models is the varied sequence nature of data. Naturally, our speech samples vary in length which leads to variable sequence vectors. However, all machine learning models require a fixed sequence as input. Therefore, we must apply some pre-processing techniques to produce MFCCs with a fixed length despite utterances with different durations. In our case it was required to apply different methods for English and Korean speakers. For Korean speakers we simply averaged each coefficient for each utterance. Since we extract 13 coefficients[4] the output ends up being a vector of length 13. While this method was sufficient and led to good results for Korean, English required a different process. For English, we averaged each coefficient to contain 5 frames[5] leading to a vector of length 65 (13*5) for each utterance.

[4] We experimented with different numbers of coefficients (see appendix table A2) and 13 was ideal.
[5] Different numbers of frames such as 3,4,6,7 were also tested.

Chapter 5. Experiment

We conducted several experiments using many different sets of prosodic features. The main experiments are dysarthria detection, severity assessment of dysarthria and cross-language assessment of dysarthria. However, for each of those experiments we also conduct several other experiments to draw comparisons. First, we evaluate the performance of models when trained on only MFCCs, then we check the performance when training on the full prosodic feature set. Then, we train on single prosodic features (e.g. only pitch, etc.). Next, we compare the prosodic features based from previous studies. Lastly, we show the performance of our feature set when applying recursive feature selection.

In most cases, machine learning models are sensitive to feature scaling. For example, an SVM model assumes that all features are centered around zero and have variances in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected. Therefore, for all prosodic measurements we center to the mean and followed by component wise scale to unit variance. We apply this method of scaling measures for all prosodic measurements, MFCC, and for all classifiers.

5.1. Model Parameters

As previously mentioned, hyperparameter tuning is an important part of building good classifiers. We implement a grid search which is an exhaustive search over specified parameter values for a given classifier. These parameters are optimized by also applying cross-validation.

Only two parameters are optimized for the random forest classifier, number of trees and depth of trees. In general, 100 trees were most optimal, while the optimal depth ranged from 30 to None, where none means all nodes are expanded until all leaves are pure or until all leaves contain less than the minimum number of samples required to split an internal node.

The SVM model had C and gamma values optimized by checking values between 10^{-4} to 10^{4}. For detection this tended to be 0.1 for gamma and 10 for C, while for assessment it was 0.01 for gamma and 10 for C. Furthermore, we tested different kernels such as poly, sigmoid and RBF, and we found RBF to provide the best results. As SVMs are inherently binary classifiers, we apply a one-versus-one approach when building the models for severity assessment. This method creates multiple binary models where n * (n - 1) / 2 classifiers are constructed and each one trains data from two classes.

Lastly, for our multi-layer perceptron we optimized the number of hidden layers, nodes, activation function, solver, learning rate, learning rate scheduler and the maximum iterations. Only 1 hidden layer with 100 nodes

was needed for detection, while assessment performed better with two hidden layers with the first containing 100 nodes and the 2nd containing 50 nodes. In both cases the ReLU activation function outperformed the logistic or tanh function. The solver for detection and assessment was Adam, which gave better results than standard stochastic gradient descent or limited-memory BFGS solvers. An initial learning rate of 0.001 was used with an adaptive learning rate. An adaptive learning rate keeps the learning rate constant as long as loss continues to decrease, otherwise the learning rate is reduced. Lastly, the optimized number of epochs until convergence was around 500 for all cases. The only difference between detection and assessment experiments relates to the activation function for the last layer. Since detection is a binary task, we use a logistic function, while the assessment is a multiclass task so we use a softmax function.

5.2. Training Procedure

5.2.1 Dysarthria Detection

The training procedure for dysarthria detection is similar for both TORGO and QoLT dataset with few differences (see figure 10). We first extract both MFCC and prosodic features. Then, we do some pre-processing of features such that we get a fixed length for MFCCs and a reduced set of prosodic features. Lastly, we concatenate both MFCCs and prosodic features and feed them into our classifier to make a prediction. All 15

speakers from the TORGO dataset were present in the training and test sets. In total, we collected 160 sentences which were split such that no sentence in the train set was in the test set. This led to 200 utterances for training and 140 for testing. Given that there are more dysarthric utterances than healthy, we balanced the dataset so there is an equal amount of utterance per group (100 for healthy 100 for dysarthric). Before validating our model on the test set, we implement a 10-fold cross-validation.

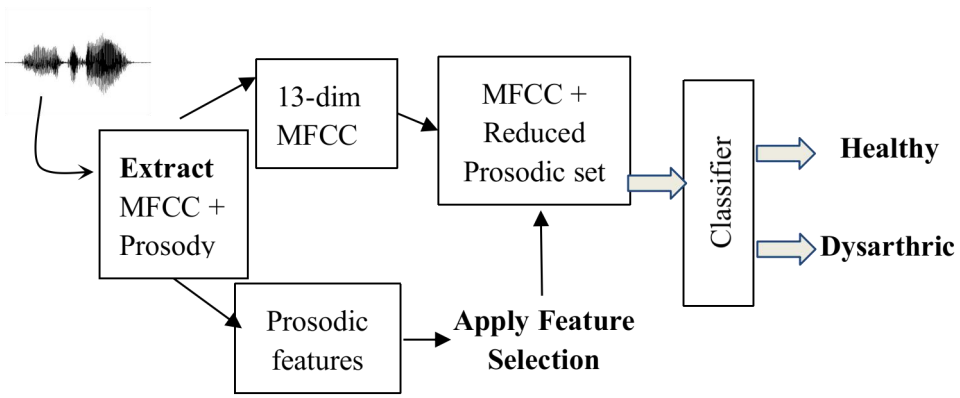

Figure 10. *Overall process for detecting dysarthria given a speech signal.*

Unlike the TORGO dataset, the QoLT dataset has a large number of speakers which allows us to test speaker-independent models such that there are no overlapping speakers in train and test sets. However, because the number of recorded sentences per speaker is 10 (5 sentences * 2), the overall amount of utterances is much lower. In total, we collected 380 utterances, 100 from healthy speakers and 280 from dysarthric speakers. For training we use the data from 6 healthy speakers and 17 dysarthric speakers,

while for testing we use 4 healthy speakers and 11 dysarthric speakers. As with the TORGO experiments, we balance the data so that both groups have an equal amount of total utterances.

5.2.2 Severity Assessment

The training procedure for assessment is similar to detection except now we have different levels of severity (see figure 11). We divided severity into four levels, healthy, mild, moderate and severe. This was based on the assessments of speech therapists during data collection of each dataset. For the TORGO database each group has an average of 85 utterances for training and 60 utterances each for testing[6]. The QoLT data has 60 utterances in each group for training and 35 for testing.

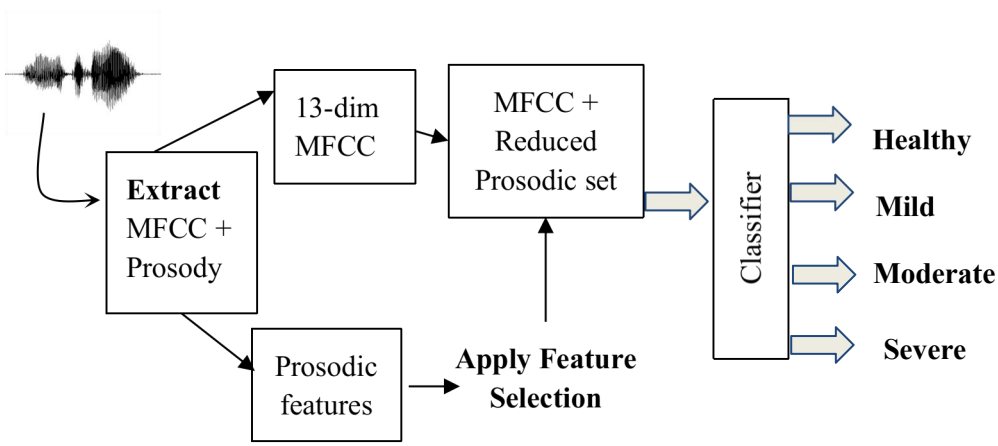

Figure 11. *Overall process for assessing the severity level of a given speech signal.*

[6] Each speaker in the TORGO dataset has a different number of recordings.

5.2.3 Cross-Language

The training procedure is again very similar to the assessment process from the previous section. However, during training we include data from both languages while testing on one language. For example, when testing on Korean data we train with both English and Korean data. Cross-language experiments were only conducted with the severity assessment task as it's a more difficult task than detection. Furthermore, when testing on a specific language we only included data from dysarthric speakers of the other language during training. For example, when testing with Korean data we include all Korean data (healthy and dysarthric) for training, plus English dysarthric data.

Chapter 6. Results

6.1. TORGO

6.1.1 Dysarthria Detection

As seen from Table 9, including our feature selected prosodic set improves on all classifiers compared to the baseline models which only use MFCCs. A relative accuracy increase of 1.84% and 1.82% was seen for SVM and MLP models respectively. In particular, we see a higher recall 96.7% to 100% which shows our model is correctly predicting all utterances coming from dysarthric speakers (zero false negatives). Recall is an important metric, as we want to correctly diagnose utterances that come from speakers with dysarthria.

Table 9. *Evaluation of baseline model and feature selected prosodic features.*

Feature set	Classifier	Accuracy %	Precision %	Recall %	F1-score %
MFCC	RF	91.6	89.1	95	92
	SVM	92.4	90.5	95	92.3
	MLP	**93.3**	**90.6**	**96.7**	**93.5**
MFCC+ prosody	RF	92.4	88	98.3	92.9
	SVM	94.1	89.6	**100**	94.5
	MLP	**95**	**90.9**	**100**	**95.2**

To compare the prosodic features selected from recursive feature selection, we also evaluated the performance of our MLP classifier when only trained on a single prosodic group. Results from table 10 suggest that

our selected features outperform any individual prosodic group. While pitch (92.4%) and voice quality (93.2%) came close to the 95% accuracy of RFE selected features, they still had a lower recall which means the presence of false negatives.

Table 10. *A comparison of MLP results when trained on individual prosodic measures.*

Feature set	Accuracy %	Precision %	Recall %	F1-score %
RFE selected features	**95**	**90.9**	**100**	**95.2**
Pitch	92.4	89.2	96.7	92.8
Voice Quality	93.2	90.6	96.7	93.5
Speech rate	90.8	86.6	96.7	91.3
Rhythm	89.1	86.2	93.3	89.6

We also compare results when using the features sets from previous studies utilizing prosodic features. Table 11 shows the results of our selected features from those proposed in previous studies. As the exact features are not known from all studies we approximate based on description of prosodic features. For example, Kim et al. (2015) only used F0 and duration measures as a representation of prosody, while Brocket et al. (2011) used F0, duration and voice quality to represent prosody. Kadi et al.

(2013), and Dahmani et al. (2013) were more specific with their feature choice and we were able to test the exact features used in their studies.

Table 11. *A comparison of other prosodic representation from previous studies.*

Feature set	Features	Accuracy %	Recall %	F1-score %
RFE selected features	f0_std, f0_quantile25, f0_quan_75, # of pauses, full duration, jitter, delta-V, varco-V	**95**	**100**	**95.2**
Kim et al. (2015)	All F0 and duration measures	90.76	95	91.2
Kadi et al. (2013)	%V, AR, mean F0, std F0, voice break, HNR, Jitter, Shimmer	93.3	96.7	93.5
Dahmani et al. (2013)	%v, delta-V	90.7	91.7	90.9
Martens et al. 2013	SR, AR, # of pauses, # of syllables	89.9	93.3	90.3
Bocket et al. (2011)	All F0, duration, pauses, jitter, shimmer	92.4	95	92.7

Lastly, there have been several other studies using a variety of features which may or may not include prosodic measures. We compare the accuracy of more recent studies on dysarthric speech detection using the same TORGO database seen in table 12. The previous study with the best results comes from Narendra et al. (2018) who reach an accuracy of 94.29% when using over 6,500 features including MFCC, prosody and glottal

features. However, our feature set produces better results while only using 8 specific features.

Table 12. *Comparison from other studies using the TORGO dataset.*

Study	Results (accuracy)	# Features used
Current Study	**95%**	8
Narendra, N. P., & Alku, P. (2018)	94.29%	6552+ glottal features (open-smile2)
Kim et al. (2015)	93.4%	11
Millet, J., & Zeghidour, N. (2019)	82.4% UAR	32+
Narendra, N. P., & Alku, P. (2020)	82.12 %	6744
Jung & Kim (2017)	89.5 %	16

6.1.2 Severity Assessment

Severity assessment is a much more difficult task as there are more ambiguities between mild/healthy, mild/moderate, and moderate/severe classes. As expected, overall lower results are achieved compared to detection, however, we see better improvements when including prosody measures (see table 13). When including prosodic features we see a relative accuracy improvement of 6.87%, 11.64% and 20.16% with RF, SVM and MLP classifiers respectively. From figure 12, we see that a majority of improvement with our MLP model is with better classification of mild

utterances. Common mistakes are predicting moderate when an utterance is from a severe speaker or predicting mild when an utterance comes from a moderate speaker.

Table 13. *Results from severity assessment task.*

Classifier	Accuracy (%) MFCC only	Accuracy (%) MFCC+prosody	Relative accuracy increase
RF	58.2	62.2	6.87 %
SVM	**63.6**	71	11.64 %
MLP	61.5	**73.9**	**20.16 %**

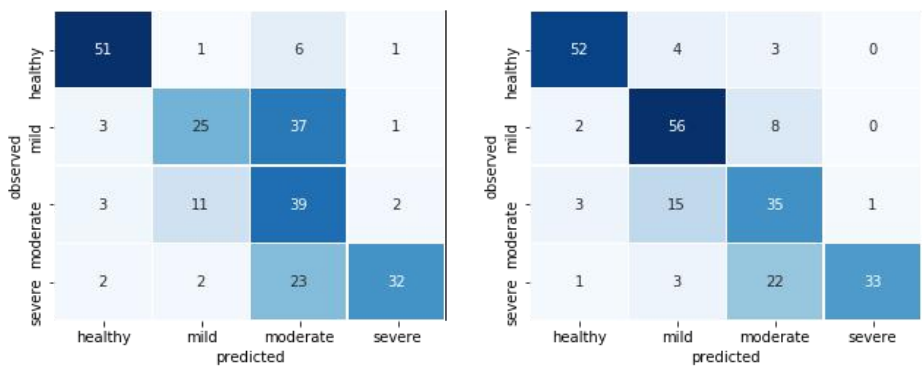

Figure 12. *Confusion matrix of MLP predictions for baseline (left) and proposed feature set (right).*

6.2. QoLT

6.2.1 Dysarthria Detection

The RFE selected features for the QoLT dataset differed from the TORGO dataset. The full set of features are seen in table 14. The features in bold are those that were also in the feature set for TORGO. There were no

major differences between the features sets for TORGO and QoLT datasets. One noticeable difference for detection is that TORGO only uses 2 voice quality (jitter & shimmer) and 1 speech rate (# of pauses) feature, while the QoLT dataset uses 4 voice quality measures and 3 speech rate features. Similarly, for assessment TORGO used 6 rhythm and pitch features while QoLT only used 4. However, speech rate was again utilized more in the QoLT dataset, 6 features, compared to the TORGO dataset (4 features). A deeper investigation regarding these differences are discussed in chapter 7.

Table 14. *RFE selected features for QoLT dataset.*

Features	Detection	Assessment
Pitch	**f0_mean**, **f0_quan_75**	**f0_mean, f0_median, f0_max, f0_quantile25**
Voice Quality	# of voice breaks, % of voice breaks, **Jitter**, Mean HNR	# of voice breaks, **Degree of voice breaks, Jitter, Mean HNR**
Speech Rate	# of syllables, **# of pauses**, rate of speech	# of syllables, **# of pauses**, rate of speech, **speaking duration, original duration, balance**
Rhythm	Crpvi	**delta-V, Vrpvi, Crpvi, Vnpvi**
Total # of features	10	18

6.2.1 Severity Assessment

Results from the severity assessment are more promising and show improvement over baseline models for all classifiers (see table 15). A relative accuracy improvement of 8.47%, 12.5% and 20.24% was seen for

MLP, SVM and RF classifiers respectively. Most improvements were seen with mild and moderate utterances. A recall increase from 33.3% to 63% was seen for mild utterance while an increase from 56.7% to 66.7% was seen for moderate utterance. Healthy utterances were also more accurate when including prosodic features (see figure 13). A precision and recall increase to 100% was seen for healthy utterances. This implies no healthy utterance was misdiagnosed as dysarthric and no dysarthric utterance was misdiagnosed as healthy.

Table 15. *Results for severity assessment in the QoLT dataset.*

Classifier	Accuracy (%) MFCC only	Accuracy (%) MFCC+prosody	Relative accuracy increase
RF	58.3	**70.1**	**20.24 %**
SVM	**60**	67.5	12.5 %
MLP	56.7	61.5	8.47 %

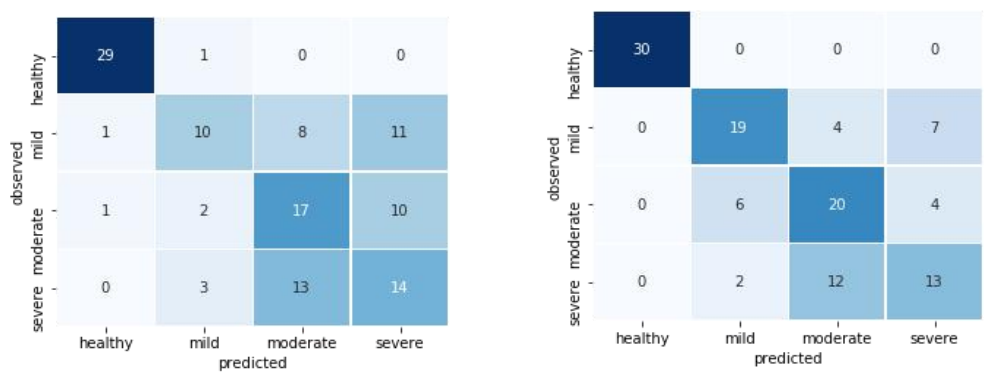

Figure 13. *Confusion matrix of random forest predictions for baseline (left) and proposed feature set (right).*

6.3. Cross-Language Assessment

The last group of experiments are to determine whether we can supplement low-data training sets with data from an outside language. In this case, this means train with both Korean and English data. Depending on what language we tested on, we made sure to balance the groups when training. Also, we found that including specific groups outperformed including all data. For example, when testing with Korean data if we include all the data from English, then mild utterances were almost always incorrectly predicted as healthy. Therefore, we only included utterances from dysarthric speakers when training with a Korean test set. Results can be seen in table 16. In general overall improvements are seen for all models when including English data. A higher relative increase was seen for the RF classifier (4.12%) but a higher accuracy was achieved when using an SVM.

Table 16. *Results for cross-language experiment when testing with QoLT.*

Classifier	Accuracy (%) Korean only data	Accuracy (%) Korean + English	Relative accuracy increase
RF	58.3	60.7	**4.12 %**
SVM	**65.9**	**68.2**	3.49 %
MLP	58.1	59.8	2.93 %

When looking at the selected features from table 17 we see that speech rate measures are very important, and all 7 features were selected. Compared to the features selected when only training on Korean data, we

see voice quality being used less, but rhythm being used more. It's difficult to determine how the cross-language features compare to the Korean and English selected features on their own, but as expected there is a trend towards using features more helpful for Korean. The added features not in table 17. were also not in the feature set for English severity assessment (articulation rate and %V).

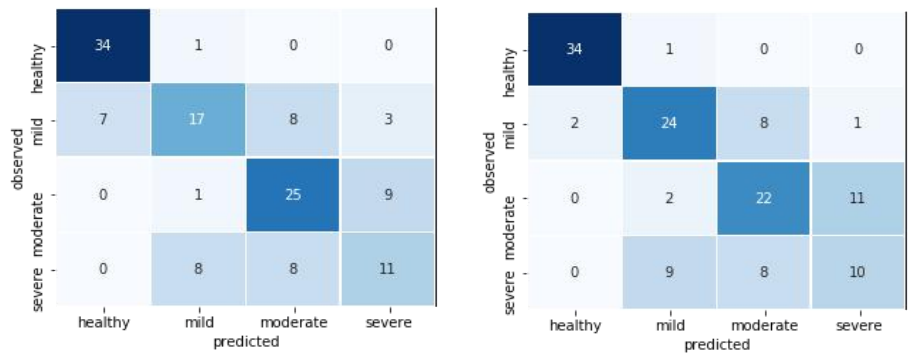

Figure 10. *Confusion matrix for cross-language assessment when only using Korean training data (left) and when using both Korean and English (right).*

Table 17. *RFE selected features for cross-language assessment.*

Features	Test with Korean	Test with English
Pitch	f0_mean, f0_median, f0_quan_75	f0_mean, f0_median, f0_quan_75, f0_quan 25, f0_std
Voice Quality	# of voice breaks, Degree of voice breaks, Shimmer	Degree of voice breaks, Shimmer, jitter, mean HNR
Speech Rate	# of syllables, # of pauses, rate of speech, articulation rate, speaking duration, original duration, balance	# of pauses, speaking duration, original duration, balance
Rhythm	%V, Vrpvi, Crpvi, Vnpvi, Cnpvi	Delta-V, delta-C, varco-C, Vrpvi, Crpvi, Vnpvi

In the opposite case where we test using English data, there are less improvements overall. However, we see better performance when identifying speakers with mild dysarthria (see figure 11). A precision increase from 65.9% to 73.3% was seen for the mild group, and a recall increase from 57.6% to 66.7% was achieved. In table 18, we see that more improvements were obtained for the random forest and MLP classifier but a higher accuracy with the SVM model.

Table 18. *Results for cross-language experiment when testing with TORGO.*

Classifier	Accuracy (%) English only data	Accuracy (%) Korean + English	Relative accuracy increase
RF	58.4	63.4	**8.5%**
SVM	**66.8**	**67.2**	0.6%
MLP	63	64.7	2.70

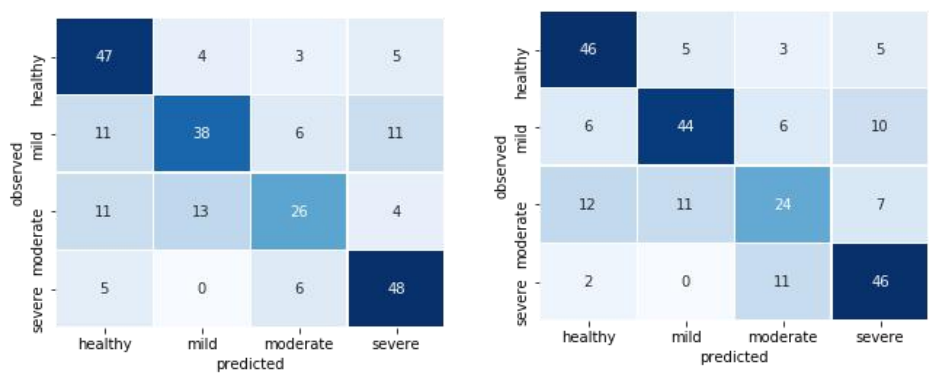

Figure 11. *Confusion matrix for cross-language assessment when only using English training data (left) and when using both Korean and English (right).*

Chapter 7. Discussion

Results from the dysarthria detection experiments suggest that prosody can help improve detection and severity assessment, but it may be dependent on the data. Detection was helpful for the TORGO dataset but minimally helpful for the QoLT dataset. Prosody is better utilized for severity assessment, as relative increases of around 20% were seen for both datasets. This is likely because prosodic impairments are severity dependent and may not generalize to all speakers with dysarthria. Lastly, based on the cross-language experiments we see that including prosodic features from a different language can help improve assessment. Features related to common prosodic impairments such as speech rate were correctly selected by the RFE algorithm. The rest of the discussion section will go over the results and specific features used by the three main experiments.

7.1. Linguistic Implications

The TORGO dataset saw about a 2% relative increase in accuracy but the QoLT dataset saw almost no improvements for detection. Given that the QoLT already had a higher accuracy with the baseline MFCC features, we cannot make any claims whether this is caused by language differences. Important distinctions exist between the datasets such as number of speakers, stimuli along with methodological differences such as the data split (speakers vs sentences) which could have contributed to the difference. As

for the selected features, there were some interesting differences when applying the RFE for each dataset. The TORGO dataset made more use of pitch (3), voice quality (2), and rhythm (2) features, while the QoLT dataset used more speech rate (3) and voice quality (4) features. The usefulness of speech rate features for Korean are supported the findings of Kim and Choi (2017) who found articulation rate to be a significant factor for predicting speech intelligibility in Korean speakers with dysarthria but not for English speakers. Furthermore, their hypothesis that the variation in the rhythm metric npvi-V would be larger for English speakers than Korean is also supported. This hypothesis was based on the fact that English speakers with hypokinetic dysarthria tend to equalize the duration of syllables despite English being a stress-timed language. In our prosodic feature set we see npvi-V along with rpvi-V for the TORGO dataset but not the QoLT which only contains the rpvi-C metric.

It is unclear why the largest group of selected prosodic measurement was voice quality for Korean speakers, but the findings of Kim et al. (1998) and Lee et al. (2000) suggest that jitter (which was selected in our feature set) is significantly increased in comparison to healthy controls. Previous studies have found other voice quality measures such as Linear Prediction residual signal (Kim & Kim, 2012) and Cepstral Peak Prominence (Seo & Seong, 2013) to also be useful for dysarthria detection. Future studies

should compare the performance of different voice quality measures to determine which type of features are most useful.

The number features selected during severity assessment were much higher than the ones selected for detection. In general, the prosodic features were more evenly distributed for both languages. The TORGO dataset utilized 6 features each from both pitch and rhythm groups, and 4 features each from the speech rate and voice quality groups. The QoLT selected 6 features from the speech rate group and 4 features each for the voice quality, rhythm and pitch groups. As seen in table 19, few differences were seen between language groups.

Table 19. *Prosodic features used in one dataset but not used in another.*

Features	Prosodic features used for English Severity assessment but not Korean	Prosodic features used for Korean Severity assessment but not English
Pitch	f0_quan_75, f0_std	None
Voice Quality	shimmer	# of voice breaks
Speech Rate	None	# of syllables, speaking rate
Rhythm	delta-C, varco-C,	None

Relevant differences were found when comparing detection and assessment. The findings of Schlenck et al. (1993) regarding severity differences seem to be apparent when looking at pitch features. Schlenck et

al. found that speakers with mild dysarthria had a lower F0 variation and speakers with severe dysarthria had a higher F0 than healthy controls. While the features F0_max and F0_std were not present in the detection feature set, they were present in the assessment set. This suggests that more relevant and refined features are needed to accurately distinguish different severities in comparison to simply distinguishing between healthy and dysarthric speakers in general. Furthermore, Ziegler, Hartmann and Hoole (1993) found the duration of syllables to be correlated with severity such that the more severe the longer the syllable duration. This is realized in our feature set for assessment which includes several durational measures in both languages. Interestingly, very few duration-based measure was selected for detection; 2 measures for detection in both languages but 9 for English and 8 for Korean when selecting for assessment. This shows the importance of duration differences when taking into account different severity groups.

7.2. Clinical Applications

Automatic detection of dysarthria has important applications in a clinical setting. We are not suggesting an automatic approach to replace speech therapists, but instead automatic methods can be used as a tool in conjunction with a therapist. An automatic approach to detection provides a more objective method of diagnosis compared to the traditional perceptual evaluation method. Furthermore, an automatic method of detection can be

more cost effective as it would be quicker to administer than the traditional approach. Future studies, however, should also incorporate an automatic approach to diagnosis that provides information on what prosodic aspects are more damaged.

Automatic severity assessment also has the same benefits of detection but has the added benefit of distinguishing different severity levels. Furthermore, being able to detect dysarthric speech from speakers with mild dysarthria is important for early detection. Perceptually, it is difficult to diagnose a speaker who has mild dysarthria as their speech is minimally affected. An automatic approach of detecting mild dysarthria can help speech therapists provide early treatment for these individuals.

Results from the cross-language experiments also show promising results for clinical applications. Individuals suffering from dysarthria who are from an underrepresented language can undergo diagnosis by using computations models trained on more represented languages. For example, we might be able to automatically diagnose dysarthria from an individual who speaks Mongolian by extracting language independent prosodic features from a model trained with Korean data. Knowledge regarding language independent impairments can also assist training models with low data. Regardless if the language is widely spoken, impaired speech is always difficult to collect and incorporating data from multiple language can alleviate the issue of low data.

Chapter 8. Conclusion

In conclusion, our study found pitch, voice quality, speech rate and rhythm measures to be useful features for severity assessment and slightly useful for detection in English. A relative accuracy increase of 2% was seen for detection in the TORGO dataset, however, no improvement was seen for the QoLT dataset. For severity assessment a relative accuracy improvement around 20% was seen for both Korean and English datasets. The results from the cross-language experiments were promising showing a relative increase of 4.12% when testing on the QoLT dataset and an increase of 8.5% when testing on the TORGO dataset.

The optimal set of prosodic features was selected by the RFE feature selection algorithm, but the exact selected features depended on the language group and task. For detection in English, pitch (standard deviation, both quantiles), speech rate (number of pauses, full duration), and rhythm (delta-V, varvoV) measures were most helpful for detection but jitter was also selected. Detection in Korean utilized more voice quality (number of voice breaks, degree of voice breaks, jitter and mean HNR) and speech rate features (number of syllables, number of pauses, speaking rate) but also had some pitch measures (mean and 75% quantile) and the Crpvi rhythm measure.

In most cases, the features selected for detection were also selected for severity assessment in both languages. In the case of English, increases

in the number of selected features was seen for all prosodic groups but in particular for voice quality and rhythm. For severity assessment in Korean, speech rate and rhythm measures contributed more measures in comparison to detection. In particular, more duration measures in general but also duration measures related to vowels.

In regards to language independent or dependent features we see that duration measures along with mean F0 tend to be useful for both Korean and English. However, pitch in general appeared to be more useful for English, while speech rate features were more helpful for Korean. Future studies with other databases in Korean and English should validate whether the previously mentioned features are truly language independent/dependent or if the patterns are limited to the databases used in our study.

Results from testing individual prosodic groups show that a holistic approach that includes multiple aspects of prosody is superior to focusing on single prosodic groups. Furthermore, the method of feature selection is very important to optimally select the most relevant features as some features may not be helpful. Lastly, future studies should further investigate the use of other prosodic features in severity assessment and detection as our study only utilized a select set of 28 features.

References

Ackermann, H., & Hertrich, I. (1994). Speech rate and rhythm in cerebellar dysarthria: An acoustic analysis of syllabic timing. *Folia Phoniatrica et Logopaedica*, *46*(2), 70-78.

An, K., Kim, M. J., Teplansky, K., Green, J. R., Campbell, T. F., Yunusova, Y., ... & Wang, J. (2018). Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks. In *INTERSPEECH* (pp. 1913-1917).

Bocklet, T., Nöth, E., Stemmer, G., Ruzickova, H., & Rusz, J. (2011, December). Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 478-483). IEEE.

Paul Boersma & David Weenink (2020): Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved March 2020 from http://www.praat.org/

Bunton, K., Kent, R. D., Kent, J. F., & Rosenbek, J. C. (2000). Perceptuo-acoustic assessment of prosodic impairment in dysarthria. Clinical Linguistics & Phonetics, 14(1), 13-24.

Darkins, A. W., Fromkin, V. A., & Benson, D. F. (1988). A characterization of the prosodic loss in Parkinson's disease. *Brain and Language*, *34*(2), 315-327.

Dahmani, H., Selouani, S. A., O'shaughnessy, D., Chetouani, M., & Doghmane, N. (2013). Assessment of dysarthric speech through rhythm metrics. Journal of King Saud University-Computer and Information Sciences, 25(1), 43-49.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential diagnostic patterns of dysarthria. Journal of speech and hearing research, 12(2), 246-269.

Darley, F. L., Aronson, A. E., & Brown, J. R. (1975). Motor speech disorders. Saunders.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, *41*(2), 385-390.

Dellwo, V., & Wagner, P. (2003). Relationships between speech rate and rhythm. In Proceedings of the ICPhS.

Dogan, M., Midi, I., Yazıcı, M. A., Kocak, I., Günal, D., & Sehitoglu, M. A. (2007). Objective and subjective evaluation of voice quality in multiple sclerosis. *Journal of Voice*, *21*(6), 735-740.

Duffy, J. R. (2013). Motor speech disorders: Substrates, differential diagnosis, and management. St. Louis, MO: Elsevier.

Choi, D. L., Kim, B. W., Kim, Y. W., Lee, Y. J., Um, Y., & Chung, M. (2012, May). Dysarthric Speech Database for Development of QoLT Software Technology. In LREC (pp.3378-3381).

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

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. Papers in laboratory phonology, 7(515-546)

Hong, S. M., Jeong, P. Y., Sim, H. S., Hong, S. M., Jeong, P. Y., & Sim, H. S. (2018). Comparison of Perceptual Assessment for Dysarthric Speech: The Detailed and General Assessments. Communication Sciences & Disorders, 23(1), 242-253.

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, *71*, 1-15.

Johns, D. F. (Ed.). (1985). Clinical management of neurogenic communicative disorders. Little, Brown.

Kadi, K. L., Selouani, S. A., Boudraa, B., & Boudraa, M. (2013). Discriminative prosodic features to assess the dysarthria severity

levels. In *Proceedings of the World Congress on Engineering* (Vol. 3).

Kang, Y., Yoon, K., Seong, C., & Park, H. (2012). A preliminary study of the automated assessment of prosody in patients with Parkinson's disease. Communication Sciences & Disorders, 17(2), 234-248.

Kang, Y., Seong, C. J., & Yoon, K. C. (2011). A study of prosodic features of patients with idiopathic Parkinson's disease. *Phonetics and Speech Sciences*, *3*(1), 145-151.

Kearns KP, Simmons NN (1988). Interobserver reliability and perceptual ratings: more than meets the ear. J Speech Hear Res. 1988;31:131–36

Kent, R. D., Kent, J. F., & Rosenbek, J. C. (1987). Maximum performance tests of speech production. Journal of Speech and Hearing Disorders, 52, 367–387.

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. Journal of communication disorders, 32(3), 141-186.

Kim, M. J., Cao, B., An, K., & Wang, J. (2018). Dysarthric Speech Recognition Using Convolutional LSTM Neural Network. In INTERSPEECH (pp. 2948-2952).

Kim, Y., & Choi, Y. (2017). A cross-language study of acoustic predictors of speech intelligibility in individuals with Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *60*(9), 2506-2518.

Kim, M. J., & Kim, H. (2012). Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Kim, H. G., Kim, W. H., Seo, J. H., Hong, K. H., Shin, H. K., & Ko, D. H. (1998). Some clinical aspects of dysarthria. *Speech Sciences*, *3*, 38-49.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*.

Kim, J., Kumar, N., Tsiartas, A., Li, M., & Narayanan, S. S. (2015). Automatic intelligibility classification of sentence-level pathological speech. *Computer speech & language*, *29*(1), 132-144.

Kodrasi, I., & Bourlard, H. (2019, May). Super-gaussianity of Speech Spectral Coefficients as a Potential Biomarker for Dysarthric Speech Detection. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6400-6404). IEEE.

Korzekwa, D., Barra-Chicote, R., Kostek, B., Drugman, T., Lajszczak, M. (2019) Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech. Proc. Interspeech 2019, 3890-3894

Le Dorze, G., Ouellet, L., & Ryalls, J. (1994). Intonation and speech rate in dysarthric speech. *Journal of communication disorders*, *27*(1), 1-18.

Lee, Z. I., Oh, S. H., Lee, Y. S., & Kim, P. T. (2000). A Study on Acoustic Characteristics of Dysarthria in Athetoid Cerebral Palsy. *Journal of the Korean Academy of Rehabilitation Medicine*, *24*(4), 678-683.

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

López, J.V.E., Orozco-Arroyave, J.R., Gosztolya, G. (2019) Assessing Parkinson's Disease from Speech Using Fisher Vectors. Proc. Interspeech 2019, 3063-3067.

Lowit-Leuschel, A., & Docherty, G. J. (2001). Prosodic variation across sampling tasks in normal and dysarthric speakers. Logopedics Phoniatrics Vocology, 26(4), 151-164.

Mairano, P., & Romano, A. (2010). Un confronto tra diverse metriche ritmiche usando Correlatore 1.0. La dimensione temporale del parlato, 427, 44.

Mayle, A., Mou, Z., Bunescu, R. C., Mirshekarian, S., Xu, L., & Liu, C. (2019, September). Diagnosing Dysarthria with Long Short-Term Memory Networks. In INTERSPEECH (pp. 4514-4518).

Menendez-Pidal, X., Polikoff, J. B., Peters, S. M., Leonzio, J. E., & Bunnell, H. T. (1996, October). The Nemours database of dysarthric speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 3, pp. 1962-1965). IEEE.

Millet, J., & Zeghidour, N. (2019, May). Learning to detect dysarthria from raw speech. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5831-5835). IEEE.

Müller, J., Wenning, G. K., Verny, M., McKee, A., Chaudhuri, K. R., ‚Jellinger, K., Werner Poewe, Litvan, I. (2001). Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders. Archives of neurology, 58(2), 259-264.

Nam H., Kwon D., (2005). Prosodic Characteristics in the Persons with Spastic and Athetoid Cerebral Palsy. Journal of speech-language & hearing disorders, 14(2), 111- 127.

Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Gonzalez-Rativa, M. C., & Nöth, E. (2014, May). New Spanish

speech corpus database for the analysis of people suffering from Parkinson's disease. In *LREC* (pp. 342-347).

Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., ... & Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. The Journal of the Acoustical Society of America, 139(1), 481- 500.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73(3), 265- 292.

Rudzicz, F., Namasivayam, A. K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation*, *46*(4), 523-541.

Sandyk, R. (1995). Resolution of dysarthria in multiple sclerosis by treatment with weak electromagnetic fields. International Journal of Neuroscience, 83(1-2), 81-92.

Schlenck, K. J., Bettrich, R., & Willmes, K. (1993). Aspects of disturbed prosody in dysarthria. Clinical linguistics & phonetics, 7(2), 119-128.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., ... & Mohammadi, G. (2012). The interspeech 2012 speaker trait challenge. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Selouani, S. A., Dahmani, H., Amami, R., & Hamam, H. (2012). Using speech rhythm knowledge to improve dysarthric speech recognition. International Journal of Speech Technology, 15(1), 57-64.

Seo, I. H., & Seong, C. J. (2012). The prosodic characteristics of dysarthria with respect to speech rate and intonation slope. *Communication Sciences & Disorders*, *17*(3), 390-402.

Seo, I., & Seong, C. (2013). Voice quality of dysarthric speakers in connected speech. *Phonetics and Speech Sciences*, *5*(4), 33-41.

Traynor, B. J., Codd, M. B., Corr, B., Forde, C., Frost, E., & Hardiman, O. M. (2000). Clinical features of amyotrophic lateral sclerosis according to the El Escorial and Airlie House diagnostic criteria: A population-based study. *Archives of neurology*, *57*(8), 1171-1176.

Tripathi, A., Bhosale S. and Kopparapu S. K, "Improved Speaker Independent Dysarthria Intelligibility Classification Using Deepspeech Posteriors," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6114-6118

Tu, M., Berisha, V., & Liss, J. (2017, August). Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In *INTERSPEECH* (pp. 1849-1853).

Wannberg, P., Schalling, E., & Hartelius, L. (2016). Perceptual assessment of dysarthria: comparison of a general and a detailed assessment protocol. *Logopedics Phoniatrics Vocology*, *41*(4), 159-167.

Wong, K. H., Yeung, Y. T., Chan, E. H., Wong, P. C., Levow, G. A., & Meng, H. (2015). Development of a cantonese dysarthric speech corpus. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Zeplin J, Kent RD. Reliability of auditory-perceptual scaling of dysarthria. In: Robin DA, Yorkston KM, Beukelman DR, editors. Disorders of motor speech: assessment, treatment, and clinical characterization. Baltimore, MD: Paul H Brookes; 1996. p. 145–54.

Ziegler, W., Hartmann, E., & Hoole, P. (1993). Syllabic timing in dysarthria. *Journal of Speech, Language, and Hearing Research*, *36*(4), 683-693.

Zyski, B. J., & Weisiger, B. E. (1987). Identification of dysarthria types based on perceptual analysis. *Journal of Communication Disorders*, *20*(5), 367-378.

Appendix

Table A1. *Selected features for severity assessment using various feature selection methods for the TORGO dataset.*

Features	Filter Method	Lasso Method	Tree-based	RFE
Pitch	f0_max, f0_min, f0_quantile25, f0_mean	f0_std, f0_min, f0_max, f0_quantile25, f0_quan_75	all	f0_mean, f0_std, f0_median, f0_quantile25, f0_quan_75
Voice Quality	% of voice breaks, Mean HNR, # of voice breaks, Shimmer	# of voice breaks, % of voice breaks, Jitter, Shimmer	Mean HNR, % of voice breaks, Shimmer, Jitter, # of voice break	% of voice breaks, Jitter, Shimmer, Mean HNR
Speech Rate	speaking duration, # of pauses, articulation rate, balance, # of syllables, full duration	# of pauses, speaking rate, articulation rate, full duration, balance	articulation rate, # of pauses, speaking duration, balance	# of pauses, speaking duration, full duration, balance
Rhythm	Cnpvi, Crpvi, Vrpvi, delta-C, delta-V, varco-C	%V, varco-V, varco-C, Vrpvi, Crpvi, Vnpvi, Cnpvi	delta-C, delta-V, Crpvi	delta-V, delta-C, varco-C, Vrpvi, Crpvi, Vnpvi
# of features	20	21	20	19

Table A2. *Results with different MFCC parameters for TORGO dysarthria detection.*

# of Mel-coefficients	Accuracy	Precision	Recall	F1-Score
12	87.4	88.1	86.7	87.4
13	**93.3**	**90.6**	**96.7**	**93.5**
20	91.5	89.1	95	91.9
13 + Δ (26-dim)	86.5	87.9	85	86.4
13 + Δ + ΔΔ (39-dim)	83.2	84.5	81.7	83.1

Table A3. *Results with different features selection for TORGO dysarthria detection.*

Feature Selection Method	Accuracy	Precision	Recall	F1-Score
Filter	**97.6**	**100**	**95.6**	**97.7**
Lasso	96.4	100	93.3	96.5
Tree-based	94.11	100	88.9	94.1
RFE	**97.6**	**100**	**95.6**	**97.7**

Table A4. *Results with different features selection for QoLT dysarthria detection.*

Feature Selection Method	Accuracy	Precision	Recall	F1-Score
Filter	91.6	87.9	96.7	92.1
Lasso	89.9	86.4	95	90.5
Tree-based	93.3	89.4	98.3	93.7
RFE	**95**	**90.9**	**100**	**95.2**

국문초록

운율 정보를 이용한 마비말장애 음성 자동 검출 및 평가

말장애는 신경계 또는 퇴행성 질환에서 가장 빨리 나타나는 증상 중 하나이다. 마비말장애는 파킨슨병, 뇌성 마비, 근위축성 측삭 경화증, 다발성 경화증 환자 등 다양한 환자군에서 나타난다. 마비말장애는 조음기관 신경의 손상으로 부정확한 조음을 주요 특징으로 가지고, 운율에도 영향을 미치는 것으로 보고된다. 선행 연구에서는 운율 기반 측정치를 비장애 발화와 마비말장애 발화를 구별하는 것에 사용했다. 임상 현장에서는 마비말장애에 대한 운율 기반 분석이 마비말장애를 진단하거나 장애 양상에 따른 알맞은 치료법을 준비하는 것에 도움이 될 것이다. 따라서 마비말장애가 운율에 영향을 미치는 양상 뿐만 아니라 마비말장애의 운율 특징을 긴밀하게 살펴보는 것이 필요하다. 구체적으로, 운율이 어떤 측면에서 마비말장애에 영향을 받는지, 그리고 운율 애가 장애 정도에 따라 어떻게 다르게 나타나는지에 대한 분석이 필요하다.

본 논문은 음높이, 음질, 말속도, 리듬 등 운율을 다양한 측면에서 살펴보고, 마비말장애 검출 및 평가에 사용하였다. 추출된 운율 특징들은 몇 가지 특징 선택 알고리즘을 통해 최적화되어 머신러닝 기반 분류기의 입력값으로 사용되었다. 분류기의 성능은 정확도, 정밀도, 재현율, F1-점수로 평가되었다. 또한, 본 논문은 장애 중증도(경도, 중등도, 심도)에 따라 운율 정보 사용의 유용성을 분석하였다. 마지막으로, 장애 발화 수집이 어려운 만큼, 본 연구는 교차 언어 분류기를 사용하

였다. 한국어와 영어 장애 발화가 훈련 셋으로 사용되었으며, 테스트 셋으로는 각 목표 언어만이 사용되었다.

실험 결과는 다음과 같이 세 가지를 시사한다. 첫째, 운율 정보를 사용하는 것은 마비말장애 검출 및 평가에 도움이 된다. MFCC 만을 사용했을 때와 비교했을 때, 운율 정보를 함께 사용하는 것이 한국어와 영어 데이터셋 모두에서 도움이 되었다. 둘째, 운율 정보는 평가에 특히 유용하다. 영어의 경우 검출과 평가에서 각각 1.82%와 20.6%의 상대적 정확도 향상을 보였다. 한국어의 경우 검출에서는 향상을 보이지 않았지만, 평가에서는 13.6%의 상대적 향상이 나타났다. 셋째, 교차 언어 분류기는 단일 언어 분류기보다 향상된 결과를 보인다. 실험 결과 교차 언어 분류기는 단일 언어 분류기와 비교했을 때 상대적으로 4.12% 높은 정확도를 보였다. 이것은 특정 운율 장애는 범언어적 특징을 가지며, 다른 언어 데이터를 포함시켜 데이터가 부족한 훈련 셋을 보완할 수 있음을 시사한다.

주제어: 마비말장애, 운율, 머신러닝, 기계 학습, 분류기, 변수 선택, 음향학

학번: 2018-23331