APSIPA Transactions on Signal and Information Processing, 2023, 12, e18

Original Paper

Automatic Analyses of Dysarthric Speech based on Distinctive Features

Ka Ho Wong* and Helen Mei-Ling Meng

Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, The People's Republic of China

ABSTRACT

Dysathria is a neuromotor disorder that causes the individual to speak with imprecise articulation. This paper presents an automatic analysis framework for dysarthric speech, using a linguistically motivated representation based on distinctive features. Our framework includes a seq2seq phonetic decoder for Cantonese dysarthric speech. The manually or automatically transcribed phones can be mapped into a representation that consists of 21 distinctive features (DF). The DFs between the transcribed phones and canonical phones are compared in order to identify articulatory error rate (AER) for each DF. This forms an AER profile for a given set of dysarthric recordings from a speaker. Experiments show that the difference between the AER profile derived from manual versus automatic phonetic transcription is relatively small – with a root mean squared error (RMSE) of 0.053 for the word-reading task and 0.085 for the sentence-reading task in CU DYS. In addition, the correlations between the AER profiles are high, at 0.97 and 0.95 for the two tasks respectively. These results reflect the viability of the proposed framework as an automated means of processing dysarthric speech to achieve articulatory analyses described by DFs. The AER profile is intuitive and interpretable, for pinpointing problem areas in articulation.

Keywords: Dysarthric, distinctive features, recognition, sequence-to-sequence.

*Corresponding author: Ka Ho Wong, khwong@se.cuhk.edu.hk.

Received 21 November 2022; Revised 23 March 2023
ISSN 2048-7703; DOI 10.1561/116.00000077

1 Introduction

Dysarthria denotes a set of speech disorders related with neurological conditions and diseases such as cerebral palsy, traumatic brain injury, stroke, Parkinson's disease, or amyotrophic lateral sclerosis, which cause disturbances in muscular control over the speech production [12]. Therefore, dysarthria may result in unnatural and unintelligible speech with unstable prosody and imprecise articulation, which engender substantial communication difficulties for dysarthric patients. It was found in phonetic analyses that dysarthric subjects with cerebral palsy had the lowest syllable-initial consonant accuracy, followed by final consonants and then vowels [43]. It was also found that for hypokinetic (a type of dysarthria) dysarthric recordings, there are characteristics of rough voice, reduced pitch, mono-loudness and imprecise consonants [44]. Imprecision is also found in tone production and format frequency transitions [3, 22, 27].

The underlying reason for these imprecise pronunciations is reduced muscular control of the articulatory movements. For example, articulatory movements of dysarthric subjects with Parkinson's disease were recorded with 3-dimensional Electromagnetic Articulatography (EMA) [47], which shows reduced velocity and distance traveled in specific articulatory movements. Although EMA can help identify articulatory undershoot as a cause of articulatory imprecision, the methodology does not offer easy accessibility to analyzing the articulatory problems for dysarthric subjects. This motivates the current investigation into articulatory analyses of dysarthric speech recordings, to gain insights into the mechanisms leading to disordered speech production. Since manual analyses of data is costly, we also aim to develop methodologies to support automatic analyses of dysarthric speech data. We choose to represent speech in terms of distinctive features (DF), which present a linguistically motivated representation that focus on the manner and place of articulation, whereby a set of compact features that essentially takes on quantized values (e.g. positive and negative) is applicable to all the languages spoken by humans. We believe that the DF representation has much to offer in terms of elucidating insights on the articulatory mechanisms that lead to how dysarthric speech deviates from normal speech. A key challenge in this investigation is the difficulty in recruiting suitable subjects for recording, and even if we do manage to recruit, recording can be difficult because the subjects' conditions may cause them to be easily fatigued.

The rest of this paper is organized as follows: Section 2 presents the background of this study, including a literature review of previous efforts in automated processing of dysarthric speech. Since our investigation is based on speech data in Cantonese Chinese, which is the predominant language in Hong Kong and surrounding regions – we will present some background knowledge about this prominent Chinese dialect. We will also elaborate on background knowledge about distinctive features (DF). Section 3 describes a Cantonese dysarthric speech corpus, CU DYS, that we have designed and collected to support the current study. Section 4 presents the proposed notion of articulatory error rate (AER) that can be derived from manual phonetic transcriptions of deviant articulations in dysarthric speech. Since manual transcription is a laborious and costly process, Section 5 presents our work in the development of a seq2seq model for automatic phonetic transcription. Section 6 presents results on the performance of automatic phonetic transcription, and how the phones may be mapped into a DF representation for assessment of DF error rates. Furthermore, if we compare the DF representation derived from the automatically transcribed phones, with the DF representation derived from the canonical phones (mapped from the textual prompts), we can obtain the articulatory error rate (AER) for each DF, across all DFs, which can directly reflect the articulatory deviations of dysarthric speech. Section 7 discusses how the AER profile of a speaker can potentially inform the design of intervention plans, and can be used to predict the severity levels of dysarthric speech from manual assessments. Finally, Section 8 presents the conclusions and future directions.

2 Background

2.1 Automated Processing of Dysarthric Speech

Previous work can be classified into three main areas: automatic speech recognition (ASR) of dysarthric recordings, automatic intelligibility ratings and automatic error analyses. Research in disordered speech face the challenge of recruiting the appropriate speakers to provide their speech recordings. Furthermore, due to their conditions, the subjects may become fatigued more easily and the recording session must be kept short. Consequently, sparse data is the dominant problem. Various techniques have been studied, such as speaker/data selection [9], speaker-specific or severity-specific modeling and adaptive models [2, 17, 18, 39, 50], use of meta-learning with model re-initialization [42], deep metric learning [38] and fusion of acoustic and articulatory features [51].

Another area is automatic intelligibility rating. Speech intelligibility has a high correlation with phoneme recognition accuracy [41] and character recognition result [40]. Automatic intelligibility can be estimated directly, with i-vector or x-vector, in English [29], in French [20, 33], and in Korean [19]. The Fisher vector is an alternative of i-vector for automatic intelligibility estimation [5]. Both the i-vector and Fisher vectors are difficult to be interpreted. Phonetic posteriorgrams (PPGs) is an alternative representation that can be used in intelligibility estimation. PPGs refer to the posterior probabilities of each phonetic class for a specific time in an utterance. PPGs derived from neural networks are shown to be useful in predicting the intelligibility of a voice disorder [25]. Other features, such as loudness, harmonicity, MFCCs, and jitters, have also been used to estimate intelligibility [32]. These interpretable variables may help speech therapists identify the problems of dysarthric speech.

The third area is automatic error analyses. This is useful in different applications. For example, phonetic error analyses can improve the performance of ASR [48]. Recurrent neural networks are applied to help recognize phonological features in dysarthric speech for visualization [16]. Articulatory features in dysarthric speech be recognized by an end-to-end automatic speech attribute transcription using a transformer model [24].

This study aims to develop an automatic approach that can analyze the deviant patterns of dysarthric speech with efficiency and interpretability. More specifically, we propose a phonologically-motivated framework for computational articulatory characterization of error patterns in dysarthric speech. We leverage the parsimonious description of speech segments afforded by distinctive features (DFs) [37], which have correlates in both articulatory and acoustic domains. We develop automatic speech recognition techniques to derive DFs from the speech signal, which also enables efficient analysis of a large amount of speech data. Corpus-based analysis and comparison between the DF-based representations of healthy and dysarthric speech will elucidate problematic articulatory gestures that cause errors and reduce intelligibility.

2.2 Cantonese Chinese

Cantonese Chinese is the predominant dialect used in Hong Kong, Macao and many overseas Chinese communities, spoken by over 60 million people worldwide. Cantonese is a tonal language with 6 tones, and dysarthric speech may exhibit deviant tonal patterns. As a first step, we focus on Cantonese syllable articulation. We adopt the Jyutping Cantonese syllable labeling (romanization) scheme, designed by the Linguistic Society of Hong Kong [26]. Each Chinese character is pronounced as a single base syllable with a lexical tone, e.g. "看" <translation: see> is pronounced as /hon3/. The base syllable (i.e. /hon/) is divided into two parts: an initial and a final. The initial is an optional consonant, which is also referred to as the syllable onset. The final consists of the syllable's vowel nucleus, which may be a monophthong or a diphthong; followed by an optional consonant referred to as the syllable coda. The lexical tones range from 1 to 6. In this work, we focus on the base syllable as the unit for analysis.

2.3 Distinctive Features

A distinctive feature (DF) is the most basic phonological unit which can differentiate between a pair of maximally close phonemes [8]. For example, the two labial phonemes [p] and [b] can be distinguished by the DF (VOICE), with [p] being [−VOICE] and [b] being [+VOICE]. Each phoneme can be represented in terms of a vector of DFs. There are two types of DFs – the first type are place features which specify the place of articulation, e.g. [LABIAL] is articulated with the lips, [ALVEOLAR] is articulated with the tongue touching near the ridge of the bone behind the teeth in the upper jaw; and [VELAR] is articulated with the back of the tongue touching near the soft palate, etc. The second type are manner features which describe the type and degree of airflow through the vocal tract. For example, [+CONTINUANT] refers to airflow being continuous but involves partial occlusion of the airway, such as vowels, glides, liquids and fricatives. [−CONTINUANT] is where there is obstruction in of airflow, e.g. for nasals, strops and affricates. [+LATERAL] refers to airflow around the sides instead of over the top of the tongue, such as for the phoneme [l]. [+NASAL] refers to the air flowing through the nasal cavity instead of the oral cavity. The complete list of 21 DFs adopted in this work is shown in Table 1.

As we attempt to map phonemes into DFs, we define four possible values for the DFs: (i) positive (‘+’), referring to the fact that the DF is present, such as the example of [+CONTINUANT] mentioned above; (ii) negative (‘-’), referring to the fact that the DF is absent, such as the example of [-VOICE] mentioned above; (iii) unspecified (‘/’) [15], referring to the fact that the feature can be either positive or negative, e.g. the feature [HIGH] specifying the tongue position does not have any effect on the production of the phoneme [m] and so we label it as [/HIGH]; and (iv) irrelevant (‘x’), which expresses that the DF has no relationship with the phoneme. For example, the feature [TENSE] describes a greater degree of constriction with a tongue body or root, but does not play a part in the articulation of the phoneme /p/ and hence we assign the value [xTENSE].

The appendix shows the mapping form phonemes to DFs used in this work. Since a stop (or plosive) consists of both a closure and a release, the two parts are mapped to two different DF vectors. Similarly, a diphthong exhibits an articulatory change and is mapped to two different DF vectors. To maintain consistency for ease of comparison, monophthongs are represented by two identical DF vectors, indicating that there is no change in articulation within the production of the phone.

DFs have articulatory correlates and as a consequence, acoustic correlates as well [37]. In order to support our investigation of the acoustic manifestations of DFs in dysarthric speech, we designed and collected the Cantonese dysarthric speech corpus, which we named CU DYS, as described in the next section.

Table 1: Brief Definitions of the 21 distinctive features (DFs) used in this work [15].

Group	Distinctive Features	Brief Meaning
Tongue	Coronal	Tongue blade is raised toward the teeth or the hard palate
	High, Low, Front, Back	Position of the tongue
	Lateral	How to the tongue manipulates the airstream flow
	Tense	Tongue configuration with a greater constriction
	Velar and Alveolar [21]	Place of obstruction made by the tongue
Lips	Labial	Constriction at the lips
	Rounded	Protrusion of the lips
Tongue/Lips	Anterior	Horizontal position of the primary constriction
Soft Palate	Nasal	Soft palate is lowered
Vocal cords	Spread glottis	Vocal cords are drawn apart
	Voiced	Vocal cords vibrate periodically
Articulator-free	Syllabic	Constitution of syllable peaks
	Consonantal	With a sustained vocal tract constriction
	Sonorant	Vocal tract configuration is open
	Continuant	Vocal tract configuration allows the airstream to flow through the centre of the oral tract
	Strident	A constriction forces the airstream to strike two surfaces
	Delayed Release [8]	Vocal tract closure released with a delay

3 Corpora

3.1 Chinese University Cantonese Dysarthric Speech Corpus: CU DYS

We have designed and collected CU DYS, a Cantonese dysarthric speech corpus [45]. We have included three reading tasks: single word-reading (61 words), short sentence-reading (23 sentences), and passage-reading (1 passage including 4 long sentences). The collection effort is conducted in collaboration with patient organizations, speech therapists and doctors. Individuals reported to have spinocerebellar ataxia (SCA) or cerebral palsy (CP) which lead to dysarthric speech were invited to participate in this study. Altogether 27 dysarthric subjects and 14 healthy subjects have been referred and invited to join the study by providing speech recordings.

Each stimulus is recorded at least twice, with the first one intended as practice for the speaker. However, if the second recording has issues with noise, mispronunciations, missing characters or other errors, then the speaker is requested to repeat until an acceptable recording is obtained. However, the passage-reading task was only recorded twice to avoid fatigue in providing longer speech recordings. Table 2 compares the durations of the speech recordings between dysarthric and healthy control speakers. It can be seen that the durations of the dysarthric recordings are around three times that of the healthy control recordings.

Table 2: The durations of acceptable speech recordings between dysarthric speakers and healthy control speakers in word-reading and sentence-reading tasks.

Dysarthric speakers	Healthy control speakers
65 minutes	23 minutes

3.2 Speech Severity and Manual Phonetic Transcriptions

The passage-reading task is mainly used for assessment of the severity of the dysarthric condition by speech therapists, because the task elicits more speech. Two speech therapists were invited to provide assessment based on a 5-point scale (where 1 has the highest speech intelligibility and 5 has the lowest speech intelligibility). A subject's speech severity rating is based on the overall average of the four utterances across the scores provided by the two speech therapists.

As regards phonetic transcriptions, we recruited undergraduate students who are linguistics majors to serve as transcribers. The students are familiar with the Cantonese Jyutping system, but they do not have prior exposure to dysarthric speech – we believe this is a suitable setup for transcription because it provides the perspective of the perception of the general listener. Each utterance is transcribed by two transcribers, to allow for some degree of perceived variability in the transcription (we will elaborate on this later). They listened with a Sennheiser PC155 headset in a soundproof room and transcribe the utterances in terms of Jyutping syllables. No other information about the textual prompts is provided to the transcribers. Also, each transcriber only transcribes the speech of at most two dysarthric speakers per week, to reduce any learning effects specific to the dysarthric speakers. The transcribers were asked to listen to the utterances for as many times as they need and provide a phonetic transcription for syllable onsets, nuclei and codas [46]. In parallel, we obtain the canonical phonetic transcriptions by first mapping the text prompts into the canonical Jyutping syllable transcriptions, which are then mapped into the Jyutping syllable onsets, nuclei and codas in the canonical phonetic transcriptions [46]. To compare the manual phonetic transcriptions with the canonical phonetic transcriptions, we first perform manual alignment between the two sequences – first mapping the syllable nuclei and then the syllable onsets and codas.

3.3 Dataset Division

We divided the CU DYS corpus into training, development and testing sets, covering word-reading and sentence-reading utterances. It was challenging to recruit dysarthric speakers and our corpus had more male speakers than female speakers. We strive to include speakers of both genders in each divided dataset. The corresponding statistics are shown in Table 3.

Table 3: Dataset division in the CU DYS corpus.

Dataset	Dysarthric speakers	Healthy control speakers
Training	10 males 3 females 1,092 utterances Severity level (average 2.5, s.d. 1.2, range 1.4 to 5)	5 males, 5 females 840 utterances
Development	4 males 2 females 504 utterances Severity level (average 2.6 s.d. 1.2, range 1.1 to 4.5)	1 male, 1 female 168 utterances
Testing	5 males 3 females 672 utterances Severity level (average 2.8 s.d. 1.5, range 1.4 to 4.9)	1 male, 1 female 168 utterances

3.4 Non-Cantonese Speech Data

In order to augment the CU DYS dataset for experimentation, we explore used some additional corpora which have phonemic transcription. Phonemic transcriptions can provide a more precise DF annotation. Therefore, we carefully select the CU SENT corpus [22] with Cantonese read speech of 80 healthy speakers, consisting of 21,600 utterances. Furthermore, dysarthric speech naturally involves deviant pronunciations which may not be found in the standard Cantonese phone set. For example, the phoneme /p/ is unvoiced [−VOICE] in Cantonese. If the deviant pronunciation of /p/ altered the DF to [+VOICE] DF, the sound cannot be found in the Cantonese phone set and will unlikely be found in a Cantonese speech corpus. However, the voiced version of /p/ is the English phoneme /b/. Hence we also augment our data by including two English corpora – (i) the TIMIT [13] training set which has 630 speakers and 6,300 utterances; and (ii) the speech recordings of healthy speakers in TORGO [36], which has 14 speakers and 11,900 utterances. This is an initial step we took in "borrowing" data from English speech recordings to support our experimentation. We believe that in the future, borrowing further across other languages would also be helpful.

As we are using both Cantonese and English data, we need to use a cross-language, machine-readable phonetic alphabet, and we have adopted SAMPA (Wells, 1997). All the Cantonese and English phoneme sets are mapped into SAMPA, devoting due consideration to articulatory contexts. For example, the Cantonese phoneme /p/ may occur in a syllable onset of coda position, and may exhibit different pronunciations. /p/ in the onset position is aspirated and is hence denoted as /p_h/, but it is unaspirated in the coda position and is denoted as /p_}/. We have adopted a total of 84 phones for Cantonese and English and these are listed in the appendix.

4 Analysis of Dysarthric, Deviant Articulations based on Manual Phonetic Transcriptions

4.1 Articulatory Error Rate (AER)

We use the articulatory error rate (AER) to represent how often a deviation occur for a specific DF based on the recordings of a dysarthric subject. Recall from Section 2.3 that the current study is based on the complete list of 21 DFs. Hence, we can compare the canonical phone transcriptions (based on the Jyutping syllable) with the manual phone transcriptions to locate the deviant pronunciations, and from those deviations we can map out the differences in DFs table lookup. Recall from the introductory section above that each DF can take on values that are basically positive or negative. However, if the subject misses a phone, then the DFs of the canonical phone will be considered to have mapped to the DF value of NULL. Conversely, if the speaker inserts a phone, then the DFs of the "non-existent canonical phone" will be treated as NULL values that are then mapped to the feature values of the inserted phone. Hence, we have three types of DF errors based on how features are replaced, namely:

- Substitution Error (S): + → - , - → +, + → x, - → x
- Deletion Error (D): + → NULL , - → NULL
- Insertion Error (I): NULL → + , NULL → -

As explained above, for each DF, we can compute the percentage of deviations occurring for *each* DF to obtain its AER. For example, if a subject's articulation shows that 20% of the [ROUND] feature differs from the canonical production, then AER for [ROUND] will be 20%. Let i be the index of the DF across the complete set of 21 features. Hence, each phone uttered by a subject may be characterized by a vector of DFs, if we examine all the recordings of a given speaker, we can then compute the articulatory error rate (AER) for each DF i, defined by Equation 1:

$$AER(DF_i) = \frac{S_i + D_i + I_i}{T_i} \tag{1}$$

where the numerator is the sum of all the error types, and T_i in the denominator denotes all the occurrences of the values of DF $i = 1, 2 \ldots 21$

Recall that each utterance is transcribed by two individuals. Our calculations can accommodate cases where there may be different transcriptions. We consider the utterance to be spoken twice when calculating AERs. For example, assuming that there is a segment for which a DF is positive. Then, if one transcriber listens and considers that a given DF is positive for that segment, but the other transcription regards that the DF is negative, the AER of the specific DF will be 0.5.

4.2 Analyses

As explained above, with canonical phones and manually transcribed phones, we calculate the AERs for each DF and for all dysarthric subjects in CU DYS. Results are shown in Table 4.

For the word-reading task, AERs range from 8.2% to 20.8% with a mean of 13.7%. DFs with the highest error rates are related to consonants, namely [ANTERIOR] (AER: 20.8%), [CORONAL] (AER: 20.0%), [ALVEOLAR] (AER: 19.8%) and [VELAR] (AER: 19.0%). DFs with the lowest error rates are [SYLLABIC] (AER: 8.2%), [DELAYED RELEASE] (AER: 8.6%), and [LATERAL] (AER: 9.4%). It is noted that [SYLLABIC] distinguishes between vowels and

Table 4: Average AERs (%) across all dysarthric subjects in word-reading task, and sentence-reading task based on manual transcriptions. The bold value is the highest value and the underline value is the lowest value in a column respectively.

	Word-reading		Sentence-reading	
DF	Mean	Standard Deviation	Mean	Standard Deviation
SYLLABIC	8.2	9.4	7.3	9.7
CONSONANTAL	10.6	12.3	9.3	12.0
SONORANT	11.6	13.3	10.1	12.8
CORONAL	20.0	19.7	16.3	19.7
ANTERIOR	**20.8**	**20.4**	**17.7**	**20.0**
LABIAL	14.7	15.6	11.4	15.0
HIGH	15.3	13.1	12.9	11.9
BACK	14.4	12.3	12.0	11.5
FRONT	13.8	11.8	11.6	10.6
LOW	14.7	12.6	11.9	11.1
ROUNDED	12.9	14.5	10.2	13.4
CONTINUANT	12.0	13.8	10.6	13.7
LATERAL	9.4	10.2	8.6	10.3
NASAL	11.3	12.6	9.7	11.8
TENSE	12.3	10	9.7	9.3
STRIDENT	10.6	12.3	9.6	12.6
SPREAD GLOTTIS	11.9	13.6	9.5	12.1
VOICED	16.0	18.3	14.6	18.3
DELAYED RELEASE	8.6	11.5	8.3	12.0
VELAR	19.0	17.7	16.5	18.2
ALVEOLAR	19.8	19.4	16.6	19.7
Average	13.7	14	11.6	13.6

consonants and the clear distinction brings a low AER. For [DELAYED RELEASE], the DF is positive only for /c/ and /z/. For [LATERAL], the DF is positive only for /l/. Hence [DELAYED RELEASE] and [LATERAL] is mostly negative and this may also lead to fewer errors. The sample standard deviation of the DF error rates ranges from [SYLLABIC] (SD: 9.4%) to [ANTERIOR] (SD: 20.4%).

We also conducted analysis between AERs of DFs against the severity of the condition. Results are shown in Figure 1. We observe that there is high correlation between AERs and severity levels across the 21 DFs, with the average at 0.90 in the word-reading task.

For the sentence-reading task, AERs range from 7.7% to 17.7% with a mean of 11.6%. The values are generally lower in sentence-reading than in the

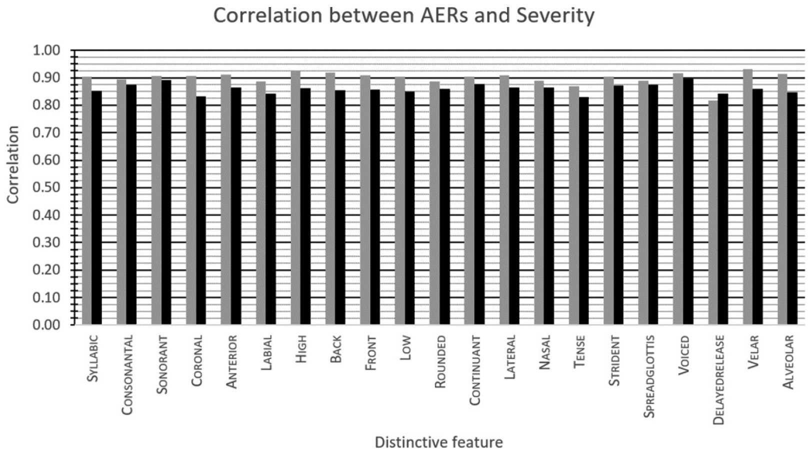

Figure 1: The correlations between AERs and severity levels for all dysarthric subjects in CU DYS based on the manual transcriptions. For each DF, the AERs are compared with severity levels across all subjects.

word-reading task. The error patterns are consistent, i.e., DFs with the highest error rates are also [anterior] (AER:17.7%), [coronal] (AER:16.3%), and [alveolar] (AER:16.6%). Also, the DFs with the lowest error rates are [syllabic] (AER:7.3%), [delayed release] (AER:8.3%), and [lateral] (AER: 8.6%). The lowest standard deviation is also [syllabic] (AER:7.3%), and the highest is also [anterior] (AER:17.7%). The sample standard deviation of the DF error rates ranges from [tense] (SD: 9.3%) to [anterior] (SD: 20%). Similar to the results of the word-reading task, we observe that the AERs of the sentence reading task are highliy correlated with the severity levels, averaging at 0.86 over all the 21 DFs (please see Figure 1).

5 Acquiring Deviant Articulations based on Automatic Transcriptions

The use of manual phonetic transcriptions to obtain articulatory errors (AERs) is time-consuming and difficult to apply at scale. In this section, we present our study in the development of DF detectors based on automatically transcribed phones using a phone recognizer.

5.1 Sequence-to-sequence (Seq2Seq) Model Trained on Healthy Speech Data

Manual phonetic transcription can be considered a sequence of phonetic labels corresponding to a sequence of speech signal. The two sequences often have different sequence lengths, hence, step-wise mapping between sequences is inapplicable. We utilize the sequence-to-sequence (Seq2Seq) model to map the sequence of acoustic frames extracted from the speech signal to the sequence of phonetic labels. In order to model the temporal dynamics, the recurrent neural networks (RNNs) [35] are used to construct the encoder of the seq2seq model. The gated recurrent unit (GRU)-based RNN is adopted [7]. Upon the RNN-based encoder, the connectionist temporal classification (CTC) [14] loss function is used to tackle the alignment between the input and output sequences, as shown in Figure 2. More specifically, we use an 80-dimensional filterbank feature with first- and second-order derivatives as input features to the encoder. The filterbank features are extracted using a 25 ms window with a window shift of 10 ms. 23 consecutive frames (the 11 previous frames, the current and the next 11 frames) are used and downsampled to a 30ms frame rate [6]. Batch normalization is applied to the input features. Three fully-connected layers of 1024 units with rectified linear activations are used. Batch normalization is applied again before the fully-connected layer outputs are fed to the bidirectional GRU layer with 1024 cells, followed by one fully-connected layer with 1024 units. The output layer is a softmax layer that predicts 87 classes (84 phones + 1 silence + 1 utterance boundary + 1 blank label).

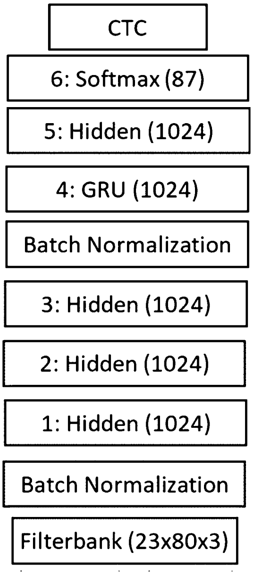

Figure 2: The DNN architecture.

The phone labels include both Cantonese and English phones because we found that using both language corpora can improve the ASR performance as discussed in the next paragraph. The blank label is used by the CTC loss to enable alignment between input and output sequences [14].

The training set consists of the CUSENT corpus (20.4 hours) [22] and the healthy speech part of the TORGO corpus (21 hours) [36], the TIMIT corpus (4.2 hours) [13], and the training set (including both the word-reading task and the sentence-reading task) of healthy speech in the CU DYS corpus (Section 3.3). We adopt an early stopping strategy [31] which terminates the training process when there is no further improvement of phone error rate (PER) on the CU DYS healthy speech development set within last N epochs ($N = 20$). Stochastic gradient descent is used as the optimizer [34]. The PER evaluated on the healthy speech of CU DYS testing set is 7.6%. If we only use the Cantonese corpora (i.e., CUSENT and CU DYS) without the extra English corpora in training, the PER is increased to 10.8%.

5.2 Phonetic Decoding of Dysarthric Speech

The trained seq2seq model is used to perform phonetic decoding of dysarthric speech. For a correctly pronounced phone, the probability outputs show a sharp distribution, as illustrated in the left plot in Figure 3, where the segment has the same label of /aa/ for both the canonical phone and the manually transcribed phone. In automatic phone transcription, the label /aa/ receives the highest probability (0.98), in stark contrast with the second

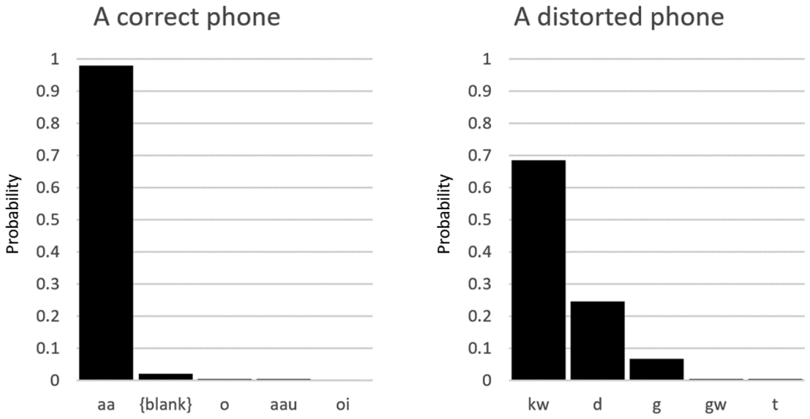

Figure 3: **Left side**: The output probabilities generated by the trained seq2seq model shows a sharp distribution for a correctly pronounced phone in the CU DYS corpus (left), where /aa/ is selected for the canonical phonetic transcription and the two manual transcriptions. {blank} refers to the case where no label is given for the acoustic frame. **Right side**: The output probabilities show much more uncertainty for a deviant, dysarthric production in CU DYS. The canonical phonetic label is /s/, which is different from the two manual phonetic transcriptions (/b/and /d/ respectively), and the phone with the highest probability is /kw/.

Note: https://www.overleaf.com/project/6355fbe1789aa541bca926f7.

highest probability (0.02) for the blank label. However, for the case of a deviant production in dysarthric speech, as illustrated in the right plot of Figure 3, there is much more uncertainty – the label /kw/ receives the highest probability (0.69), and the label /d/ receives the second highest probability (0.25). For this particular segment, the canonical label is /s/ and there is disagreement between the two manually transcribed labels (namely, /b/ and /d/). We also noted that for dysarthric productions, the manually transcribed label(s) often appear(s) among the top labels with highest probability. Based on this observation, we further align two sequences of phonetic labels (i.e. the canonical phonetic labels, and the one from seq2seq model) in the next section.

5.3 Phonetic Alignment for Computing AERs

The recognized phonetic labels output by the seq2seq model need to align with the canonical phonetic labels for calculation of the AERs. Since non-Cantonese corpora (i.e. English corpora) are used in model training, it is possible that the recognizer may output a non-Cantonese phone label for a frame because it received the highest probability. However, the manual transcriptions (treated as the ground truth) do not include any non-Cantonese phones. Hence, in preparation for the alignment, we need to identify each recognized non-Cantonese phone label and map it to the "back-off" Cantonese phone label, which has the least number of differing DF values. This mapping process leads us to assign the probability of the "back-off" phone to be the sum of the probability of the non-Cantonese phone label and the probability of the replacement (back-off) Cantonese phone label. Specifically, let α denote the non-Cantonese phone label with the highest probability for a frame; and β denote Cantonese phone label closest to α in terms of DFs; and let $p[\phi]$ denote the probability of the phone ϕ. Then, the mapping due to the "back-off" phone β will lead us to updated probability $p'(\beta)$ for the recognized phone label:

$$p'(\beta) = p(\alpha) + p(\beta). \tag{2}$$

Another point to note is that consecutive acoustic frames may often be repeated as they belong to the same phone segment. For these repeating phone labels in successive frames, we assign the {blank} label. For example, given three consecutive acoustic frames that belong to the same phone segment /w/, if the recognition outputs are the phone labels /w/-/w/-/w/, the procedure will relabel the repeated /w/ so that the sequence becomes /w/-{blank}-{blank}.

As a consequence of the above procedures, we have two sequences of phonetic labels that will be aligned using dynamic programming [1]. The two sequences are: (i) the sequence of canonical phones $c_1, c_2, \ldots, c_j, \ldots, c_M$; where M is the length of the sequence of canonical phones in the utterance; and (ii) the sequence of frame $f_1, f_2, \ldots, f_k, \ldots f_N$; where N is the total number of frames. The similarity between a pair of aligned phones c_j and $\gamma_k \in \Gamma$ in the frame f_k, may be the proportion of their common DFs values, but we may also consider a more refined "similarity" based on the frequencies of errors observed in the CU DYS training set. For example, in the case of a canonical phone bigram /a n/, we observe that the substitution of /n/ with /ng/ is much more common than with the phone /m/ (even though all are nasal phones). We believe that is important to capture such phenomenon in the similarity metric, and hence it is useful to consider the frequency of substitution errors as a reflection of similarity. For a given canonical phone pair (c_{j-1}, c_j) , if c_j is substituted with γ_k very frequently, then it suggests that they are more similar in the given context and are thus harder to discern clearly. Therefore, we propose a similarity term based on the relative frequencies of corresponding substitutions found in the training set, which is used as a correction factor for the probability of the recognized phoneme label γ_k – and this is used as the weighting factor for the DF similarity during the alignment by dynamic programming (please see Equation 3).

$$Weighting\ factor = p(\gamma_k)\frac{Count(c_j \rightarrow \gamma_k | c_{j-1}, c_j)}{Count(c_{j-1}, c_j)}. \tag{3}$$

The DF similarity between c_j and γ_k is the proportion of DF with common values between their DF vectors. If there is an insertion or deletion, then the DF distance is 42 (i.e., 21 DFs for each of 2 vectors as mentioned in Section 2.2). If recognition produces a blank output, the DF distance is arbitrarily set to (21 DFs * 1.5), to be lower than the cost of insertion/deletion.

6 Comparing Articulatory Analyses based on Manual and Automatic Transcriptions

Thus far, we have described our framework for articuatory analyses of dysarthric speech through the use of DFs, whose values are obtained from the comparison between the canonical phonetic transcription of text prompts and the manual phonetic transcription of the spoken utterances. This section aims to examine the feasibility of replacing manual phonetic transcription with automatic phonetic transcription.

6.1 Phone Error Rates from Recognition

When comparing manually transcribed phones and automatically transcribed (i.e. recognized) phones, the PER is calculated as:

$$PER = \frac{S_p + D_p + I_p}{T_p} \tag{4}$$

where S_p, D_p, I_p, and T_p, are the counts of substitutions, deletions, insertions, and total number of phones respectively.

We consider a phone error to have occurred if the recognized phone is different from the two manually transcribed phones for the same speech segment. PERs of the word-reading task is 33.0%, and the sentence-reading task is 35.6%. The performance is similar to other low-resource phone recognition systems [23].

6.2 DF Error Rates from PER

Since automatic phone transcription will inevitably have recognition errors, these areas will translate into DF errors during articulatory analyses. As mentioned earlier, there are 21 DFs in total and each phone is mapped to two DF vectors to capture possible dynamics. Therefore, in comparing an automatically transcribed phone with a manually transcribed phone, the maximum distance between their DF vectors is 42.

Analysis shows that misrecognized phones tend to be close counterparts of the canonical phone and hence their DF vectors tend to have more values in common. For example, confusable vowels like /a/ and /aa/ differ only in the DF [TENSE]; confusable nasals /m/ and /n/ differ in the 3 DFs [CORONAL], [LABIAL] and [ALVEOLAR].

The range of DF error rates (as a consequence of phonetic misrecognition) for word-reading tasks range from a minimum of 12.1% for [LATERAL] to 17.7% for [HIGH], with an average DF error rate of 14.6%. For sentence-reading tasks, the DF error rates range from a minimum of 13.7% for [SPREAD GLOTTIS] to 19.7% for [HIGH], with an average of 16.4%. Details are shown in Table 5(a).

6.3 Root Mean Squared Errors (RMSE) of AER

Recall from Section 4.1 that we can derive the DF vectors from the manually transcribed phones of a subject's read speech, and compare then with the DF vectors derived from the canonical phones of the textual prompts for reading. This comparison is used to compute the AER (articulatory error rate) for each DF, to characterize the deviant articulations in the subject's recordings. Since manual phonetic transcription is laborious, it will be more efficient if we can use automatically transcribed phones in place of manually transcribed phones. Therefore, it will be useful to examine the AERs obtained for the full set of DFs based on manual phonetic transcriptions and compare them with the automatic transcriptions. This comparison is done for each DF, and for each subject we obtain the squared error, then we compute the mean squared error across all testing subjects, and finally the square root to obtain the RMSE (root mean squared error). Results are shown in Table 5(b). In the word-reading task, RMSE values range from 0.029 (for the DF [LATERAL]) to

Table 5: (a) DF error rates computed across all testing subjects; (b) RMSE of AERs between the DFs obtained from manual versus automatic phone transcriptions, computed across all testing subjects; (c) Correlation of AERs between the DFS obtained from manual versus automatic phone transcriptions, computed across all testing subjects.

	DF error rate(%)		RMSE of AERs		Correlation of AERs	
	(a)		(b)		(c)	
DF	Word	Sentence	Word	Sentence	Word	Sentence
SYLLABIC	13.3	14.2	0.046	0.065	0.98	0.97
CONSONANTAL	12.5	14.2	0.030	0.054	0.98	0.96
SONORANT	13.7	16.8	0.033	0.073	0.99	0.96
CORONAL	14.8	16.1	0.053	0.010	0.99	0.93
ANTERIOR	14.8	16.6	0.061	0.083	0.98	0.97
LABIAL	16.3	18.4	0.036	0.076	0.99	0.94
HIGH	17.7	19.7	0.058	0.071	0.97	0.95
BACK	17.1	18.5	0.055	0.077	0.97	0.95
FRONT	16.5	18.6	0.046	0.064	0.97	0.95
LOW	17.0	18.7	0.043	0.085	0.97	0.92
ROUNDED	15.2	16.5	0.041	0.059	0.98	0.95
CONTINUANT	14.0	15.3	0.037	0.061	0.99	0.96
LATERAL	12.1	13.7	0.029	0.046	0.98	0.98
NASAL	13.9	16.9	0.037	0.071	0.99	0.94
TENSE	16.8	19.2	0.072	0.080	0.88	0.92
STRIDENT	13.2	13.9	0.034	0.052	0.99	0.96
SPREAD GLOTTIS	12.9	13.7	0.032	0.050	0.99	0.97
VOICED	13.8	16.8	0.063	0.013	0.99	0.97
DELAYED RELEASE	12.7	14.1	0.205	0.300	0.90	0.92
VELAR	14.6	15.8	0.053	0.086	0.99	0.96
ALVEOLAR	14.2	15.9	0.052	0.100	0.99	0.94
Average	14.6	16.4	0.053	0.085	0.97	0.95

0.205 (for the DF [delayed release]) with a mean of 0.053. For the sentence reading task, the range of RMSE is from 0.046 (for [LATERAL]) to 0.300 (for [DELAYED RELEASE]) with a mean of 0.085. These are new results in the current using the seq2seq model for phone recognition first, before mapping the phones to the DF representations. Our previous approach (reported in [46]) used multi-layered perceptrons (MLP) for direct DF recognition, and reported only results on correlation but not RMSE. The current work compares both approaches using RMSE and obtains superior improvements (from 26.7% to 51.5%) in reduction of RMSE. From Table 5, We observe that [DELAYED RELEASE] presents an outlier. The reason is that this DF is only applicable to /c/ (e.g. /ce/ “車” <translation: car>) or /z/, with the DF values in “+” . When the canonical phone is /c/ and the manual transcribed phone is also /c/, then it is no error of [DELAYED RELEASE]. However, the recognizer misrecognized the phone as /s/, /j/, or /h/, which [DELAYED RELEASE] is “x” , so there all become articulatory errors.

In addition to computing RMSE, we also wish to compute the correlation, for each DF, between the AERs obtained from manual versus automatic phonetic transcription across all testing speakers. Results are shown in Table 5(c). For the word-reading task, the mean correlation is 0.97. For the sentence-reading task, the mean correlation is 0.95. As an illustration, we can visualize in Figure 4 the AERs of the feature [ROUNDED] across subjects, where the mean correlation is at 0.98. These results based on DFs obtained from seq2seq phone recognition are also significantly improved, when compared with our previous approach based on direct DF recognition using MLP [46], which obtained 0.90 and 0.91 for word- and sentence-reading tasks.

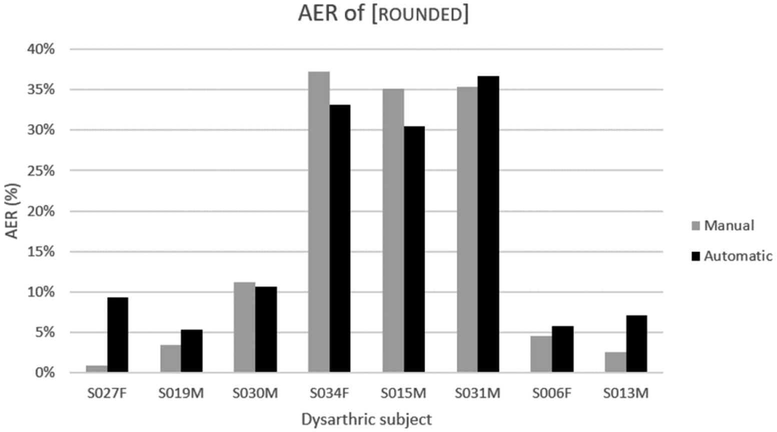

Figure 4: The AERs of the DF [ROUNDED] across dysarthric subjects in the word-reading task, comparing AERs derived from automatic phonetic transcription with those from manual phonetic transcriptions.

Overall, with the lower RMSE and high correlation values, we establish that the use of automatic phonetic transcriptions for DF-based analysis is a viable alternative to the use of manual phonetic transcriptions.

7 Applications in Analysis of Dysarthric Speech

This section discusses potential applications of DF-based AERs for the analyses of dysarthric speech. In this section, the AERs are derived from automatic phonetic transcriptions.

7.1 Comparing Healthy and Dysarthric Speech using AERs

Average AERs for each DF, computed from the recordings of healthy and dysarthric subjects in the CU DYS test set are shown in Figure 5 (for word-reading task) and Figure 6 (for sentence-reading task). Healthy subjects have lower AERs in all 21 DFs, compared with dysarthric subjects. In the word-reading task, the range of AERs from healthy subjects is from 0.8% (for

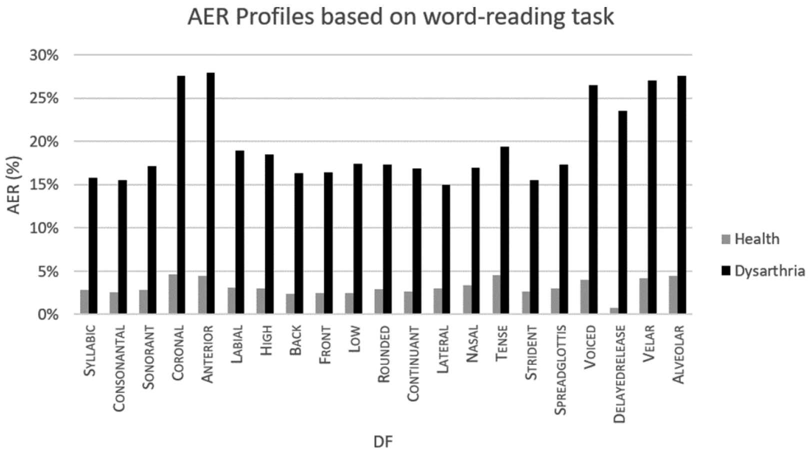

Figure 5: Comparison between healthy and dysarthric speech from the word-reading task of the CU DYS test set, in terms of AERs across DFs.

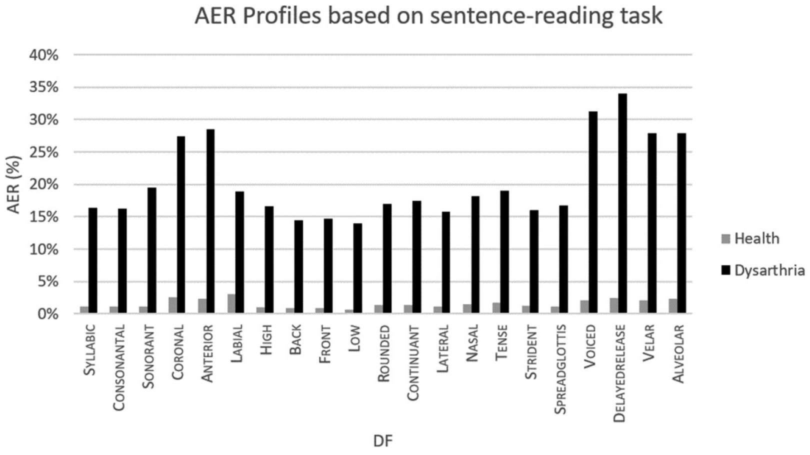

Figure 6: Comparison between healthy and dysarthric speech from the sentence-reading task of test set of CU DYS in terms of the AERs across DFs.

[DELAYED RELEASE]) to 4.7% (for [CORONAL]), compared with corresponding figures in the range for dysarthric subjects, i.e., from 15.0% (for [LATERAL]) to 27.9% (for [ANTERIOR]) in the word-reading task. A similar trend is observed for the sentence-reading task – ranging from 0.7% (for [LOW]) to 3.1% (for [labial]) from healthy subjects, 14.0% in [LOW] to 34.1% in [DELAYED RELEASE] from dysarthric subjects.

The patterns of AERs across the two tasks are also similar. DFs with the AERs higher than 25% include [CORONAL], [ANTERIOR], [VOICED], [VELAR], and [ALVEOLAR]; while the corresponding AERs in healthy speech are lower than 5%. All [CORONAL], [VELAR], and [ALVEOLAR] are related to articulation with the tongue in consonants. This finding is consistent with Whitehill and Ciocca [43], which also indicates that consonants have lower accuracies than vowels.

7.2 AER Profile of Individual Dysarthric Subject

Figure 7 illustrates the AER profiles of two dysarthric subjects (S015M, male with cerebral palsy; and S027F, female with spinocerebellar ataxia) based on the word-reading task. The average AER of S015M is 35.5%. He has difficulties pronouncing consonants like [CORONAL] (49.3%), [ANTERIOR] (50.0%), [VELAR] (49.0%), and [ALVEOLAR] (49.7%). The average AER of the other dysarthric subject, S027F, is 8.7%. S027F has difficulties in pronouncing vowels such as [high] (12.9%), [TENSE] (14.6%). Distinct AER profiles may inform the design of more personalized articulation practice.

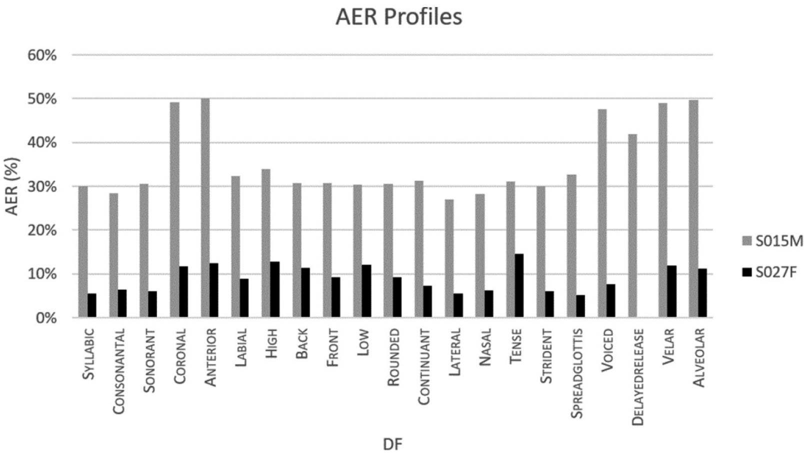

Figure 7: The AER profiles of dysarthric subjects: S015M, and S027F.

7.3 Relationship between AERs and Speech Severity Levels

Figure 8 shows the correlations between the automatically derived AERs and speech severity levels based on the test set for both tasks. All the correlations are higher than 0.9 – with an average of 0.95 (standard deviation 0.02) in the word-reading task, and 0.96 (standard deviation 0.02) in the sentence-reading task. Among the DFs, [TENSE] has the highest correlations for both tasks, which may potentially offer some indication of severity.

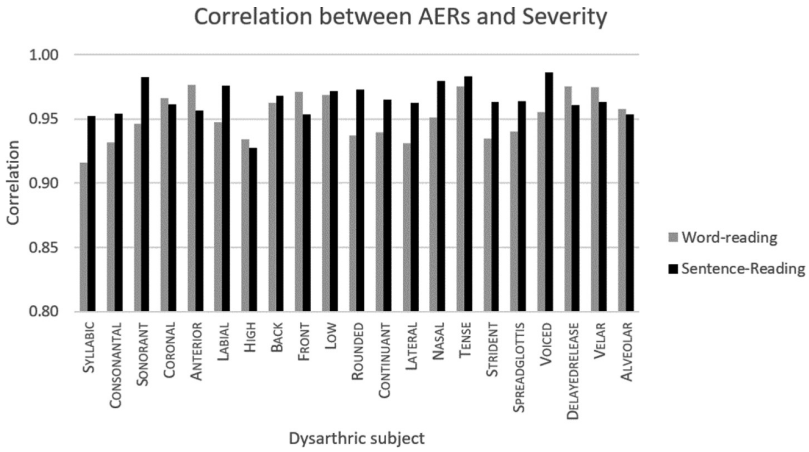

Figure 8: The correlations between AERs and severity levels across the dysarthric subjects in the test set.

Next, we explored the possibility of estimating the speech severity level based on AERs based on the data from a subject. We applied Random Model Trees [30] for regression, implemented with Weka [11]. Modeling is based on the training and development data sets of CU DYS. Estimation of severity level based on AER is conducted on the test set. An illustration is provided in Figure 9 based on the word-reading task. The correlation between estimated severity levels and clinician-assessed severity levels is 0.998 (with RMSE 0.085) for the word-reading task; and 0.921 (with RMSE 0.598) for the sentence-reading task.

8 Conclusions and Future Work

This paper presents our approach in the use of distinctive features (DFs) as descriptors of articulation as a representation of dysarthric speech (i.e., a type of disordered speech) that can reflect its deviations from normal speech. The main contribution of this study is unlocking the potential research area of

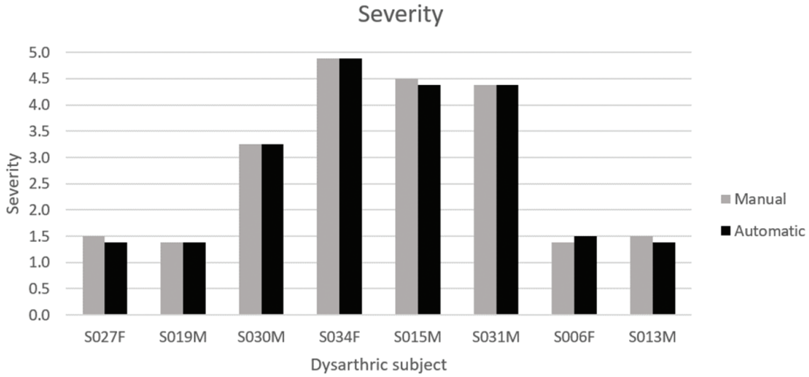

Figure 9: The estimated speech severity levels using AERs from the word-reading task, in comparison with the manually labeled severity levels. 1 is the lowest severity (i.e. the highest speech intelligibility), and 5 is the most severe.

articulatory analyses in dysarthria through the use of a distinctive feature (DF) representation. The paper demonstrates various articulatory analyses, including the identification of potential articulatory problems, profiling personal articulatory problems, and using the articulatory problems to estimate severity.

The study is conducted using our home-grown speech corpus known as CU DYS. We need to overcome the challenge of recruiting dysarthric speakers who are willing and capable of attending speech recording sessions. The linguistically-motivated DF representation is mapped from the phonetic representation obtained through manual transcription, to obtain a DF vector representation for each phone. Human perceptual variability between the two transcribers is accommodated through treating the two transcriptions as two phone occurrences. To alleviate the costly process of manual transcription, we have also developed a seq2seq model for automatic phonetic recongition. Any phonetic transcription errors from automatic recognition will lead to DF errors. While the PER (phone error rates) are at 33.0% and 35.6% respectively for the word-reading and sentence-reading tasks, the average DF error rates are at 14.6% and 16.4% respectively.

Next, we used the DF representation obtained from automatic phonetic transcriptions and compare them with those obtained from the canonical phonetic labels mapped from the text prompts. This comparison leads to the AERs (articulatory error rates) and we found that the RMSE is low at 0.053 with a high correlation at 0.97 for the word-reading task, and the corresponding values for the sentence reading task are 0.085 (RMSE) and 0.95 (correlation). These low RMSE and high positive correlations demonstrate the feasibility of automatic DF-based analyses for identifying articulatory problems in dysarthric speech. Automatic Speech Recognition (ASR) technologies have been proposed for speech analyses, e.g., in Xiong *et al.* [48], and phone posteriors are often used to support severity analysis [25]. In this work, we present the use of AER (articulatory error rates) derived from the automatic phone transcriptions, to elucidate the articulatory difficulties shown in the dysarthric speech signal from individual subjects. The AER profile of a subject may potentially help inform the design of intervention plans.

Future work will be devoted to further improvement of phone recognition performance to lower DF error rates and AER. We note that previous efforts in neurological research and clinical practice have been dedicated to categorizing speech characteristics and underlying neuro-pathophysiology of dysarthria [10] because the type of dysarthria can provide a clue to identify the causal disorder and suitable treatment plans [28, 49]. It has also been shown that phonological features can reveal some dysarthria types [4]. We believe that it will be useful to investigate the use of the AER profiles to inform categorization of dysarthric type, as well as the design of personalized intervention plans for dysarthric subjects.

Acknowledgements

The authors would like to thank Professor Edwin Chan, Professor Patrick Wong, Professor Raymond Tong, and Mr. Borislav Dzodzo from The Chinese University of Hong Kong for their referral and comments, and Ms. Wing-Sum Yeung for her assistance in data collection. This research is partially sponsored by a grant from the Hong Kong SAR Government General Research Fund (Reference No. GRF415513), and Microsoft Research Asia Regional Seed Grant.

Appendix

Table A1.a: Mapping from the phonemic inventory to DFs (part 1).

IPA DFs	pʰ		tʰ		kʰ		b		d		g		m	m̩	n	n̩	ŋ	ŋ̍	f	v	θ	ð	s	z	ʃ	ʒ	ʧʰ	ʤ	h	ɦ	l	l̩	ɻ	r̃	ɾ	ʔ	w	j
Syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
Consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-
Sonorant	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+
Coronal	-	-	+	+	-	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	+	+	+	+	+	+	+	+	/	/	+	+	+	+	+	/	-	+
Anterior	+	+	+	+	-	-	+	+	+	+	-	-	+	+	+	+	-	-	+	+	+	+	+	+	-	-	-	-	/	/	+	+	-	+	+	/	-	-
Labial	+	+	-	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	/	/	-	-	-	-	-	/	+	-
High	/	/	/	/	+	+	/	/	/	/	+	+	/	/	/	/	+	+	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	+	+
Back	/	/	/	/	+	+	/	/	/	/	+	+	/	/	/	/	+	+	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	+	-
Front	/	/	/	/	-	-	/	/	/	/	-	-	/	/	/	/	-	-	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	-	+
Low	/	/	/	/	-	-	/	/	/	/	-	-	/	/	/	/	-	-	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	-	-
Rounded	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	/	/	-	-	-	-	-	/	+	-
Continuant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	+	+
Lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	/	/	-	-	-	+	-	/	-	-
Tense	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	+	+
Strident	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-
Spread Glottis	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	+	-	-	-	-	-	-	-	-
Voiced	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	+	-	+	-	+	-	+	+	+	+	+	+	-	+	+
Delayed Release	x	-	x	-	x	-	x	-	x	-	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	+	+	x	x	x	x	x	x	x	-	x	x
Dental	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Retroflex	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
Velar	-	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-
Alveolar	-	-	+	+	-	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	+	+	-	+	+	-	-	-

Table A1.b: Mapping from the phonemic inventory to DFs (part 2).

IPA DFs	i	ɪ	ɛ	æ	ɑ	ɔ	ʌ	ʊ	u	ʉ	ɝ	ə	ɨ	ɚ	ə̥	eɪ		aɪ		aʊ		ɔɪ		oʊ		e	o
Syllabic	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Consonantal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sonorant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Coronal	x	x	x	x	x	x	x	x	x	x	+	x	x	+	x	x	x	x	x	x	x	x	x	x	x	x	x
Anterior	x	x	x	x	x	x	x	x	x	x	-	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x
Labial	-	-	-	-	-	+	-	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	+
High	+	+	-	-	-	-	-	+	+	+	-	-	+	-	-	-	+	-	+	-	+	-	+	-	+	-	-
Back	-	-	-	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	+
Front	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-	-	+	-	-	+	-
Low	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	+	-	+	-	-	-	-	-	-	-
Rounded	-	-	-	-	-	+	-	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+	+	-	+
Continuant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tense	+	-	-	/	+	-	-	-	+	+	+	-	+	-	-	+	-	+	-	+	-	-	-	+	-	+	+
Strident	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Spread Glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
Voiced	x	x	x	x	x	x	x	x	x	x	x	+	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x
Delayed Release	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Dental	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Retroflex	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
Velar	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Alveolar	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Table A1.c: Mapping from the phonemic inventory to DFs (part 3).

IPA DFs	p		t̪		t̪ʰ		k		n̪	s̪	t̪s	t̪sʰ	ʈʂ	ʈʂʰ	tɕ	tɕʰ	ʂ	ʐ	ɕ	x	ɹ̠	ɥ	y	ɯ	ɤ	a	t		kʷ		kʷʰ		ts	tsʰ	œ	ɵ	ɐ
Syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+
Consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-
Sonorant	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+
Coronal	-	-	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	x	x	x	x	+	+	-	-	-	-	+	+	x	x	x
Anterior	+	+	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	+	-	x	x	x	x	+	+	-	-	-	-	+	+	x	x	x
Labial	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	+	+	+	-	-	+	+	-
High	/	/	/	/	/	/	+	+	/	/	/	/	/	/	/	/	/	/	/	+	/	+	+	+	-	-	/	/	+	+	+	+	/	/	-	-	-
Back	/	/	/	/	/	/	+	+	/	/	/	/	/	/	/	/	/	/	/	+	/	-	-	+	+	-	/	/	+	+	+	+	/	/	-	-	-
Front	/	/	/	/	/	/	-	-	/	/	/	/	/	/	/	/	/	/	/	-	/	+	+	-	-	-	/	/	-	-	-	-	/	/	+	-	-
Low	/	/	/	/	/	/	-	-	/	/	/	/	/	/	/	/	/	/	/	-	/	-	-	-	-	+	/	/	-	-	-	-	/	/	-	-	+
Rounded	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	+	+	+	-	-	+	+	-
Continuant	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	-	-	-	-	-	-	+	+	+	+	+
Lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Nasal	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tense	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	+	+	+	+	+	x	x	x	x	x	x	x	x	-	+	-
Strident	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
Spread Glottis	-	-	-	-	+	+	-	-	-	-	-	+	-	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	+	-	-	-
Voiced	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	+	+	x	x	x	x	-	-	-	-	-	-	-	-	x	x	x
Delayed Release	x	-	x	-	x	-	x	-	x	x	+	+	+	+	+	+	x	x	x	x	x	x	x	x	x	x	x	-	x	-	x	-	+	+	x	x	x
Dental	-	-	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	+	-	x	x	x	x	-	-	-	-	-	-	-	-	x	x	x
Retroflex	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Velar	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	x	x	x	x	-	-	+	+	+	+	-	-	x	x	x
Alveolar	-	-	+	+	+	+	-	-	+	+	+	+	-	-	-	-	-	-	-	-	+	-	x	x	x	x	+	+	-	-	-	-	+	+	x	x	x

References

[1] R. E. Bellman, *Dynamic Programming*, Princeton, New Jersey: Princeton University Press, 1957.

[2] C. Bhat, B. Vachhani, and S. Kopparapu, “Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-taper Spectral Estimation,” in *Interspeech*, 2016.

[3] F. J. van Brenk, K. Tjaden, and A. Lowit, “Variability of F2 transitions in dysarthria,” in *Boston Speech Motor Control Symposium*, Boston, United States, 2019, Poster session.

[4] H. M. Chandrashekar, V. Karjigi, and N. Sreedevi, “Breathiness Indices for Classification of Dysarthria Based on Type and Speech Intelligibility,” in *International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, Chennai, India, 2019, 266–70.

[5] H. Chandrashekar, K. Veena, and N. Sreedevi, “Speech Intelligibility Assessment of Dysarthria using Fisher Vector Encoding,” *Computer Speech & Language*, 77, 2023, 10.

[6] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, “State-of-the-art Speech Recognition with Sequence-to-sequence Models,” in *the International Conference on Acoustics, Speech, and Signal Processing*, Calgary, Alberta, Canada, 2018.

[7] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” in *Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.

[8] N. Chomsky and M. Halle, *The Sound Pattern of English*, Harper & Row, 1968.

[9] H. Christensen, I. Casanueva, S. Cunningham, P. Green, and T. Hain, “Automatic Selection of Speakers for Improved Acoustic Modelling : Recognition of Disordered Speech with Sparse Data,” in *Spoken Language Technology Workshop*, 2014.

[10] J. R. Duffy, *Motor Speech Disorders: Substrates, Different Diagnosis, and Management*, St. Louis: Elsevier, 2013.

[11] E. Frank, M. A. Hall, and I. H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan kaufmann, 2016.

[12] D. B. Freed, *Motor Speech Disorders: Diagnosis and Treatment*, Clifton Park, Delmar: Cengage Learning, 2012.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, N. L. Dahlgren, and V. Zue, “TIMIT Acoustic-Phonetic Continuous Speech Corpus (LDC93S1),” 1993.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *the International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[15] M. Halle and G. N. Clements, *Problem Book in Phonology: A Workbook for Introductory Courses in Linguistics and in Modern Phonology*, MIT Press, 1983.

[16] Y. Jiao, V. Berisha, and J. M. Liss, "Interpretable Phonological Features for Clinical Applications," in *International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2017.

[17] N. M. Joy, S. Umesh, and B. Abraham, "On Improving Acoustic Models For TORGO Dysarthric Speech Database," in *Interspeech*, 2017.

[18] M. J. Kim, J. Yoo, and H. Kim, "Dysarthric Speech Recognition Using Dysarthria-Severity-Dependent and Speaker-Adaptive Models," in *Interspeech*, 2013.

[19] M. Kim and H. Kim, "Automatic Assessment of Dysarthric Speech Intelligibility Based on Selected Phonetic Quality Features," in *The 13th International Conference on Computers Helping People with Special Needs*, Vol. 2, 2012.

[20] I. Laaridh, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard, "Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers," in *Interspeech*, Hyderabad, India, 2018.

[21] P. Ladefoged and K. Johnson, *A Course in Phonetics*, Boston: Wadsworth, Cengage Learning, 2009.

[22] T. Lee, W.-K. Lo, P.-C. Ching, and H. Meng, "Spoken Language Resources for Cantonese Speech Processing," *Speech Communication*, 3-4 36, 2002, 327–42.

[23] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal Phone Recognition with a Multilingual Allophone System," in *ICASSP 2020*, 2020.

[24] Y. Lin, L. Wang, J. Dang, S. Li, and C. Ding, "End-to-end Articulatory Modeling for Dysarthric Articulatory Attribute Detection," in *International Conference on Acoustics, Speech, and Signal Processing*, Barcelona: IEEE, 2020.

[25] Y. Liu, T. Lee, P. Ching, T. K. T. Law, and K. Y. S. Lee, "Acoustic Assessment of Disordered Voice with Continuous Speech Based on Utterance-level ASR Posterior Features," in *Interspeech*, Stockholm, Sweden, 2017.

[26] LSHK, "The Jyutping Scheme," 1993, https://www.lshk.org/jyutping.

[27] J. K. Y. Ma, "Perception of Tones Produced by Cantonese Dysarthric Speakers," *Bachelor's Thesis*, University of Hong Kong, 2000.

[28] J. Marchant, M. McAuliffe, and M.-L. Huckabee, “Treatment of Articulatory Impairment in a Child with Spastic Dysarthria Associated with Cerebral Palsy,” *Developmental Neurorehabilitation*, 1st ser., 2008.

[29] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, “Intelligibility Assessment and Speech Recognizer Word Accuracy Rate Prediction for Dysarthric Speakers in a Factor Analysis Subspace,” *ACM Transactions on Accessible Computing*, 6(3), 2015, 10.

[30] B. Pfahringer, *Random Model Trees: An Effective and Scalable Regression Method*, New Zealand: University of Waikato, Department of Computer Science, 2010.

[31] L. Prechelt, “Early Stopping - But When?” In *Neural networks: Tricks of the trade*, ed. G. Montavon, G. B. Orr, and K.-R. Müller, Berlin, Germany: Springer-Verlag Telos, 1999, 57–69.

[32] B. AI-Qatab and B. M. Mustafa, “Classification of Dysarthric Speech According to the Severity of Impairment: an Analysis of Acoustic Features,” *IEEE Access*, 9, 2021, 18183–94.

[33] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, “Automatic Prediction of Speech Intelligibility based on X-Vectors in the Context of Head and Neck Cancer,” in *Interspeech*, Shanghai: IEEE, 2020.

[34] H. Robbins and S. Monro, “A Stochastic Approximation Method,” *The Annals of Mathematical Statistics*, 3rd ser., 1951, 400.

[35] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “Dynamic Recurrent Neural Networks: Theory and Applications,” *Transactions of Neural Networks*, 5, 1994, 153–6.

[36] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The TORGO Database of Acoustic and Articulatory Speech from Speakers with Dysarthria,” *Language Resources and Evaluation*, 4th ser. 46, 2012, 523–41.

[37] K. N. Stevens, *Acoustic Phonetic*, Cambridge, MA: MIT Press, 1998.

[38] Y. Takashima, R. Takashima, T. Takiguchi, and Y. Ariki, “Dysarthric Speech Recognition Based on Deep Metric Learning,” in *Interspeech*, Shanghai: IEEE, 2020.

[39] J. Tobin and K. Tomanek, “Personalized Automatic Speech Recognition Trained on Small Disordered Speech Datasets,” in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore: IEEE, May 2022.

[40] A. Tripathi, S. Bhosale, and S. K. Kopparapu, “A Novel Approach for Intelligibility Assessment in Dysarthric Subjects,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona: IEEE, May 2020.

[41] P. Vijayalakshmi, T. Nagarajan, and M. R. Reddy, “Assessment of Articulatory and Velopharyngeal Sub-systems of Dysarthric Speech,” *International Journal of Biomedical Soft Computing and Human Sciences,*

special issue on Biosensors: Data acquisition, Processing and Control, 14(2), 2009, 87–94.

[42] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu, and H. Meng, "Improved End-to-End Dysarthric Speech Recognition via Meta-learning Based Model Re-initialization," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Hong Kong: the International Speech Communication Association (ISCA), 2021.

[43] T. L. Whitehill and V. Ciocca, "Speech Errors in Cantonese Speaking Adults with Cerebral Palsy," *The Clinical Linguistics and Phonetics*, 2000, 111–30.

[44] T. L. Whitehill, J. K.-y. Ma, and A. S.-Y. Lee, "Perceptual Characteristics of Cantonese Hypokinetic Dysarthria," *The Clinical Linguistics and Phonetics*, 2003, 265–71.

[45] K.-H. Wong, Y.-T. Yeung, E. H.-Y. Chan, P. C.-M. Wong, G.-A. Levow, and H. Meng, "Development of a Cantonese Dysarthric Speech Corpus," in *Interspeech*, Dresden, Germany, 2015.

[46] K.-H. Wong, W.-S. Yeung, Y.-T. Yeung, and H. Meng, "Exploring Articulatory Characteristics of Cantonese Dysarthric Speech Using Distinctive Features," in *the 41th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2016)*, Shanghai, China, 2016.

[47] L.-T. Wong, "Kinematic and Correlational Analyses on Labial and Lingual Functions during Syllable Repetitions in Cantonese Dysarthric Speakers with Parkinson's Disease of Varying Severity using Electromagnetic Articulography (EMA)," *Bachelor's Science Thesis*, University of Hong Kong, 2014.

[48] F. Xiong, J. Barker, and H. Christensen, "Phonetic Analysis of Dysarthric Speech Tempo and Applications to Robust Personalised Dysarthric Speech Recognition," in *International Conference on Acoustics, Speech and Signal Processing*, Brighton, United Kingdom: IEEE, 2019.

[49] K. M. Yorkston and D. R. Beukelman, "Ataxic Dysarthria: Treatment Sequences Based on Intelligibility and Prosodic Considerations," *Speech and Hearing Disorders*, 1981, 398–404.

[50] J. Yu, X. Xie, S. Liu, S. Hu, M. W. Y. Lam, X. Wu, X. Liu, and H. Meng, "Development of the CUHK Dysarthric Speech Recognition System for the UASpeech Corpus," in *Interspeech*, Hyderabad, 2018, 2938–42.

[51] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multimodel Acoustic-Articulatory Feature Fusion For Dysarthric Speech Recongition," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore: IEEE, May 2022, 7372–6.