Article

Not peer-reviewed version

Discriminating Children with Speech Sound Disorders from Children with Typically Developing Speech Using the Motor Speech Hierarchy Probe Words: A Preliminary Analysis

Linda Orton , Richard Palmer , Roslyn Ward * , Petra Helmholz , Geoffrey Strauss , Paul Davey , Neville W Hennessey

Posted Date: 6 May 2025

doi: 10.20944/preprints202505.0225.v1

Keywords: speech sound disorders; assessment; kinematic; digital biomarkers; Motor Speech Hierarchy Probe Words

 doi:10.20944/preprints202505.0225.v1

Article

Discriminating Children with Speech Sound Disorders from Children with Typically Developing Speech Using the Motor Speech Hierarchy Probe Words: A Preliminary Analysis

Linda Orton [1], Richard L. Palmer [2], Roslyn Ward [1,*], Petra Helmholz [2], Geoffrey R. Strauss [1], Paul Davey [1] and Neville W. Hennessey [1]

[1] School of Allied Health, Curtin University, Perth, WA 6845 Australia

[2] School of Earth and Planetary Sciences, Curtin University, Perth, WA 6845 Australia

* Correspondence: r.ward@curtin.edu.au

Abstract: Purpose: The Motor Speech Hierarchy (MSH) Probe Words (PW) have yet to be validated as an effective tool to discriminate between impaired and typically developing speech motor control. The first purpose of this preliminary study was, therefore, to examine the effectiveness of the mandibular control subtest of the MSH-PW in classifying typically developing (TD) and speech sound disordered (SSD) children aged between 3 years 0 months and 3 years 6 months. Secondly, we compared automatically derived kinematic measures of jaw range and control with MSH-PW consensus scoring to assist in identifying deficits in mandibular control. **Methods**: Forty-one children with TD speech and 13 with SSD produced the 10 words of the mandibular stage of the MSH-PW. A consensus team of speech pathologists observed video recordings of the words to score motor speech control and phonetic accuracy, as detailed in the MSH-PW scoring criteria. Specific measures of jaw and lip movements during speech were also extracted to derive the objective measurements, with agreement between the perceptual and objective measures of jaw range and jaw control evaluated. **Results**: A significant difference between TD and SSD groups was found for jaw range ($p = .006$), voicing transitions ($p = .004$) and total mandibular scores ($p = .015$). SSD and TD discrimination was significant (at alpha = 0.01) with a balanced classification accuracy of 0.79. Initial analysis indicates objective kinematic measures using facial tracking show good agreement with perceptual judgements of jaw range and jaw control. **Conclusions**: Preliminary data indicate the MSH-PW can discriminate TD speech from SSD at the level of mandibular control and can be used by clinicians to assess speech motor control. Further investigation of objective measures to support perceptual scoring is indicated.

Keywords: speech sound disorders; assessment; kinematic; digital biomarkers; Motor Speech Hierarchy Probe Words

1. Introduction

Speech Sound Disorder (SSD) refers to difficulties producing and using speech sounds and speech segments, resulting in reduced accuracy and clarity of speech production. They are the most prevalent of all childhood communication difficulties [1] affecting 3.4%-5.6% of pre-school aged children [2] and comprising more than 70% of a Speech-Language Pathologist's (S-LP) caseload [3]. Children with SSD are more likely to experience adverse social, educational and psychological outcomes than children without SSD [4,5]. These difficulties may further limit employment opportunities throughout the lifespan [6]. Minimizing the impact of SSD is contingent on providing accurate diagnosis to direct intervention approaches.

The causes of SSD can be organic or functional; organic SSDs arise from an underlying structural (e.g., cleft palate), motor/neurological or sensory/perceptual cause; while functional SSDs, which are

 doi:10.20944/preprints202505.0225.v1

more prevalent [5], are idiopathic and include articulation and phonological disorders [7]. Diagnosis of functional SSDs seeks to identify the underlying contribution of speech difficulties. For example, identifying whether the child is having difficulties learning the linguistic-phonological rules of the target language (i.e., a phonological impairment) or/and difficulties with the motor aspects of speech production. The diagnosis of SSD, however, is challenging due to the nature of SSDs and the limitations of current clinical practices [8–10]. McCabe, Korkalainen & Thomas [9] highlight that the "overlap between the symptoms of different disorders with the same speech features … from multiple different breakdowns…" complicates SSD, while Littlejohn and Maas [8] note that diagnosis is impacted by a "poor understanding of, and limited focus on the underlying impairment(s)" (p. 2).

As part of the assessment process S-LPs are encouraged to conduct a comprehensive case history; assessment of oral structures and hearing function; an error analysis from a connected speech sample to identify a child's phonetic inventory and phonological error patterns or processes; as well as obtain measures of phonological mean length of utterance and quantify intelligibility [11]. In practice, however, time and ease of use are key factors influencing clinical decisions [12], with S-LPs routinely using standardized single-word naming tasks to evaluate speech sound inventory and error patterns [12,13]. Measures of phoneme accuracy within the single-word naming tests, including percent consonant correct (PCC), percent vowels correct (PVC), and percent phonemes correct (PPC), are frequently used to determine the presence and severity of SSDs [14].

The determination of speech sound error patterns typically relies on auditory-perceptual analysis using International Phonetic Alphabet (IPA) transcription [14,15]. While this is a fundamental part of diagnosis, using auditory-perceptual assessment alone is limiting [16–18], and there is no gold standard validation of perceptual measures that can discriminate SSD from TD speech. Auditory perceptual assessments do not allow clinicians to determine the contribution of speech motor control to the production difficulties of a child [19]. Further, the reliance on single-word assessment tools, framed predominantly on linguistic models of speech production, is problematic for differential diagnosis of SSD because these tools tend to focus on investigating phonological deficits [20], as opposed to the speech movement patterns associated with underlying constraints at the level of speech motor control.

McCauley and Strand's [21] 2008 review of standardized tests that evaluate speech motor performance of children concluded that "clinicians are in the position of having no tests that can be considered well developed for use with children with motor speech disorders" (p. 89). While new standardized assessment tools have been developed since this review, for example, the Dynamic Evaluation of Motor Speech Skill (DEMSS) [22], a recent review of assessment and intervention approaches for SSD identified 37 published assessment tools for SSD with the majority focusing on specific skills and only four assessing combined articulatory, phonetic and motor based development [4]. In their 2024 review of tools and approaches supporting diagnosis of childhood motor speech disorders, McCabe, Korkalainen and Thomas [9] state, "there are not yet validated tools for comprehensively assessing all speech production processes" (p. 9).

A tool recently developed to measure speech motor control is the Motor Speech Hierarchy - Probe Words (MSH-PW)[23]. The MSH [26] comprises seven stages that reflect the hierarchical (i.e., increasing motor complexity) and interactive development of speech motor control: Stage I: Tone: Stage II: Phonatory Control; Stage III: Mandibular Control; Stage IV: Labial-Facial Control; Stage V: Lingual Control; Stage VI: Sequenced Movements and Stage VII: Prosody. The Probe Words (PWs) cover stages III to VI, with 10 words and one phrase, in each stage. The probe words are scored visually and auditorily by observing a child say the target word and judging whether the speech movements look and sound appropriate/inappropriate, based on specific criteria. Criteria for stage III, Mandibular Control are, for example: appropriate jaw range; appropriate jaw control stability; appropriate close-open phase; appropriate voicing transitions; and correct syllable structure. In 2021, Namasivayam and colleagues [23] reported key measures of validity and reliability of the MSH-PW. Their data indicate high content, construct and criterion-rated validity, as well as high reliability on measures of internal consistency and intra-rater reliability, and moderate agreement on interrater

 doi:10.20944/preprints202505.0225.v1

reliability. This validation study, however, did not undertake a fine-grained analysis of individual scoring criteria (e.g., jaw range, jaw control), and the MSH-PW assessment tool was only validated on children aged 3 to 10 years with moderate-to-severe motor speech disorder. Therefore, construct validity in the form of distinguishing between the speech motor skills of TD children and SSD children, and scoring criteria involved in diagnosing impaired speech motor control, has yet to be established for the MSH-PW.

Furthermore, despite perceptual measures being used to judge articulatory control, the gold standard for evaluating speech motor control is based on instrumental analysis. Researchers have long advocated the need to combine perceptual analysis with instrumental analysis of kinematic (i.e., the study of motion: displacement, velocity and acceleration) and acoustic measures [15,18,21]. The use of kinematic analysis of speech has progressed with the development of new video and motion-tracking technologies [29], including the use of video-based tracking of jaw movements (e.g. [24]). In recent years, the enormous potential of Machine Learning (ML) in the support of a diagnosis of SSD has been recognized [25].

Computer vision-based approaches to measuring facial movements offer an objective and well-defined standard for detecting atypical speech patterns [26,27] and have demonstrated the potential for detecting facial movements associated with disorders [27–29]. While promising, the features used by these systems are not well grounded in the existing clinical understanding of which facial movements are involved in which aspects of speech motor control. As such, the decisions made by these systems lack explicative transparency.

This study focused on the assessment of mandibular control in young 3-year-old children using the MSH-PW. First, we investigated whether MSH-PW mandibular control scores obtained from expert clinicians could distinguish TD children from children diagnosed with a SSD. The aim was to validate the perceptual scoring of MSH-PW as being sensitive to individual differences in speech motor skill at the level of mandibular control, and to identify MSH criteria that could be predictive of disordered mandibular control, relative to TD children. The mandibular stage was chosen based on existing literature that indicates jaw control and stability may be a useful marker for determining SSD subtypes [30].

Second, we employed a state-of-the-art facial mesh detection and tracking algorithm [36] to extract measurements of facial movements identified as clinically salient in the assessment of speech motor control from recorded video of children speaking words from the mandibular stage of the MSH-PW. We evaluated how well these extracted facial movement measurements agree with perceptual scores for the jaw range and jaw control criteria of the MSH-PW. We selected these two criteria because clinicians rely predominantly on the child's facial movements in order to score jaw range and control.

This preliminary study, therefore, sought to answer the following questions:

Q1. Do the MSH-PW criteria, using expert consensus scoring, discriminate TD children from those with SSD? We predicted TD children would score more highly than SSD children in relation to the MSH-PW mandibular control criteria and that mandibular control criteria could be predictive of whether a child was TD or had an SSD.

Q2. Can kinematic measurements derived from automated facial tracking accurately predict expert the consensus perceptual scores of the MSH-PW jaw range and jaw control criteria? We expected good agreement between objective measures obtained from facial tracking and expert clinician judgements of appropriate and inappropriate jaw range and jaw control as indicated by the predictive accuracy in logistic regression classification models with objective measures as predictors and clinician judgements as the outcome.

 doi:10.20944/preprints202505.0225.v1

2. Materials and Methods

2.1. Participants

Participants were 54 children aged between 3 years and 0 months and 3 years and 6 months who were recruited from the Perth Metropolitan area and surroundings between December 2019 and December 2024 as part of a larger ongoing study on children's speech development. Recruitment for this larger study sought children with typical speech development, however, during the assessment process a number of children with characteristics of SSD were identified. All participants had complete facial tracking and speech and language assessment data available for analysis. Forty-one of the participants (21 male; 20 female) had typical speech development (TD) while 13 (5 male; 8 female) presented with SSD. Standardized assessments, including the GFTA-3 Sounds-in-Words subtest [31] standard score and measures of phoneme accuracy (PPC, PVC, PCC), and Verbal Motor Production Assessment for Children (VMPAC [32]) Focal Oral Motor Control and Sequencing subtests, parent measure of intelligibility using the Intelligibility in Context Scale (ICS [33]) and clinical observations were used to determine allocation into TD and SSD groups.

The mean age was 37.9 months for children in the TD group (range = 36 months and six days to 41 months and 27 days) and 37.3 months for those in the SSD group (range = 36 months and 499 days to 41 months and 23 days). There was no significant difference in age between the two groups, $t(52) = 1.179$, $p = .244$, nor were there significant differences in gender between the two groups, $\chi^2(N = 54) = 0.64$, $p = .432$. All participants were identified as having age-appropriate language and fine and gross motor development based on parent report via the Ages & Stages Questionnaires®, Third Edition (ASQ®-3) [34] and standardized language assessment using the Clinical Evaluation of Language Fundamentals Preschool- 2nd Edition, Australian and New Zealand Standardized Edition (CELF-P2 [35]), as seen in Table 1. All TD children scored within the normal range on the Goldman-Fristoe Test of Articulation, 3rd Edition (GFTA-3) Sounds in Words subtest (SS > 85) [40].

The SSD group comprised children who scored below average on the GFTA-3 (SS < 85, n = 9), or who met two or more of the following criteria (n = 4): scores below the 5th percentile on the VMPAC for oromotor and/or sequencing subtests; a PCC, PPC or PVC, calculated from the GFTA-3 Sounds in Word subtest, greater than 2 standard deviations below the total sample mean; the presence of atypical speech errors as identified by Morgan et al. [36], and Dodd et al. [37]. Group allocation was confirmed by single-case t-tests using the Singlims_ES.exe program [38], with each participant in the SSD group showing statistically significant differences between their score and the TD sample across at least two inclusion criteria measures.

Children with structural deficits (e.g., cleft lip/palate), hearing loss, English as a second language, a diagnosed language, cognitive, neuro-developmental and/or psychological disorder (e.g. cerebral palsy, autism spectrum disorder) and/or motor disorder (developmental coordination disorder, Ehlers Danlos or hypermobility) were excluded from this study. Using the GCSI 39 auto tympanometer, participants demonstrated hearing threshold levels of 20 dB or lower across each frequency of 1000Hz, 2000Hz and 4000Hz, which is consistent with the ASHA Childhood Hearing Screening protocol [39]. Children who were unable to engage in assessment activities due to attentional and behavioral difficulties were also excluded from participation.

Table 1. Mean (*SD*) Participant Characteristics for 3-Year-Old Typically Developing and Speech Sound Disordered Children.

Participant Characteristics	TD ($n = 41$)	SSD ($n = 13$)
Age (Months)	37.90 (1.61)	37.31 (1.65)
ASQ-3		
Communication[a]	56.39 (5.02)	52.78(6.18)
Personal Social[a]	54.03 (5.58)	52.22 (4.41)

 doi:10.20944/preprints202505.0225.v1

Problem Solving[a]	56.67 (5.61)	56.67 (4.33)
Fine Motor[a]	49.17 (10.79)	53.33 (6.12)
Gross Motor[a]	55.83(5.79)	55.00 (4.33)
CELF-P2		
Core Language SS[b]	106.97 (9.32)	101.45 (10.40)
Core Language PR[b]	65.15 (20.59)	58.00 (28.77)
GFTA-3		
Sounds in Words SS	**103.48** (8.96)	**83.92** (7.24)
Sounds in Words PR	**57.83** (20.61)	**16.54** (12.69)
PCC	**80.81** (9.85)	**53.58** (9.05)
PVC	**99.20** (1.52)	**94.25** (6.24)
PPC	**87.10** (6.12)	**69.03** (4.55)
VMPAC		
Focal Oral Motor[c]	**61.53** (14.46)	**40.29** (13.28)
Sequencing[c]	**52.36** (14.68)	**32.00** (15.00)
ICS (Total Score)	**23.37** (2.38)	**21.83** (1.03)

Note: SS = Standard Score; PR = Percentile Rank. ASQ-3= Ages & Stages Questionnaires®, Third Edition; CELF-P2 = Clinical Evaluation of Language Fundamentals Preschool Australian and New Zealand Standardized 2nd Edition; GFTA-3 = Goldman-Fristoe Test of Articulation, 3rd Edition; VMPAC = Verbal Motor Production Assessment for Children. PCC = percentage of consonants correct. PVC = percentage of vowels correct. PPC = percentage of phonemes correct. Focal oral motor and sequency subtests are percentage scores. Means in bold indicate the difference between the TD and SSD groups was statistically significant ($p < .05$). [a]$n = 36$ & $n = 9$ for TD and SSD, respectively, due to missing data. [b]$n = 39$ & $n = 11$ for TD and SSD, respectively, due to missing data. [c]$n = 27$ & $n = 11$ for TD and SSD, respectively, due to missing data.

2.2. Procedure

2.2.1. Data Collection

The CELF-P2, GFTA-3 Sounds-in-Words subtest and VMPAC assessments were completed at a home visit and in accordance with administration guidelines outlined in respective manuals. Participants attended a laboratory at Curtin University to complete a hearing screen and the MSH-PW task. The laboratory room was selected to minimize noise and vibration and staged to be child-friendly using posters of popular children's characters and toys (e.g., Bluey). Testing took place on weekends to further minimize background noise from on campus activities close to and within the building. The average ambient noise level prior to participant arrival, using the Protech QM1589 sound level meter, was 35 dBA, which is below the 48 dB minimum sound level Rusz et al. [40] recommend for the recording of speech.

Participants were provided with free play time to enable researchers to develop rapport and to familiarize the child with the laboratory room. For completion of the MSH-PW, participants were seated on a custom-built chair designed for optimal head position and safety. A Blackmagic Pocket Cinema Camera 4K video camera recorded full HD (1920 x 1080p) at a frame rate of 60 frames per second. The 45mm camera lens was placed central to the child's chair on a Sirui SH15-CN video tripod., as detailed in Palmer et al. [41]

The researcher explained the task to the participant; that they would be shown pictures and asked to copy the investigator saying the target words. Participants were told it was important to

 doi:10.20944/preprints202505.0225.v1

remain seated and to keep looking at the picture and were reminded of this as needed during the task.

A wireless Rodelink LAV microphone (RodeFilmaker Kit) was attached to the participant's clothing, with the receiver connected to one channel of the stereo microphone input of the camera for digital audio recording. With participants who refused to wear the microphone, or for those who repeatedly fiddled with the microphone, it was clamped to the front of the back cushion on the chair at mouth height. A Sennheiser Me66 shotgun microphone attached to the camera and pointing at the speaker was connected to the remaining microphone input channel and served as a second or backup audio recording if required.

The image of the MSH-PW target word was cast onto a 93cm Samsung television directly in the child's line of sight, via a HP EliteBook laptop computer. The order of administration was randomized using an online program [42]. Participants were required to name each of the ten target pictures in response to the instruction, "say X". All responses were videorecorded to enable facial tracking to undertake kinematic analysis and phonetic transcription. Participants were asked to repeat words on occasions they did not repeat the target word or if their body and/or head movements would have compromised accurate facial tracking. Participants were given general feedback and encouragement for their performance and engagement in the tasks. Breaks between tasks were offered as required. The testing session lasted no more than 60 minutes, with variability around participant attention and the need for play breaks.

2.2.2. Data Preparation

The video files from the camera were imported into Adobe Premier Pro to identify the speech movement boundaries for each target word. These onsets and offset boundaries (timestamps) were subsequently used to mark the acoustic word boundaries for each word on text grids within the PRAAT program [43]. A Python script was developed to automate the generation of TextGrids, utilizing the PraatIO library [44] for efficient data processing.

2.2.2.1. Phonetic Transcription

Narrow phonetic transcription and phonological error analysis using the Khan-Lewis Phonological Analysis approach [45] was completed by three experienced S-LPs. Transcriptions were completed in PRAAT and the S-LPs used a combination of the acoustic information, and the videos of target word production for each participant to support their transcription. Prior to commencing transcription, the S-LPs underwent a period of review where, firstly, the S-LPs collaboratively completed example case study transcriptions. Protocols were established for specific PRAAT settings, the use of specific diacritics and other coding variabilities (e.g., coding a final devoicing error as /p/, rather than the voiceless /b/). Second, S-LPs then independently transcribed case study data and points of difference were discussed. Following this calibration, participant transcriptions were randomly allocated between the three S-LPs. After every eight to ten samples were completed, a participant sample was randomly selected for transcription by each S-LP to ensure transcription consistency was maintained. Inter-rater reliability of broad transcription was 86.1%. Once transcription was completed, PPC, PVC and PCC were calculated from broad transcription.

2.2.2.2. Perceptual Scoring

Consensus scoring of five features: appropriate mandibular range, mandibular control/stability, open-close or close-open (phase) movements, voicing transitions, and syllable structure, was undertaken by three certified PROMPT Instructors. A binary scoring system of appropriate (1) or inappropriate (0) was assigned to each word for each feature, according to the scoring criteria, detailed in the MSH-PW manual [46]. Before commencing the scoring of the study sample, the three instructors met with Ms. Deborah Hayden (DH), a co-developer of the MSH-PW, on three occasions to discuss the scoring criteria definitions, calibrate, and collaboratively assess an example case study,

 doi:10.20944/preprints202505.0225.v1

separate to the data set reported in this paper. Given the multidimensionality of features within mandibular range, mandibular control and phase movements, definitions were further refined from those outlined in the MSH-PW manual to allow for consistency in scoring. Three different cases, not related to this study, were subsequently scored independently for reliability analysis. The interclass correlation coefficient, using a mixed model with absolute agreement, showed good agreement (ICCs > .85) between DH and the consensus scorers. After establishing inter-rater reliability, participants were independently scored in sets of five. The scores were collated by an independent research assistant, with items of difference identified. These items were resolved by reviewing the video footage and discussion, until the point of difference was resolved. This process was repeated until all participants were scored.

2.2.2.3. Selection of Kinematic Measurements

Of the five specific criteria evaluated perceptually in Stage III: Mandibular Control, two facial movements, jaw range and jaw control/stability, were further evaluated using measures derived from computer vision-based approaches to measuring facial movements. The specific measurements included mouth opening, which was measured as the ratio of mouth width (between cheilion) to mouth height (between stomion superius/inferius), and lateral deviation of the tip of the chin (pogonion) from rest. The landmarks involved in these measurements are shown in Figure 1. Lines connecting the cheilion landmarks (C_R and C_L) and stomion superius/inferius (S_S and S_I) show the relative distances involved in calculating the mouth opening ratio. Landmark P represents pogonion and the arrows its recorded shift laterally from midline. Measurements of both displacement and velocity were taken. For lateral deviation of pogonion, the extracted distance was normalized by a measurement of facial width derived from the upper face using distances between bilateral landmarks at zygion, tragion, and exocanthion.

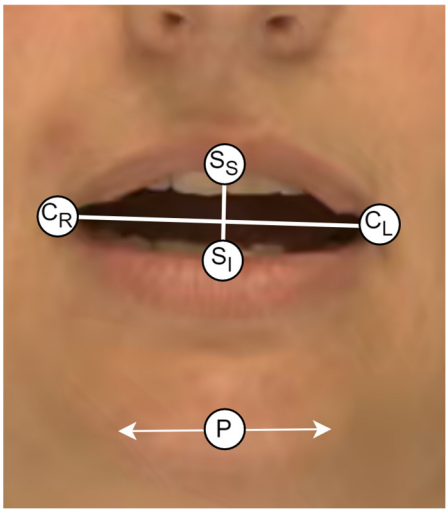

Figure 1. The landmarks used in the extraction of the kinematic measurements.

The flowchart in Figure 2 details the data collection and processing procedures. The analysis elements relating to Research Question 1 and 2 are shown as diamond symbols and are described in more detail in Section 2.3.

 doi:10.20944/preprints202505.0225.v1

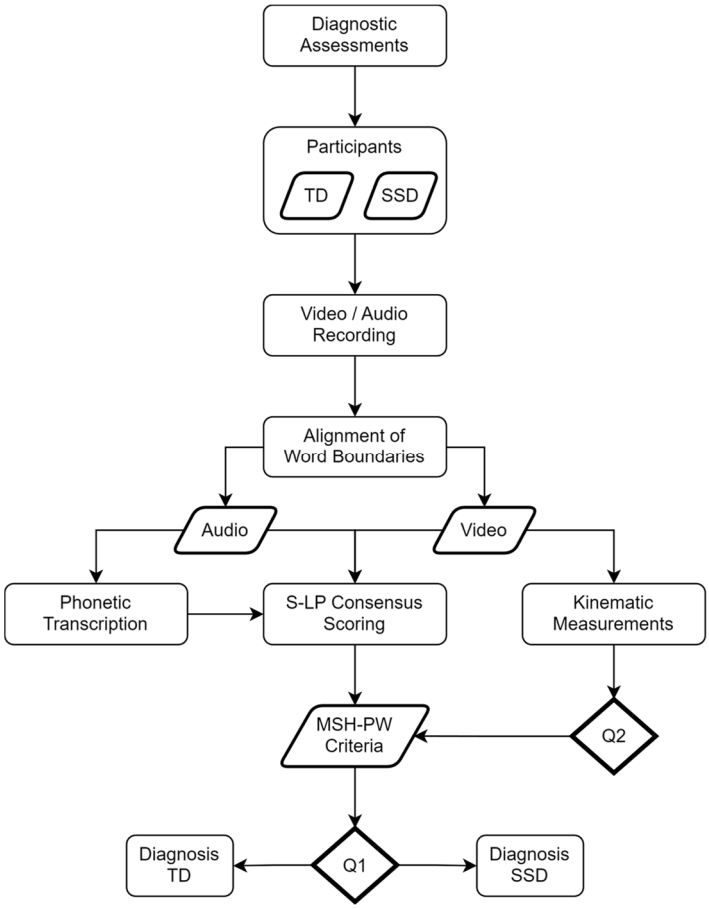

Figure 2. Flowchart showing data collection and processing procedures.

2.3. Data Analysis

2.3.1. Analysis Related to Research Question 1

2.3.1.1. Perceptual Analysis

IBM SPSS Statistics (Version 29) was used to conduct independent samples t-tests or Mann-Whitney U tests (if normality of data was violated) to compare TD and SSD groups in MSH-PW mandibular measures of jaw range, jaw control, phase, voicing transitions, syllable structure and total mandibular percentage score. The normality of the data distributions was assessed using the Shapiro-Wilks test ($p > .05$) and z score skewness and kurtosis values within +/- 1.96. The PVC, PCC and PPC obtained from the MSH-PW phonetic transcriptions were included as outcome measures.

2.3.1.2. Classification Analysis

For each of the five MSH-PW criteria, the scores were encoded as separate ten-valued feature vectors (corresponding to the binary appropriate/inappropriate scores for each of the ten words) and associated with the diagnostic label TD or SSD for training a logistic regression model (with implementation provided by the Scikit-learn Python library [47]) specific to that criterion. A meta-classifier (also a logistic regressor) was then trained on the individual outputs of the five criterion models to predict the final diagnostic label. Leave-One-Out Cross-Validation (LOOCV) was then performed to evaluate the overall scheme's classification performance. A grid search approach was used to tune the models at both the individual criterion level, and at the meta-model level in terms of the ratio, strength, and type of regularization used (lasso, or ridge).

 doi:10.20944/preprints202505.0225.v1

2.3.2. Analysis Related to Research Question 2

A state-of-the-art facial detection and tracking algorithm was used to extract time and space normalized measurements of facial movements from video recorded of the 41 TD participants speaking the ten Probe Words from the mandibular stage of the MSH-PW. The TD sample was used to maximise the data available and to initially assess prediction of scoring with a normative or TD sample, without the confound of some children having a SSD. The SSD sample size was not sufficient for this analysis. Measurements were captured at a sampling rate matching the recording rate of 60 frames per second and linearly interpolated across 1000 timepoints to give sufficient granularity for spoken productions to be aligned at corresponding relative timepoints.

For each word, an expert S-LP was consulted to identify which of the facial movements, over what sub-period intervals of a word's production ought to best characterize it in terms of either the jaw range or the jaw control criteria of the MSH-PW. These rules were then programmatically encoded as parameters to the classification model. The mouth opening measurement was identified for the jaw range criterion as the sole measurement to be evaluated. For jaw control, two different measurements were used: the first derivative of mouth opening (i.e., velocity), and the lateral deviation of pogonion from rest. For classification, an input feature was defined as the average of the absolute value of the Z-score difference of a single participant's range of motion in a measurement subtracted from its mean movement for the corresponding word over all other participants scored as having appropriate jaw range/control (where appropriate) in the spoken sub-period interval of the word shown in Table 3. For each word, LOOCV was performed using logistic regression to predict a label for jaw range/control as appropriate or inappropriate. A grid search approach was used to tune the model's hyperparameters to maximize classification performance. This included balancing the strength and relative proportions of lasso and ridge regularization. Class weightings were also adjusted in the model to account for dataset imbalance.

Table 3. Measurement Sub-Period Intervals for MSH-PW Criteria Classification.

Probe Word	MSH-PW Criteria	Sub-Period Interval (%)
Ba	Jaw Range	[20, 60]
	Jaw Control	[10, 40]
Eye	Jaw Range	[20, 60]
	Jaw Control	[10, 40]
Map	Jaw Range	[10, 50]
	Jaw Control	[5, 70]
Um	Jaw Range	[0, 70]
	Jaw Control	[0, 70]
Ham	Jaw Range	[10, 60]
	Jaw Control	[10, 60]
Papa	Jaw Range	[10, 40], [45, 80]
	Jaw Control	[10, 40], [45, 80]
Bob	Jaw Range	[10, 50]
	Jaw Control	[10, 50]
Pam	Jaw Range	[10, 60]
	Jaw Control	[10, 60]
Pup	Jaw Range	[10, 50]
	Jaw Control	[10, 50]
Pie	Jaw Range	[15, 50]

 doi:10.20944/preprints202505.0225.v1

Jaw Control	[15, 50]

3. Results

3.1. Research Question 1

3.1.1. Mean Differences Between SSD and TD Groups

Mean MSH-PW mandibular subtest scores and total scores for TD and SSD groups are presented in Table 4, along with PVC, PCC and PPC.

Table 4. Means, Standard Deviations and Cohen's d for MSH-PW Mandibular Scores for Typically Developing and Speech Sound Disordered Children.

	TD	SSD	d (95% CI)
Jaw Range	**8.29** (1.79)	**6.46** (2.22)	0.96 (0.31-1.61)
Jaw Control	6.95 (2.96)	5.38 (2.84)	0.53 (-0.10-1.16)
Phase	6.54(3.29)	5.23 (2.74)	0.41 (-0.22 - 1.04)
Voicing Transitions	**9.07** (1.17)	**7.92** (1.38)	0.94 (0.28 - 1.59)
Syllable Structure	9.76 (0.54)	9.54 (0.52)	0.41 (-0.22 - 1.03)
Mandibular Percent Total	**81.22** (15.33)	**69.08** (14.37)	0.80 (0.16 - 1.44)
PVC	**91.23** (8.30)	**79.50** (16.17)	1.10 (0.44 - 1.75)
PCC	**89.93**(10.22)	**80.18** (13.84)	0.83 (0.18 - 1.47)
PPC	**89.93** (7.31)	**79.75** (9.93)	1.27 (0.60- 1.94)

Note. The group means in bold are significantly different.

Results showed significantly higher scores for TD compared to SSD children for jaw range, $U(N = 54) = 133.5$, $p = .006$, voicing transitions, $U(N = 54) = 132.0$, $p = .004$, and total mandibular scores, $t(52) = 2.524$, $p = .015$. There was no statistically significant difference between groups on jaw control, $U(N = 54) = 183.5$, $p = .089$, open-close or close-open phase, $U(N = 54) = 197.0$, $p = .156$, or syllable structure, $U(N = 54) = 201.5$, $p = .085$. Cohen's d effect size was large for jaw range (.96), voicing transitions (.94) and total mandibular score (.80). There was a medium Cohen's d effect sizes for jaw control (.47), and small to medium effect size for phase (.37) and syllable structure (.35).

For phonetic accuracy using the MSH-PW mandibular word set, the TD group showed significantly higher PVC, $U(N = 54) = 125.0$, $p = .003$, PCC, $U(N = 54) = 137.0$, $p = .008$, and PPC, $U(N = 54) = 115.0$, $p = .002$, than the SSD group. Cohen's d showed a large effect size for all three measures: PVC (1.101), PCC (.833) and PPC (1.273). These results indicate the mandibular word set provides speech samples that are sensitive to the speech production difficulties of the SSD children when compared to TD children.

3.1.2. Classification Analysis

The confusion matrix of Table 5 shows the results of classifying the perceptually scored MSH-PW criteria to the binary diagnostic labels TD or SSD. The balanced accuracy score derived as the average of the sensitivity (recall on the SSD class) and specificity (recall on the TD class) is used here to report performance due to its robustness to imbalanced datasets. Similarly, the balanced precision (also known as the macro-averaged precision) calculated as the average of the positive predictive value (PPV) and the negative predictive value (NPV) is used to report overall precision performance.

 doi:10.20944/preprints202505.0225.v1

Table 5. Confusion matrix and statistics for the classification of TD or SSD participants given expert consensus perceptual scoring of the MSH-PW.

		Predicted			
		TD	**SSD**	**Recall**	**Precision**
True	**TD**	30	11	0.73	0.94
	SSD	2	11	0.85	0.50
		Bal. Acc. / Prec.		0.79	0.72

At an alpha of 0.01, Monte Carlo simulation over one million runs estimated the significance threshold for the class balanced accuracy and precision statistics as approximately 0.729 and 0.661 respectively. Since the evaluated balanced accuracy and precision scores exceed these thresholds, it is concluded that the null hypothesis of no association between the mandibular perceptually scored MSH-PW criteria and the diagnostic class is rejected. The precision-recall plot of Figure 3 summarizes the classification performance for the SSD class showing how precision/specificity varies with increasing recall/sensitivity. A typical precision-recall curve shows a relatively smooth tradeoff between recall and precision with precision decreasing to chance levels as recall increases. The atypical profile of this plot likely indicates a lack of generalizability in the trained model.

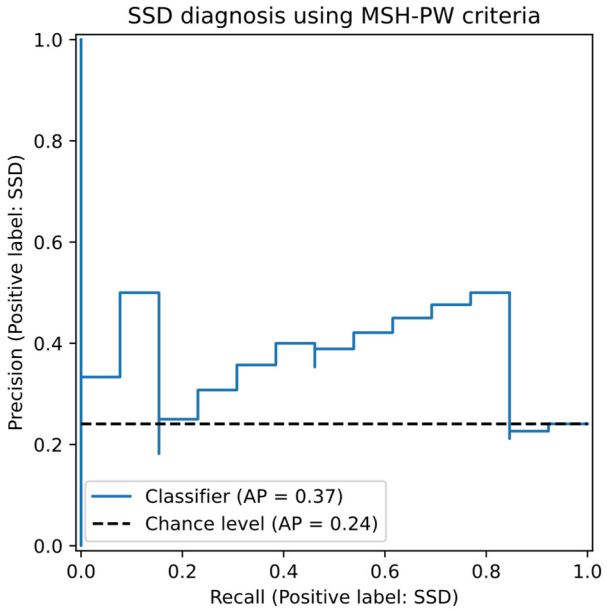

Figure 3. Precision-Recall curve for SSD Classification using the MSH-PW Criteria.

The classification scheme was trained over the whole dataset using the hyperparameters derived through LOOCV to find the weights used by the criterion models and by the meta-classifier. After normalizing, these weights (which act as coefficients for the respective features) reflect the contribution of the features in supporting determination of the final diagnostic labels. Under the LOOCV derived hyperparameters tuned to maximise classification performance, only the classifier trained against the Jaw Range scores contributed meaningfully to the final diagnostic label. Lasso regularization in the meta-classifier minimized the weights of the other criterion classifiers to zero, indicating that their contributions did not improve overall performance. Within the Jaw Range classifier itself, lasso regularization minimized weightings against all words except the following: “Ba”, “Eye”, “Um”, “Pam”, and “Pie”. Measurements taken from the productions of the other words in the set were ignored by the Jaw Range classifier.

 doi:10.20944/preprints202505.0225.v1

3.2. Research Question 2

The confusion matrix in Table 6 shows the resulting classification performance for appropriate or inappropriate jaw range given optimal tuning of the model's hyperparameters via LOOCV for the evaluated sample.

Table 6. Confusion matrix and statistics for the classification of appropriate / inappropriate Jaw Range using the objectively measured "Mouth Opening" facial feature.

		Predicted			
		Inapp.	**App.**	**Recall**	**Precision**
True	**Inapp.**	71	45	0.61	0.38
	App.	117	307	0.72	0.87
		Bal. Acc. / Prec.		0.67	0.62

The overall balanced accuracy was calculated as 0.67 and balanced precision as 0.62. Both these statistics are higher than the respective significance thresholds of 0.583 for accuracy and 0.542 for precision derived via Monte Carlo simulation for their occurrence by chance at alpha = 0.01. It is therefore concluded that the null hypothesis of no correlation between the jaw range criterion and the objective mouth opening facial feature is rejected. The associated precision-recall plot is shown in Figure 4 which shows that precision stays relatively high before smoothly decreasing past approx. 75% sensitivity.

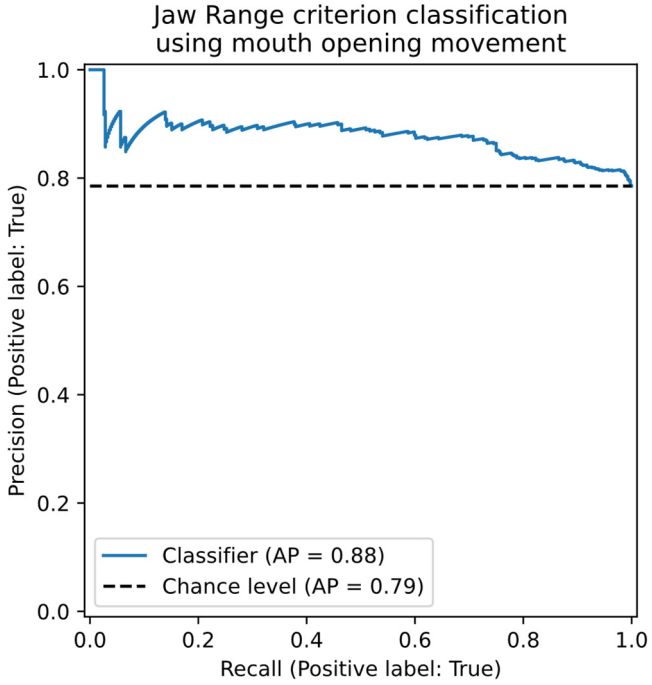

Figure 4. Precision-Recall plot for the classification of appropriate / inappropriate Jaw Range using the objective mouth opening facial feature.

For classification of appropriate/inappropriate jaw control using the velocity of mouth opening, classification performance was significant at an alpha of 0.01. The significance thresholds for balanced accuracy and precision were estimated to be 0.574 and 0.547 respectively for the jaw control criterion. The confusion matrix and statistics for the optimally tuned model hyperparameters in Table 7 shows

 doi:10.20944/preprints202505.0225.v1

classification performance above these thresholds meaning that the null hypothesis of no correlation between the measurement and the jaw control score is rejected.

Table 7. Confusion matrix and statistics for the classification of appropriate / inappropriate Jaw Control using the objectively measured "Mouth Opening velocity" facial feature.

		Predicted			
		Inapp.	**App.**	**Recall**	**Precision**
True	**Inapp.**	85	100	0.46	0.50
	App.	84	271	0.76	0.73
		Bal. Acc. / Prec.		0.61	0.62

The associated precision-recall plot is shown in Figure 5. Precision starts high before plateauing around 50% sensitivity, tailing off from 80% sensitivity.

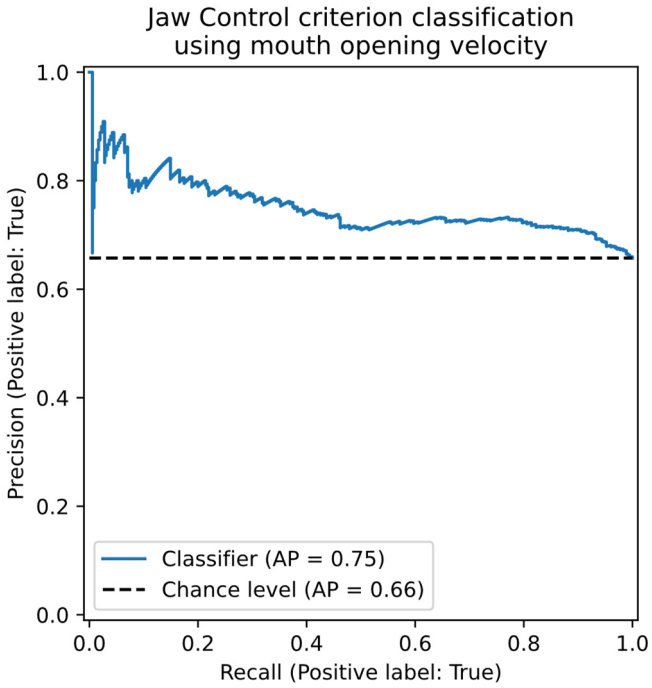

Figure 5. Precision-Recall plot for the classification of appropriate / inappropriate Jaw Control using the objective "Mouth Opening velocity" facial feature.

The confusion matrix showing results for the classification of jaw control using the objective measurement of lateral displacement of Pogonion is shown in Table 8. Again, the results for balanced accuracy and precision are significant at alpha = 0.01 though the relative incidence of type I and type II errors is reversed. This potentially indicates that the features are characterizing different aspects of the data. The associated precision-recall plot is shown in Figure 6. Classification performance tracks similarly to that obtained using the velocity of mouth opening feature.

 doi:10.20944/preprints202505.0225.v1

Table 8. Confusion matrix and statistics for the classification of appropriate / inappropriate Jaw Control using the objective measurement of lateral displacement of pogonion.

		Predicted			
		Inapp.	**App.**	**Recall**	**Precision**
True	**Inapp.**	109	76	0.59	0.46
	App.	127	228	0.64	0.75
		Bal. Acc. / Prec.		0.62	0.61

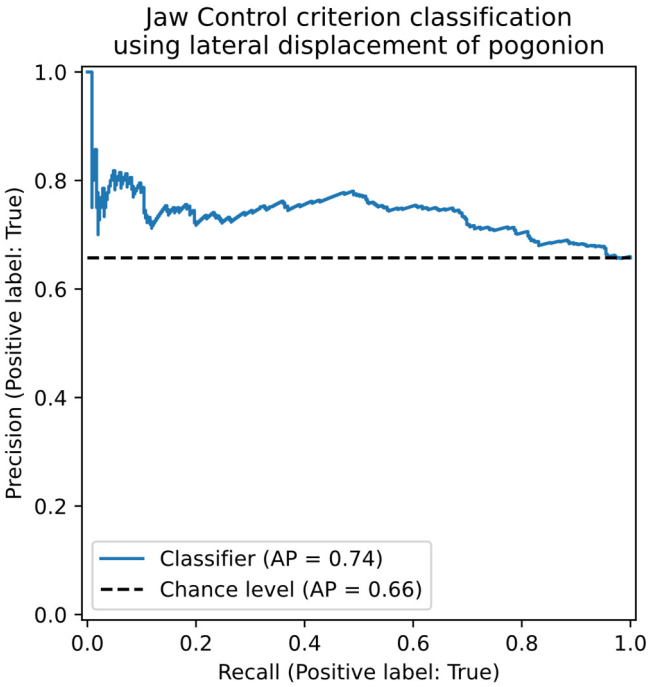

Figure 6. Precision-Recall plot for the classification of appropriate / inappropriate Jaw Control using the objective "Mouth Opening velocity" facial feature.

Finally, the confusion matrix of Table 9 shows classification performance when combining both the mouth opening velocity and the lateral displacement of pogonion features. In this case it was possible to tune the model's hyperparameters to improve classification performance over that obtained by either feature alone offering evidence of the multidimensional nature of the jaw control criterion. The associated precision-recall plot shown in Figure 7 shows improved precision at all levels of sensitivity.

Table 9. Confusion matrix and statistics for the classification of appropriate / inappropriate jaw control using the combined objective facial features.

		Predicted			
		Inapp.	**App.**	**Recall**	**Precision**
True	**Inapp.**	88	97	0.48	0.55
	App.	71	284	0.80	0.75
		Bal. Acc. / Prec.		0.64	0.65

 doi:10.20944/preprints202505.0225.v1

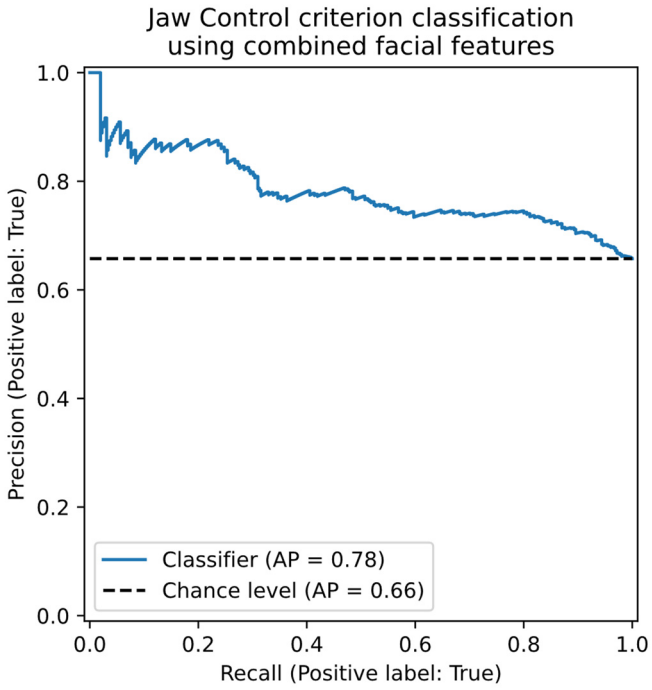

Figure 7. Precision-Recall plot for the classification of appropriate / inappropriate Jaw Control using the combined objective facial features.

4. Discussion

In this paper, we report on a preliminary study aimed at exploring the potential of the MSH-PW to discriminate the speech of children with SSD from TD speech, through evaluation of the Stage III: Mandibular Control word set. We aimed to determine, firstly, whether the observations of speech pathologists on five features: appropriate mandibular range, mandibular control/stability, open-close or close-open movements, voicing transitions and syllable structure, could accurately classify children with TD speech from those with an SSD. Secondly, we evaluated the agreement between the subjective visual observations of the consensus scores completed by three S-LPs and those derived through kinematic measurements, extracted from a state-of-the-art facial mesh detection and tracking algorithm. These research questions were selected to inform the development of norms for the MSH-PW for the purpose of diagnosing impaired speech motor control in children and evaluate the feasibility of supporting S-LPs with objectively acquired measurements of motor speech control, framed within the MSH-PW scoring criteria. Each research aim will be discussed in turn.

4.1. Classification of Children Based on Perceptual Scoring

The results of this study found there were significant differences in MSH-PW jaw range, voicing transitions and mandibular total score between children with TD speech and those with SSD when scored perceptually by S-LPs. Discrimination analysis indicated a significant correlation between perceptually scored MSH-PW mandibular criteria and diagnostic class as determined by a battery of diagnostic tools used in current clinical practice. This suggests perceptual scoring of the MSH-PW mandibular subtest can discriminate between children with TD speech and those with SSD, with potential for the MSH-PW to be used by S-LPs in diagnosing impaired speech motor control.

The finding of no significant difference in jaw control contrasts with existing literature that has established the significance of jaw control and stability in speech sound production [48–52]. Development of jaw control is a key feature of speech development. For example, jaw movement velocities are slower and more variable in young children [54] with children with typical development refining multidimensional jaw movements by around 6-8 years of age [55,56]. Wilson and Nip [55] highlighted the importance of jaw control and stability in supporting lip and tongue

 doi:10.20944/preprints202505.0225.v1

movements, noting its involvement in nearly all articulatory positions [53]. Jaw control has also been identified as a feature of SSD. For example, Mogren and colleagues found children with SSD had larger lateral jaw movements than children with TD speech [53]. Similarly, Terband et al. [54] identified clear deviances in lateral jaw movement within their SSD group compared to their sample of TD participants. Reasons for the contrast in these findings with this current study were considered.

Firstly, previous studies examining motor control tend to feature participants with clearly identified motor speech disorders (e.g., [55]) , including childhood apraxia of speech (e.g., [56]) and/or may differentiate between various subtypes of SSD. For example, Terband et al. [54] differentiated children with phonetic articulation disorder, phonological disorder and childhood apraxia of speech. Participants in the SSD group of this current study had not been identified as having SSD prior to their participation in the study (while some parents expressed uncertainty over whether their child's speech was developing within age expectation, concerns had not been sufficient to seek S-LP assessment). As such, it is possible participants in the SSD group demonstrated more mild SSD features and/or that children within this group primarily had difficulties at the phonological level, rather than motor-based difficulties. This suggestion is supported by Terband et al.'s finding that a participant with a phonetic articulation disorder had "very normal values" on lateral jaw movement. As such, is it plausible that the results of this current study may be reflective of the characteristics of participants in the SSD group. Further analysis of children with diagnosed motor speech difficulties may yield classification differences across a wider range of MSH-PW criteria. Secondly, the mandibular word set items were intentionally selected to include only bilabial consonants and low vowels with targets achieved through open-close (e.g., Um), close-open (e.g., Ba), close-open-close (e.g., Map) and close-open-close-open (e.g., Papa) jaw movements. With the vowel target determining jaw height, differences in PVC and PPC between the TD and SSD groups indicate there are differences in speech production accuracy from an auditory perceptual perspective that may not be evident in perceptual movement analysis over the ten mandibular target items. Furthermore, difficulties in jaw control, specifically, may not be evident until motor complexity increases in the higher MSH-PW stages. Further investigations are required to determine the impact on jaw control as young children are required to integrate jaw stability with labial facial and lingual movements, and sequencing these movements in multisyllabic and phrase level speech.

4.2. Agreement between Perceptual Scoring and Kinematic Measures

Our second research question sought to explore the agreement between the perceptual scores of jaw range with the kinematic measure labelled mouth opening; and jaw control with two kinematic measures labelled mouth opening velocity and lateral movement of the pogonion. We found good agreement for both jaw range and jaw control.

The rating of jaw range required consensus raters make a binary judgement of appropriate or inappropriate for age, where a judgement of inappropriate was given for movements considered restricted or over extended, as defined for each vowel height position, for each word, in the MSH-PW scoring manual. Our preliminary results suggest the objective measure of mouth opening could be used to support speech pathologists in their assessment of jaw range.

It is proposed that the agreement between the consensus raters with the mouth opening measurement resulted from their already established internal representation of jaw height, and that this representation was aligned with the objective kinematic measures, resulting in good agreement. This proposal is based on the knowledge that the consensus scorers are familiar with the vowel quadrilateral that describe jaw and tongue positions, as well as an established body of literature that specifies jaw height adjustments contribute to the production of vowels [57,58]. It is, therefore, conceivable that the raters utilized this knowledge, along with their experience, to inform their decision making. That is, when a child produced a vowel error, the associated jaw height position could be evaluated as too high or too low, with respect to the intended target.

Similarly, there was good agreement between the consensus scores and kinematic measures for jaw control. The rating of jaw control required the consensus raters to make a binary judgement of

 doi:10.20944/preprints202505.0225.v1

appropriate or inappropriate for age, based on velocity and midline or anterior-posterior stability of the jaw. The finding that the combined measures showed greater agreement than the individual measure is likely reflective of the multidimensionality of the jaw control criterion and jaw control movements in general. Multidimensionality is an essential feature of jaw movement and the integration and balancing of vertical, lateral, and rotational movements with precise timing and velocity enable speakers to adapt to the acoustic and articulatory demands of different phonemes and provide a dynamic scaffold for tongue and lip movements [52,58–61]. As outlined, the rating for jaw control is based on several criteria reflective of controlled, smooth speech movements within the vertical plane. The velocity of mouth opening and lateral displacement of the pogonion are both key metrics of jaw control and the interplay of each movement contributes to the production of fluent, intelligible speech.

Agreement between consensus scorers and kinematic measurements derived from automated facial tracking was likely aided by the high level of experience the consensus scorers had in the assessment of speech motor control. Further research should explore the level of agreement in ratings of jaw range and jaw control with S-LPs who have less experience in the assessment of speech motor control; and determine whether kinematic measures can support the clinical judgements of S-LPs of differing levels of experience when scoring jaw range and jaw control criterion of the MSH-PW.

4.3. Clinical Application

Perceptual, single word speech assessments such as the GFTA-3 are a critical component in the assessment of children's speech and the diagnosis of SSD [12,49], providing timely and convenient measures of speech development and accuracy. As highlighted, there are limitations with current perceptual assessments, including their focus on identifying phonological deficits, rather than also assessing for underlying speech motor control difficulties [50]. The MSH-PW was designed to measure inappropriate speech motor control through the perceptual visual and auditory assessment of single word productions. The findings of this preliminary study of the Stage III: Mandibular Control level of the MSH-PW indicate the assessment tool can identify perceptual differences in the appropriateness of jaw range and voicing transitions between children with TD speech compared to those with SSD, and support further research into the additional levels of the MSH-PW; Stage IV: Labial-Facial Control, Stage V: Lingual Control and Stage VI: Sequenced movements. The generation of normative data for children's performance at these levels of speech motor control, along with the MSH-PW total scores, would be beneficial.

5. Limitations

This research analyzed data from a small sample of children comprising 41 TD participants and 13 SSD participants. All children were within the age range of 3;0 to 3;5 years. This small, limited age sample limits generalization of the findings to a larger population and those younger, or older, than this target age. Children aged between 3;0 and 3;5 years frequently present with speech sound errors (e.g.,[59]), and it is possible that participants in the TD sample scored within age expectations on standardized assessments at the time of their participation in the study but may be identified with SSD as they get older and error patterns that are currently developmentally appropriate, persist [37]. Similarly, the SSD group comprised children who had not previously been identified as having SSD suggesting their SSD features may have been less severe and not a broad representation of the severity of SSD in children seeking speech pathology assessment and intervention. Additionally, the SSD group were not diagnosed according to subtype using a classification framework (e.g., phonological disorder, childhood apraxia of speech)[60]. That the SSD children as a group were significantly lower than TD controls on the VMPAC, indicating poorer oromotor and sequencing skills, suggests some speech motor involvement within the SSD group. Future research with a larger sample size is needed to investigate the role of the MSH-PW word sets in differentially diagnosing subtypes of SSD.

 doi:10.20944/preprints202505.0225.v1

A further limitation of this study is that the analysis is restricted to the MSH-PW mandibular word set. These 10 words contain a limited set of consonants (e.g., m, p, b) which may be insufficient to accurately perceive differences in jaw control, phase, and syllable structure. Work is in progress to analyze the remaining stages (thirty words and four phrases) of the MSH-PW.

This study also sought to discover if it was possible to associate facial measurements extracted from recorded video with the mandibular range and control criteria of the MSH-PW. Three different measurements were tested, with further investigation of other facial measurements warranted, especially given the likely multi-dimensional nature of the MSH-PW criteria. Finally, the accuracy of extracted facial measurements in distinguishing between disordered and typically developing mandibular control was not directly tested in the present study. Future research is, therefore, needed to investigate whether objective measures obtained through facial tracking can support the differential diagnosis of SSD.

6. Conclusion

The data in this paper provides preliminary evidence that children with speech sound disorders (SSD) showed significant differences compared to children with typically developing speech in measures of jaw range, voicing transitions, and total mandibular score, as scored on the MSH-PW. Future work is focused on undertaking analyses of the stages IV, V and VI of the MSH-PW. The findings of these analyses will help inform the potential of the MSH-PW to identify issues with motor speech control in children with SSD and as such provide further validation of the MSH-PW for use in the diagnosis of SSD [23]. For young 3-year-old children, jaw range and voicing transitions may serve as relevant markers of an underlying deficit in speech motor control that could impact articulation accuracy and limit a child's speech intelligibility.

Further, the high agreement between consensus raters and the objective measures of speech-related mouth movements, obtained using a state-of-the-art facial mesh detection and tracking algorithm, suggests objective measures of motor speech control are clinically feasible. Future investigation will explore the relationships between other extracted facial measurements and criteria of the MSH-PW in addition to those reported in this paper. We will utilize a data driven approach using the extracted measurements from video and evaluate the relationship to diagnosis.

Author Contributions: Conceptualization, R.L.P., R.W., P.H. and G.R.S.; Methodology, L.O., R.L.P., R.W., P.H., G.R.S., P.D. and N.W.H; Software, R.L.P., P.E., G.R.S., P.D., N.W.H.; Validation, : Formal Analysis, L.O., R.W., R.L.P. and N.W.H.; Investigation, L.O., R.L.P, and R.W.; Resources, : Data Curation, : R.W, G.R.S, P.D. and N.W.H.; Visualization, : Supervision, R.W., P.H. and N.W.H.; Project Administration, R.W. : Funding Acquisition, R.W.

Funding: This research was supported by the Health Department of Western Australia, grant WANMA/Ideas2023-24/9 and REA20222, as well as the PROMPT Institute grant number 59592.

Institutional Review Board Statement: This research was conducted in accordance with the National Statement on Ethical Conduct in Human Research (2007). Ethics approval was obtained from the Curtin University Human Research Ethics Committee (Approval number: HRE2020-0327).

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: Data are available subject to the ethical considerations of this study.

Acknowledgments: The authors would like to acknowledge the participants and their families; Kaelee Koprowicz for assistance with data collection, preparation and ground truth scoring; Elizabeth Barty, Kathryn Daniels, Anne Walker and Deborah Hayden for their assistance with data preparation and ground-truth scoring.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

 doi:10.20944/preprints202505.0225.v1

References

1. Dodd, B., Differential diagnosis of pediatric speech sound disorder. Current Developmental Disorders Reports, 2014. 1: p. 189-196.
2. Eadie, P., et al., Speech sound disorder at 4 years: Prevalence, comorbidities, and predictors in a community cohort of children. Developmental Medicine & Child Neurology, 2015. 57(6): p. 578-584.
3. Mullen, R. and T. Schooling, The national outcomes measurement system for pediatric speech-language pathology. 2010.
4. Harding, S., et al., Outcome measures for children with speech sound disorder: an umbrella review. BMJ open, 2024. 14(4): p. e081446.
5. McCormack, J., et al., A systematic review of the association between childhood speech impairment and participation across the lifespan. International Journal of Speech-Language Pathology, 2009. 11(2): p. 155-170.
6. Felsenfeld, S., P.A. Broen, and M. McGue, A 28-year follow-up of adults with a history of moderate phonological disorder: Educational and occupational results. Journal of Speech, Language, and Hearing Research, 1994. 37(6): p. 1341-1353.
7. Association, A.S.-L.-H. Speech Sound Disorders:Articulation and Phonology. 2024 [cited 2024 August 3]; Available from: https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/.
8. Littlejohn, M. and E. Maas, How to cut the pie is no piece of cake: toward a process-oriented approach to assessment and diagnosis of speech sound disorders. International Journal of Language & Communication Disorders, 2023.
9. McCabe, P., J. Korkalainen, and D. Thomas, Diagnostic Uncertainty in Childhood Motor Speech Disorders: A Review of Recent Tools and Approaches. Current Developmental Disorders Reports, 2024: p. 1-8.
10. Stringer, H., et al., Speech sound disorder or DLD (phonology)? Towards a consensus agreement on terminology. International Journal of Language & Communication Disorders, 2023.
11. Fabiano-Smith, L., Standardized tests and the diagnosis of speech sound disorders. Perspectives of the ASHA special interest groups, 2019. 4(1): p. 58-66.
12. Diepeveen, S., et al., Clinical reasoning for speech sound disorders: Diagnosis and intervention in speech-language pathologists' daily practice. American Journal of Speech-Language Pathology, 2020. 29(3): p. 1529-1549.
13. Skahan, S.M., M. Watson, and G.L. Lof, Speech-language pathologists' assessment practices for children with suspected speech sound disorders: Results of a national survey. 2007.
14. Green, J.R., Mouth matters: Scientific and clinical applications of speech movement analysis. Perspectives on Speech Science and Orofacial Disorders, 2015. 25(1): p. 6-16.
15. Kent, R.D., Research on speech motor control and its disorders: A review and prospective. Journal of Communication disorders, 2000. 33(5): p. 391-428.
16. Frisch, S.A. and R. Wright, The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. Journal of Phonetics, 2002. 30(2): p. 139-162.
17. Pollock, K.E. and M.C. Berni, Transcription of vowels. Topics in Language Disorders, 2001. 21(4): p. 22-40.
18. Pouplier, M., Tongue kinematics during utterances elicited with the SLIP technique. Language and Speech, 2007. 50(3): p. 311-341.
19. Nip, I.S., J.R. Green, and D.B. Marx, The co-emergence of cognition, language, and speech motor control in early development: A longitudinal correlation study. Journal of Communication Disorders, 2011. 44(2): p. 149-160.
20. Shriberg, L.D., J. Kwiatkowski, and H.L. Mabie, Estimates of the prevalence of motor speech disorders in children with idiopathic speech delay. Clinical linguistics & phonetics, 2019. 33(8): p. 679-706.
21. McCauley, R.J. and E.A. Strand, A review of standardized tests of nonverbal oral and speech motor performance in children. 2008.
22. Strand, E.A. and R.J. McCauley, Dynamic evaluation of motor speech skill (DEMSS) manual. 2019: Paul H. Brookes Publishing, Company.
23. Namasivayam, A.K., et al., Development and validation of a probe word list to assess speech motor skills in children. American Journal of Speech-Language Pathology, 2021. 30(2): p. 622-648.

 doi:10.20944/preprints202505.0225.v1

24. Bandini, A., A. Namasivayam, and Y. Yunusova. Video-Based Tracking of Jaw Movements During Speech: Preliminary Results and Future Directions. in Interspeech. 2017.
25. Bhardwaj, A., et al., Transforming Pediatric Speech and Language Disorder Diagnosis and Therapy: The Evolving Role of Artificial Intelligence. Health Sciences Review, 2024: p. 100188.
26. Garg, S., et al., ADFAC: automatic detection of facial articulatory features. MethodsX, 2020. 7: p. 101006.
27. Guarin, D.L., et al., Video-based facial movement analysis in the assessment of bulbar amyotrophic lateral sclerosis: clinical validation. Journal of Speech, Language, and Hearing Research, 2022. 65(12): p. 4667-4678.
28. Nöth, E., et al., Automatic evaluation of dysarthric speech and telemedical use in the therapy. Phonetician, 2011. 103(1): p. 75-87.
29. Bandini, A., et al. Automatic detection of amyotrophic lateral sclerosis (ALS) from video-based analysis of facial movements: speech and non-speech tasks. in 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). 2018. IEEE.
30. Vick, J.C., et al., Data-driven subclassification of speech sound disorders in preschool children. Journal of Speech, Language, and Hearing Research, 2014. 57(6): p. 2033-2050.
31. Goldman, R. and M. Fristoe, Goldman-Fristoe Test of Articulation Manual. 2015, Minnesota, USA: Pearson.
32. Hayden, D. and P. Square, VMPAC Verbal Motor Production Assessment for Children Examiner's Manual. 1999, USA: PsychCorp.
33. McLeod, S., L.J. Harrison, and J. McCormack, The Intelligibility in Context Scale: Validity and reliability of a subjective rating measure. 2012.
34. Squires, J., D.D. Bricker, and E. Twombly, Ages & stages questionnaires. 2009: Paul H. Brookes Baltimore.
35. Wigg, E.H., W.A. Secord, and S. Eleanor, Clinical Evaluation of Language Fundamentals Preschool Second Edition Australian and New Zealand Examiner's Manual. 2004, Sydney, Australia: Person Clinical and Talent Assessment.
36. Morgan, A., et al., Who to refer for speech therapy at 4 years of age versus who to "watch and wait"? The Journal of pediatrics, 2017. 185: p. 200-204. e1.
37. Dodd, B., et al., Phonological development: a normative study of British English-speaking children. Clinical linguistics & phonetics, 2003. 17(8): p. 617-643.
38. Crawford, J.R., P.H. Garthwaite, and S. Porter, Point and interval estimates of effect sizes for the case-controls design in neuropsychology: rationale, methods, implementations, and proposed reporting standards. Cognitive neuropsychology, 2010. 27(3): p. 245-260.
39. American Speech Language and Hearing Association. Childhood hearing screening [Practice portal]. n.d. [cited 2019 October 25]; Available from: https://www.asha.org/practice-portal/professional-issues/childhood-hearing-screening/.
40. Rusz, J., et al., Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. Movement Disorders, 2021. 36(4): p. 803-814.
41. Palmer, R., et al., Facial Movements Extracted from Video for the Kinematic Classification of Speech. Sensors, 2024. 24(22): p. 7235.
42. Social Psychology Network. Research Randomizer. 2019.
43. Boersma, P. and D. Weenink, Praat: doing phonics by computer. 2019.
44. Mahrt, T., PraatIO. 2016.
45. Khan, L.M. and N. Lewis, Khan-Lewis Phonological Analysis Manual. 2015, Minnesota, USA: Pearson.
46. The Primpt Institute. Probe Words Scoring Manual American English. 2022, Sante Fe, New Mexico: The PROMPT Institute.
47. Pedregosa, F., et al., Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011. 12: p. 2825-2830.
48. Hayden, D.A. and P.A. Square, Motor speech treatment hierarchy: A systems approach. Clinics in Communication Disorders, 1994. 4(3): p. 162-174.
49. Moore, C.A., Physiologic development of speech production, in Speech motor control in normal and disordered speech, B. Maassen, et al., Editors. 2004, Oxford University Press: Oxford, UK. p. 191-211.
50. Grigos, M.I., Changes in articulator movement variability during phonemic development: a longitudinal study. Journal of Speech, Language, and Hearing Research, 2009. 52(1): p. 164-177.

 doi:10.20944/preprints202505.0225.v1

51. Smith, A. and L. Goffman, Stability and patterning of speech movement sequences in children and adults. Journal of Speech, Language, and Hearing Research, 1998. 41(1): p. 18-30.
52. Namasivayam, A.K., et al., Relationship between speech motor control and speech intelligibility in children with speech sound disorders. Journal of communication disorders, 2013. 46(3): p. 264-280.
53. Mogren, Å., A. McAllister, and L. Sjögreen, Range of motion (ROM) in the lips and jaw during vowels assessed with 3D motion analysis in Swedish children with typical speech development and children with speech sound disorders. Logopedics Phoniatrics Vocology, 2021.
54. Terband, H., Y. van Zaalen, and B. Maassen, Lateral jaw stability in adults, children, and children with developmental speech disorders. Journal of Medical Speech-Language Pathology, 2013. 20(4): p. 112-118.
55. Namasivayam, A.K., et al., Relationship between speech motor control and speech intelligibility in children with speech sound disorders. Journal of communication disorders, 2013. 46(3): p. 264-280.
56. Moss, A. and M.I. Grigos, Interarticulatory coordination of the lips and jaw in childhood apraxia of speech. Journal of Medical Speech-Language Pathology, 2012. 20(4): p. 127-132.
57. Mooshammer, C., P. Hoole, and A. Geumann, Jaw and order. Language and Speech, 2007. 50(2): p. 145-176.
58. Ladefoged, P., K. Johnson, and P. Ladefoged, A course in phonetics. Vol. 3. 1997, Boston: Harcourt Brace Jovanovich College Publishers.
59. Grunwell, P., Natural phonology, in The new phonologies: Developments in clinical linguistics, M.J. Ball and R.D. Kent, Editors. 1997, Singular Publishing Group, Inc.: San Deigo, CA.
60. Rvachew, S. and T. Matthews, Considerations for identifying subtypes of speech sound disorder. International Journal of Language & Communication Disorders, 2024. 59: p. 2146-2157.