Uncertainty analysis of clinically relevant distances derived using photogrammetric intersection for the assessment of motor-speech-control in children

Liam Boyle[1], Petra Helmholz[1], Roslyn Ward[2], Richard Palmer[2], Derek Lichti[3]

[1] School of Earth and Planetary Sciences, Curtin University, Kent St, Bentley, Australia (Liam.Boyle; Petra.Helmholz)@curtin.edu.au
[2] Curtin School of Allied Health, Curtin University, Kent St, Bentley, Australia (R.Ward; R.L.Palmer)@curtin.edu.au
[3] Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (ddlichti@ucalgary.ca)

Keywords: Speech Sound Disorders (SSDs), Speech-Language Pathologists, BlazeFace, Photogrammetry, Least Squares Adjustment

Abstract

Perceptual analysis is the current benchmark standard used by speech-language pathologists in diagnosing speech sound disorders (SSDs). Yet, related research indicates access to objective measures could improve the assessment process. Recent technological advances have contributed to developing AI-based methods to provide clinicians access to objective measures for speech-motor control by calculating inter-landmark distances of anatomically relevant facial landmarks. However, landmarks placed by AI-based methods extract the landmarks' coordinates without associated uncertainties. Consequently, inter-landmark distances extracted as objective measurements also lack uncertainty information, potentially compromising their suitability for assessment purposes. In contrast, photogrammetry can predict facial inter-landmark distances and their uncertainties through intersection and variance propagation. In this paper, we use a combination of the markerless BlazeFace algorithm and photogrammetry to examine how different weightings of the image observations, introduced for the photogrammetric intersection, impact the assessment of whether the calculated inter-landmark distances significantly change during the production of spoken words. We selected 16 inter-landmark distances to assess jaw movement. We analysed the movements of five children saying 10 words. Overall, four different weightings and two different camera setups were tested. Setup 1 used 2 cameras, and setup 2 used 3 cameras. The weightings based on comparing the BlazeFace landmarks to a reference were too large when applied to setup 1. They did not allow the reliable determination of inter-landmark distance changes as predicted by current literature depending on the camera setup used. Smaller weights were able to be statistically tested for jaw movements correctly. For setup 2, all weights could detect inter-landmark distances reliably.

1. Introduction

Speech sound disorders (SSDs) are a prevalent communication difficulty in young children, affecting speech intelligibility, with potentially lifelong consequences. Clinical assessment is required to establish a diagnosis and facilitate timely access to targeted intervention (Daniel & McLeod, 2017). Related research indicates access to objective measures could support the assessment process (Murray et al., 2021). However, speech-language pathologists currently have limited access to objective measures of speech-motor control (Rebernik et al., 2021). This paper focuses on objective measures of jaw movement and control assessed using the inter-landmark distances of anatomically relevant facial landmarks.

Recent technological advances have contributed to the development of Machine Learning (ML) methods that provide clinicians with access to objective measures. The workflow to attain a set of anatomical facial landmarks can be manual (physical markers placed on the face) (Deli et al., 2010), semi-automatic (digitisation) (Aynechi et al., 2011), or automated (prediction by an algorithm) (Bandini et al., 2017; Berends et al., 2024). Markerless approaches allow for greater practical implementation in clinical practice, and these are the focus of this paper.

To date, several markerless tracking systems for speech have been reported in the literature. For example, Bandini et al. (2017) used a video-based system to quantify jaw movements by measuring Euclidean distances from the nose tip to landmarks on the right and left jawline. The overall accuracy of jaw tracking was approximately 2 mm, which was considered acceptable. However, accuracy decreased during faster movements, highlighting a limitation of the workflow. Mogren et al., (2022) compared movement patterns of the lips and jaw in lateral, vertical and anteroposterior directions, determining their range of motion using the SmartEye Pro tracking system but no information related to the geometric accuracy of the detected motion was reported. The SMAAT (Speech Movement and Acoustic Analysis Tracking, www.smaat.org) assessment tool provides clinically relevant objective measures of jaw and lip movements to assist speech-language pathologists in assessing speech motor control (Palmer et al., 2024). The SMAAT workflow utilises the convolutional neural network (CNN) based BlazeFace algorithm (Bazarevskyet et al., 2019) to detect the 3D coordinates of a 478-point facial mesh from single 2D images of human faces. The BlazeFace algorithm was trained using 2D images together with their associated 3D facial models to learn how to fit the facial mesh onto a face detected within a single 2D image, thus enabling it to estimate the 3D spatial coordinates of each point including depth.

The accuracy of marker placement using these methods can be evaluated against an independent reference. For instance, Aynechi et al. (2011) compared traditional anthropometric measurements (calliper measurement) with the 3dMD stereophotogrammetric camera system. Differences were observed in measurements involving ears and soft tissue landmarks without distinct edges. The accuracy of the 3dMD-derived distances was reported to be < 2mm. Palmer et al. (2020) performed a similar comparison, using 3D facial images captured by both the 3dMD and the VectraH1 stereophotogrammetric camera systems. The measurements were again compared to traditional anthropometric measurements (acquired using calliper and measuring tape) performed by clinicians and non-experts. Most reported differences were in the sub-mm range, however, the results showed that measuring bias can be introduced depending on the method of measurement used. Hence, reported

accuracies, such as in Aynechi et al. (2011) could be argued as a systematic error in performing the measurements but not a clear indicator of the accuracy of the photogrammetric method compared to traditional anthropometric measurements (such as calliper measurements). Another limitation of this analysis is that only still images (single frames) can be analysed.

Many ML-based methods provide the extracted landmarks' coordinates only without their associated uncertainties. Consequently, the inter-landmark distance measurements that are derived from these landmarks, such as might be used in the assessment of jaw movement, can suffer from large error bounds degrading their utility for assessment purposes.

Photogrammetry uses the image observation of facial landmarks from multiple stationary cameras to predict their 3D coordinates and uncertainties through intersection and variance propagation. Consequently, the uncertainties of the image observations can be propagated to the 3D intersected facial landmarks and, hence, to any derived inter-landmark distances. Statistical methods can then be applied to evaluate how significantly the inter-landmark distances change. This approach builds on the advantage of markerless facial landmark detection with photogrammetry and least squares adjustment (LSA) to estimate placement uncertainties.

In this paper, we used a combination of the markerless BlazeFace algorithm and photogrammetry. BlazeFace was used to infer the positions of eight clinically relevant facial landmarks used for tracking jaw movement in video frames and generate the image observations of these landmarks. Five landmarks were located on the jaw (Bandini et al., 2017; Mogren et al., 2022), three landmarks were located on the upper facial region (Palmer et al., 2024). Photogrammetric intersection was used to derive head-centred 3D coordinates of these eight facial landmarks. Next, sixteen clinically relevant inter-landmark distances across ten words were calculated, and their uncertainty propagated from the results of the photogrammetric intersection. The selected inter-landmark distances were not independent and follow similar approaches in other related studies, such as Bandini et al. (2017), Mogren et al. (2022) and Palmer et al. (2024). Chi-square tests at the 5% significance level were used to detect significant changes in the inter-landmark distances per word per participant.

In this paper we show how different weightings applied to the image observations and network designs used for the photogrammetric intersection influence the significance of change in the calculated inter-landmark distances. We first assumed that the landmarks less susceptible to soft tissue deformation and skin tissue artifacts would remain stable. Secondly, based on established literature, we assumed the upper facial region would not be directly associated with the speech tasks (Sarhan et al., 2023), and therefore also likely to remain stable across all conditions (for each word, for each participant and each weighting).

The paper is organised as follows. Background to this work is provided in Section 2. Section 3 outlines the theories and methodologies used. Section 4 reports on the experiments performed. Section 5 details the results, and finally, Section 6 concludes the paper.

2. Background

In surveying, it is well established that any observation contains errors. Systematic errors are usually modelled and corrected for, and gross errors are usually eliminated. Only random errors can be adjusted in the LSA. Random errors are characterised by their probability density function, usually assumed to be Gaussian. Their behaviour is quantified by the mean, assumed to be zero, and the variance σ^2. The covariance matrix $\mathbf{C_l}$ contains the variance σ^2. If all observations have the same precision, σ^2, then $\mathbf{C_l}$is a scalar matrix

$$\mathbf{C_l} = \sigma^2 \cdot \mathbf{I}. \quad (1)$$

A cofactor is related to variance and co-variance by the variance factor σ_0^2 with $q_{ii} = \frac{\sigma_i^2}{\sigma_0^2}$ leading to the co-factor matrix using

$$\mathbf{Q_l} = \frac{1}{\sigma_0^2} \cdot \mathbf{C_l} \quad (2)$$

and consequently the weight matrix **P** which is the inverse of the co-factor matrix with

$$\mathbf{P} = \mathbf{Q_l^{-1}} = \frac{1}{\sigma^2}\mathbf{C_l^{-1}}. \quad (3)$$

The cofactor matrix of the parameters $\mathbf{Q_x}$ is a by-product of the parameter calculations using LSA and is defined as

$$\mathbf{Q_x} = (\mathbf{A^T P A})^{-1}. \quad (4)$$

The challenge is that the variance factor σ_0^2 and the best approximation of the observation variances σ^2, are not known a priori and their magnitudes are often only vaguely understood. Furthermore, the design matrix **A** which depends on the network design also impacts $\mathbf{Q_x}$. In summary, the three possibilities for influencing $\mathbf{Q_x}$ are:

1. The variance factor σ_0^2 can be controlled by the selection of the instruments and the repetition number of multiple observations.
2. The matrix **A** depends on the network's geometry, i.e., the relative position of the images connecting to the object points to the observations.
3. The weight matrix **P** contains the a priori weightings, which are functions of the type of the observables and the relative precision.

The $\mathbf{Q_x}$ matrix impacts statistical testing for detecting significant changes between points over multiple epochs. This is also true for LSA applied in Photogrammetry.

2.1 Variance Factor

The variance factor σ_0^2 is often seen only as a suitable scalar value for the weight matrix and is a global measure of the variance of image coordinate measurements. In photogrammetry, the variance of the image observation of manual point picking is well established and can be seen as prior knowledge, such as applied in (Fraser, 1984). Other methods are manual assessment of the variance of image coordinate measurements (e.g. used in Barone et al., 2020) or knowledge of the precision of the method used, such as image matching (e.g. used in Fraser, 2000).

The challenge is that the BlazeFace (Bazarevskyet et al., 2019) algorithm which is used to extract the image observations in our research is deterministic. This means that the same input will always result in the same output. Hence, selecting the variance factor by repeating multiple observations (i.e., reusing the same input) is inappropriate.

2.2 Network Design

Several papers have been published analysing the network design of close-range photogrammetry approaches, including Fraser (1984). The network design aspect of our research was

investigated by Boyle et al., 2024. For this publication, two network designs have been applied. The first design was investigated by Boyle et al., 2024, and the second design is a slight adaptation of the first to account for constraints during data capture. Details are provided in section 3.

2.3 Weight matrix

In their work on the network design considerations for non-topographic photogrammetry, Fraser (1984) states that the solution to the weight problem (estimation of the observation variances) is not straightforward. In their paper, equal weighting in the form of a scalar matrix is used. In subsequent papers, the variance of image observation is based on the precisions achievable through image-matching approaches of 0.3 – 0.5 pixels (Fraser, 2000). In other work, the variance is estimated by manually assessing the confidence region (Barone et al., 2020).

2.4 Conclusion

None of the above mentioned methods can be used when using an ML-based approach to extract clinically significant landmarks in single images due to the lack of uncertainty measures provided by the ML-based method. Instead, studies focus on the accuracy of the 3D facial landmarks compared to a reference, such as images captured with the Vectra H1 camera (Palmer et al., 2020). In contrast to previous studies, we use this information not for the accuracy assessment of a single frame image in object space but to project the 3D Vectra coordinates of facial landmarks back into the image plane. It enables us to derive variances of the image space observations introduced as uncertainty measurements for the 2D BlazeFace facial landmarks into a photogrammetric intersection. A range of these predicted uncertainty measures are used to assess their impact on the uncertainties of inter-landmark distances during speech production to assist with the analysis of speech sound disorders.

3. Methodology

Using the BlazeFace algorithm, observations of facial landmarks were detected from 2D image frames of video. The frames were extracted from time-synchronised cameras with known interior orientation parameters (IOPs) and exterior orientation parameters (EOPs). Through photogrammetric intersection, 3D coordinates with uncertainties were derived. Finally, inter-landmark distances were derived, propagating the errors from the photogrammetric intersection to these distances. Different weightings of the 2D image observations were applied and their impact on the uncertainties of the 3D facial landmark coordinates and the derived inter-landmark distance uncertainties were analysed. Statistical testing of distances of the first frame compared to all following frames was performed to determine which inter-landmark distances do not significantly change and hence can be labelled as "stable". The impact of the weightings contributing to the decision of stable/significant changing inter-landmark distances for speech production was then analysed.

3.1 Camera calibration and extraction of image observations for each camera frame

Each camera j was subjected to an individual calibration adjustment using a general photogrammetry process applying the Brown camera model (Brown, 1971) solving for principal distance (c) and the principal point offset (x_p and y_p) and distortions (Δx, Δy). The calibration was performed using a hand-held 3D frame that was moved through three arcs, held approximately square, tilted at ±30º, and rotated 90º to the left and right to ensure the best possible geometry could be implemented.

Using BlazeFace, the facial landmarks' locations (x_i, y_i) were detected within each image frame. In this paper, BlazeFace was operated in single face-tracking mode, and a 15-frame (quarter-second) buffer was added to the onset time to allow BlazeFace to settle on the detected face. This buffer was removed for further processing. The BlazeFace landmark coordinates were standardised to the width of the input image frame. Hence, they could be easily converted into an image coordinate system with its origin in the center of the image. The IOP corrections were applied before the image coordinates (x_{ij}, y_{ij}) are further processed.

The EOPs of a camera (j), which include the perspective centre location (X^C_j, Y^C_j, Z^C_j) and rotation (ω_j, φ_j, κ_j) were calculated using a resection of pre-established Ground Control Points (GCPs) (Boyle et al., 2014).

3.2 Photogrammetric intersection

The intersection process calculates the 3D coordinates (X_i, Y_i, Z_i) for each facial landmark i from the 2D image (x_i, y_i) observations obtained using BlazeFace; this can be done by applying the collinearity equations, assuming that the calculated IOPs and EOPs of each camera j are known:

$$fx_{ij} = x_{ij} + \hat{v}_{x_{ij}} = x_{p_j} - c_j \frac{\mathrm{U_{ij}}}{\mathrm{W_{ij}}} + \Delta x_{ij}, \quad (5)$$

$$fy_{ij} = y_{ij} + \hat{v}_{yx_{ij}} = y_{p_j} - c_j \frac{\mathrm{V_{ij}}}{\mathrm{W_{ij}}} + \Delta y_{ij} \quad (6)$$

where $\hat{v}_x$, $\hat{v}_y$ are the image point residuals. Furthermore, (U,V,W) was formulated as

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix} = \mathbf{M}\left(\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_i - \begin{pmatrix} X_c \\ Y_C \\ Z_C \end{pmatrix}_j^C\right) \quad (7)$$

where $\mathbf{M}$ is the matrix used for the rotation from object space to image space that can be parameterised as

$$\mathbf{M} = \mathbf{R}_3(\kappa_j)\mathbf{R}_2(\varphi_j)\mathbf{R}_1(\omega_j). \quad (8)$$

The parametric (or Gauss–Markov) adjustment model was used for the intersection. The linearised collinearity equations for all image points in a block can be grouped into the standard parametric model

$$\mathbf{v} = \mathbf{A} \cdot \Delta\mathbf{x} - \mathbf{l}. \quad (9)$$

With $\mathbf{A}$ being the design matrix of partial derivatives taken with respect to the object space point coordinates, $\mathbf{l}$ being the vector of image point observations, $\Delta\mathbf{x}$ being the vector of corrections for the unknowns $\mathbf{x}$ (3D object space coordinates) and $\mathbf{v}$ being the residuals. $\mathbf{A}$ is formulated as

$$\mathbf{A} = \begin{bmatrix} \frac{\partial f_{x_{ij}}}{\partial X_i} & \frac{\partial f_{x_{ij}}}{\partial Y_i} & \frac{\partial f_{x_{ij}}}{\partial Z_i} \\ \frac{\partial f_{y_{ij}}}{\partial X_i} & \frac{\partial f_{y_{ij}}}{\partial Y_i} & \frac{\partial f_{y_{ij}}}{\partial Z_i} \end{bmatrix}. \quad (10)$$

The associated weight matrix for this group of observations was denoted as $\mathbf{P}$. The parameter correction δ and the corresponding covariance matrix $\mathbf{C_x}$ were obtained as

$$\delta = -(\mathbf{A^T P A})^{-1}\mathbf{A^T P} w = -\mathbf{Q}_x\mathbf{A^T P} w \quad (11)$$

$$\mathbf{C_x} = \sigma_0^2\, \mathbf{Q_x} \quad (12)$$

where $\mathbf{Q_x}$ is the cofactor matrix of the parameters and w is the misclose.

3.3 Weighting of image observations

Equation 9 shows the impact of **P** on $\mathbf{Q_x}$ if **A** is fixed because the same network design is used. The most practical solution of assuming **P** in the form of $\sigma^2\mathbf{I}$ is only used in the initial solution to find an approximation for **x.** Otherwise, it cannot be used, as BlazeFace does not extract all image points with the same accuracy.

The image coordinates of facial landmarks of a single frame showing a neutral face using the BlazeFace method were compared to a reference to find an appropriate weight. The reference was based on measurements captured with the Vectra H2 camera, and these images were also captured of a neutral face. The Vectra H2 object space coordinates of the facial landmarks were transformed into the same coordinate system as **x** using a 3D rigid body with the scale factor fixed to 1. Using the collinearity equations, the image coordinates of all points and for all camera stations of the Vectra H2 camera x_{vectra} were calculated. Consequently, it was possible to calculate the difference in image coordinates from both solutions for all points using

$$\Delta x_{Vectra_{ij}} = x_{iVectra} - x_{ij},$$
$$\Delta y_{Vectra_{ij}} = y_{iVectra} - y_{ij}. \quad (13)$$

This method was applied to all participants, all landmarks, and all cameras, allowing the variance of the observations to be estimated from the set of image coordinate differences.

Based on the derived variance, the covariance matrix $\mathbf{C_l}$ and the variance factor σ_0^2 can be calculated with **P** using

$$\mathbf{C_l} = \sigma_0^2\,\mathbf{P^{-1}} \quad (14)$$

allowing us to determine the $\mathbf{Q_x}$ needed for the statistical testing of inter-landmark distances between frames. The same **P** will be applied to all processed frames, including those showing a non-neutral face, under the assumption that the variance does not change between frames.

3.4 Inter-landmark distances and statistical testing

Inter-landmark distances D_{kl} were calculated from the object space coordinates between the landmarks k and l as:

$$D_{kl} = \sqrt{\Delta X_{kl}^2 + \Delta Y_{kl}^2 + \Delta Z_{kl}^2}\,. \quad (15)$$

Its standard deviations was calculated through variance propagation using:

$$\mathbf{A} = \left[\frac{\partial D_{kl}}{\partial X_k}\ \frac{\partial D_{kl}}{\partial Y_k}\ \frac{\partial D_{kl}}{\partial Z_k}\ \frac{\partial D_{kl}}{\partial X_l}\ \frac{\partial D_{kl}}{\partial Y_l}\ \frac{\partial D_{kl}}{\partial Z_l}\right], \quad (16)$$

$$\sigma_{Dkl}^2 = \left(\mathbf{A}\begin{bmatrix} q_{xx_k} & q_{xy_k} & q_{xz_k} & 0 & 0 & 0 \\ & q_{yy_k} & q_{yz_k} & 0 & 0 & 0 \\ \vdots & & q_{zz_k} & 0 & 0 & 0 \\ & & & q_{xx_l} & q_{xy_l} & q_{xz_l} \\ & & & & q_{yy_l} & q_{yz_l} \\ & & & \cdots & & q_{zz_l} \end{bmatrix}\mathbf{A^T}\right) \quad (17)$$

The inter-landmark distances D_{kl} and their standard deviations were calculated per frame F_i over the duration of a word. The first frame, F_1 for each word, showed the participant with a neutral expression and was used as the reference frame. To test for movement during speech production, the changes in the distances between the same two points k and l between frames and the first frame were calculated using

$$\Delta D_{klF_i} = D_{kl_{F_1}} - D_{kl_{Fi}}\,, \quad (18)$$

and its variance with

$$\sigma_{Dkl}^2 = \sigma_{DklF1}^2 + \sigma_{DklFi}^2. \quad (19)$$

Per distance differences from the reference frame to all following frames, a statistical test with a degree of freedom of m = 1, was calculated and compared against a chi-square upper-tail statistical threshold with a significance level of 5%.

$$\Pr\{C < \chi_{m,1-\alpha}^2\} = 1 - \alpha \quad (20)$$

$$\text{where } C = \Delta D_{kl_{F_i}}^{T} \cdot \sigma_{Dkl}^{-2} \cdot \frac{\Delta D_{klF_i}}{m} \quad (21)$$

If the test statistic $\chi_{m,1-\alpha}^2$ exceeds the threshold, the distance was assumed unstable, meaning it had changed significantly during word production. In this case, the distance was allocated the value of $f = 1$.

$$f = \begin{cases} 1 & if\ C > \chi_{m,1-\alpha}^2 \\ 0 & if C \le \chi_{m,1-\alpha}^2 \end{cases} \quad (22)$$

The sum of f was calculated upon the statistical testing of distances for a particular word. This sum must be zero for inter-landmark distances that did not change for the duration of a word.

4. Experiment

4.1 Camera specifications

The data capture system utilised two to three stationary Blackmagic (BM) Pocket Cinema 4K cameras, each fitted with an Olympus Digital 45mm (f1.8) lens. The camera's full 4/3" sensor resulted in a narrow field-of-view setup. This configuration ensures that the head occupies nearly the full view from a secure distance. Secure so that no participant can manipulate the camera. The BM cameras are capable of recording 4096×2160 resolution at 60 frames per second (FPS); however, for data capture, the resolution is set to 1920×1080 (HD) and 60 FPS, resulting in an actual principal distance of approximately 120 mm. The highest resolution is unnecessary as the BlazeFace algorithm down-samples images to 256×256 pixels internally before processing.

4.2 Camera setup

The data capture setup consisted of three tripod-mounted BM cameras placed approximately 3.0m from the participant. This distance varied depending on the physical limitations of the environment during the data capture for each participant. The primary camera was placed in front of the participant, creating an artificial centreline. For the remainder of the paper, this camera is named BMC. The other cameras were positioned on either side of BMC to capture the face's left and right sides (BML and BMR respectively). Two different camera setups were used. The cameras were located either approximately 30º or 45º to each side, achieving a horizontal convergence angle of approximately 60º or 90º (Figure **1**). The elevation angles were dictated by the specific participant's seated height position but were approximately 5º above the horizontal plane.

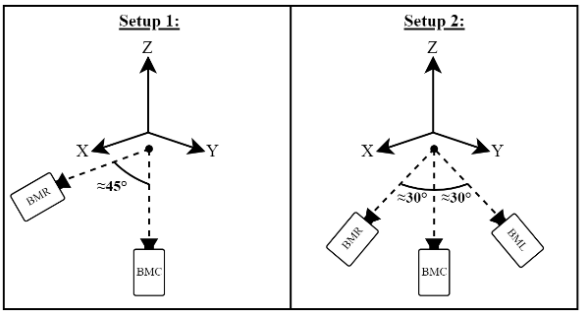

Figure 1: Visualisation of the camera setup 1 and setup 2.

4.3 Data capture

For this work, data from five children with a mean age of 3 years 5 months (±4 months) were used. Each child was seated on a chair, as shown in Figure 2, with a target field in the background used for resecting the cameras. The children were asked to name pictures from the Motor Speech Hierarchy Probe-Word list (Namasivayam et al., 2021), with the words cast onto a screen directly in the child's line of sight. The ten words associated with stage III mandibular control of the Motor Speech Hierarchy Probe-Word list inform the dataset of this paper (Table 1).

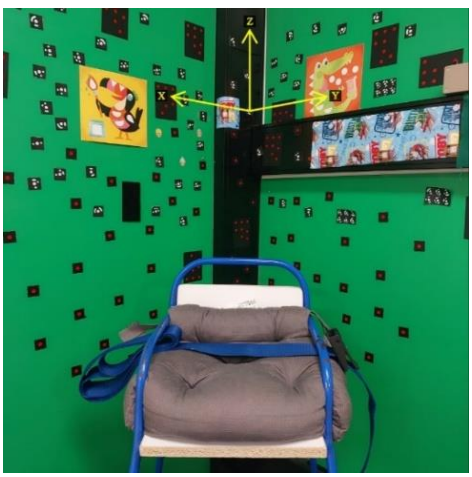

Figure **2**: Target field for the resection of the cameras.

	1	**2**	**3**	**4**	**5**
Baa	58	48	42	44	39
Bob	66	38	36	29	34
Eye	50	60	41	43	37
Ham	55	45	24	30	30
Map	46	52	43	27	30
Pam	68	68	38	42	51
Papa	60	48	39	39	45
Pie	70	65	34	45	42
Pup	42	42	31	33	34
Um	47	53	23	34	24
Sum	562	519	351	366	366

Table 1: Frames per word and participants (row 1) processed.

Several pre-processing steps were performed. These included (a) time-synchronising recordings from different stationary video cameras, (b) determining the onset and offset time points for each word, and (c) extracting the frames from the video sequences for each word. Table 1 details the number of processed frames per word. In total, 2164 frames were processed.

4.4 Landmark locations and inter-landmark distances

BlazeFace Mesh ID points were matched with clinically relevant landmarks (BlazeFaceMeshID, 2025). Two additional landmarks (GNR and GNL) were added by using BlazeFace points to best estimate their positions. An overview of the landmarks is provided in Table 2 and visualised in Figure 3.

BlazeFace Mesh IDs	**Clinical Landmarks**	**Abb.**
10	Metopion	M
9	Glabella	G
168	Sellion	S
199	Pogonion	P
150	Mid-Mandibular Border (right)	MMBR
379	Mid-Mandibular Border (left)	MMBL
176	Medial Gnathion (right)	GNR
400	Medial Gnathion (left)	GNL

Table 2: Clinically relevant distances used in this research.

The inter-landmark distances are listed in Table 3. Based on our previous assumptions regarding skin tissue deformation and limited association with movement during speech tasks, we expected seven of 16 distances to remain stable (Trotman and Faraway, 1998), whilst five would change due to the vertical movement associated with the speech task. Four inter-landmark distances were expected to vary due to changes associated with the development of motor speech control in young children (Green et al., 2000) and are highlighted with (*) in Table 3.

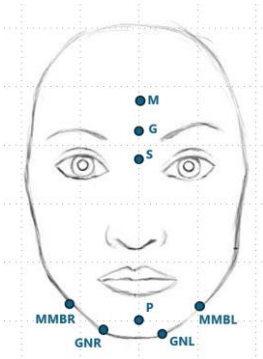

Figure 3: Placement of the clinically relevant landmarks.

#	**Clinical Landmarks**		**Distance d_{kl}**	**Stable**
1	M	G	M.G	Y
2	M	S	M.S	Y
3	G	S	G.S	Y
4	MMBR	MMBL	MMBR.L	Y
5	GNR	GNL	GNR.L	Y
6	P	GNR	P.GNR	Y
7	P	GNL	P.GNL	Y
8	M	MMBR	M.MMBR	N*
9	M	GNR	M.GNR	N*
10	M	MMBL	M.MMBL	N*
11	M	GNL	M.GNL	N*
12	S	MMBR	S.MMBR	N
13	S	GNR	S.GNR	N
14	S	MMBL	S.MMBL	N
15	S	GNL	S.GNL	N
16	S	P	S.P	N

Table 3: Inter-landmark distances and if they are assumed to be stable (Y) or not (N). Distances indicated with (*) can display a small amount of movement in children.

5. Results

5.1 IOPs and EOPs

The IOPs associated with principal distance (c) and the principal point offset (x_p and y_p) were successfully estimated. Due to each camera's very narrow field of view (principal distance of approximately 120 mm), the lens distortion (Δx, Δy) at the extremities of the image were negligible (Boyle et al., 2014).

The self-calibration of each camera was successful. The number of images and GCPs varied between the datasets. The estimated 3D accuracy RMS values of each calibration are presented in Table 4.

The EOPs were calculated using the resection method with one image per camera position. The GCPs used for the resection varied based on their visibility (the participants occluded some). The RMS values of the resection are presented in Table 5.

Participant	BMR	BMC	BML
1	0.2951	0.0833	-
2	0.0507	0.1046	-
3	0.0862	0.0813	-
4	0.1160	0.1279	-
5	0.1579	0.1461	0.1256

Table 4: RMS of object point coordinates (mm) of self-calibration.

Participant	BMR	BMC	BML
1	0.41	0.72	-
2	0.39	0.83	-
3	1.56	1.21	-
4	0.72	0.79	-
5	0.68	0.53	0.84

Table 5: RMS of image coordinate residuals (pixels) for resection.

5.2 Weighting of image observations

The distances between the image coordinate from BlazeFace and the reprojected image coordinates of the Vectra H2 camera were calculated using equation 13. All differences (for all participants, words, and x and y coordinates) are presented in Figure 4; statistics are presented in Table 6. The mean has a small negative bias, and the standard deviation is 0.141 mm, which is equivalent to approximately 11.72 pixels.

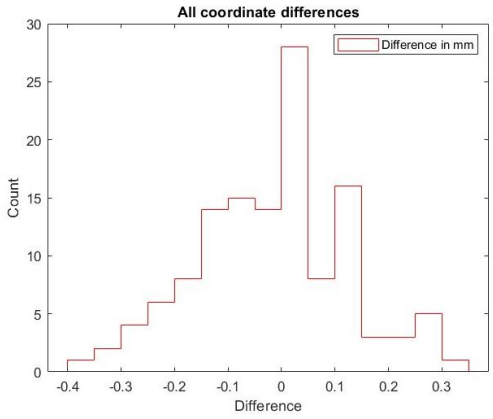

Figure 4: Image coordinate differences between BlazeFace and Vectra H2.

A similar small negative bias in the individual means of the x and y coordinates is seen in Figure 5. The standard deviation for the x image coordinate is 0.096mm (approximately 7.96 pixels), about half that of the y image coordinate's standard deviation of 0.176mm (approximately 14.59 pixels).

Differences	Mean [mm]	σ [mm]	σ [pixel]
All (x and y)	-0.013	0.141	11.72
x coordinates	-0.006	0.096	7.96
y coordinates	-0.021	0.176	14.59

Table 6: Mean values and standard deviations of the differences between BlazeFace and Vectra in image space.

Overall, the standard deviations are relatively large compared to the photogrammetric value of around 0.4 pixels, which is usually applied. There are several reasons for this. The placement of some reference landmarks in the Vectra images is very difficult using images only i.e., landmark localisation cannot be inferred through palpation of the bony extrusions under softer tissues. In addition, some landmarks are intrinsically less precisely defined. For instance, Farkas and Schendel (1995) define the placement of Metopion as "the most anterior (or most convex) midline point on the frontal bone. If the forehead region is relatively flat, place this landmark vertically at the midpoint between the superior facial border and glabella." (p. 112).

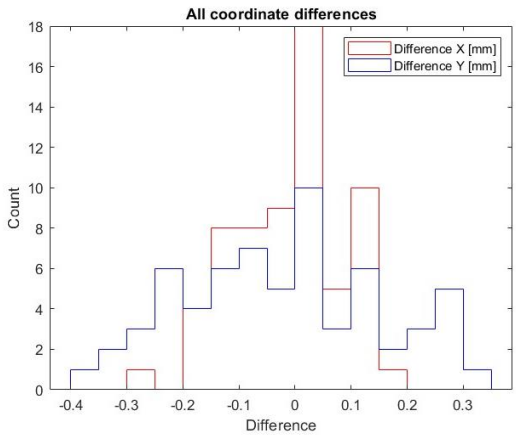

Figure 5: Image coordinate differences of x and y observations between BlazeFace and VectraH2.

We processed the data using a standard deviation of 3 and 5 pixels. It is assumed that a standard deviation of 3 pixels can be achieved for a clinician for well-defined landmarks in video sequences and 5 pixels for trained non-experts. A summary of all weights processed for the statistical testing of the inter-landmark distances is presented in Table 7.

	σ_x	σ_y
W1	0.141 mm / 11.72 pixels	
W2	0.096 mm / 7.96 pixels	0.176 mm / 14.59 pixels
W3	5 pixels (0.060 mm)	
W4	3 pixels (0.036 mm)	

Table 7: Weighting to be used for the statistical testing of inter-landmark distances.

5.3 Statistical tests of inter-landmark distances

As shown in Figure **1**, two camera setups were used. As the design of the network impacts **A** and consequently $\mathbf{Q}_{\boldsymbol{x}}$, the results for the two different camera setups are presented separately.

The results for Participants 1-4 (setup 1) are presented in Table 8, and the results for Participant 5 (setup 2) are in Table 9. The

percentages indicate the number of frames where the calculated inter-landmark distances changed significantly from the first frame denoted by an *f*-value of 1 (equation 22). A value of "0" indicates that none of the inter-landmark distances changed from the first frame. The highlighted cells indicate where the results do not fit the state-of-the-art prediction.

The results outlined in Table 8 and Table 9 show a general agreement that all weights confirm our expectation based on the current literature of stable inter-landmark distances. Of the seven stable distances (rows 1-7) only the distance 'MMBR.L' in weight 4 showed a significant change in 1.1% of all frames for setup 1 (Table 8). In Table 8, the five unstable distances also indicated four incorrect predictions (rows 8-16): S.GNL, S.MMBR, and S.MMBL using weighting 1 and S.MMBL for weighting 2. This trend is not visible in the second setup Table 9). While the percentage of those landmarks is generally lower, all weightings agree with the expert assessment.

Notably, the percentage of significant distance changes linked to GNR and GNL is always larger than those of MMBR and MMBL. This is true for both setups. This is expected due to the positioning of the respective landmarks such that the GNR and GNL experience more relative movement than the MMBR and MMBL due to the pivot point of the jaw and placement along the jaw and chin.

Distance D_{kl}	Expert	W1	W2	W3	W4
M.G	0	0	0	0	0
M.S	0	0	0	0	0
G.S	0	0	0	0	0
MMBR.L	0	0	0	0	1.1
GNR.L	0	0	0	0	0
P.GNR	0	0	0	0	0
P.GNL	0	0	0	0	0
M.MMBR	Change*	6.0	14.5	32.6	48.0
M.GNR	Change*	12.1	21.3	40.8	56.5
M.MMBL	Change*	2.5	7.3	24.6	43.7
M.GNL	Change*	7.8	16.0	36.2	52.0
S.MMBR	Change	0	1.6	18.2	35.8
S.GNR	Change	1.3	6.5	29.1	45.2
S.MMBL	Change	0	0	8.5	29.4
S.GNL	Change	0	2.8	22.8	39.8
S.P	Change	0.7	5.0	29.5	43.5

Table 8: Percentage of frames of the statistical differences of inter-landmark differences for network setup 1 and weights W1 to W4. "0" stands for stable, so no statistical differences are detected. All highlighted cells contain results that disagree with the expert opinion.

Next, it can be observed for setup 1 (Table 8) that distances linked to the right facial landmark have a higher percentage of change than to the left facial landmarks. For instance, the percentage of changed distances for all weighting for M.GNR is approximately 5% larger than M.GNL. This could be due to the camera setup used. This camera setup contains the front and right cameras (BMC and BMR), which together mostly capture the right part of the face. Hence, the predicted coordinates for the right face side could be better, and more significant movement could be detected. The results of camera setup 2 (Table 9) can confirm this assumption. Including a third camera also capturing the left part of the face could remove the bias of larger percentage changes.

There is an overall trend that the percentage of changed distances increases from weight 1 to 4 for both setups. For instance, for setup 1 (Table 8) the percentages increase for M.GNR from 12.1% to 21.3%, to 40.8% and to 56.5%. And the percentages increase for M.GNL from 7.8%, to 16.0%, 36.2%, and 52.0%. The reason is that the standard deviations used become more strict from weight 1 ($\sigma_x = \sigma_y$ =11.72 pixels), to weight 2 (σ_x = 7.96 pixels, σ_y = 14.59 pixels), to weight 3 ($\sigma_x = \sigma_y$ = 5 pixels) and to weight 4 ($\sigma_x = \sigma_y$ =3 pixels). Consequently, a smaller standard deviation will lead to the detection of a greater number of significant distance changes. For instance, focusing on the percentage changes for S.P in Table 8, the percentage of changed distances for weighting 1 is only 0.7 % and for weighting 2 only 5.0%. It increases to 29.5% and 43.5% for weighting 3 and 4. Regarding weighting 2, 5% of changes correspond to 87 frames from 1758 frames overall. This is equivalent to an average of approximately 2 frames per person per word, and a word cannot be said in 2 frames. Hence, we must assume that there are no significant inter-landmark changes, which is conflicting with the expert assumption. When applying a 5% threshold to all percentage changes in Table 8, most results for weighting 1 and 2 disagree with the expert assumption.

For setup 2 (Table 9), the lowest values are 7.4% and 10.7% associated with weighting 1 for S.MMBR and S.MMBL, respectively, exceeding the threshold of 5%. Introducing the third camera meant that all tested weights could reliably detect inter-landmark distance.

Distance D_{kl}	Expert	W1	W2	W3	W4
M.G	0	0	0	0	0
M.S	0	0	0	0	0
G.S	0	0	0	0	0
MMBR.L	0	0	0	0	0
GNR.L	0	0	0	0	0
P.GNR	0	0	0	0	0
P.GNL	0	0	0	0	0
M.MMBR	Change*	40.8	48.5	81.3	90.2
M.GNR	Change*	47.9	56.4	86.5	93.6
M.MMBL	Change*	37.4	48.2	81.0	90.5
M.GNL	Change*	48.8	56.4	86.2	93.6
S.MMBR	Change	7.4	21.8	60.7	82.2
S.GNR	Change	23.9	38.3	69.0	89.0
S.MMBL	Change	10.7	22.1	63.5	82.8
S.GNL	Change	22.1	35.6	70.2	87.4
S.P	Change	19.9	36.5	66.6	87.1

Table 9: Percentage of frames of the statistical differences of inter-landmark differences for network setup 2 and weights W1 to W4. "0" stands for stable, so no statistical differences are detected.

6. Conclusion

This paper aimed to investigate the effects of image coordinate weighting and network design when statistically testing 16 facial inter-landmark distances associated with jaw movement and control. A total of ten individual words were tested across five children, all processed by the BlazeFace algorithm to extract 2D image coordinates. The 3D coordinates were determined by photogrammetry intersection and the inter-landmark distances, and their variances were consequently calculated using variance propagation. Chi-square tests at a significance level of 5% were used to detect significant changes in the inter-landmark distances per word per participant.

Overall, four different weightings and two camera setups were tested. Camera setup 1 included 2 cameras and setup 2 included 3. Weightings 1 and 2 were based on comparing the 3D object space coordinates derived from BlazeFace to a reference created using the stereo-photogrammetric camera Vectra H2. Both weightings were too pessimistic for camera setup 1 and did not allow for the reliable determination of inter-landmark distance changes. In contrast, for setup 2 where a third camera was introduced, all inter-landmark distance changes were reliably detected irrespective of the weightings used. In contrast to this,

the two camera setup required a more precise setting of the weights to detect the distance changes.

Weightings 3 and 4 were based on an expert's ability to manually place landmarks in the images. Both are suitable for the detection of inter-landmark distances. The strictest weight (weight 4) created the most promising values for both camera setups.

For further analysis of jaw movement, it is important to examine not only how an inter-landmark distance changes but also when these changes occur. Future research will analyse this temporal aspect rather than solely determining the percentage of frames that have changed. Further, to enable meaningful comparison between participants, time must be normalised. This comparison will further quantify whether weight 3 or weight 4 is more appropriate.

Acknowledgements

Research reported in this publication was supported by the Medical Research Future Fund under grant number 2016518, and completed as part of a doctoral research program. It was also supported by the Canadian Mitacs Globalink Research Award - for research in Canada (IT38799).

The study was approved by the Human Research Ethics Committee of Curtin University (protocol code 2020-0327; 15 February 2024) and conducted in accordance with the National Statement on Ethical Conduct in Human Research. Informed consent was obtained from all participants. We would like to express our sincere gratitude to all the participants who volunteered their time for this research.

References

Aynechi, N., Larson, B. E., Leon-Salazar, V., & Beiraghi, S. ,2011. Accuracy and precision of a 3D anthropometric facial analysis with and without landmark labeling before image acquisition. *The Angle Orthodontist*, *81*(2), 245–252.

Bandini, A., Namasivayam, A., Yunusova, Y., 2017. Video-Based Tracking of Jaw Movements During Speech: Preliminary Results and Future Directions. *Interspeech 2017*, 689-693.

Barone, B., Marrazzo, M., Oton, C.J., 2020. Camera Calibration with Weighted Direct Linear Transformation and Anisotropic Uncertainties of Image Control Points. *Sensors* 2020, 20(4), 1175; https://doi.org/10.3390/s20041175.

Bazarevskyet, V., Kartynnik, Y., Vakunov, A., Raveendran, K., Grundmann, M., 2019. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. *Computer Vision and Pattern Recognition*. CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Long Beach, CA, USA, 2019

BlazeFaceMeshID, 2025. Last access Feb 2025, See https://storage.googleapis.com/mediapipe-assets/documentation/mediapipe_face_landmark_fullsize.png

Boyle, L., Helmholz, P., Lichti, D.D., Ward, R., 2024. Validation of Camera Networks Used for the Assessment of Speech Movements. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVIII-2-2024, 41–48, https://doi.org/10.5194/isprs-archives-XLVIII-2-2024-41-2024

Berends, B., Bielevelt, F., Schreurs, R., Vinayahalingam, S., Maal, T., de Jong, G., 2024. Fully automated landmarking and facial segmentation on 3D photographs. *Scientific Reports*, 14. https://doi.org/10.1038/s41598-024-56956-9

Brown., 1971. Close range camera calibration. *Photogramm Eng*, 37(8), 855–866.

Daniel, G. R., & McLeod, S., 2017. Children with speech sound disorders at school: Challenges for children, parents and teachers. *Australian Journal of Teacher Education (Online)*, 42(2), 81-101. https://doi.org/10.14221/ajte.2017v42n2.6

Deli, R., Di Gioia, E., Galantucci, L., Percoco, G., 2010. Automated Landmark Extraction for Orthodontic Measurement of Faces Using the 3-Camera Photogrammetry Methodology. *The Journal of craniofacial surgery*. 21. 87-93.

Farkas, L.G., Schendel, S.A., 1995. Anthropometry of the Head and Face. *American Journal of Orthodontics and Dentofacial Orthopedics* 547, 1995, 107, 112–112.

Fraser, C., 1984: Network design considerations for non-topographic photogrammetry. *Photogrammetric Engineering and Remote Sensing*, Vol 50. No 8, August 1984, pp 1115-1126.

Fraser, C., 2000: High-resolution Satellite Imagery: A review of metric aspects. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* Vol. XXXIII, Part B7. Amsterdam 2000.

Green, J. R., Moore, C. A., Higashikawa, M., & Steeve, R. W., 2000. The physiologic development of speech motor control: Lip and jaw coordination. *Journal of Speech, Language, and Hearing Research*, 43(1), 239-255.

Mogren, Å., McAllister, A., & Sjögreen, L., 2022. 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*, 47(4), 219–229.

Murray, E., Iuzzini-Seigel, J., Maas, E., Terband, H., & Ballard, K. J., 2021. Differential diagnosis of childhood apraxia of speech compared to other speech sound disorders: A systematic review. *American Journal of SpeechLanguage Pathology*, 30(1), 279-300. https://doi.org/10.1044/2020_AJSLP-20-00063

Namasivayam, A.K.; Huynh, A.; Bali, R.; Granata, F.; Law, V.; Rampersaud, D.; Hard, J.; Ward, R.; Helms-Park, R.; van Lieshout, P., 2021. Development and Validation of a Probe Word List to Assess Speech Motor Skills in Children. *Am. J. Speech-Lang. Pathol*. 2021, 30, 622–648.

Palmer, R.L., Helmholz, P., Baynam, G., 2020: Cliniface: phenotypic visualisation and analysis using non-rigid registration of 3d facial images, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLIII-B2-2020, 301–308.

Palmer, R., Ward, R., Helmholz, P., Strauss, G.R., Davey, P., Hennessey, N., Orton, L., Namasivayam, A., 2024. Facial Movements Extracted from Video for the Kinematic Classification of Speech. *Sensors 2024*. 24(22), 7235

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M., 2021. A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, 12(1), 1-24. https://doi.org/10.5334/labphon.237

Sarhan, F. R., Olivetto, M., Ben Mansour, K., Neiva, C., Colin, E., Choteau, B., ... & Dakpé, S., 2023. Quantified analysis of facial movement: A reference for clinical applications. *Clinical Anatomy*, *36*(3), 492-502.

Trotman, C. A., & Faraway, J. J., 1998. Sensitivity of a method for the analysis of facial mobility. II. Interlandmark separation. *The Cleft palate-craniofacial Journal*, 35(2), 142-153.