CymruFluency - A fusion technique and a 4D Welsh dataset for Welsh fluency analysis

Arvinder Pal Singh Bali[1], Gary K.L. Tam[1], Avishek Siris[1], Gareth Andrews[1], Yukun Lai[2], Bernie Tiddeman[3], and Gwenno Ffrancon[4]

[1] Department of Computer Science, Swansea University, UK
[2] School of Computer Science and Informatics, Cardiff University, UK
[3] Department of Computer Science, Aberystwyth University, UK
[4] Academi Hywel Teifi, Swansea University, UK

Abstract. Welsh is a linguistically rich yet under-resourced minority language. Despite its cultural significance, automated fluency assessment remains largely unexplored due to limited datasets and tools. Existing models focus on high-resource languages, leaving Welsh without sufficient multi-modal resources. To address this, we introduce CymruFluency, the first 4D dataset for Welsh fluency assessment, capturing both audio and 3D lip movements with expert-annotated fluency scores. Building on this, we propose a multi-modal fluency classification framework that combines audio features (mel spectrograms) and manually annotated 3D lip landmarks. Our fusion approach significantly improves fluency prediction over unimodal models, emphasizing the critical role of 3D lip dynamics in Welsh learning. This research advances minority language processing by integrating articulatory features into fluency evaluation, offering a powerful tool for Welsh language learning, assessment, and preservation. Project page: `https://github.com/arvinsingh/CymruFluency`

1 Introduction

Welsh is a culturally significant language spoken in Wales, UK, with historical and linguistic ties to other Celtic languages, such as Scottish Gaelic and Irish. Despite its rich heritage and approximately 851,700 speakers [35], Welsh is considered a challenging language for new learners due to its unique phonetic structure, consonant mutations, and complex pronunciation rules. These linguistic features, while integral to the language's identity, create barriers for learners, making effective educational resources and assessment tools essential. Moreover, applications such as AI-driven language tutoring, speech therapy, and automatic fluency evaluation could play a crucial role in preserving and revitalizing Welsh by providing accessible learning support and standardized assessment methods.

However, Welsh faces a significant challenge due to the limited availability of resources, including data, speech, and fluency assessment tools. Unlike high-resource languages such as English, French, and German, which benefit from extensive datasets and commercial support, spoken Welsh is underrepresented in both technology and research. Existing speech datasets—such as the Avalinguo Audio Dataset [3], MPS Dataset [14], Speechocean [37], PSCPSF [20],

Fluent speakers Non-fluent speakers

Fig. 1: Mouth and lip shapes of fluent and non-fluent Welsh speakers while pronouncing 'Gwybodaeth angenrheidiol'.

UCLASS [17], and LibriStutter [18]—primarily focus on widely spoken languages, leaving Welsh without comparable resources. Additionally, disfluency corpora like FluencyBank [6], which compiles various disfluency datasets primarily in English, further emphasize this imbalance. Visual datasets like LRW (Lip Reading in the Wild) [11] incorporate lip movements but lack synchronized fluency annotations. Although recent initiatives in the Welsh Tech Action Plan [36], including efforts on voicebank, text translation, and name recognition, are emerging, commercial speech recognition systems like Apple's Siri still do not support Welsh. The Common Voice Welsh dataset offers open-source audio samples [5], but it does not include visual data or detailed fluency labels. The recent MuAViC [4] and MultiTalk [32] datasets represent significant advancements in multilingual speech recognition and 3D talking head generation. MuAViC provides a robust audio-visual corpus across nine languages, while MultiTalk offers a comprehensive multilingual video dataset designed to enhance 3D talking head models. To our knowledge, there are no visual datasets that include Welsh. This gap in visual and audio resources further limits the ability of educators and speech therapists, who often have to rely on informal or English-based assessment tools [25, 10].

Despite Welsh's linguistic importance, research on automated Welsh fluency assessment remains scarce. Speech fluency evaluation is an essential component of language learning, aiding educators and learners in assessing proficiency and identifying areas for improvement. Recent advancements in fluency assessment for high-resource languages utilize techniques such as Support Vector Machines, Random Forest, Multilayer Perceptron [7], Gaussian Mixture Models, and Convolutional Neural Networks [28] to analyze spectral features (e.g., MFCCs, jitter), while 3D-CNNs have been used to extract spatiotemporal lip movement patterns [13]. While recent studies explore multi-modal fusion, e.g., late fusion of audio and lip embeddings in English speech recognition [30], these methods are not optimized for fluency prediction or tested on languages with strong viseme-phoneme correspondence, such as Welsh.

Recent studies in multi-modal disfluency detection [24, 27] emphasize the effectiveness of combining audio and facial cues to enhance prediction accuracy. While primarily focusing on English disfluencies, these findings underscore the potential of visual features, particularly beneficial for languages like Welsh, which rely on consonant mutations and distinct phonemic variations. Existing

techniques in audio/video disfluency detection [12, 27] often utilize modalities such as EEG, facial muscle encoding, and textual annotations like 'repetition', 'prolongation', 'block', and 'pauses'. However, these methods rely on data that is difficult or labor-intensive to obtain.

Welsh presents a unique challenge for fluency learning and assessment due to the distinct lip and mouth movements required for accurate pronunciation. The language features complex phonemes — such as the rolled 'R,' 'CH,' 'DD,' and 'LL' — which can be challenging for both novice speakers and speech recognition systems. Many Welsh phonemes involve articulatory gestures not commonly found in English, making visual cues—such as lip protrusion and tongue movement—crucial for speech analysis (see Figure 1 for examples of these distinct mouth shapes in the 3rd and 4th columns when speaking a Welsh phrase). These factors highlight the potential value of a fluency evaluation and learning app with 3D visual feedback. Discussions with speech therapists suggest that such an app could provide valuable feedback for children born with hearing impairments, as well as for Parkinson's patients, supporting home rehabilitation and tracking speech progress as it changes over time. However, developing this learning feedback and rehabilitation app would require a dataset to capture Welsh phonetics, its associated 3D facial movements, and their correlation.

To this end, we introduce CymruFluency, the first 4D Welsh-speaking facial dataset, capturing both audio and high-resolution 3D lip movements with annotated landmarks. The dataset provides a snapshot of contemporary Welsh, particularly from the southern region, offering insights for Welsh language analysis. We also develop a multi-modal fluency classification technique that integrates audio features (mel spectrograms) and 3D lip landmarks, using a fusion strategy that combines embeddings from audio and landmark subnetworks. Our results show that this multi-modal approach outperforms unimodal models based on audio or mouth cues, supporting our hypothesis that lip and mouth movements are useful in Welsh learning and fluency assessment. Our contributions include:

1. We present CymruFluency, the first 4D Welsh fluency dataset, incorporating synchronized audio, 3D lip movement, and expert-assigned fluency scores.
2. We develop a novel multi-modal fluency classification framework that fuses audio and landmark data for improved Welsh fluency machine assessment.
3. Our results show that multi-modal analysis enhances fluency prediction accuracy for Welsh, a phonetically complex and under-researched language.
4. We contribute to the broader study of minority language processing by showing how articulatory features enhance automatic fluency evaluation.

2 Dataset CymruFluency

2.1 Participants & Data Collection

The CymruFluency dataset includes recordings from 33 speakers, each producing 10 predetermined Welsh phrases (Table 1), selected by a Welsh expert teacher with increasing difficulty. The anonymized participants were openly recruited via university emails, including Welsh learners of varying proficiency levels,

Table 1: Increasingly difficult Welsh phrases

Phrase	Welsh	Meaning in English
V1	Eisteddfod yr Urdd	Welsh Youth Music Competition
V2	Prynhawn da bawb	Good afternoon everyone
V3	Dyn busnes yw e	It's a businessman
V4	Papur a phensil	Paper and pencil
V5	Ardderchog	Excellent / Superb
V6	Llwyddiant ysgubol	Great success
V7	Yng nghanol y dref	In the town center
V8	Dwy neuadd gymunedol	Two community halls
V9	Llunio rhestr fer	Shortlisted
V10	Gwybodaeth angenrheidiol	Necessary information

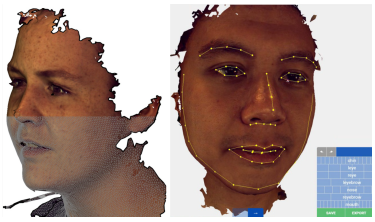


Fig. 2: 3D mesh quality and landmarking in progress.

native/fluent speakers, and international participants (from the EU, Africa, China, India, etc.) with no prior Welsh knowledge. After a research briefing and consent, participants watched prerecorded videos by a Welsh native speaker of each phrase three times before recording. All participants were incentivized with Amazon vouchers for their participation. This resulted in 327 valid sequences (three excluded due to capturing and reconstruction issues), with each sample lasting between 1 and 3 seconds, averaging 2.47 seconds, for a total duration of 13.46 minutes. A sample video for V10 is available at {`https://tinyurl.com/5yaxxn3n`}.

Fluency was rated on a scale of 0 to 5 by the same Welsh expert teacher. Table 2 presents the fluency scoring rubric used to evaluate spoken Welsh phrases (V1–V10) from the 33 speakers. The criteria include intelligibility, error frequency, and accent strength, with scores as follows: 5=native, 4=fluent with 1–2 issues, 3=more than 2 issues, 2=intelligible with significant issues, 1=barely intelligible, and 0=unintelligible. For our experiments, we treat fluency prediction as a binary classification, categorizing speakers as fluent (score ≥ 4) - as all fluent speakers scored 4 or above - or non-fluent (score < 4). Table 3 shows statistics for fluent and non-fluent speakers.

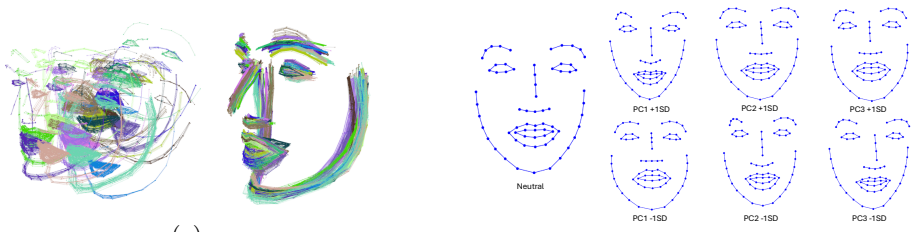
Table 2: Fluency scoring rubric (0–5) for spoken Welsh phrases

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	Avg
1	5	5	5	5	5	5	5	5	5	5	5.0
2	5	-	5	5	-	5	5	5	5	-	5.0
3	4	5	4	4	4	5	5	4	4	5	4.4
4	2	3	3	3	3	2	4	1	2	3	2.6
5	4	5	4	5	4	4	3	4	3	3	3.9
6	3	4	4	3	3	4	4	3	3	2	3.3
7	0	1	3	2	3	2	3	2	3	3	2.2
8	4	4	4	5	5	5	4	3	4	2	4.0
9	3	4	3	2	3	3	4	3	4	2	3.1
10	1	2	3	4	4	3	3	3	3	3	2.9
11	4	4	4	5	5	4	4	4	4	4	4.2
12	2	1	2	2	2	1	3	0	2	0	1.5
13	1	2	3	2	2	1	2	2	2	3	2.0
14	5	5	5	5	5	4	5	5	5	5	4.9
15	5	5	5	5	5	5	5	5	5	5	5.0
16	2	4	3	4	2	3	2	3	4	3	3.0
17	5	5	5	5	5	5	5	5	5	5	5.0
18	1	2	2	3	3	3	1	0	3	0	1.8
19	0	0	1	1	0	0	0	0	0	0	0.2
20	5	5	4	5	5	5	5	5	5	5	4.9
21	2	0	1	1	2	0	0	2	0	1	0.9
22	1	2	3	3	2	1	4	4	2	1	2.3
23	0	1	0	1	1	1	2	1	1	1	0.9
24	4	4	4	4	3	4	5	4	4	4	4.0
25	4	4	4	4	4	4	4	4	4	4	4.0
26	5	5	5	5	5	5	5	5	5	5	5.0
27	5	4	4	4	4	4	4	4	4	4	4.1
28	1	2	2	3	3	1	1	3	2	0	1.8
29	5	5	5	5	5	5	5	5	5	5	5.0
30	1	3	3	2	2	1	3	1	2	2	2.0
31	4	4	4	4	4	4	4	5	3	4	4.0
32	2	4	3	3	1	1	3	0	4	1	2.2
33	2	2	2	3	2	1	3	1	1	1	1.8

Table 3: Fluency statistics for fluent and non-fluent subjects.

Class	Sequences	Min	Max	Avg
Fluent	157	4	5	4.6
Non-fluent	170	0	3	2.2

Each phrase was assessed based on intelligibility, frequency of errors, and accent strength, with consideration for regional dialect variations that could influence scoring. The fluency scores ranged from 0 to 5. The recordings were



(a) (b)

Fig. 3: (a) Variations in the 3D facial landmark configuration along the top three principal components. The left contour (neutral face) represents the mean facial shape across all samples. The contours labeled "PC1 +1 SD" and "PC1 -1 SD" illustrate the dominant mode of variation, showing how the face deviates from the mean when this component is increased or decreased by one standard deviation. Similarly, "PC2 ±1 SD" and "PC3 ±1 SD" reveal the secondary and tertiary modes of variation, respectively. (b) Visualization of facial landmark alignment across multiple sequences using nine nose points.

captured using a medical-grade 3dMD system [1]. Calibration was performed before every session, and sequences were recorded at 48 fps with synchronized audio and 3D facial data. The system uses six stereo cameras (four infrared sensors for geometry reconstruction and two color sensors for texture capture) to obtain high-fidelity data. Raw image data (∼5MB per bitmap, ∼1.4GB/sec) were stored locally and processed offline to reconstruct detailed 3D geometry and texture, with each mesh containing ∼17K vertices.

2.2 Two Modalities

Audio was captured concurrently with the 3D recordings and saved in the '.wav' format at a sampling-rate of 16KHz. Each audio sample duration lasts from 1-3 seconds and for each sample, a set of features is extracted, including Mel Frequency Cepstral Coefficients (MFCC), Zero-Crossing Rate (ZCR), Root Mean Square Energy (RMSE), and Spectral Flux (SF)[7]. These features are computed for each individual time frame across the entire duration of the audio sample.

3D Landmarks Each facial sequence was manually annotated per frame using Landmarker.io [23] with the iBug68 template [29]. All annotations were cross-checked by two annotators. Manual annotation is crucial because automated techniques often struggle to accurately capture 3D lip shape, such as protrusion, which is essential for detailed speech articulation analysis, especially for Welsh. By manually annotating 20 key lip landmarks, we ensure high accuracy and consistency, providing a strong foundation for future applications such as speaking head synthesis, real-time streaming, and medical rehabilitation. These annotations not only essential for our experiments but also serve as a valuable prior for advancing techniques in computer vision and speech analysis. For our pilot study, we utilize these annotated landmarks for fluency analysis to demonstrate their usefulness and validate our hypothesis.

Natural speech involves subtle head and body movements. To stabilize facial landmarks while preserving their shape and size, we apply rigid alignment using

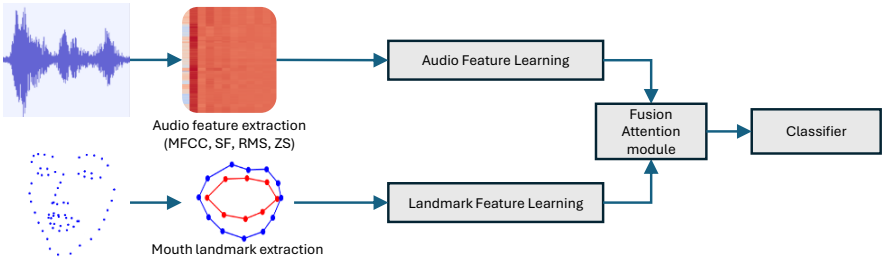


Fig. 4: Complete pipeline of the classification architecture with Audio/Landmark learning module and Fusion module.

nine nose points [33], leveraging the nose's stability [26]. Figure 3a shows landmark sequences of multiple subjects speaking phrase V2, before and after alignment. Figure 3b shows the mean facial landmarks (the neutral pose baseline). PCA [2] applied to the aligned data captures the primary modes of variation, with standard deviations (SD) quantifying deviations from this baseline.

3 Methodology

3.1 Network Architecture Overview

In this section, we present the network architecture designed to integrate audio and landmark features for fluency analysis using the CymruFluency dataset. Existing fluency prediction techniques for major languages require extensive annotations [19, 12] and often lack accessible source code, making them less applicable to our setting. Given the relatively small size of our dataset and the short duration of Welsh phrases, we focus on exploring and designing compact models. We begin by investigating various audio and landmark feature learning techniques (Section 3.2) for unimodal fluency classification. Section 3.3 further explores multi-modal fusion strategies for evaluating our hypothesis. Finally, we summarize our findings in Section 4.

3.2 Feature Learning

Audio To extract meaningful representations from short phrases, we explore two feature-learning approaches inspired by prior work in speech processing. First, following compact feature-based models used in fluency research [7], we use static fixed-length audio feature vector, later concatenated with landmark features before being processed through a multilayer perceptron (MLP) to serve as a ***baseline*** approach. This approach is effective when key acoustic characteristics have already been distilled into a compact representation. Second, to model ***fine-grained*** temporal patterns in speech, we leverage recurrent neural networks designed for sequence learning [15, 9]. Specifically, long short-term memory (LSTM) and gated recurrent units (GRU) capture subtle fluency cues by modeling timing variations and acoustic transitions, offering a richer fluency representation.

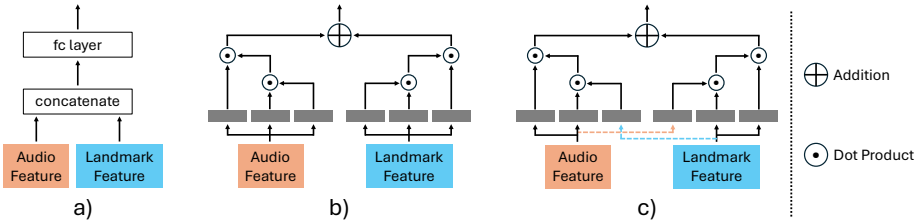


Fig. 5: Visualization of three audio-landmark fusion strategies: (a) Concatenation followed by a fully connected layer; (b) Modality-specific self-attention captures intra-modal dependencies; (c) Cross-attention models inter-modal interactions (blue and orange dotted arrows indicate feature conditioning from a different modality).

Landmarks To capture articulatory dynamics from 3D mouth and lip landmark data, we explore several feature learning approaches, drawing on prior work in temporal data analysis: As a simple baseline, we apply mean pooling across the landmark sequence to produce a compact representation, treating it as a static feature. While computationally efficient, static pooling serves only as a baseline and may miss fine-grained motion details. Our main focus is on models that exploit the full temporal sequence of 3D landmarks to capture the dynamic articulatory patterns crucial to fluency. To this end, we employ a LSTM network, which has proven effective in sequential gesture and speech-related tasks [31]. In some configurations, we enhance the LSTM with an attention mechanism to focus on key facial movements indicative of fluency. Transformer-based architectures have shown strong performance in action recognition [22], but their high data requirements make them prone to overfitting on small datasets like ours. Similarly, we avoid using 3D convolutional neural networks (3D CNNs) on landmark sequences, as their large model size and complexity also require large-scale datasets for stable training [8, 21]. Therefore, inspired by advancements in skeleton based motion analysis and the topological structure of mouth and lip landmarks, we apply a spatio-temporal graph convolutional network (ST-GCN [16]) to model both spatial and temporal dependencies. The inner and outer lip landmarks form two loops. This topology may help ST-GCN capture lip movements and identify fluency cues in short phrases.

3.3 Fusion Strategy

Baseline Approach: We use fixed-length audio features concatenated with mean-pooled landmark features. The 83-dimensional input (23 audio + 60 landmarks) is passed to an MLP with one hidden layer of 128 units, followed by batch normalization, ReLU, dropout, and a final layer for binary classification.
Fine-grained Approach: After extracting modality-specific features (using GRU, LSTM, ST-GCN), we explore fusion strategies. Early fusion risks losing modality details, while late fusion may overlook important dependencies [38]. We adopt intermediate fusion to strike a balance between modality preservation and cross-modal interaction, considering both concatenation and attention-based fusion. For attention-based fusion, we investigate two approaches: self-attention

and cross-attention. Self-attention enhances feature representation within a single modal by re-weighting features based on contextual relationships and capturing intra-modal dependencies [34]. Cross-attention enables inter-modal fusion by computing attention scores between audio and 3D landmarks, allowing one modality to selectively focus on the most relevant aspects of the other for effective integration. In contrast, simple concatenation lacks explicit weighting, relying on subsequent layers to infer feature importance, which may be less effective when modalities vary in relevance. However, given our small dataset and potential redundancy between audio and temporal landmarks, concatenation offers a simple and effective approach for our exploration. After fusion, the resulting multi-modal representation is passed through a classifier to produce logits for binary prediction. In attention-based fusion (self or cross-attention), the classifier is a single fully connected layer that outputs probabilities for the two classes: fluent or non-fluent. In concatenation-based fusion, the classifier refines the fused features using a fully connected layer with ReLU and dropout, followed by a final output layer.

Figure 5 visualizes three fusion strategies: a) concatenation of audio and landmark features, b) self-attention on each feature, and c) cross-attention, where one modality conditions the other. After attention in b) and c), the features are combined and passed through a fully-connected layer for fusion.

4 Experiment

4.1 Experiment Setup

We evaluate our method on a PC with a 2.6GHz CPU, 8GB RAM, and 1080Ti GPU. The CymruFluency dataset is split using a stratified random approach: 70% (229 samples) for training and 30% (98 samples) for testing, ensuring balanced fluent and non-fluent classes. This speaker-independent split is based on fluency scores from expert annotators, avoiding data biases and leakage. We use the Adam optimizer with cross-entropy loss, a learning rate of 1e-5 to 1e-3, and a batch size of 16 for 500 epochs.

4.2 Results and Analysis

Table 4 presents a comparison between our model and the state-of-the-art SVM baseline [7]. The SVM baseline was originally evaluated on the Avalingo dataset - a larger English dataset with 1,732 samples and a three-class fluency annotation scheme (high, medium, low). In contrast, we focus our evaluation on the CymruFluency dataset, as it includes both audio and visual modalities, whereas Avalingo is audio-only and thus not directly comparable to our multimodal approach. To enable a fair comparison on CymruFluency, we adapt the SVM to binary classification. While it achieves 92% accuracy, our model outperforms it with 99%, highlighting its robustness in low-resource, short-utterance conditions.

CymruFluency consists of 327 short utterances (1–3 seconds) labeled with binary fluency classes (Fluent vs. Non-Fluent), requiring models to make precise predictions based on limited input. The SVM's limited performance on

this dataset reflects its reduced adaptability in constrained settings. In contrast, our model achieves high accuracy, suggesting better suitability for short, low-resource, multimodal fluency assessment.

Table 4: Comparison with existing approaches on fluency detection on the CymruFluency dataset.

Model	**Modality**	**Classes**	**Accuracy**
SVM [7]	Audio	2	92%
Our approach	Audio + 3D Landmarks	2	**99%**

Table 5: Ablation study comparing network variations: Audio-only (top), Landmark-only (middle), and Audio + Landmark (bottom).

Audio	**Landmark**	**Fusion**	**Accuracy**
GRU	-	-	82%
LSTM	-	-	89%
-	ST-GCN	-	79%
-	LSTM	-	93%
Baseline			
Static Feat	Mean pool	Concatenation	80%
Fine-grained			
GRU	ST-GCN	Concatenation	86%
GRU	ST-GCN	Self-Attention	92%
GRU	ST-GCN	Cross-Attention	90%
LSTM	LSTM	Concatenation	93%
LSTM	LSTM	Self-Attention	**99%** (97/98)
LSTM	LSTM	Cross-Attention	96%

Table 5 evaluates modality-specific and fusion architectures. Unimodal systems - audio-only GRU (82%), LSTM (90%), and landmark-only ST-GCN (79%), LSTM (93%) - show moderate performance, confirming that both vocal and lip dynamics provide useful fluency cues. Notably, landmark-only LSTM (93%) outperforms audio-only models. Simple feature concatenation (MLP-baseline) performs poorly (80%) due to limited fusion capacity. In contrast, self-attention fusion improves results: GRU (audio) + ST-GCN (landmark) with self-attention achieves 92%, outperforming unimodal models by 12 and 17 points. The LSTM-LSTM self-attention model reaches 99% accuracy (97/98), surpassing the best unimodal baselines by 6 and 10 points.

Combining audio and landmark data via self-attention yields near-perfect accuracy. Cross-attention and concatenation underperform, highlighting the need for better temporal alignment. The 17.7% gap between landmark-only LSTM and ST-GCN suggests that LSTM's simpler architecture generalizes better in low-data settings, while ST-GCN is more prone to overfitting.

These results collectively establish that:

1. Our dataset's 3D landmarks capture useful dynamics of lip movements, providing visual fluency cues unavailable to prior audio-only methods.
2. Attention-based fusion is highly effective to combine audio and landmark features for modeling complex phonetic structure and pronunciations for Welsh fluency detection.
3. Our proposed method achieves strong performance despite limited training data, showing its suitability for low-resource scenarios – a necessity for minority languages like Welsh.

Limitations While CymruFluency provides a valuable multi-modal resource for Welsh fluency analysis, several limitations remain. First, as the first study of its kind, fluency annotations were made by a single expert. Future work will involve multiple raters to assess inter-annotator reliability. Second, with only 33 speakers and 10 fixed phrases, the dataset's size and diversity are limited, which may affect generalizability. This also constrained the performance of more complex models,

architectures, and fusion designs, limiting us to basic fusion strategies. We plan to expand the dataset in future work. Lastly, although the study uses 4D facial and audio inputs to explore scientific hypotheses and demonstrate the value of 3D lip movement for both learners and machine evaluation, such hardware is not widely available in real-world or consumer settings. Adapting the system for video- or audio-only inputs, with 4D priors, will require further fine-tuning.

5 Conclusion

This study takes a first step toward overcoming a major challenge in Welsh research and technology: the lack of data for language teaching, learning, and fluency assessment. To support this effort, we introduce CymruFluency, the first Welsh-language dataset integrating audio and 4D facial geometry across fluency levels. By capturing a contemporary snapshot of spoken Welsh, this dataset lays the groundwork for advancing Welsh-language technologies.

Our research introduces multi-modal techniques for Welsh fluency prediction, emphasizing the role of articulatory dynamics like lip and mouth movements. Results show that combining 3D lip shape cues with audio significantly enhances fluency predictions. It highlights the usefulness of the CymruFluency dataset. We also explore various fusion strategies, and demonstrate that self-attention mechanisms improve performance.

Future work will expand the CymruFluency dataset with more data, longer sequences, and finer annotations, while developing a smartphone app for two user groups: Welsh language learners and speech therapy patients. The app for learners will include fluency assessments, video feedback with 3D head synthesis, and tools to improve pronunciation. For speech therapy patients, the app will offer home-based rehabilitation exercises and progress monitoring. Although stereo capture is not widely available, we aim to adapt 3D features for 2D or audio-only systems to enhance accessibility. Additionally, we will explore automated video-based techniques and 3D priors for advanced speech analysis. We will release the dataset and source code for non-commercial research, supporting adaptive language education and speech therapy both within and beyond the Welsh community. We hope this study inspires further research on under-resourced minority languages to advance language technologies.

Acknowledgements This research was supported by Coleg Cymraeg Cenedlaethol Small Grant 2017, Cherish-DE (EP/M022722/1) Escalator Fund 2019, 2021(1RR, 52E), Swansea University SPIN fund, Wales Network Innovation Small Grant 2023 and EPSRC IAA Fund 2024. We thank Sam, Jiali, Shuyu, LLY Chan, Mart, Olabayoji, Vijay, Ananth, Jishma, Mark, Iestyn and all anonymized participants for their contributions to this project. For the purpose of open access the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

1. 3dMD: 3D Face Imaging Systems, `https://3dmd.com/`

2. Active shape models-their training and application. Computer Vision and Image Understanding **61**(1), 38–59 (1995)
3. Agrija, contributors: Avalinguo audio set. `https://github.com/agrija9/Avalinguo-Audio-Set`, accessed: 2025-02-12
4. Anwar, M., Shi, B., Goswami, V., Hsu, W.N., Pino, J., Wang, C.: Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. In: Proc. INTERSPEECH (2023)
5. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). pp. 4211–4215 (2020)
6. Bernstein Ratner, N., MacWhinney, B.: Fluency bank: A new resource for fluency research and practice. Journal of Fluency Disorders **56**, 69–80 (2018)
7. Brena, R., Zuvirie, E., Preciado Grijalva, A., Valdiviezo, A., Gonzalez-Mendoza, M., Zozaya-Gorostiza, C.: Automated evaluation of foreign language speaking performance with machine learning. International Journal on Interactive Design and Manufacturing (IJIDeM) **15**, 1–15 (09 2021)
8. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733 (2017)
9. Cho, K., Merriënboer, B.V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734 (2014)
10. Chondrogianni, V., John, N.: Tense and plural formation in welsh–english bilingual children with and without language impairment. International Journal of Language & Communication Disorders **53**(3), 495–514 (2018)
11. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Proc. ACCV. pp. 87–103 (2017)
12. Das, A., Mock, J., Huang, Y., Golob, E., Najafirad, P.: Interpretable self-supervised facial micro-expression learning to predict cognitive state and neurological disorders. Proc. AAAI **35**(1), 818–826 (2021)
13. Devi, T.M., Keerthana, S., Santhi, P., Pravallika, P., Rajeshwari, S.: Silent speech recognition: Automatic lip reading model using 3d cnn and gru. In: Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 1. pp. 827–832 (2025)
14. Gothi, R., Kumar, R., Pereira, M., Nayak, N., Rao, P.: A dataset and two-pass system for reading miscue detection. In: Proc. INTERSPEECH. pp. 123–130 (2024)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
16. Huang, J., Kang, H.: 3d skeleton-based human motion prediction using spatial-temporal graph convolutional network. International Journal of Multimedia Information Retrieval **13**, 33 (2024)
17. Kourkounakis, T.: UCLASS Stutter Annotations (2021), `https://doi.org/10.5683/SP3/1FIUHM`
18. Kourkounakis, T., Hajavi, A., Etemad, A.: Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 2986–2999 (2021)
19. Lea, C., Mitra, V., Joshi, A., Kajarekar, S., Bigham, J.P.: Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In: Proc. ICASSP. pp. 6798–6802 (2021)

20. Liu, J., Wumaier, A., Fan, C., Guo, S.: Automatic fluency assessment method for spontaneous speech without reference text. Electronics **12**(8) (2023)
21. Martin, P.E., Benois-Pineau, J., Péteri, R., Morlier, J.: Fine grained sport action recognition with twin spatio-temporal convolutional neural networks. Multimedia Tools and Applications **79**, 20429–20447 (2020)
22. Mazzeo, P.L., Spagnolo, P., Fasano, M., Distante, C.: Human action recognition with transformers. In: Sclaroff, S., Distante, C., Leo, M., Farinella, G.M., Tombari, F. (eds.) Image Analysis and Processing – ICIAP 2022 (2022)
23. Alabort-i Medina, J., Antonakos, E., Booth, J., Snape, P., Zafeiriou, S.: Menpo: A comprehensive platform for parametric image alignment and visual deformable models. In: ACM International Conference on Multimedia. pp. 679–682 (2014)
24. Mohapatra, P., Likhite, S., Biswas, S., Islam, B., Zhu, Q.: Missingness-resilient video-enhanced multimodal disfluency detection. In: Proc. INTERSPEECH. pp. 5093–5097 (2024)
25. Mulgrew, L., Duffy, O., Kennedy, L.: Assessment of minority language skills in english–irish-speaking bilingual children: A survey of slt perspectives and current practices. International Journal of Language & Communication Disorders **57**(1), 63–77 (2022)
26. Nair, P., Cavallaro, A.: 3-D face detection, landmark localization, and registration using a point distribution model. Trans. on Multimedia **11**(4), 611–623 (2009)
27. Nie, L., Kadiri, S.R., Agrawal, R.: Mmsd-net: Towards multi-modal stuttering detection. In: Proc. INTERSPEECH. pp. 5113–5117 (2024)
28. Panda, A., Acharya, R., Kopparapu, S.K.: Oral fluency classification for speech assessment. In: The 31st European Signal Processing Conference (EUSIPCO) (2023)
29. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE ICCV-Workshop, 300 Faces in-the-Wild Challenge (300-W) (2013)
30. Sayed, H.M., ElDeeb, H.E., Taie, S.A.: Multimodal data fusion architectures in audiovisual speech recognition. In: Information Systems and Technologies. pp. 655–667 (2024)
31. Srivastava, S., Singh, S., Pooja, Prakash, S.: Continuous sign language recognition system using deep learning with mediapipe holistic. Wireless Personal Communications **137**, 1455–1468 (2024)
32. Sung-Bin, K., Chae-Yeon, L., Son, G., Hyun-Bin, O., Ju, J., Nam, S., Oh, T.H.: Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. In: Proc. INTERSPEECH. pp. 1380–1384 (2024)
33. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE TPAMI **13**(4), 376–380 (1991)
34. Waswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
35. Welsh Government: Welsh language data from the annual population survey (Oct 2023 - Sep 2024). Tech. rep., Welsh Government (2024)
36. Welsh Government: Welsh language technology action plan: Final report 2018-2024. Tech. rep., Welsh Government (2024)
37. Zhang, J., Zhang, Z., Wang, Y., Yan, Z., Song, Q., Huang, Y., Li, K., Povey, D., Wang, Y.: speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In: Proc. INTERSPEECH (2021)
38. Zhao, F., Zhang, C., Geng, B.: Deep multimodal data fusion. ACM Computing Survey **56**(9) (2024)