Interspeech 2024
1-5 September 2024, Kos, Greece

Missingness-resilient Video-enhanced Multimodal Disfluency Detection

Payal Mohapatra[*1], *Shamika Likhite*[*1], *Subrata Biswas*[2], *Bashima Islam*[2], *Qi Zhu*[1]

[1]Northwestern University, USA
[2]Worcester Polytechnic Institute, USA

(payalmohapatra,shamikalikhite,qzhu)@northwestern.edu, (sbiswas,bislam)@wpi.com

Abstract

Most existing speech disfluency detection techniques only rely upon acoustic data. In this work, we present a practical multimodal disfluency detection approach that leverages available video data together with audio. We curate an audio-visual dataset and propose a novel fusion technique with unified weight-sharing modality-agnostic encoders to learn the temporal and semantic context. Our resilient design accommodates real-world scenarios where the video modality may sometimes be missing during inference. We also present alternative fusion strategies when both modalities are assured to be complete. In experiments across five disfluency-detection tasks, our unified multimodal approach significantly outperforms Audio-only unimodal methods, yielding an average absolute improvement of 10% (i.e., 10 percentage point increase) when both video and audio modalities are always available, and 7% even when video modality is missing in half of the samples.

Index Terms: speech disfluency; multimodal learning.

1. Introduction

Speech disfluency, encompassing hesitations, repetitions, or prolongations, affects about 1% of the global population [1]. Disfluent speech impacts communication, interpersonal skills, social connections [2], and access to voice-assisted technologies [3]. Thus, advancing research in disfluency is critical for enhancing accessibility and inclusivity in technology [4].

Speech disfluency research particularly faces the challenge of lack of public datasets for benchmarking due to high annotation cost, the clinical nature of the task, and the usage of proprietary datasets [5, 6]. Certain studies tackle this scarcity by artificially introducing disfluencies into typical speech [7]. However, this method may oversimplify the nuances of naturally occurring disfluency and might not effectively translate to real-world contexts [8]. UCLASS [9], TORGO [10], and FluencyBank [11] are some of the disfluency-specific speech corpora collected from monologues and interviews with people who stutter ranging from children to dysarthric patients, but they lack labels for disfluencies. Past researchers have conducted private annotations [6, 12] for these databanks, adhering to an inconsistent taxonomy in the process. To mitigate these shortcomings, Lea et. al [13] have released an annotated dataset, SEP28k, and also provided labels for FluencyBank audio clips with five different classes of disfluencies. All these datasets primarily consist of unimodal audio data, with a few works exploring manual text transcriptions as an additional modality [14, 15].

*Equal contribution. This work is supported by the National Science Foundation grants 2038853, 2328973, 2324936, and 2347692.

Most existing works in disfluency detection also heavily rely on audio data, employing handcrafted features [13, 16], pretrained language model embeddings [17], or custom disfluency foundation models with unlabeled atypical speech [18]. These are often paired with classifiers like statistical models [19], support vector machines [8], or deep neural networks [13, 17, 18, 16]. Some recent works incorporate manual text transcriptions [20] as an additional modality or use off-the-shelf foundation models for untranscribed audio [15, 14] to improve disfluency detection. Despite outperforming the audio-only features, audio-text multimodal studies recognize the limited expressiveness of disfluency markers through text and the inherent errors in transcribing disfluent speech.

In this work, we offer a novel perspective in enhancing disfluency detection by incorporating visual cues. The first challenge in developing audio-visual multimodal frameworks for disfluency detection is the lack of paired annotated audio-visual disfluency datasets. Second, due to various practical issues such as sensor malfunctions, resource limitations, environment constraints (poor lighting, obstructions, etc.), and privacy concerns, it is common to have missing modalities, particularly video, during inference. Finally, unlike mainstream speech enhancement tasks, the paradigms of audio-visual multimodal learning may not seamlessly extend to the assessment of paralinguistic tasks in atypical speech. Hence, designing multimodal learning frameworks tailored to disfluent speech, resilient to modality-missingness, is crucial to improving disfluency detection.

Our Approach. To design a novel multimodal fusion framework resilient to arbitrary missingness of video modality, we first curate a custom audio-visual dataset leveraging metadata [13] and raw datasets [11] from past works. Our architecture features a weight-sharing encoder for both modalities, allowing it to operate on samples where one of the modalities is missing during training or inference, as shown in Figure 1. (I). We utilize a temporal decimator to project the densely sampled audio input's feature embedding from a pretrained foundation model to the same vector space as the features extracted from the full facial region of the speaker in video samples. Subsequently, both modalities share a learnable encoder, which learns the corresponding temporal context. The encoder then collapses it before incorporating the dynamic scaling factors for the modalities and feeding to a classifier head. Our empirical results show that commonly-used strategies like concatenating features in a lower dimension for fusion and using cropped lip regions as typically used for speech recognition tasks [21, 22] do not work very well in this case. Hence, we also demonstrate other fusion strategies when the availability of both modalities is always guaranteed, as shown in Figure 1. (II, III).

In summary, our contributions include: (1) a 3.3-hour paired and annotated audio-visual disfluency dataset with

10.21437/Interspeech.2024-1458

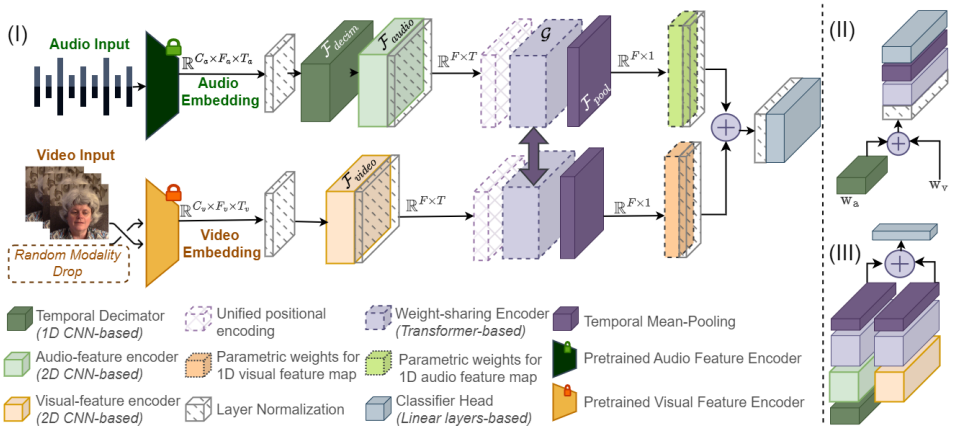

Figure 1: *Illustration of our multimodal learning framework for speech disfluency detection. (I) Unified Modality Fusion Network resilient to missing video modalities, (II) Modality-specific early fusion, and (III) Modality-specific late fusion.*

benchmarking splits (Section 2.1), (2) a novel fusion technique using weight-sharing encoders for disfluency learning despite missing video modalities (Section 2.3), (3) alternative fusion strategies for end-to-end disfluency training without concern for missing modalities (Section 2.3), and (4) comprehensive evaluation against unimodal and multimodal baselines, demonstrating the superiority of our audio-visual learning for disfluency detection (Section 3.1). Our open-source code is available[1].

Related Works. *Emotion understanding* is another paralinguistic task where multimodal learning is conducted beyond audio modality, through visual cues, physiological signals, hand gestures, text, etc. [23, 24], utilizing extensive datasets for modality-dependent fusion [25], cross-modal interaction structures [26], and collaborative learning [24]. General *multimodal learning* for addressing modality-missingness applies reconstruction in latent space [27], statistical imputation [28], encoding missingness [29], or reinforcement learning-based policy design to weigh incomplete input vectors [30] dynamically. In contrast, our work introduces a unified encoder scheme featuring weight-sharing and additive fusion for a small-scale end-to-end trainable disfluency detection task.

2. Approach

2.1. Dataset Curation

The lack of multimodal disfluency detection frameworks from an audio-visual perspective is largely due to scarce public paired and annotated audio-visual datasets. We take efforts to curate an audio-visual dataset by leveraging meta-data and open-source databases from past works [13, 11] and adhere to their taxonomic recommendations. FluencyBank contains video recordings from 32 individuals with disfluent speech in the English language. Lea et. al [13] released an annotated dataset of 3-second audio segments for the FluencyBank database containing 4,144 audio clips. They recruit 3 annotators to systematically label each audio segment as one of the five disfluencies – Blocks (Bl), Word repetition (WP), Sound repetition (SnD), Interjection (Intrj), and Prolongation (Pro) or as fluent speech. One of the challenges with this annotation schema is the inter-annotator disagreement and past works have highlighted its detrimental impact on performance [17]. To address this, we conduct a distillation of the dataset by only retaining samples with a majority vote similar to our previous work [17] and constructing a corpus for FluencyBank containing 4,004 annotated audio clips. Next, we leverage the natural temporal alignment of audio and video databases from FluencyBank and segment the available videos based on the meta-data for the start and end times of a continuous recording from the audio annotations to construct a paired multimodal dataset for FluencyBank. Note that although the FluencyBank corpus originally also offered text transcriptions, past works [14] have highlighted its temporal misalignment and inaccuracies in capturing the frame-level disfluencies to a degree that can be leveraged effectively for overall disfluency categorization. Hence, we focus only on the audio-visual multimodal dataset in this work. All the audio data are sampled at 44.1kHz and the video data at 30 frames per second. The summary of the amount of data available for each type of speech disfluency and their descriptions are given in Table 1. To facilitate benchmarking for multimodal speech disfluency research, we release this curated audio-visual annotated dataset of approximately 3.3 hours along with the training and evaluation splits for reproducibility[1].

Table 1: *Audio-Visual dataset from the FluencyBank corpus.*

Category	Description	Size
Blocks(Bl)	Long unnatural pauses	469
Word Repetition(WP)	Repetition of any word	337
Sound Repetition(SnD)	Intra-word phoneme repetition	428
Interjection(Intrj)	Filler words or non-words	701
Prolongation(Pro)	Extended sounds within a word	407
NoStutteredWords	Fluent speech	1660

2.2. Preprocessing

Audio Embeddings. Past works [17, 31, 32] consistently highlight the effectiveness of features extracted from large-scale pretrained networks over traditional acoustic features. We leverage large-scale foundation models trained on 1000s of hours of typical speech data to extract meaningful features. In this case, we utilize the base architecture of wav2vec 2.0 [33] to support a small-scale end-to-end training with limited labeled data. The 3-second audio data is resampled to 16kHz and fed to 12 frozen layers of transformers in wav2vec 2.0. In Section 3.2 we also show that one can easily replace this extractor as suitable for their task. The audio input, $\mathrm{x_a} \in \mathbb{R}^{1\times T_a}$ where T_a is the temporal length of each sample, is transformed to $\mathrm{w_a} \in \mathbb{R}^{C_a\times F_a\times T_a}$. Audio signal is sampled at a much higher rate than video modality. To facilitate projection to

[1] https://github.com/payalmohapatra/Multimodal-Speech-Disfluency.git

a common latent space eventually, we implement an embedding decimator, $\mathcal{F}_{decim} : \mathbb{R}^{C_a \times F_a \times T_a} \rightarrow \mathbb{R}^{C_a \times F_a \times T_{decim}}$, where $T_{decim} < T_a$ using 1D convolution (CNN) layers along the temporal axis. This is followed by an audio-feature encoder, $\mathcal{F}_{audio}$ designed using two layers of 2D CNNs with max-pooling to further summarize the embeddings to a common latent space, $\mathbb{R}^{F \times T}$.

Video Embeddings. The video input $\mathrm{x_v} \in \mathbb{R}^{3 \times P \times P \times T_v}$, where P is the pixel resolution of the image stream and T_v is the total number of frames in the video segment. To extract meaningful video features from the full facial region, we leverage Ma et. al's [34] preprocessing pipeline for automatic speech recognition, which is well-established for conventional end-to-end visual speech recognition tasks. This pipeline consists of 3D CNNs with a ResNet-18 encoder which transforms $\mathrm{x_v}$ to $\mathrm{w_v} \in \mathbb{R}^{C_v \times F_v \times T_v}$. After this stage a vision encoder, $\mathcal{F}_{video}$, symmetric to $\mathcal{F}_{audio}$ is employed for feature summarization.

2.3. Unified Modality Fusion Network

Different from most multimodal frameworks, we propose a novel weight-sharing encoder $\mathcal{G}$, which is common to multiple modalities. Through $\mathcal{F}_{audio}$ and $\mathcal{F}_{video}$ encoders, both modalities are projected to a common latent space. Their respective outputs are normalized and added with a positional encoding using different frequencies of sinusoids similar to [35]. $\mathcal{G}$ is achieved using a multi-head (16 heads in our case) attention transformer encoder layer [35] realized using three input streams $\mathrm{Q}, \mathrm{K}, \mathrm{V} \in \mathbb{R}^{T \times F}$ from learnable networks to compute the representative attention, $\mathrm{A} \in \mathbb{R}^{T \times F}$, such that,

$$\mathrm{S} = \mathrm{Q}.\mathrm{K}^{\mathrm{T}} \in \mathbb{R}^{T \times T}, \mathrm{P} = \mathtt{softmax}(\mathrm{S}), \mathrm{A} = \mathrm{PV} \in \mathbb{R}^{T \times F}.$$

Following the multi-headed attention, a temporal mean pooling $\mathcal{F}_{pool}$ is employed, where $\mathcal{F}_{pool}(\mathrm{A}) : \mathbb{R}^{T \times F} \rightarrow \mathbb{R}^{1 \times F}$. The intuition for this design is that the relevant temporal dependencies are learned by $\mathcal{G}$ and now it is important to focus on the semantic differentiability of the features. The common latent embedding dimensions for both modalities facilitate the design of this weight-sharing encoder: $\mathrm{r_a}, \mathrm{r_v} = \mathcal{G}(\mathcal{F}_{audio}(\mathrm{w_a})), \mathcal{G}(\mathcal{F}_{video}(\mathrm{w_v}))$, where $\mathrm{r_a}, \mathrm{r_v}$ are the results of individual forward passes of both modalities through $\mathcal{G}$. This design is tolerant of the missingness of one of the modalities due to this shared-weights-based fusion. It can support the partial availability of modalities during both the training and inference stages. To learn the dynamic importance of the feature maps from these modalities, we incorporate learnable scalars $\mathrm{c_a}, \mathrm{c_v} \in \mathbb{R}^F$, such that $\mathrm{r} = \mathrm{c_a} \cdot \mathrm{r_a} + \mathrm{c_v} \cdot \mathrm{r_v}$. The overall design of this missingness-resilient audio-visual disfluency detection through unified fusion (**DAV-unified**) network is shown in Figure 1. (I). For scenarios where both modalities are guaranteed to be available, we also propose two variants of multimodal learning framework, early fusion (as shown in Figure 1. (II)) and late fusion (Figure 1. (III)).

Early Fusion (DAV-early). We empirically found that a mere concatenation of features from the two domains, $\mathrm{w_a}$ and $\mathrm{w_v}$, does not perform well. So we propose a similar design as the unified fusion network, by adding the projections of $\mathrm{w_a}$ and $\mathrm{w_v}$ and then feeding the result's normalized version to the feature encoder with positional encoding. It is designed similarly to $\mathcal{G}$.

Late Fusion (DAV-late). In this case, most of the initial pipeline of data transformation is identical to the unified fusion network. However, instead of a shared encoder, it contains two dedicated encoders $\mathcal{G}_a$ and $\mathcal{G}_v$ to learn modality-specific weights. The resulting embeddings from these two encoders are then added to fuse deeper in the network.

Modality Dropout Augmentation. Often in disfluency applications, audio is the primary modality and video data is missing at times. To support such scenarios during inference and training, we adopt an augmentation strategy of dropping the video data randomly during training to allow the model to learn to switch states when one of the modalities is missing. Following the Bernoulli distribution, we sample a binary masking variable m, for the mini-batch of size B. $m_i \sim p_i^m * (1-p)^{(1-m_i)} \forall i \in (0, B-1)$, where p is the probability of dropping set to 0.5 in this case. Although we apply modality dropout to only the video domain, this is a versatile augmentation that can be extended to all modalities as suitable. $\mathrm{x_v}$ is dropped if $\mathrm{m_i}$ is 1.

2.4. Classifier Head

The output after the fusion (in case of early fusion schema, the output of $\mathcal{G}$), r, is given to the classifier head, $\mathcal{C}$, which is realized using a few fully-connected layers. Note that incorporating $\mathcal{F}_{pool}$ also significantly helps in reducing the number of trainable parameters for $\mathcal{C}$. To achieve semantic distinction between the fluent and disfluent speech segments, we optimize the function $\mathcal{L}_{\mathrm{CE}} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{y}_i \log \mathcal{C}(\mathrm{r})$, where B is the size of a batch in the mini-batch training and y_i is the one-hot form of the label y_i. As a side note, we explored a joint training approach using a cosine similarity loss function, treating the shared encoders as a siamese network [18] with $\mathrm{r_a}$ and $\mathrm{r_v}$ along with $\mathcal{L}_{\mathrm{CE}}$, but did not find significant gains.

2.5. Training Setup

All experiments are performed on an Ubuntu OS server equipped with NVIDIA TITAN RTX GPU cards using the PyTorch framework. During model training, we use Adam optimizer with a learning rate from 1e-6 to 1e-4, batch size of 512, and the maximum number of epochs is set to 500 based on the suitability of each setting. Since our class distribution is imbalanced, we used a custom sampler to re-weight the samples per class in every minibatch. We tune these optimization-related hyperparameters for each setting and save the best model checkpoint using early exit based on the minimum value of $\mathcal{L}_{\mathrm{CE}}$ achieved on 10% of training data held out for validation.

3. Results

3.1. Experimental Setup

We evaluate the effectiveness of our multimodal disfluency detection approach, including variants **DAV-united**, **DAV-early** and **DAV-late**, against state-of-the-art multimodal and unimodal disfluency, and generic speech recognition methods. We use metrics F1 $= \frac{2 \cdot \mathrm{TP}}{\mathrm{TP} + \frac{1}{2}(\mathrm{FP}+\mathrm{FN})}$ and Balanced Accuracy (BA) $= \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)$, where TP is true positive, FP is false positive, TN is true negative and FN is false positive, to fairly assess imbalanced datasets. We hold out 20% of the data for each class for evaluation and provide the exact train test splits for the curated dataset. Every experiment is carried out with 3 different seeds (123, 456, 789) and the mean and standard deviations are reported. We report performance gains as the absolute difference from the baseline. We conduct experiments for five different disfluency-detection tasks and maintain uniform feature extraction pipelines for audio and video inputs across all models unless specifically required by the model. Specifically, the

Table 2: *Comparison of three variants of our multimodal approach (DAV-early, DAV-late, DAV-unified) and four unimodal and multimodal baselines on balanced accuracy and F1-score across five disfluency tasks. The best results are highlighted in* ***bold****.*

Task	Blocks		Word Repetition		Sound Repetition		Interjection		Prolongation		Avg.
	BA	F1	BA	F1	BA	F1	BA	F1	BA	F1	Acc.
Audio-Only	$0.64_{\pm 0.01}$	$0.44_{\pm 0.01}$	$0.62_{\pm 0.00}$	$0.35_{\pm 0.01}$	$0.65_{\pm 0.02}$	$0.44_{\pm 0.02}$	$0.62_{\pm 0.02}$	$0.49_{\pm 0.03}$	$0.66_{\pm 0.02}$	$0.44_{\pm 0.04}$	0.64
Video-Only	$0.71_{\pm 0.01}$	$0.53_{\pm 0.02}$	$0.75_{\pm 0.02}$	$0.54_{\pm 0.01}$	$\mathbf{0.75_{\pm 0.03}}$	$\mathbf{0.58_{\pm 0.010}}$	$\mathbf{0.79_{\pm 0.01}}$	$\mathbf{0.71_{\pm 0.01}}$	$0.67_{\pm 0.01}$	$0.45_{\pm 0.01}$	0.73
AT-Dsflnt [15]	NA	NA	0.57	0.71	0.37	0.55	0.77	0.76	NA	NA	0.57
Auto-AVSR [21]	$0.57_{\pm 0.04}$	$0.33_{\pm 0.09}$	$0.56_{\pm 0.02}$	$0.54_{\pm 0.01}$	$0.55_{\pm 0.02}$	$0.3_{\pm 0.06}$	$0.57_{\pm 0.02}$	$0.43_{\pm 0.03}$	$0.55_{\pm 0.02}$	$0.31_{\pm 0.01}$	0.56
DAV-early	$0.70_{\pm 0.04}$	$0.53_{\pm 0.04}$	$\mathbf{0.77_{\pm 0.02}}$	$\mathbf{0.6_{\pm 0.04}}$	$0.73_{\pm 0.02}$	$0.56_{\pm 0.02}$	$0.8_{\pm 0.01}$	$0.72_{\pm 0.01}$	$0.68_{\pm 0.01}$	$0.47_{\pm 0.01}$	0.73
DAV-late	$0.71_{\pm 0.03}$	$0.54_{\pm 0.02}$	$0.71_{\pm 0.04}$	$0.55_{\pm 0.05}$	$0.70_{\pm 0.02}$	$0.54_{\pm 0.03}$	$\mathbf{0.79_{\pm 0.01}}$	$\mathbf{0.71_{\pm 0.01}}$	$0.67_{\pm 0.01}$	$0.46_{\pm 0.04}$	0.72
DAV-unified	$\mathbf{0.71_{\pm 0.03}}$	$\mathbf{0.55_{\pm 0.04}}$	$0.75_{\pm 0.03}$	$0.58_{\pm 0.04}$	$\mathbf{0.75_{\pm 0.01}}$	$\mathbf{0.58_{\pm 0.02}}$	$0.79_{\pm 0.02}$	$0.70_{\pm 0.03}$	$\mathbf{0.69_{\pm 0.02}}$	$\mathbf{0.51_{\pm 0.01}}$	**0.74**

baselines include the following.

Unimodal Models. In the **Audio-Only** model, we use the wav2vec 2.0 BASE feature extractor with 2D CNN layers, a transformer-based encoder, and a classifier head (upper branch in Figure 1). For the **Video-Only** model, we follow a comparable embedding extraction pipeline as outlined in Section 2.2, employing a transformer-based encoder and a classifier head (lower branch in Figure 1).

AT-Dsflnt [15]. We leverage a state-of-the-art audio-text-based multimodal disfluency detector (AT-Dsflnt) [15] with frozen layers, which uses untranscribed audio input, generates text through a custom automatic speech recognition pipeline, and models a multimodal input space for subsequent analysis. Notably, AT-Dsflnt employs a distinct training dataset and disfluency taxonomy. We establish correspondence for three disfluency tasks: filled pauses as interjections, repetitions and word repetitions, and partial words and repetitions as sound repetition; and present results while acknowledging differences in training data and taxonomy.

Auto-AVSR [21]. We leverage a lip-reading-focused audio-visual speech classification state-of-the-art model, Auto-AVSR [21], for disfluency detection applications. We follow the necessary preprocessing steps for Auto-AVSR by cropping the lip-region of the videos and conducting end-to-end training for the five disfluency tasks.

3.2. Effectiveness of Audio-Visual Multimodal Learning

Table 2 summarizes the performance of four baseline models and the three variants of our multimodal disfluency detection models. We observe that **our approach boosts the average accuracy of the disfluency detection across all tasks by 10% and up to 17% in some cases**. It is evident that visual data significantly enhances the accuracy of disfluency detection. In fact, in two of the cases, the Video-only model can perform as well as the audio-visual multimodal frameworks. This can be attributed to the paralinguistic nature of speech disfluency [36], which presents itself distinctly through facial gestures allowing the model to learn semantic disfluency properties accurately.

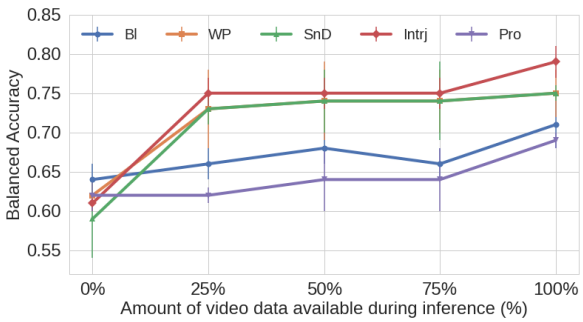

Figure 2: *Performance of our DAV-unified approach on varying availability of visual modality during inference.*

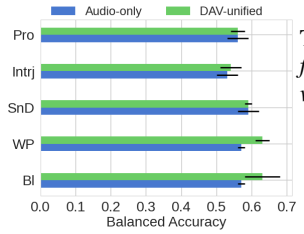

Figure 3: *Performance on zero-shot transfer to a different acoustic dataset (SEP28k).*

Table 3: *Balanced Accuracy (BA) for different types of audio and video features for Blocks.*

Features	Model	BA
Full Face	Video-only	$0.71_{\pm 0.01}$
+ wav2vec2.0	DAV-unified	$0.71_{\pm 0.03}$
Lip ROI	Video-only	$0.57_{\pm 0.02}$
+ wav2vec2.0	DAV-unified	$0.59_{\pm 0.03}$
Hubert	Audio-only	$0.58_{\pm 0.01}$
+ Full Face	DAV-unified	$0.68_{\pm 0.02}$
+ Lip ROI	DAV-unified	$0.57_{\pm 0.04}$

Resilience to missing video modality. To demonstrate the resilience of our approach to the missing of video modality, we conduct inference on test sets with varying amount of available video data as shown in Figure 2. Even with 50% of the visual data missing, our **DAV-unified offers 7% average boost in performance** over the unimodal acoustic model.

Additional Insights. We conduct a zero-shot cross-dataset transfer, evaluating models trained on the FluencyBank on SEP28k, consisting solely of audio samples. In the absence of video input for this task, our DAV-unified model matches the performance of an Audio-Only model, with up to a 6% boost in balanced accuracy, highlighting its generalization ability.

We examine the impact of various audio-visual features on speech disfluency tasks in Table 3. Visual features are extracted from both the full face and the cropped lip region of interest (ROI) [34], and audio features are from the base versions of Hubert and wav2vec 2.0. Unlike mainstream audio-visual speech recognition trends, focusing on the lip region diminishes performance, with an average degradation of 14% and 11.5% in Video-Only and DAV-unified, respectively, compared to full facial features. Reasons for this drop could be the 18% less paired data available for training the lip ROI cases due to the inability to detect landmarks and video quality issues for some samples. Moreover, the paralinguistic nature of disfluency implies potential advantages derived from a comprehensive facial context. While these findings are intriguing, it is essential to recognize that larger databases and variations in pretrained networks for visual features could impact and shape this observation.

4. Conclusion

We present a novel audio-visual multimodal learning approach for speech disfluency detection, curating a FluencyBank dataset and developing a unified weight-sharing multimodal design that is resilient to missing video modality. Our findings highlight the substantial performance gains from video, even when present partially in a multimodal framework, surpassing the commonly used acoustic-only modality for disfluency detection.

5. References

[1] E. Yairi and N. Ambrose, "Epidemiology of stuttering: 21st century advances," *Journal of fluency disorders*, vol. 38, no. 2, 2013.

[2] A. Kirkland, J. Gustafson, and E. Székely, "Pardon my disfluency: The impact of disfluency effects on the perception of speaker competence and confidence," in *24th International Speech Communication Association, Interspeech 2023, Dublin, Ireland, Aug 20 2023-Aug 24 2023*. International Speech Communication Association, 2023, pp. 5217–5221.

[3] L. Clark, B. R. Cowan, A. Roper, S. Lindsay, and O. Sheers, "Speech diversity and speech interfaces: Considering an inclusive future through stammering," in *Proceedings of the 2nd Conference on Conversational User Interfaces*, 2020.

[4] P. Mohapatra, S. Preejith, and M. Sivaprakasam, "A novel sensor for wrist based optical heart rate monitor," in *2017 IEEE international instrumentation and measurement technology conference (I2MTC)*. IEEE, 2017.

[5] O. Shonibare, X. Tong, and V. Ravichandran, "Enhancing asr for stuttered speech with limited data using detect and pass," *arXiv preprint arXiv:2202.05396*, 2022.

[6] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[7] T. Kourkounakis and A. Hajavi et. al., "Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, 2021.

[8] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Machine learning for stuttering identification: Review, challenges and future directions," *Neurocomputing*, 2022.

[9] P. Howell, S. Davis, and J. Bartrip, "The university college london archive of stuttered speech (uclass)," 2009.

[10] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, 2012.

[11] N. B. Ratner and B. MacWhinney, "Fluency bank: A new resource for fluency research and practice," *Journal of fluency disorders*, vol. 56, 2018.

[12] I. Esmaili, N. J. Dabanloo, and M. Vali, "An automatic prolongation detection approach in continuous speech with robustness against speaking rate variations," *Journal of medical signals and sensors*, vol. 7, no. 1, 2017.

[13] C. Lea, V. Mitra, A. et al.Joshi, S. Kajarekar, and J. P. Bigham, "Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[14] A. Romana and K. Koishida, "Toward a multimodal approach for disfluency detection and categorization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[15] A. Romana, K. Koishida, and E. M. Provost, "Automatic disfluency detection from untranscribed speech," *arXiv preprint arXiv:2311.00867*, 2023.

[16] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Stutternet: Stuttering detection using time delay neural network," in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021.

[17] P. Mohapatra, A. Pandey, B. Islam, and Q. Zhu, "Speech disfluency detection with contextual representation and data distillation," in *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, 2022.

[18] P. Mohapatra, B. Islam, M. T. Islam, R. Jiao, and Q. Zhu, "Efficient stuttering event detection using siamese networks," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[19] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob, "Mfcc based recognition of repetitions and prolongations in stuttered speech using k-nn and lda," in *2009 IEEE student conference on research and development (SCOReD)*. IEEE, 2009.

[20] J. C. Rocholl, V. Zayats, D. D. Walker, N. B. Murad, A. Schneider, and D. J. Liebling, "Disfluency detection with unlabeled data and small bert models," *arXiv preprint arXiv:2104.10769*, 2021.

[21] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[22] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, 2020.

[23] D. Dadebayev, W. W. Goh, and E. X. Tan, "Eeg-based emotion recognition: Review of commercial eeg devices and machine learning techniques," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, 2022.

[24] X. Zhang and Y. Li, "A dual attention-based modality-collaborative fusion network for emotion recognition." INTERSPEECH, 2023.

[25] S. Ghosh, U. Tyagi, S. Kumar, M. Suri, and R. R. Shah, "A novel multimodal dynamic fusion network for disfluency detection in spoken utterances," *arXiv preprint arXiv:2211.14700*, 2022.

[26] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang, "Multimodal emotion recognition with temporal and semantic consistency," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, 2021.

[27] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021.

[28] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[29] P. Mohapatra, A. Pandey, S. Keten, W. Chen, and Q. Zhu, "Person identification with wearable sensing using missing feature encoding and multi-stage modality fusion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[30] Q. Gao, D. W. et al.and Joshua David Amason, S. Yuan, C. Tao, R. Henao, M. Hadziahmetovic, L. Carin, and M. Pajic, "Gradient importance learning for incomplete observations," in *International Conference on Learning Representations*, 2022.

[31] S. P. Bayerl, G. R. et al.and Shammur Absar Chowdhury, T. Ciulli, M. Danieli, K. Riedhammer, and G. Riccardi, "What can Speech and Language Tell us About the Working Alliance in Psychotherapy," in *Proc. Interspeech 2022*, 2022.

[32] P. Mohapatra, A. Pandey, Y. Sui, and Q. Zhu, "Effect of attention and self-supervised speech embeddings on non-semantic speech tasks," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[33] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, 2020.

[34] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, 2022.

[35] A. Vaswani, N. Shazeer, J. Parmar, Niki et al.and Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[36] W. H. Perkins, R. D. Kent, and R. F. Curlee, "A theory of neuropsycholinguistic function in stuttering," *Journal of Speech, Language, and Hearing Research*, 1991.