Dataset Curation for Visual Speech Recognition

Claire Chen[1*] Maya Krolik [1*]
[1] Massachusetts Institute of Technology
cluchen@mit.edu mayaak@mit.edu

Abstract

Motivated by challenges in dataset curation, where inconsistencies and lack of diversity limit model benchmarking and progress, we introduce a scalable, customizable pipeline for curating high-quality datasets for visual speech recognition (VSR). As part of the pipeline, our proposed metric provides real-time feedback by identifying demographic gaps and guiding targeted augmentation during the dataset curation process. More specifically, the pipeline consists of 4 iterative steps: 1) Collect and crop videos to focus on lip regions, and generate transcripts, 2) Extract demographic features, 3) Apply metric to evaluate coverage, and 4) Assess dataset quality and perform targeted data acquisition if needed. This process produces well-distributed data and is applicable beyond lip-reading. Applying our metric to existing datasets reveals major gaps at certain demographic intersections (e.g. lack of Black, Female, senior data). Using this pipeline, we curated a dataset with roughly 2x better coverage than baseline dataset. By improving quality, consistency, and representation, we aim to democratize access to robust datasets and advance the development of more accurate, generalizable models for real-world applications.

1. Introduction

Our project addresses the difficulties of curating high-quality datasets for visual speech recognition transformer models, commonly known as lip-reading models. As detailed in section 2, the current process of collecting, processing, and filtering training data remains highly tedious, often forcing researchers to devise creative methods to compensate for the shortage of large, high-quality datasets. Furthermore, model performance has been shown to improve significantly with increased data volume [3]. However, the absence of standardized quality control metrics makes it difficult to determine whether a model's success stems from genuine architectural innovation or simply from the quantity and quality of its training data.

In this work, we introduce a scalable and customizable pipeline for curating high-quality, diverse datasets for visual speech recognition. The proposed metric, which provides immediate feedback during dataset curation by identifying gaps in demographic representation, enables targeted data augmentation to address underrepresented categories. This ensures that our pipeline produces consistent, well-distributed training data and can also be applied to domains beyond lip-reading. By improving the quality, consistency, and representativeness of training data, we aim to advance the development of more accurate, generalizable models for real-world applications.

2. Relevant Work

We identified three key papers that outline existing datasets and lip-reading models, and the associated gaps:

- **LRS3-TED** [9]: one of the largest lip-reading datasets consisting of 407 hours of TED talk videos. Although the dataset provides cropped face tracks and transcripts, it lacks an objective metric to evaluate the quality of the data, including demographic coverage. It is also unavailable to the public.

- **Auto-labeling** [3]: first trained a model on labeled data to then label unlabeled datasets such as VoxCeleb2 and AVSpeech. They used automatically generated transcriptions to train various lip-reading models. As a result, they decreased the Word Error Rate (WER) in their models, stressing the importance of data availability in their training process as key to their success.

- **Contactless Lip Reading** [2]: raise the issue of insufficient data on children and certain races resulting in models whose WER do not generalize well to the general population. They tried solving this problem by focusing on and learning physical motions of the lips using laser sensors.

Across the papers, we find the following shortcomings:

*Both authors contributed equally.

- None of the three papers creates any metric to statistically determine the diversity and quality of their datasets, relying instead on pure volume.
- [3] attempted to overcome the lack of training data by generating their own data and [2] learned ways of talking with physical equipment to overcome bias in existing data. These works show the importance of having a large, diverse dataset.

Our work bridges these gaps by providing a scalable pipeline for automnomous dataset curation. We also create a metric that evaluates a dataset's coverage to ensure that generated dataset is representative of the real world.

3. Data

Since the premise of our project is to create a pipeline from raw videos to labeled data ready for training, the data we use originate from publicly available sources. Web scraping sources such as YouTube is legal, given the original functionality of the website is not affected. We will focus on finding sources from various speakers including all racial groups, ages, and gender to address the concerns raised in [2] and [8].

3.1. Data Collection

In order to collect data, we scrape videos on Youtube using python library pytubefix. This allows us to extract the visual and audio components of a video separately. We then use Google's mediapipe and pre-trained shape and facial recognition models from dlib to locate a bounding box around the face in a video. Then, we crop the video to a square that is centered on the speaker's lips. To create a transcript, we use the base model of OpenAI's Whisper [5]. The base model is light and could process data overnight on our personal devices, with roughly the same compute time as input length.

4. Methods

Our project can be broken down into the following four sections. The first two sections are the pre-processing pipeline and the last two sections introduce the dataset evaluation metric and filtering techniques. A visual schematic of these sections can be found below in Figure 1. The entire pipeline is then wrapped inside a larger loop for automated data collection in Figure 2. Given an explanation of the desired purpose of a dataset, we use an LLM to create a YouTube search query that kickstarts the scrapping, filtering, and evaluation pipeline. It then refines the search given the missing demographic classes (as provided by our metric calculation; we used OpenAI's 4o model and less than $1 of credits). We also vector embed new data points to filter for noisy data acquisition. This process is also illustrated below. All of our code has been made publicly available on github.

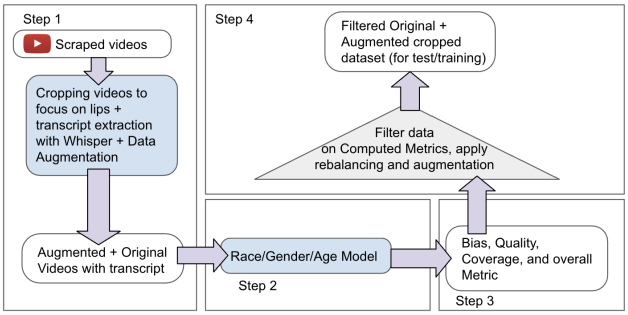

Figure 1. Project Pipeline for Data Scrapping and Filtering

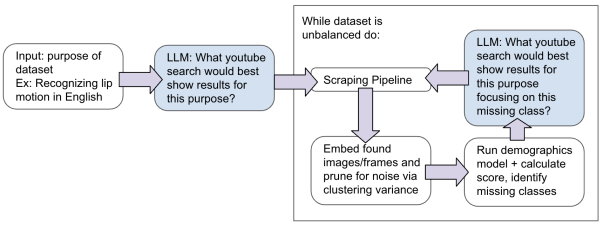

Figure 2. Automatic Data Collection Pipeline with LLM

4.1. Automated Lip Tracking and Transcripts

We use mediapipe [4] and dlib to extract the lip region commonly used as inputs to lip-reading models [3] and run OpenAI's Whisper [5] locally to generate synchronized transcripts, as detailed in subsection 3.1. All videos and audios are broken down into 30 second chunks to ensure consistency and to simplify our quality score calculation.

4.2. Feature Extraction and Segmentation

To extract and process key visual features that are critical for determining the diversity and quality of a dataset, we refer to models that determine race, age, and gender [1]. We carefully chose a model called FairFace that was intentionally trained on balanced data, as to not bias our labels. These features (race, age, gender) are the inputs for our metric outlined below.

4.3. Quality Assessment Metric

The goal of the metric is to ensure balanced representation across demographic categories, as well as a sufficient number of samples in each subcategory (list of categories and subcategories are shown in the table below). Hence, the metric is composed of a coverage score (CS) and a minimum representation constraint.

4.3.1 Cross-Category Coverage Score (CS)

A challenge when evaluating dataset diversity is that simple per-category metrics — such as averaging scores across individual demographic dimensions (e.g., race, gender, or age separately) — can mask severe underrepresentation in cross-sections of the population. For instance, a dataset might contain a large number of Asian males and White females, which would result in high marginal diversity scores when considered independently for race or gender. However, such a dataset could still lack groups like Asian females or White males; simple averages would yield deceptively high scores, masking these critical gaps.

Category	Example Sub-Category
Race	White, Hispanic, Black, Asian, Other
Gender	Male, Female, Non-binary
Age	Child, Adolescent, Adult, Senior

Figure 3. Proposed Classes and their Sub-Categories

To address this, we introduce the **Cross-Category Coverage Score (CS)**. This metric systematically evaluates all possible demographic intersections (i.e. every combination of race, gender, and age).

The CS is computed as follows:

- Define all possible cross-category groups G as the Cartesian product of demographic categories (e.g., Race $\times$ Gender $\times$ Age).
- For each group, compute the raw number of samples (allowing for counts of zero).
- Normalize each group's count by dividing by the maximum group count across all combinations:

  $$\text{Normalized coefficient for group } g = \frac{\text{count}(g)}{\max_{g'} \text{count}(g')}$$

- Finally, compute CS as a weighted combination of:
  - the minimum normalized coefficient (worst-case coverage)
  - and the average normalized coefficient (overall coverage quality)

$$CS = 0.5 \times \min_{g \in G} \left(\frac{\text{count}(g)}{\max_{g'} \text{count}(g')} \right)$$
$$+0.5 \times \frac{1}{|G|} \sum_{g \in G} \left(\frac{\text{count}(g)}{\max_{g'} \text{count}(g')} \right)$$

Weighing both equally ensures that:

- Datasets are penalized if any group is severely underrepresented (through the minimum term).
- Datasets are also rewarded/penalized for good/bad overall distribution (through the average term).

To make the metric actionable:

- If the overall CS falls below a user-defined threshold (e.g. 0.6), the dataset is flagged for insufficient cross-category coverage.
- Any specific cross-category groups with normalized coefficients below a lower threshold (e.g. 0.2) are explicitly listed. These low-coverage groups can then be used to prompt a data augmentation system (e.g. our proposed LLM) to target data collection for the most underrepresented subpopulations.

Example calculation: Consider the following observed raw counts of a simplified dataset with only Race (white, Asian) and Gender (male, female):

Category	**Raw Count**	**Normalized Coeff**
(White, Male)	2	0.4
(White, Female)	5	1
(Asian, Male)	3	0.6
(Asian, Female)	0	0

The minimum normalized coefficient is 0 since we are completely missing data from the (Asian, Female) category, while the average normalized coefficient is 0.5. So overall CS is 0.25. Since $0.25 < 0.7$, the dataset would be flagged. Furthermore, (Asian, Female) would be flagged as a low-coverage group.

4.3.2 Minimum representation constraint

While the coverage score (CS) accounts for diversity and fairness, it does not guarantee that the dataset is sufficiently large for training. For example, a dataset with very few samples (e.g., 1 or 2 samples) that are evenly distributed among cross-categories could still achieve a high CS, but would not be suitable for robust training and evaluation.

Hence, we impose a minimum representation constraint to ensure that every cross-category group (i.e., combinations of race, gender, and age group) has a sufficient number of samples:

$$\min_{(i,j,k)} (\text{count}_{ijk}) \geq T$$

- count_{ijk} is the number of samples in the group defined by the ith race, jth gender, and kth age category.
- T is a predefined minimum threshold for each cross-category group.

A dataset that does not satisfy this constraint is flagged as underrepresented and would require additional data collection. This ensures that the dataset is not only diverse and fair across demographic categories, but also sufficiently large across all cross-category intersections for proper training and evaluation.

4.4. Dataset Filtering and Evaluation

We use the computed coverage score (CS) to systematically assess the diversity and fairness of a dataset. A dataset is flagged for low coverage if its score $CS < T_{CS}$, where T_{CS} is a predefined threshold reflecting acceptable diversity standards set by the user. An additional feature is that, using normalized coefficients, it identifies specific underrepresented cross-categories with low scores, such as (Asian, Female), making it easy to pinpoint which demographic combinations require remediation. We also inspect individual group representations; a dataset is additionally flagged if any cross-category group fails to meet a minimum number of samples required for reliable training and evaluation.

4.4.1 Data Augmentation

If the dataset is flagged, we recommend the following techniques to address the issue:

- **Data augmentation** for underrepresented categories by rotating, flipping videos.
- **Rebalance** the dataset to include more samples from underrepresented categories.
- **Targeted data acquisition** to find data that fills in identified demographic gaps using an LLM.

Our pipeline LLM integrates OpenAI's GPT 4o to automate the data augmentation process by taking in as input the objective of the dataset and the missing classes and acting as a "data scientist" to design search queries for YouTube to be fed into our scrapping pipeline. This reduces manual curation efforts and enables scalable dataset improvement over multiple cycles.

We perform a vector embedding step here to assess image quality by checking whether our scraped information is consistent and what we are actually looking for semantically. For example, if the goal were to be to collect videos of people running, the embedding of people running should cluster rather close together, allowing us to identify out of distribution examples of, for example, people eating. To do this, we run our cropped images through the encoding layers of FaceNet [7] using Euclidean distance as a proxy for semantic similarity. Additionally, we run a clustering algorithm (with two clusters: faces and non-faces) to identify all frames that do not contain faces in them. As an example, Figure 4 shows UMAP projection of these embeddings isolating non-faces.

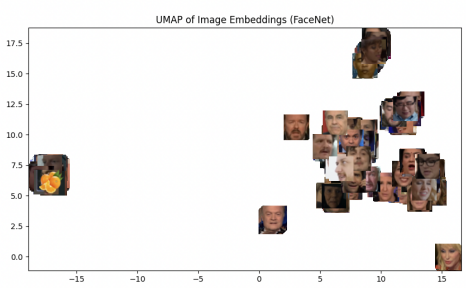

Figure 4. Vector Embedding of Images using FaceNet

4.4.2 Performance Evaluation of the Pipeline

To quantitatively assess the impact of our pipeline, we track the CS after each iteration of dataset augmentation and rebalancing. By monitoring improvements in CS and checking satisfaction of the minimum representation constraint at each step, we systematically evaluate whether the quality, diversity, and fairness of the dataset improve over time. Note that we do not assess the quality of individual preprocessing stages (such as trimming or demographic categorization), as these components rely on existing, prevalidated models. Our focus is on the effectiveness of the entire dataset curation pipeline and improvement process.

We applied the metric and evaluation techniques to the lip-reading data we scrapped to demonstrate performance. However, the general pipeline can be easily applied to any other dataset to ensure a high overall coverage.

5. Results

A pipeline bug in our progress report caused us to underestimate curation time—each video took roughly its full duration to process due to sequential steps: download, lip crop, transcription, and demographic analysis. Our dataset now includes roughly 2 hours of video (results below).

We applied our coverage metric to a preliminary dataset of 80 minutes of videos scraped from YouTube. Before any improvements using LLM-guided data acquisition, the coverage score (4.3.1) was notably low, at 0.07. Some of the most severely underrepresented groups included (Black, Female, Senior) and (Hispanic, Male, Adult). These results are consistent with findings from 4.3.2 on minimum sample constraints, where (Black, Female, Senior) and (Hispanic, Male, Adult) had 0 samples. Figure 5, Figure 6, Figure 7 are matrices visualizing the normalized coefficient for the dataset (note: since visualizing three categories simultaneously in 3D is difficult, we present cross-sectional 2D matrices for each category pair.) The heatmaps reveal significant underrepresentation of Black and Hispanic speakers relative to White speakers for both genders. Additionally, the dataset is overall biased towards female speakers, containing very little male adults and children.

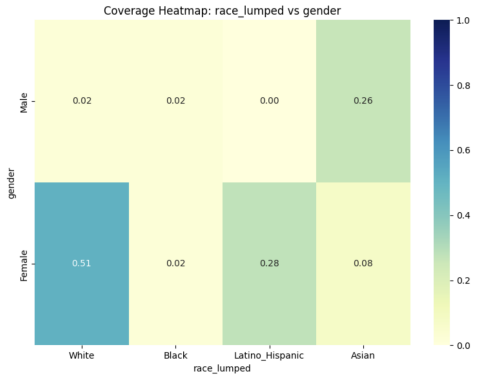

Figure 5. Race vs Gender Heatmap of Initial Dataset

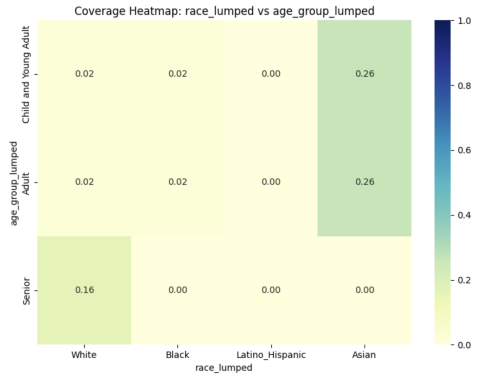

Figure 6. Race vs Age Heatmap of Initial Dataset

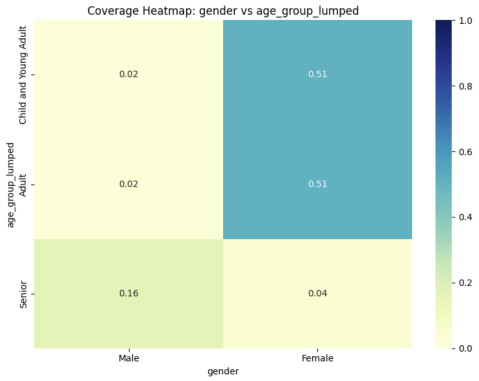

Figure 7. Gender vs Age Heatmap of Initial Dataset

To address these deficiencies, we used our LLM-assisted data retrieval process to scrape additional targeted samples for underrepresented cross-categories. After 3 iterations and using 120 minutes of Youtube videos, the coverage score increased from 0.07 to 0.11. (Black, Female, Senior) and (Hispanic, Male, Adult) still had low normalized coefficients, but increased from 0 to 0.003 and 0.015. The number of samples in those categories also increased from 0 to 1 and 5 samples. The updated heatmaps Figure 8, Figure 9, and Figure 10 show improvement from the previous heatmaps, where coefficients have increased to show greater coverage across the cross-categories. Note that there are still more data in some categories than others, likely due to the imbalanced distribution of the demographic categories on Youtube; running more iterations and applying data augmentation could mediate this issue.

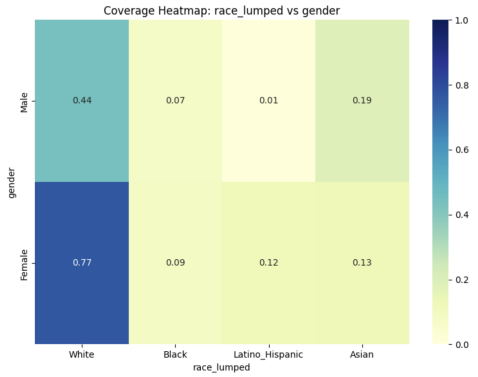

Figure 8. Race vs Gender Heatmap of Updated Dataset

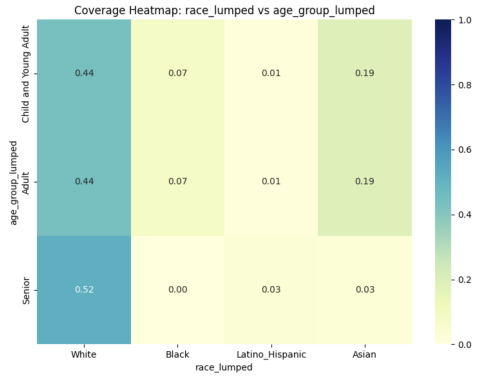

Figure 9. Race vs Age Heatmap of Updated Dataset

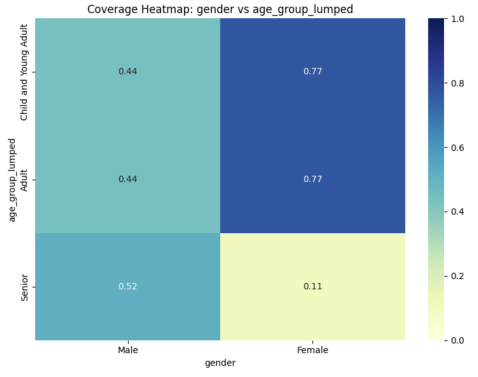

Figure 10. Gender vs Age Heatmap of Updated Dataset

Additionally, we applied our coverage metric to the MIRACL-VC1 dataset [6], the largest publicly available lip-reading dataset we could access. MIRACL-VC1 achieved a coverage score of 0.06, even lower than our preliminary dataset. The corresponding normalized coefficient matrices are shown in Figure 11, Figure 12, and Figure 13.

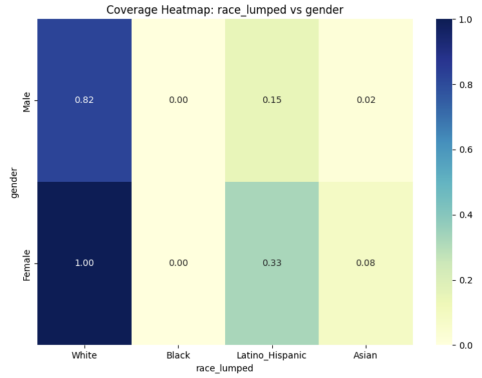

Figure 11. Race vs Gender Heatmap of MIRACL-VC1

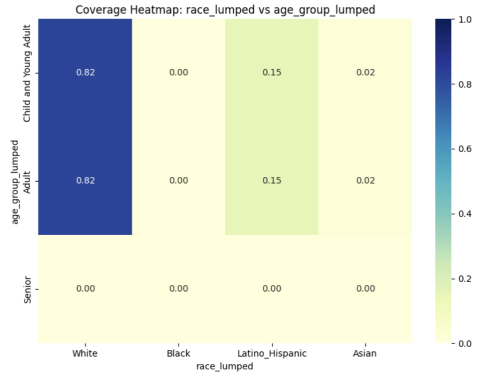

Figure 12. Race vs Age Heatmap of MIRACL-VC1

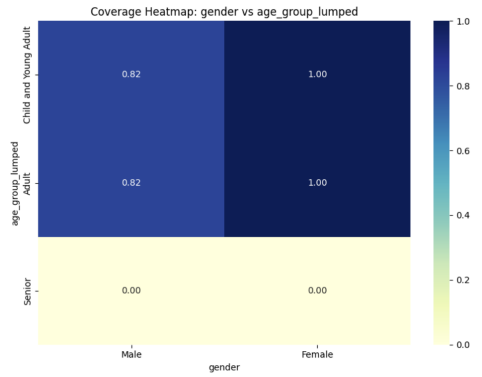

Figure 13. Gender vs Age Heatmap of MIRACL-VC1

The score and heatmaps clearly show that MIRACL-VC1 entirely lacks data for Black speakers and senior individuals, with extremely low coefficients for Asians as well. In contrast, the dataset is dominated by samples of White females, children, and adults. This imbalance is reflected in the low coverage score and highlights the importance of enforcing diversity metrics during dataset curation. Through visualizations, we can also see that the categories are much more concentrated/less spread out than the LLM-updated dataset.

Our final curated dataset has been made publicly available on kaggle.

6. Practical Implications

Our initial project motivation was to train a model for VSR tasks since VSR tasks are applicable towards topics ranging from espionage to accessibility and improved speech-to-text models. However, as the project progressed, we realized that the availability of the data was much worse than anticipated, demonstrating a demand for high-quality datasets that could contribute to the training of lip-reading models. Even beyond VSR, the problem of balanced and high-quality datasets is core to all of computer vision and machine learning. Our pipeline helps address this problem by creating an automatic web crawler to collect and filter diverse data. By lowering the barrier to consistent, high-quality dataset creation, we help standardize model evaluation and enable more meaningful comparisons across architectures. This also leads to models that are more generalizable in real-world settings with great demographic diversity.

7. Conclusion

In this project, we developed a fully automated pipeline for curating class-balanced, high-quality datasets tailored to user-specified tasks. By simply specifying a target objective (e.g., "recognizing lip motion in English" or "analyzing human running patterns"), users can generate real-world datasets with minimal manual intervention and without sacrificing diversity or quality. Our modular design allows individual components—such as detectors, cropping models, or transcription tools—to be easily swapped for different domains. Additionally, our cross-category coverage metric provides a scalable way to evaluate dataset balance across a wide range of computer vision applications, not just for lipreading. Given the central role of data quality in VSR and machine learning more broadly, we hope this work enables us to separate the effects of data quality from those of model architecture, and supports the creation of more generalizable, equitable models for real-world deployment.

8. Individual Contribution

Our tasks in this project were completed roughly according to the split below.

Claire: Designed project and worked on writeup. Designed quality assessment metrics and dataset filtering. Implemented metric calculation and result visualization.

Maya: Designed project and worked on writeup. Implemented pre-processing pipeline and LLM automation, integrating with metric to automate the process.

References

[1] Elvina Ardelia, Jericho Thenando, Alexander A S Gunawan, and Muhammad E Syahputra. Predicting age and gender across different races using convolutional neural networks: A deep learning approach. In *2024 International Conference on Information Technology and Computing (ICITCOM)*, pages 150–154, 2024. 2

[2] Yao Ge, Chong Tang, Haobo Li, Zikang Chen, Jingyan Wang, Wenda Li, Jonathan Cooper, Kevin Chetty, Daniele Faccio, Muhammad Imran, and Qammer H. Abbasi. A comprehensive multimodal dataset for contactless lip reading and acoustic analysis. *Scientific Data*, 10(1), Dec. 2023. 1, 2

[3] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-avsr: Audio-visual speech recognition with automatic labels. *arXiv:2303.14307*, 2023. 1, 2

[4] Jinwoo Nam. Dlib-lip-detection. https://github.com/skaws2003/Dlib-lip-detection, 2018. [Accessed 03-04-2025]. 2

[5] OpenAI. Introducing Whisper. https://openai.com/index/whisper/, 2022. [Accessed 03-04-2025]. 2

[6] Apoorv Patn. Lip reading image dataset. https://www.kaggle.com/datasets/apoorvwatsky/miraclvc1, 2020. [Accessed 12-05-2025]. 6

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. 2015. 4

[8] Changchong Sheng, Gangyao Kuang, Liang Bai, Chenping Hou, Yulan Guo, Xin Xu, Matti Pietikäinen, and Li Liu. Deep learning for visual speech analysis: A survey, 2022. 2

[9] Andrew Zisserman Triantafyllos Afouras, Joon Son Chung. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv:1809.00496v2*, 2018. 1