Chinese Lip Reading Based on Audio-Visual Hidden Unit BERT

指導教授：陳冠文
組別：第7組
組員：110550165何存益 110550085房天越

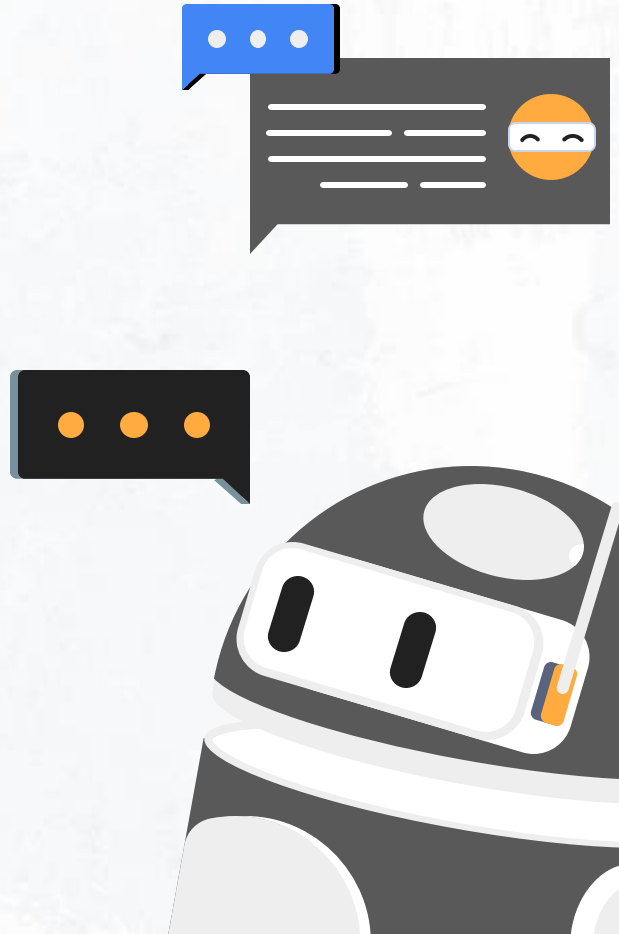
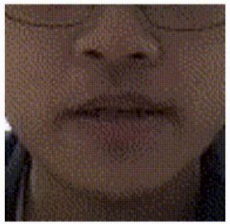


Table of Content

Introduction

- Traditional lip-reading models are typically trained using only visual datasets.
- AV-HuBERT[1] employs self-supervised learning to cluster **audio and visual features** into hidden units, which significantly enhances lip-reading accuracy.
- In our project, we adapted AV-HuBERT for the **Chinese language** by training it on the CMLR dataset[2], a Chinese audio-visual dataset.
- Our model demonstrated its ability to distinguish test videos without relying on audio input.



AV-HuBERT → 交通大學將會取得勝利

Method

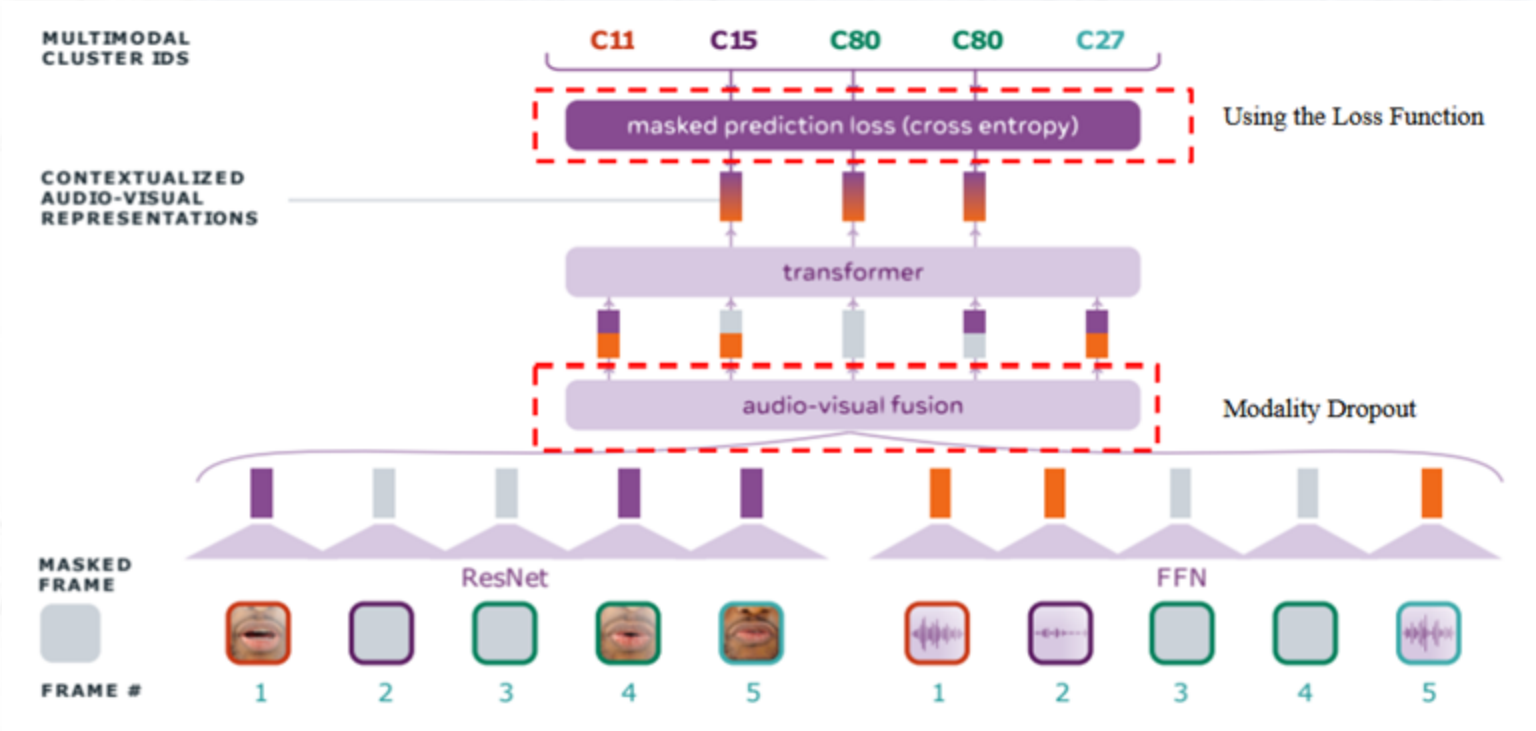


image from Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction

Method

Modality Dropout

$$\mathbf{f}_t^{av} = \begin{cases} \texttt{concat}(\mathbf{f}_t^a, \mathbf{f}_t^v) & \text{with } p_m \\ \texttt{concat}(\mathbf{f}_t^a, \mathbf{0}) & \text{with } (1-p_m)p_a \\ \texttt{concat}(\mathbf{0}, \mathbf{f}_t^v) & \text{with } (1-p_m)(1-p_a) \end{cases}$$

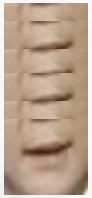
Training Loss Function

$$L = -\sum_{t \in M^a \cup M^v} \log p_t(z_t) - \alpha \sum_{t \notin M^a \cup M^v} \log p_t(z_t)$$

Method - CMLR Preprocessing

Video

Detect landmark
Extract ROI

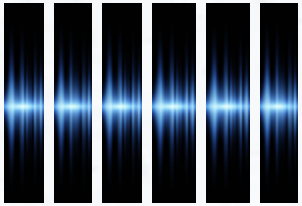


Label	Text
0	推動
1	社會
2	的
3	經濟
4	穩定
5	發展

Input Video Frame

Audio

Turn to 16000Hz
and mono



Label 0 1 2 3 4 5

Text 推動 社會 的 經濟 穩定 發展

Input Audio Frame

Method - Fine-tune

Finally, we fine-tuned our model, enabling it to predict Chinese content.

```
14340.3s    8070      Reference: 歷史悠久 參與 可以 夏天
14340.3s    8071      Prediction: 歷史人 的 第 上海
14340.3s    8072      CER: 61.54%
14340.3s    8073    Sample 27:
14340.3s    8074      Reference: 發達國家 正視 部門 國家
14340.3s    8075      Prediction: 發展中國家 劇 了 貿電
14340.3s    8076      CER: 61.54%
14340.3s    8077    Sample 28:
14340.3s    8078      Reference: 抗震救災 渡過 堅強 幹部
14340.3s    8079      Prediction: 抗震救災 工作 溫家寶 總
14340.3s    8080      CER: 53.85%
14340.3s    8081    Sample 29:
14340.3s    8082      Reference: 和平 主要 印
14340.3s    8083      Prediction: 和平 也 更多
14340.3s    8084      CER: 57.14%
14340.3s    8085    Sample 30:
14340.3s    8086      Reference: 哥本哈根 出版 企業 利益
14340.3s    8087      Prediction: 哥本哈根 會議 即將 召開
14340.3s    8088      CER: 46.15%
14340.3s    8089    Sample 31:
14340.3s    8090      Reference: 山西省 在 加大 反
14340.3s    8091      Prediction: 山西省 結合 的合
```

▲Part of the results

Method - Generate Vocabulary Table

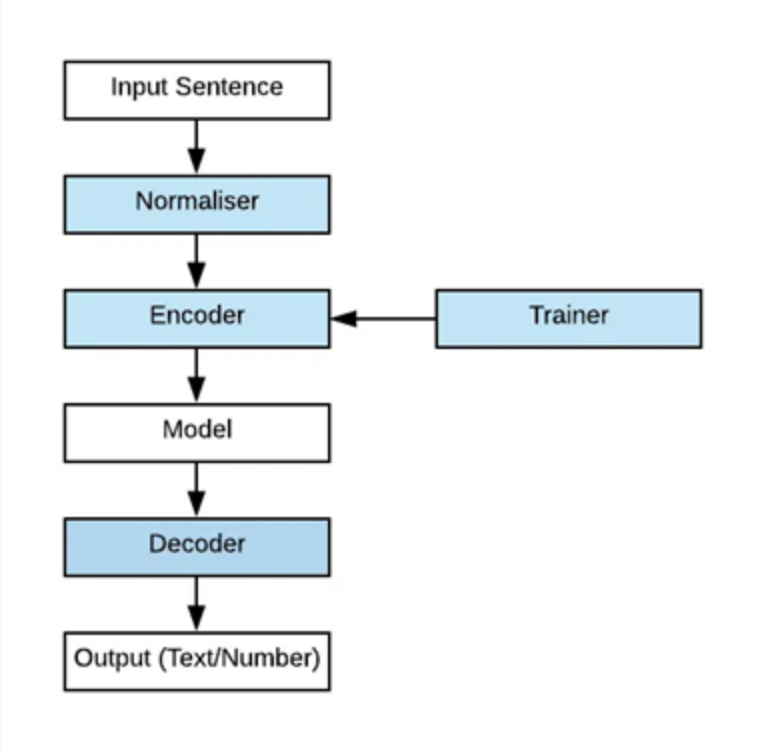
- Unlike English, where words are naturally separated by spaces and can be easily tokenized, Chinese lacks such delimiters, making segmentation more challenging.
- We need to find a more suitable tokenization method for Chinese.

SentencePiece

Method - SentencePiece

- SentencePiece processes all input by converting it into Unicode characters, enabling unified handling across multiple languages.

```
"\u6539\u9769 \u628a \u515a\u4e2d",
"\u6539\u9769 \u9886\u57df \u9ad8",
"\u4e60\u8fd1\u5e73 \u53d1\u5c55 \u4eca\u5929",
"\u65b0\u95fb\u8054\u64ad \u8bda \u5c31 \u97e9 \u64a4\u56de",
"\u4e60\u8fd1\u5e73 \u8fdb\u884c \u96e8 \u5b66",
```



Result - Generate Vocabulary Table

Due to the special structure of Chinese language, we use Character Error Rate (CER) instead of Word Error Rate (WER) that is mentioned in the paper.

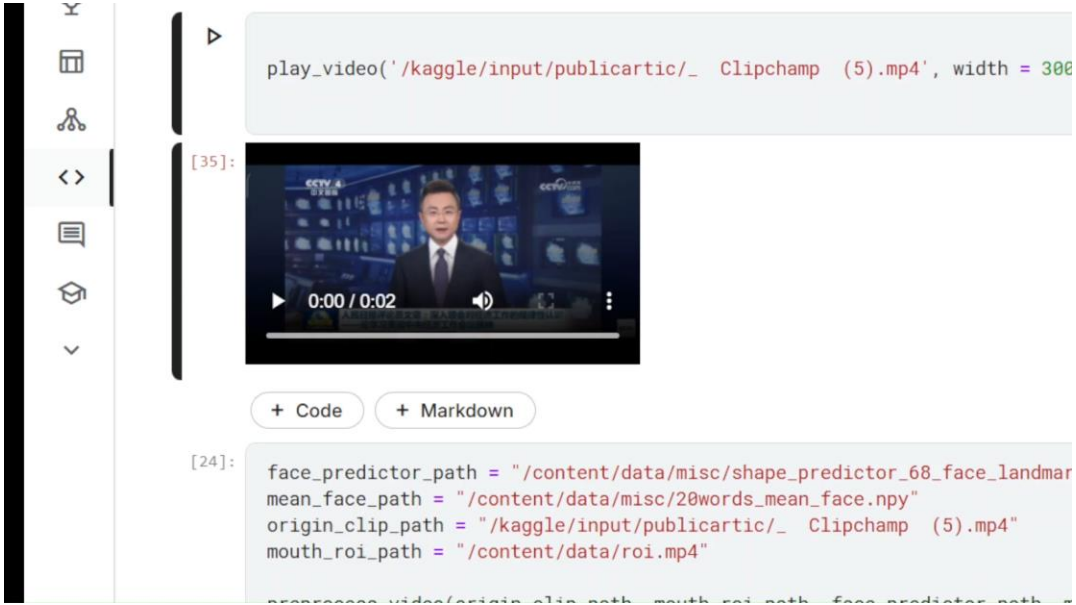
	Sentence Tokenize (Unigram)	the word in CMLR	Word tokenize
CER↓	83.27%	65.07%	**56.35%**
Part of Vocab	深切的哀悼 1 发 1 分才 1 猛 1 理念新思想新战略 1 的十八大代表 1 碑 1 第三届 1 筋 1 简政放权放管结合优化服务改革 1 纠 1 组合拳 1 经历了 1 网络攻击 1 美丽乡村 1	具有 1 法律 1 交通 1 主要 1 发展的 1 造成 1 _他们 1 可 1 引 1 时代 1 增强 1 严 1 访 1 兵 1	_国家 1 _各 1 _人民 1 _全 1 会 1 区 1 大 1 _以 1 _地 1 _社会 1 _要 1 部 1

Result - Video vs. Audio-Visual

Fine-tune	With video only	With audio and video
Character Error Rate (CER)↓	79.11%	**56.35%**

- The experiment demonstrates that incorporating both audio and video modalities during fine-tuning significantly improves performance compared to using video alone.

Result - Demo Video

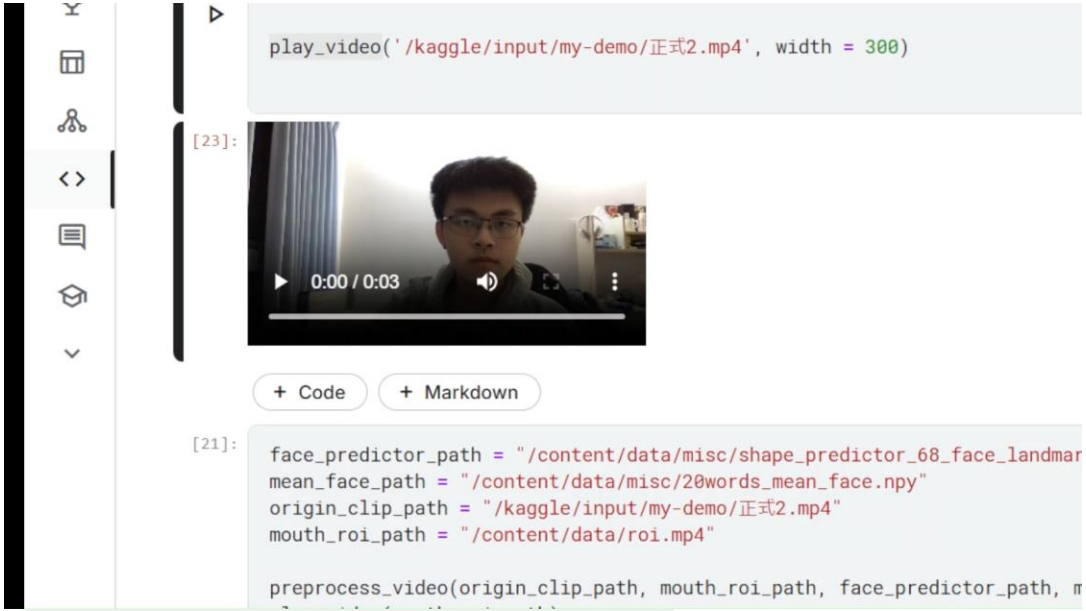


[Link]

Prediction - 人民日报将发表评论员

Ground Truth - 人民日报发表评论员文章

Result - Demo Video



Link

Prediction - 澳门特别行政区发风各一

Ground Truth - 澳門特別行政區改革開放

Result - Some Problems #1

- Chinese has more similar-sounding words and homophones than English, making it more difficult for the model to distinguish.

Reference: 數量
Prediction: 樹量

Reference: 智利
Prediction: 治理

Reference: 戰線
Prediction: 展現

Reference: 留守
Prediction: 留首

Reference: 涼山
Prediction: 良善

Reference: 合理
Prediction: 合力

Result - Some Problems #2

- The model may adjust the prediction word order for context, which results in higher CER.

Reference: 清華大學 信息 家

Prediction: 清華大學 交 信息

Reference: 製造 力量 農

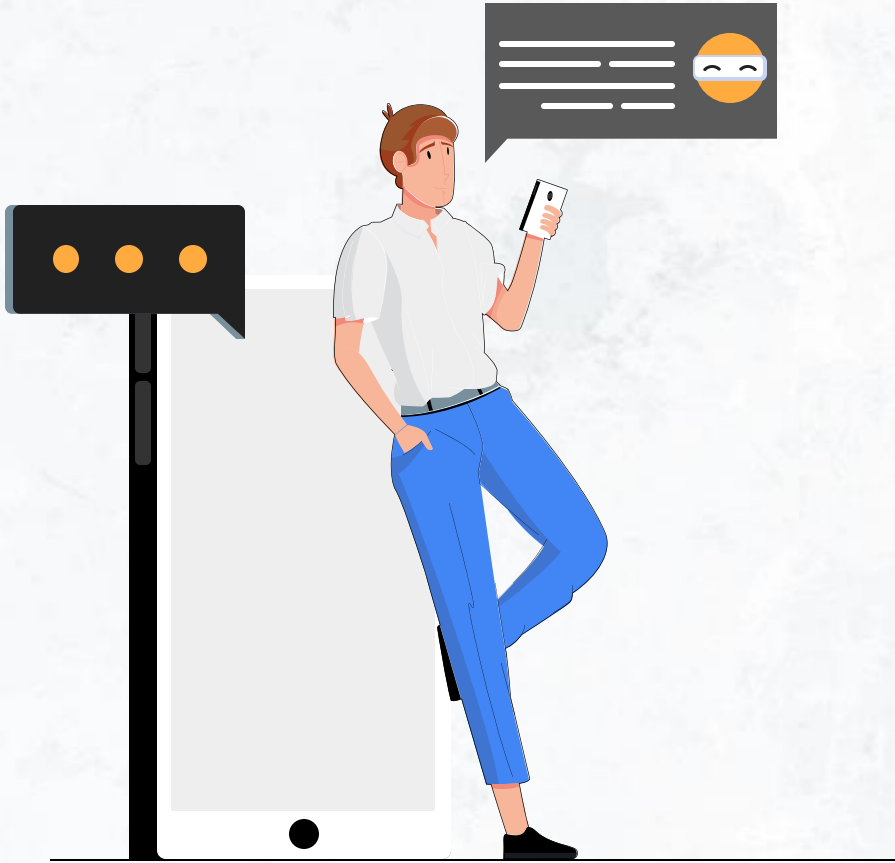
Prediction: 製造 的 力量

Conclusion

- We trained the AV-HuBERT model based on CMLR dataset to predict the Chinese language.
- We discovered that different methods of constructing vocabulary lists can affect the accuracy.
- We showed that using Audio-Visual instead of only visual data can significantly decrease the character error rate for Chinese language.

Thank you!

CREDITS: This presentation template was created by **Slidesgo** and includes icons by **Flaticon**, infographics & images by **Freepik** and content by **Eliana Delacour**



Reference

[1] AV-HuBERT (Audio-Visual Hidden Unit BERT)

https://github.com/facebookresearch/av_hubert

[2] Ya Zhao, Rui Xu, and Mingli Song. A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading. ACM International Conference on Multimedia in Asia 2019

https://dl.acm.org/doi/pdf/10.1145/3338533.3366579

[3] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, Abdelrahman Mohamed. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction, In ICLR 2022

https://arxiv.org/pdf/2201.02184