VALLR: Visual ASR Language Model for Lip Reading

Supplementary Material

1. Introduction

This supplementary material provides additional qualitative results from our proposed network architecture, evaluated on the LRS3 dataset [1]. We include video examples referenced as exampletranslation1.mp4 and exampletranslation2.mp4 to demonstrate the effectiveness of the method in performing silent video captioning. These examples overlay the predicted captions on the original videos to offer an intuitive understanding of the predictions. Furthermore, we provide detailed analysis of common errors in phoneme-to-word reconstruction to identify limitations and strengths of the model. The additional results in this document are organized in three parts:

Tab. 1: Common errors in phoneme → word predictions, such as homophones (too vs. to) or under-represented words (sunflower misinterpreted as son and flower)

Tab. 2: Challenging cases, including unseen names (e.g., Kofi Annan), demonstrating areas for future improvement with larger-scale pretraining.

Tab. 3: Correct phoneme and word predicitions.

These results complement the main paper and showcase the robustness of our approach while highlighting areas for refinement.

2. Qualitative Results

In this section, we provide additional qualitative results in two parts. Errors in the phoneme-to-word reconstruction process are highlighted in red to draw attention to specific areas where the network requires improvement.

2.1. Part 1: Common Errors

In Tab. 1, we observe that common errors include substitutions of homophones like too and to, or there and their. Similarly, in examples involving the word sunflower, the model predicts son and flower as separate words. These errors are likely due to the absence of the compound word sunflower in the fine-tuning dataset, while the individual words son and flower are well-represented. Despite these mistakes, the resulting text remains logical and understandable.

2.2. Part 2: Challenging Cases

In Tab. 2, we highlight challenging cases such as the omission of Kofi Annan. This demonstrates the model's difficulty in reconstructing previously unseen names during phoneme-to-word mapping. Such issues could be mitigated with additional pretraining on larger and more diverse text datasets.

2.3. Part 3: Correct Predictions

In Tab. 3, we showcase examples where the model successfully predicted the phoneme → word mappings without errors. These results demonstrate the model's capability to reconstruct accurate text from silent video inputs in scenarios with strong phoneme-word correlations and sufficient representation in the fine-tuning dataset.

References

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

File	Video → Phoneme (Errors in red)	Phoneme → Word (Errors in red)	Ground Truth (GT)
1.mp4	[’W’, ’AH’, ’N’, ’D’, ’EY’, ’AH’, ’Y’, ’NG’, ’B’, ’OY’, ’K’, ’AH’, ’M’, ’S’, ’AH’, ’P’, ’AA’, ’DH’, ’AH’, ’S’, ’AH’, ’N’, ’F’, ’L’, ’AW’, ’ER’, ’W’, ’AY’, ’L’, ’V’, ’IH’, ’S’, ’IH’, ’T’, ’IH’, ’NG’, ’DH’, ’AH’, ’G’, ’AA’, ’R’, ’D’, ’AH’, ’N’, ’AH’, ’N’, ’D’, ’HH’, ’IY’, ’N’, ’OW’, ’T’, ’AH’, ’S’, ’AH’, ’Z’, ’HH’, ’AW’, ’W’, ’IY’, ’K’, ’IH’, ’T’, ’L’, ’UH’, ’K’, ’S’]	ONE DAY A YOUNG BOY COMES UP ON THE SON FLOWER WHILE VISITING THE GARDEN AND HE NOTICES HOW WEEK IT LOOKS	ONE DAY A YOUNG BOY COMES UPON THE SUNFLOWER WHILE VISITING THE GARDEN AND HE NOTICES HOW WEAK IT LOOKS
2.mp4	[’JH’, ’AH’, ’S’, ’T’, ’L’, ’AY’, ’K’, ’R’, ’IY’, ’CH’, ’IH’, ’NG’, ’AW’, ’T’, ’T’, ’UW’, ’DH’, ’AH’, ’S’, ’AH’, ’N’, ’F’, ’L’, ’AW’, ’ER’, ’M’, ’AY’, ’P’, ’R’, ’AH’, ’V’, ’AY’, ’D’, ’IH’, ’NG’, ’S’, ’AH’, ’M’, ’W’, ’AH’, ’N’, ’HH’, ’UW’, ’IH’, ’Z’, ’K’, ’AH’, ’G’, ’L’, ’EH’, ’K’, ’T’, ’AH’, ’D’, ’AY’, ’S’, ’AH’, ’L’, ’EY’, ’T’, ’AH’, ’D’, ’AO’, ’R’, ’F’, ’ER’, ’G’, ’AA’, ’T’, ’AH’, ’N’]	JUST LIKE REACHING OUT TO THE SON FLOWER BY PROVIDING SOME ONE WHO IS NEGLECTED ISOLATED OR FOR GOT TEN	JUST LIKE REACHING OUT TO THE SUNFLOWER BY PROVIDING SOMEONE WHO IS NEGLECTED ISOLATED OR FORGOTTEN

Table 1. Common errors in phoneme → word predictions, including homophones and compound words.

File	Video → Phoneme (Errors in red)	Phoneme → Word (Errors in red)	Ground Truth (GT)
6.mp4	[’K’, ’OW’, ’F’, ’IY’, ’AE’, ’N’, ’AH’, ’N’, ’S’, ’EH’, ’D’, ’DH’, ’IH’, ’S’, ’W’, ’IH’, ’L’, ’B’, ’IY’, ’M’, ’EH’, ’N’, ’AH’, ’F’, ’IH’, ’SH’, ’AH’, ’L’, ’T’, ’UW’, ’M’, ’AY’, ’T’, ’R’, ’UW’, ’P’, ’Z’, ’AA’, ’N’, ’DH’, ’AH’, ’G’, ’R’, ’N’, ’D’]	NAN SAID THIS WILL BE BENEFICIAL TOO MY TROOPS ON THE GROUND	KOFI ANNAN SAID THIS WILL BE BENEFICIAL TO MY TROOPS ON THE GROUND
7.mp4	[’AH’, ’L’, ’T’, ’AH’, ’M’, ’AH’, ’T’, ’L’, ’IY’, ’DH’, ’AE’, ’T’, ’S’, ’W’, ’AH’, ’T’, ’IH’, ’T’, ’Z’, ’AH’, ’AW’, ’T’]	ULTIMATE LEE THATS WHAT ITS ABOUT	ULTIMATELY THAT’S WHAT IT’S ABOUT

Table 2. Challenging cases, including omissions of names and complex phoneme → word mappings.

File	Video → Phoneme	Phoneme → Word	Ground Truth (GT)
3.mp4	[’IH’, ’N’, ’M’, ’AY’, ’F’, ’EY’, ’TH’]	IN MY FAITH	IN MY FAITH
4.mp4	[’AY’, ’TH’, ’IH’, ’K’, ’DH’, ’AH’, ’K’, ’AE’, ’M’, ’ER’, ’AH’, ’IH’, ’S’]	I THINK THE CAMERA IS	I THINK THE CAMERA IS
5.mp4	[’DH’, ’IH’, ’S’, ’M’, ’AH’, ’S’, ’T’, ’B’, ’IY’, ’K’, ’R’, ’IY’, ’EY’, ’T’, ’D’]	THIS MUST BE CREATED	THIS MUST BE CREATED
8.mp4	[’AE’, ’K’, ’CH’, ’UW’, ’AH’, ’L’, ’IY’, ’Y’, ’UW’, ’ER’]	ACTUALLY YOU ARE	ACTUALLY YOU ARE
9.mp4	[’DH’, ’EY’, ’IH’, ’N’, ’V’, ’EH’, ’N’, ’T’, ’AH’, ’D’, ’DH’, ’AE’, ’T’, ’T’, ’R’, ’AH’, ’D’, ’IH’, ’SH’, ’AH’, ’N’, ’F’, ’AO’, ’R’, ’DH’, ’EH’, ’R’, ’ER’, ’AY’, ’V’, ’AH’, ’L’, ’HH’, ’IY’, ’R’]	THEY INVENTED THAT TRADITION FOR THEIR ARRIVAL HERE	THEY INVENTED THAT TRADITION FOR THEIR ARRIVAL HERE
10.mp4	[’T’, ’AY’, ’D’, ’IY’, ’M’, ’UW’, ’T’, ’S’, ’IH’, ’S’, ’V’, ’EH’, ’R’, ’IY’, ’F’, ’AH’, ’S’, ’IY’, ’AH’, ’B’, ’AW’, ’T’, ’HH’, ’IH’, ’Z’, ’F’, ’UH’, ’T’, ’W’, ’EH’, ’R’]	TIDY BOOTS IS VERY FUSSY ABOUT HIS FOOTWEAR	TIDY BOOTS IS VERY FUSSY ABOUT HIS FOOTWEAR

Table 3. Examples of correct phoneme → word predictions. These results showcase the model’s ability to caption silent videos with high accuracy.