Do Slides Help? Multi-modal Context for Automatic Transcription of Conference Talks

Anonymous ACL submission

Abstract

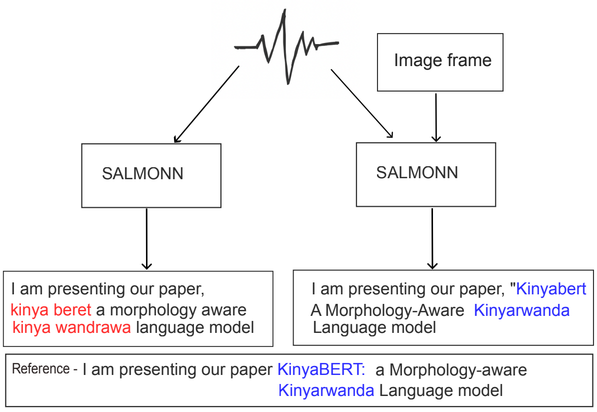
State-of-the-art (SOTA) Automatic Speech Recognition (ASR) systems mainly rely on acoustic information while disregarding additional multimodal context. However, visual information are essential in disambiguation and adaptation.

While most work focuses on speaker images to handle noise conditions, this work also focuses on integrating presentation slides for the use cases of scientific presentation.

In a first, we create a benchmark for multimodal presentation including an automatic analysis of transcribing domain-specific terminology. Next, we explore methods for augmenting speech models with multi-modal information. We mitigate the lack of datasets with accompanying slides by a suitable approach of data augmentation. Finally, we train a model using the augmented dataset, resulting in a relative reduction in word error rate of approximately 49%, across all words and 15%, for domain-specific terms compared to the baseline model.

1 Introduction

Automatic Speech Recognition (ASR) like many other NLP tasks are currently solved by using pre-trained models rather than learning models from scratch (Han et al., 2021). Although modern ASR systems have an overall similar to human performance on general data yet one important challenges remain in accurately transcribing specialized vocabulary for example, in academic settings. The Figure 1 illustrates the challenge for ASR systems. An ASR system relying on only audio (i.e. SALMONN (Tang et al., 2023)) is not able to correctly transcribe the domain-specific terms Kenya-Birth and Kenya Rwandan (highlighted by red).



I am presenting our paper, kinya beret a morphology aware kinya wandrawa language model

I am presenting our paper, "Kinyabert A Morphology-Aware Kinyarwanda Language model

Reference - I am presenting our paper KinyaBERT: a Morphology-aware Kinyarwanda Language model

Figure 1: An example of SALMONN transcription before and after using multi-modal input. Left: SALMONN baseline makes mistakes (in red) for multiple words. Right: SALMONN correctly transcribes words (in blue) using multi-modal inputs.

As conference talks and lectures often include presentation slides, humans often can correctly spell these words by using this additional context. Therefore, we propose to integrate visual context (slides) into existing state-of-the-art ASR system to enable them to also exploit this context. As shown on the right side, the final model is able to properly transcribe these words as Kinyabert and Kinyarwanda (highlighted in blue) when the correct words are presented to the model in the additional information provided from the accompanying slides of the talk.

In a first step, we extended an existing benchmark, the ACL dataset (Salesky et al., 2023) with additional slide context as well as a target, automatic evaluation for domain-specific terms to evaluate this assumption. Furthermore, we verify our assumption that these terms are challenging for SOTA models like Whisper (Radford et al., 2023) and SALMONN (Tang et al., 2023).

When integrating visual context into ASR model to handle domain-specific words, we want to keep the strong SOTA performance of current large-scale models. Therefore, we focus on approach that can add this ability to existing models. One interesting aspect of current models is there ability to handle zero-shot task. Therefore, we first propose a zero-shot integration that already is able to exploit the visual context.

In a second step, we investigate methods to train the model to better integrate the contextual information. This gives rise to the challenge that we need dedicated training data for this scenario. We address this problem by using large language models (LLMs) to augment ASR training data with presentation slides.

The primary contributions of this paper are:

- Analysing the ability of ASR to transcribe domain-specific words, particularly from scientific talks.
- Integration of multi-modal information into existing pre-trained models.
- Application of training approaches with augmented data to improve the transcription on domain-specific terms.

2 Related Work

There has been multiple work where model performance has been improved by additional information integration. Authors of the paper (Maergner et al., 2012) create a lecture specific vocabulary, based on the content of the related documents of the lectures. Construction of a vocabulary with relevant content helps the model to produce a reduced word error rate up to 25 percent.

Additionally, combining modalities for the improvement of ASR is also considered. Starting from Hidden Markov model for speech recognition and manually created features represented visual components, combining modalities were also considered for the task of establishing relation between words and non-linguistic context (Fleischman and Roy, 2008) to compensate data deficiency. Later extraction of visual feature from videos using deep learning architectures was incorporated into ASR models on open-domain videos (Miao and Metze, 2016). These approaches are extended with SOTA sequence to sequence model (Gupta et al., 2017) which helped to extract relevant context information from the videos for ASR.

Automatic speech recognition had made a significant progress in recent years by generating accurate transcriptions. With the advent of Whisper, we are now able to generate better transcriptions on unseen datasets. However, transcribing domain-specific datasets or low resource datasets, abbreviations, disfluencies still posses challenge for the SOTA ASR models(Ma et al., 2023). Many approaches focus on fusing audio and visual modalities to address challenges such as proper name transcription, error correction , noisy environments, and multimodal context (Peng et al., 2023),(Kumar et al., 2023).

In recent works, the integration of presentation slides into Multimodal Automatic Speech Recognition (ASR) has gained attention due to the potential benefits of leveraging visual information to improve transcription. The SLIDESPEECH dataset (Wang et al., 2024b) is a large scale audio-visual corpus enriched with slide. However a only a part of their dataset is transcribed and synchronized with the slides.

slideAVSR (Wang et al., 2024a) uses presentation slides in addition to speech for enhancing audio visual speech recognition. The FQ ranker in this work selects prompt words based on their frequency. In contrast, we focus on domain-specific words regardless its frequency in the dataset. Additionally, we present an approach to automatically generate relevant time synchronized slides for an existing large dataset for training in contrast to their manual approach. Other work such as LCB-Net (Yu et al., 2024) propose a novel long-context biasing network for AVSR to leverage the long context information. Methods of data augmentation has also been considered to create synthetic data with variations of audio and visual modality for the purpose of enhanced speech recognition(Oneață and Cucu, 2022).

In this work, we perform data augmentation by generating slides to mitigate the lack of relevant data. Leveraging the augmented data we perform ASR, thereby enhancing model performance.

3 Multimodal Scientific Presentation Benchmark

In this section we analyse two baseline models on the ability to transcribe on domain-specific words. The models are evaluated using an evaluation dataset. We describe the dataset in Section 3.1 and give details of model performance on the dataset in Section 3.4.

3.1 Benchmark defintion

For evaluating the model performances we use the ACL 60/60 dataset (Salesky et al., 2023). This dataset consists of a development (*dev*) and evaluation (*eval*) data each with audio recordings and transcripts of technical presentations from ACL 2022 conference. Both the dev and eval sets consist of five recordings each. Each of these datasets has a duration of approximately one hour. The dataset has been segmented using three approaches. Of which we use only sentence wise segments created manually for our task. In contract to the other two segments, the dataset consists of aligned text and audio segments only for the manually created segmented approach.

3.2 Metric

The traditional word error rates (WER) is used to evaluate ASR model performances by giving equal importance to every word present in the transcript. Our interest lies specifically on ASR performances on words commonly existing in scientific domain. To this end, we go beyond WER and perform a domain-specific word only WER. The exact strategies to select domain-specific words are described in Section 3.4, Section 4.2 and Section 5.2. For analysing baseline model performance, we only consider the domain-specific words in the manually transcribed sentences and count the cases where these words are either deleted or substituted in the model transcriptions. We aggregate the deletion and the substitution counts and divide it by the total occurrence of domain-specific words in the manual transcript. In this paper, we refer the WER on domain-specific words to as WER-terms.

$$\text{WER-terms} = \frac{\text{deletions} + \text{substitutions}}{|\text{domain-specific words}|}$$

3.3 Baselines

To study the ability of ASR models to transcribe domain-specific words we use two models, whisper and SALMONN.

Whisper : Whisper is a transformer based encoder decoder model created by OpenAI, mainly to perform the task of automatic speech recognition and translation (Radford et al., 2023). It is trained on about 680k hours of speech data collected from the internet. Whisper encodes the input speech and generates audio features in its encoder part, which is eventually forwarded to the decoder. The decoder takes in the audio features along with positional encoding and produces transcription for the input audio. Additionally, Whisper also takes help from its previous transcriptions for generating the current transcription.

Salmonn : The SALMONN model, developed at Tsinghua University and ByteDance (Tang et al., 2023), empowers Large Language Models (LLMs) like Vicuna (Chiang et al., 2023) with the ability to directly perceive and understand general audio inputs. This enables them to achieve competitive performance on various speech and audio processing tasks. The model employs a window-level Q-Former (Zhang et al., 2024) module to integrate the outputs from two encoders: Whisper (Radford et al., 2023) for speech and BEATs (Chen et al., 2022) for general audio. These combined outputs, referred to as augmented audio tokens, are then aligned with the LLM's internal representation.

3.4 Analysis

Table 1: Word error rate on all words (WER) and on domain-specific words (WER-terms).

Model	ACL dev		ACL eval	
	WER	WER-terms	WER	WER-terms
Whisper Large V2	8.20	12.91	12.95	26.08
SALMONN 13B v1	17.42	38.44	20.31	57.97

We evaluate the models on their ability to transcribe the ACL dataset. We find that for all models, the word error rate (WER-terms) on domain-specific words is significantly higher compared to WER on all words. We select the domain-specific words by removing all the common words from the ACL dataset. The common words are obtained from a general purpose dataset (Di Gangi et al., 2019) and we filter such words from the transcripts. The remaining words in the transcript are considered as domain-specific words for the purpose of this analysis.

The results of model performance on the ACL dataset are summarized in Table 1. We find that for Whisper the WER-terms is approximately 1.5 and 2 times higher on ACL dev and eval datasets respectively. While evaluating SALMONN on the ACL dataset, we observe that it generates approximately 2.2 times as many mistakes for the domain-specific words on the ACL dev dataset and approximately 2.9 times on the ACL eval dataset. This shows that although the performance of both models are different, they consistently make more mistakes while transcribing domain-specific words.

Table 2 gives the statistics on the domain-specific words extracted from the dataset with this approach. The count of special words in the ACL

Table 2: Statistics of domain-specific words

			Whisper		SALMONN	
Data	Unique special words	Total special words	Times recognised	Times not recognised	Times recognised	Times not recognised
ACL dev	130	333	290	43	204	129
ACL eval	115	276	204	72	116	160

dev dataset is 333 of which 130 are unique special words. Similarly, there are in total 276 special words in the ACL eval dataset of which 115 are unique. We also present the number of times both models are able to recognize the special words. Columns *Times recognized* and *Times not recognized* of Table 2 show the details of how many of the domain-specific words are recognized and not recognized by both Whisper and SALMONN models respectively.

We clearly see that the domain-specific words pose a difficult challenge to state-or-the-art ASR systems. This motivates the integration of additional context like presentation slides.

4 Multi-modal Context Extraction and Integration

Our analysis on Section 3.4 shows that the current automatic speech recognition models make up to three times more mistakes while transcribing domain-specific words. To this end, we propose a multi-modal context extraction and integration system. We build our system on top of an existing ASR model and enrich it through multi-modal information.

We follow a cascaded approach of context integration through a three step process. The first step is to generate images similar to presentation slides. Next, we obtain text from the images and finally, augment ASR model with the extracted multi-modal information. Figure 3 provides an overview of the approach. The following section provides the details on our approach to extract the information and integration.

4.1 Image Frame Extraction

For obtaining the relevant context, we start from the corresponding video recordings of the scientific talks of the ACL dataset and extract aligned image frames. Given video recording of a presentation with slides and the audio of the presentation, our first component is extraction of image frames from the video (denoted by 1 in Figure 3). In general, video recordings of presentations are not accompanied by their respective slides. As a result, we extract the image frames from the recorded video presentation. For our case, the audio segments are always less than 30 sec and therefore we assume that while demonstrating the content of a particular segment, the presenter uses only one single slide.

For each of the audio files, at first, we use the available audio segments. We derive the information of the segment duration, which is the length of each segment and an offset timestamp indicating its starting timestamp with respect to the full audio file. Using these information we then map the audio segments to the original video recording to obtain the respective video segments. From each such video segments, we extract one image frame corresponding to the timestamp in the middle of the segment duration. We use these image frames in the next steps to generate prompts for the pre-trained models.

4.2 Text Extraction

In the second component, (denoted by 2 in Figure 3) we perform text extraction on the obtained frames from the previous step (Section 4.1). To perform this task, we use LLaVA-NeXT (Liu et al., 2024) (referred to as Llava in rest of this paper), due to its ability of better visual reasoning and optical character recognition (OCR) capability. We provide the model with previously extracted image frames and a suitable prompt as input (explained in Appendix 8), in order to generate information for each provided frame. Figure 2 shows one example input pair for the model and the generated output text from the model given the input pair.

The Llava method results in a large number of extracted texts, which needs to be filtered further (denoted by 3 in Figure 3). The primary motivation behind this is to obtain only domain-specific words. To this end, we filter the extracted text by removing

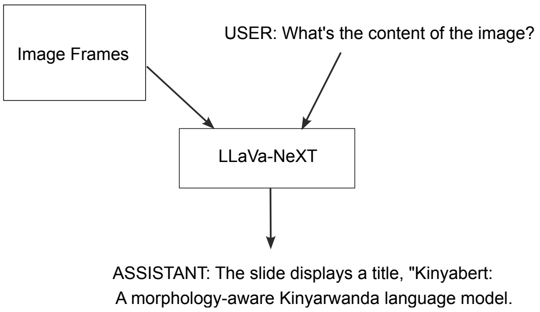


Image Frames
USER: What's the content of the image?
LLaVa-NeXT
ASSISTANT: The slide displays a title, "Kinyabert:
A morphology-aware Kinyarwanda language model.

Figure 2: Text generation with LLaVA-NeXT. The model is provide with a instruction prompt along with an image. LLaVA-NeXT generates text based on the provided inputs.

all common words. This is done by discarding all words present in a general presentation dataset (Di Gangi et al., 2019), resulting in a collection of only domain-specific words.

4.3 Context Integration

The extracted information is then provided to an existing multi-modal ASR model (denoted by 4 in Figure 3). Such ASR systems include an LLM which can be prompted with text to perform the required transcription task. In this work, we focus on improving ASR performance by integrating the context as part of such prompts.

In particular, we use the additional information to enrich the input to SALMONN. By default, there exists a text prompt used in SALMONN that provides instruction (Details provided in the Appendix (explained in Appendix 8) to the integrated LLM Vicuna (Chiang et al., 2023) about the task to be performed. We modify the default text prompt with the information extracted from the previous step (Section 4.2).

We observe that other existing datasets containing speech and the corresponding transcriptions are not accompanied by the video recording or slides of the talk. For our purposes, we require a visual modality for context integration to the selected ASR model. To this end, we generate such modality and use it to improve ASR performance.

5 Data Augmentation

ASR systems with integrated LLMs can be prompted in a zero-shot manner. Existing work (Wei et al., 2021) has shown that compared to zero-shot, fine-tuning of models can be useful to achieve further improvements. To this end, we first perform a zero-shot prompting and further enhance the capability of the ASR model to generate accurate transcriptions by incorporating and training with additional information.

Enhancing ASR using visual modality, a dataset comprising both visual (e.g. images or slides) and speech data is essential. To address the lack of required relevant multi-modal data, this work synthesizes a dataset by augmenting an existing dataset. For our purpose, we augment images to an existing dataset where we generate images that corresponds to presentation slides. This generated image is then added to the dataset lacking inherent similar multi-modal content. This novel strategy of automatically generating and augmenting a visual modality allows us to use any existing speech dataset.

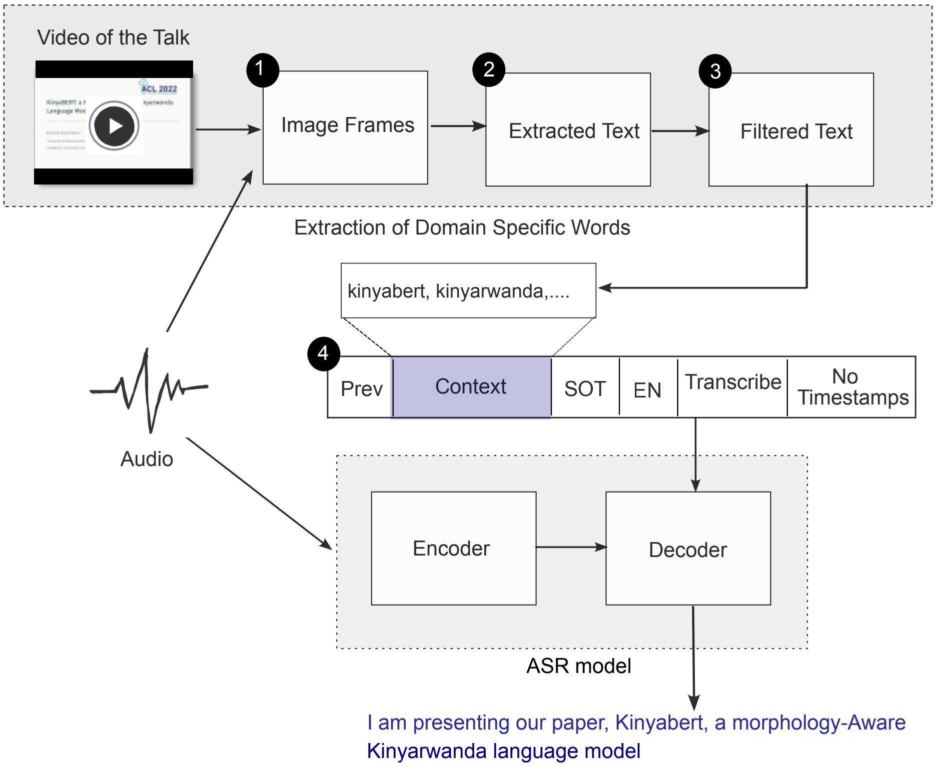
5.1 Generation of Presentation Slides

In this approach, we generate presentation slides for existing speech content through a series of steps. First, we segment the speech transcript into smaller textual units, selecting a chunk size of eight sentences. Our choice of chunk size results in approximately 15–20 slides for a 20–30 minutes speech, ensuring an allocation of 60–90 seconds of speech per slide.

Next, we employ LLaMA 3 to generate LaTeX code for these text chunks. For our case, we use LLaMa 3 and guide it with a pair of instructions consisting of a high level system prompt and a more task specific prompt to generate latex code based on the text chunks (explained in Appendix 8). In the final stage, we convert the generated LaTeX code into images. This involves first compiling the LaTeX code into PDFs and subsequently extracting images from the generated PDF files. We adopt a methodology where images are generated from PDFs rather than directly utilizing the PDFs, as such resources are often unavailable in standard datasets. Conversely, presentation videos are typically accessible, which allows us to extract time-aligned slides corresponding to the speech, as described in Section 4.1.

5.2 Text Extraction

After obtaining the images from the generated slides, we follow the approach of text extraction as outlined in Section 4.2. Since the target dataset for augmentation of information is a general purpose dataset, we adopt a separate text filtration approach. To this end, we first collect all text corresponding to the talks in the dataset. Next, for each talk, we only keep the text relevant to that particular talk and fil-



Video of the Talk
1
Image Frames
2
Extracted Text
3
Filtered Text
Extraction of Domain Specific Words
kinyabert, kinyarwanda,....
4
Prev
Context
SOT
EN
Transcribe
No Timestamps
Audio
Encoder
Decoder
ASR model
I am presenting our paper, Kinyabert, a morphology-Aware Kinyarwanda language model

Figure 3: Overview of our approach

ter out the remaining text. We consider such words that are unique to each talk as the domain-specific words.

6 Experimental Setup and Results

This section provides details on our experimental setup in Section 6.1. Information about the dataset used for training is included in Section 6.2, followed by a detailed description of the results in Section 6.3.

6.1 Experimental Setup

We adopt the SALMONN model, SALMONN 13B v1, developed by Tsinghua University and ByteDance, as our baseline. We perform multi-modal ASR on SALMONN 13B v1 model, considering images to provide additional information to the model.

For extracting text from the images with LLaVA-NeXT, we use llava-v1.6-mistral-7b model which uses CLIP-ViT-L-336px (Radford et al., 2021) as image encoder and LLaMa (Touvron et al., 2023) for language understanding. We provide the model with an image as well as a suitable prompt to generate the text from the image.

For generation of slides we use LLaMa 3 (Dubey et al., 2024) to create latex code. Next we use the python library *subprocess* to execute the shell commands *pdflatex* and *pdftoppm* respectively to generate latex code to PDF and image.

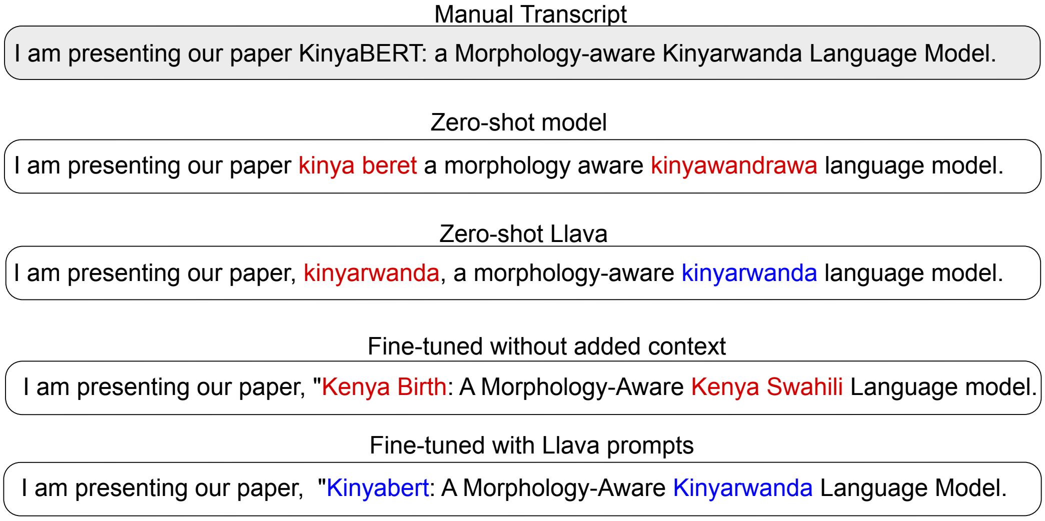
6.2 Dataset

For training the ASR model we use MUSTC (Multilingual Speech Translation Corpus) (Di Gangi et al., 2019) which is primarily designed as a speech translation data. The dataset consists of around 400 hours of audio recordings from English TED Talks speech, transcription and translated transcripts in multiple languages, which are applicable to train model for speech recognition and speech translation tasks.

Since MUSTC does not contain any visual modality, we augment it with the generated images as described in Section 5. Based on the text extraction and filtration approach described in Section 5.2, we obtain 16,830 domain-specific words for 2551 talks present in the dataset.

6.3 Results

In this section we first analyse the quality of the text extracted using Llava in Section 6.3.1. Next we describe the zero-shot performances of the model on the extracted text Section 6.3.2 and finally we compare the zero-shot performance of the model to a model fine-tuned using the extracted information



Manual Transcript

I am presenting our paper KinyaBERT: a Morphology-aware Kinyarwanda Language Model.

Zero-shot model

I am presenting our paper kinya beret a morphology aware kinyawandrawa language model.

Zero-shot Llava

I am presenting our paper, kinyarwanda, a morphology-aware kinyarwanda language model.

Fine-tuned without added context

I am presenting our paper, "Kenya Birth: A Morphology-Aware Kenya Swahili Language model.

Fine-tuned with Llava prompts

I am presenting our paper, "Kinyabert: A Morphology-Aware Kinyarwanda Language Model.

Figure 4: Example of transcriptions generated by different models with respect to the manual transcript. The figure shows that the best possible transcript is generated while fine-tuning the ASR model with Llava prompts.

Table 3: Word error rate on all words and on domain-specific words in zero-shot approach.

Model	ACL dev		ACL eval	
	WER	WER-terms	WER	WER-terms
SALMONN pre-trained	17.42	38.44	20.31	57.97
+ LLaVA-NeXT prompts	10.31	28.62	16.54	48.33
+ Ref prompts	10.93	17.12	14.09	35.87

elucidated in Section 6.3.3.

6.3.1 Quality of the Llava extracted text

We perform an analysis to check the quality of extracted text using Llava from the images of the video frames. For this, we compare the special words that are present in reference text with the Llava extracted text. Table 5 summarizes this result. We find that Llava produces a large number of unique special words of which 81 and 60 corresponds to the special words present in the reference of ACL dev and ACL eval dataset respectively. This represents an overlap of 62% and 52% with the reference text. The reason for extracting large number of unique special words is because an image that corresponds a slide usually contains additional text that is not uttered by the speaker and as a result not present in the transcript.

We also measure how the selected ASR model performs on the Llava extracted text in comparison to the reference. We find that for ACL dev, 81 unique special words overlap with the reference text and is present in total 269 times of which 172 times is recognized by SALMONN while 97 is not recognized. Similarly, for ACL eval, the total number of the unique special words is 180, of which 73 times it is recognized by the ASR model while 107 times it is not.

6.3.2 Zero-shot performance of the ASR model on the extracted data

We evaluated the zero-shot performance of SALMONN while providing the extracted domain-specific words as prompts and compare it to the model without any additional prompts.

Table 3 shows the results of this experiment. The first row of the table shows the WER of both ACL dev and ACL eval dataset of SALMONN without any prompts. In comparison, we find that when prompted with special words obtained either using Llava or from the reference, the ASR model outperforms the model without any additional prompts. Both WER and the WER-terms decrease by around 30% to 40% for ACL dev set when prompted with Llava extracted text and around 37% and 55% when prompted with reference text as shown in last row of Table 3. Similar improvements on WER and WER-terms are also observed for the ACL eval dataset. For this experiment, we use the special words obtained from the reference as prompts to show the model performance when prompted in the best possible setting.

We observe that by using special words as prompts, the performance of the models improve significantly.

Table 4: WER and WER-terms scores of different setup using SALMONN, the pre-trained model, zero-shot with Llava prompts, Fine-tuned with no additional context and the Fine-tuned model with additional information from Llava. The row Fine-tuned with ref shows the best possible setup where the model is fine-tuned using domain-specific words from the reference transcript.

Model	ACL dev		ACL eval	
	WER	WER-terms	WER	WER-terms
Fine-tuned with ref	9.67	10.51	14.63	29.35
Zero-shot	17.42	38.44	20.31	57.97
Zero-shot Llava	10.31	28.62	16.54	**48.33**
Fine-tuned	10.9	30.33	15.74	51.45
Fine-tune with Llava	**10.24**	**19.33**	**14.85**	48.89

Table 5: Statistics of domain-specific words from Llava and ref approaches. Counts of recognized and non recognized words for SALMONN baseline.

Data	Text source	Unique special words	Common with reference	Times recognised	Times not recognised
ACL dev	ref	130	-	204	129
	Llava	367	81	172	97
ACL eval	ref	115	-	116	160
	LlaVa	669	60	73	107

6.3.3 Fine-tuning performance using augmented data

For this experiment, our goal is to check if the performance of the ASR model can be improved further by fine-tuning compared to zero-shot performance. To this end, we fine-tune SALMONN using the augmented dataset obtained in Section 5 and compare it to four other setups. Table 4 summarizes the results of our experiment in the last two rows of the table.

The first row of the Table 4 (denoted by fine-tuned with ref) shows the best possible setup where the ASR model is fine-tuned with domain-specific words from the reference transcript. The second and third setup that corresponds to the second and third row of the table is described in Section 6.3.2. The fourth setup shown as fine-tuned in the table is performance of SALMONN when fine-tuned using the MUSTC dataset without any augmentation. The last row of the table shows our default setup of fine-tuning the model using the augmented dataset with Llava prompts.

Fine-tuning SALMONN for ASR tasks require a task description as an instruction to the integrated Llama model. For the fourth setup, the model is fine-tuned using the configurations used by the model authors i.e., no changes are made to the task description. For our default configuration, we modify this task description with additional special words and change the instruction to consider the special words while transcribing (explained in Appendix 8). Additionally, we make sure that during extraction of special words as outlined in Section 5, there exists no overlap between special words from training and evaluation datasets.

We find that our default setup achieves the best overall WER and WER-terms score for both dataset. Compared to the zero-shot model, WER-terms improves by about 49% and 15% respectively for both the datasets. Our results demonstrate that the overall model performance improves on transcribing special words that are not present in the training dataset which shows that the model is not merely remembering the words.

Figure 4, shows an example prediction by the model with each setup described earlier. Considering both the Zero-shot model and the Fine-tuned model without context, we find that the models makes mistake on both words *KinyaBERT* and *Kinyarwanda*. The zero-shot with Llava model improves but unable to transcribe correctly. Whereas the Fine-tuned model with LLava generates the correct transcription likely due to its acquired ability to incorporate from the additional information.

7 Conclusion and Future work

Current Automatic Speech Recognition (ASR) systems exhibit challenges in accurately transcribing domain-specific words. This limitation hinders their effectiveness in various applications. We present an analysis of the model performance on transcribing domain-specific words to demonstrate this. This paper investigates the potential of augmenting ASR models with information extracted from slides to improve performance. We explore the use of visual information extracted from video recordings of slides as prompts. When trained with prompts, the model develops ability to generate better transcription on domain-specific terms. This shows the effectiveness of multi-modal information in enhancing ASR performance.

The promising results presented in Section 6.3 highlight the potential for further advancements. We propose future work that integrates image representations into the model and further investigate models performances on end-to-end approaches.

Limitations

While our augmented data approach proves effective and results in significant improvements in model performance, it is not without limitations, presenting opportunities for future research.

In our work we consider slides to extract domain-specific words that can be used as additional information for context integrated ASR. Slides often contains summarized, bullet-pointed information which may lead to omit domain-specific words to some extend which may effect the models ability to recognize them correctly. Speakers often elaborate the slides with their own words introducing mismatch between speech and the slide content which also creates similar problem.

Apart from that, the ASR model in this work integrates a pre-trained LLM. LLMs are heavily dependent on the quality and diversity of their training data. Although we achieve improved model performance with our augmented data there remains further scope of improvement. When integrating additional information to the LLM, it may fail to effectively combine these sources of information, leading to misaligned predictions for some cases. Incorporating LLMs into the ASR pipeline for context integration introduces substantial computational overhead, which can slow down the processing time. On the other the LLM might misinterpret the contextual information for the speech and lead to produce incorrect transcription.

References

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Fleischman and Deb Roy. 2008. Grounded language modeling for automatic speech recognition of sports video. In *Proceedings of ACL-08: HLT*, pages 121–129.

Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. 2017. Visual features for context-aware speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5020–5024. IEEE.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Vanya Bannihatti Kumar, Shanbo Cheng, Ningxin Peng, and Yuchen Zhang. 2023. Visual information matters for asr error correction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Rao Ma, Mengjie Qian, Mark JF Gales, and Kate M Knill. 2023. Adapting an asr foundation model for spoken language assessment. *arXiv preprint arXiv:2307.09378*.

Paul Maergner, Alex Waibel, and Ian Lane. 2012. Unsupervised vocabulary selection for real-time speech recognition of lectures. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4417–4420. IEEE.

Yajie Miao and Florian Metze. 2016. Open-domain audio-visual speech recognition: A deep learning approach. In *Interspeech*, page 3.

Dan Oneață and Horia Cucu. 2022. Improving multimodal speech recognition by data augmentation and speech representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4579–4588.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the hidden talent of web-scale speech models for zero-shot task generalization. *arXiv preprint arXiv:2305.11095*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hao Wang, Shuhei Kurita, Shuichiro Shimizu, and Daisuke Kawahara. 2024a. Slideavsr: A dataset of paper explanation videos for audio-visual speech recognition. *arXiv preprint arXiv:2401.09759*.

Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, Shiliang Zhang, and Ming Li. 2024b. Slidespeech: A large scale slide-enriched audio-visual corpus. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11076–11080. IEEE.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Fan Yu, Haoxu Wang, Xian Shi, and Shiliang Zhang. 2024. Lcb-net: Long-context biasing for audio-visual speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10621–10625. IEEE.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. 2024. Vision transformer with quadrangle attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

8 Appendix

Textual Context Integration We instruct SALMONN by providing text prompts to Vicuna that ask questions about the processed audio. The LLM then responds with textual answers based on its understanding. The model is trained for various speech related tasks with suitable prompt structure, as follows

USER: [Auditory Tokens] Can you transcribe the speech into a written format? \n ASSISTANT:

Here, *[Auditory Tokens]*are the output tokens of the window-level QFormer, followed by user prompts in the form of questions with respect to the task performed by the model on the given audio.

Our extracted domain-specific terms from accompanying slides are included in prompts with the following structure

USER: [Auditory Tokens] Please can you transcribe the speech referring to the following tokens wherever needed: kinyarwanda, kinyabert, nlp, pre-trained, ...? \n ASSISTANT:

Here, domain-specific words like *Kinyarwanda, Kinyabert, NLP, and pre-trained* are included in the user prompt. The overall prompt is designed to emphasize both these special words and the task itself.

Model Instruction for text extraction To exhibit LLaVa-Next models OCR quality an extract text from slides we provide the model with an image and a suitable text prompt. the structure of the instruction is given as follow:

"[INST] <image>\nUSER: Extract the text from the sides? [/INST]"

the *<image>* tag is replaced with the image input for LLaVa-Next following with the user prompt. The instruction should always start with the *[INST]* tag and end with *[/INST]* tag.

Model Instruction for data augmentation For creating the multi-modal context for data augmentation, we use LLaMa 3 and guide it with a pair of instructions consisting of a high level system prompt and a more task specific prompt to generate latex code based on text chunks. This consists of a system prompt and a user prompt as follows:

"role": "system", "content": "you are a presenter who wants to inform and inspire",

"role": "user", "content": generate one presentation slide with the main points and concepts in latex, from the following text:<chunk>

The *chunk* in the user prompt is replaced by the parts of talk for which we want to generate the latex code.