Multi-Modal Retrieval For Large Language Model Based Speech Recognition

Anonymous ACL submission

Abstract

Retrieval is a widely adopted approach for improving language models leveraging external information. As the field moves towards multi-modal large language models, it is important to extend the pure text based methods to incorporate other modalities in retrieval as well for applications across the wide spectrum of machine learning tasks and data types. In this work, we propose multi-modal retrieval with two approaches: kNN-LM and cross-attention techniques. We demonstrate the effectiveness of our retrieval approaches empirically by applying them to automatic speech recognition tasks with access to external information. Under this setting, we show that speech-based multi-modal retrieval outperforms text based retrieval, and yields up to $50\,\%$ improvement in word error rate over the multi-modal language model baseline. Furthermore, we achieve state-of-the-art recognition results on the Spoken-Squad question answering dataset.

1 Introduction

The wide adoption of large language models (LLMs) has driven new application areas leveraging this technology. One such direction is jointly modeling multi-modal inputs and outputs with a single generative LLM model. In the speech domain, models such as those proposed by Rubenstein et al. (2023), jointly model text and audio by tokenizing speech signals into discrete units. With an expanded vocabulary encompassing tokens of multiple modalities, this modeling approach has been used in both single task (Xue et al., 2023) and multi-task settings, with Maiti et al. (2023) arguing that the multi-task training of speech-LLMs improves overall generalization of the model through synergies across tasks and modalities.

With generalization capabilities, multi-modal LLMs such as AudioPalm (Rubenstein et al., 2023) and Seamless (Barrault et al., 2023), have targeted many tasks, including Automatic Speech Recognition (ASR). These models rely on the decoder language model (LM) to generate the output text transcription when consuming tokenized speech as a prompt. Because the multi-modal LLM approach can leverage the use of a large text-only corpus for training and multi-tasking, the approach has an advantage compared to traditional ASR models such as recurrent neural network transducers (RNN-T) (Makino et al., 2019) or whisper-like architectures (Radford et al., 2023), which rely primarily on paired audio-transcription data. However, enterprise-grade ASR systems often include further advances, such as functionality to incorporate auxiliary information to assist decoding accuracy, which has yet to be fully addressed with these new multi-modal LLM-based approaches.

Perhaps the two most common approaches for incorporating auxiliary information for ASR have been shallow fusion with an external LM (Gourav et al., 2021; Zhao et al., 2019; Le et al., 2021) and contextual biasing (Sathyendra et al., 2022; Liu et al., 2021; Chang et al., 2021). Shallow fusion with an external LM is a modular way to bias the ASR model at inference time by interpolating the probability distribution of the ASR model with that of the external LM. Shallow fusion though, can suffer from a loss of generality since the external LM does not have direct access to acoustic information. Neural biasing, meanwhile, resolves this issue by ingesting the auxiliary information directly into the acoustic model training (Sathyendra et al., 2022). Yet, neural biasing is ultimately limited in the number of contextual documents it can ingest since attention over an ever larger number of documents renders the method less effective through dilution. We aim to address both of these limitations in this work through *retrieval augmentation*.

Retrieval augmentation is a well known approach to improve existing LMs for ingesting additional information (Mialon et al., 2023; Khandelwal et al., 2019; Guu et al., 2020; Borgeaud et al., 2022; Wang et al., 2023a; Zhong et al., 2022; Wu et al., 2020; Zhang et al., 2022a; Karpukhin et al., 2020; Yasunaga et al., 2022). The overarching idea of all retrieval augmented models is to use an external knowledge base, i.e. retrieval corpus, to improve LM performance. During inference, the retrieval corpus is queried for relevant context and information. The query usually consist of a key computed using an encoder model followed by a search step to find the closest neighbors to the key - typically in a cosine similarity or Euclidean distance sense. The retrieved neighbors are provided to the retrieval augmented LM as additional inputs, which can be used as prompts or cross-attended over.

1.1 Contributions

In this work, we show that using multi-modal retrieval can improve results significantly over canonical text based retrieval. Specifically, we demonstrate our method for speech recognition tasks in two settings 1) Ingesting dynamic multi-modal information; and 2) Domain adaptation of the multimodal LLM. We propose and detail two retrieval approaches to achieve this result: a kNN-LM and a cross-attention based neural model. Experimentally, we compare each retrieval approach using two model sizes: a small model with 300 million parameters (Zhang et al., 2022b); and a larger model with 7 billion parameters. We ultimately demonstrate that while both approaches are capable of significant reduction of word error rate (WER) for domain adaptation, only the cross-attention model improved consistently speech recognition performance for the dynamic information task. We also show that multi-modal LLMs can be used effectively as key encoders for nearest neighbour search, removing the need to use an external neural model as encoder. This result leads to a deployable, application-friendly simplification which has compelling savings of compute resources.

1.2 Related Work

One of the first retrieval augmented LM was kNN-LM that used retrieved results to directly augment token softmax probabilities (Khandelwal et al., 2019). Since the kNN-LM did not use a neural network to ingest the dynamic information, the method is easy to apply on existing models, but limited the performance compared with more involved models.

Subsequent models such as RETRO (Borgeaud et al., 2022) or REALM (Guu et al., 2020) used a cross-attention based mechanism to incorporate the retrieved context into the causal and masked-LMs. RETRO devised a chunked cross-attention to retrieve text continuations which also allowed it to scale to a very large knowledge base, when compared with the kNN-LM.

For speech-recognition applications, Zhou et al. (2023) used a modified kNN-LM. In this work, the authors used a Connectionist Temporal Classification (CTC) decoder to create retrieval keys as opposed to an LM used in the standard kNN-LM (Khandelwal et al., 2019). This change enabled the keys to have acoustic information, but limited the training data to consist only of transcribed speech compared with multi-modal LMs which can utilize both modalities independently. The output probabilities were computed in the same manner as in the standard kNN-LM. Further, retrieval methods have also been successfully applied in cold fusion (Yusuf et al., 2023). With a pre-trained LM as the key encoder, partial hypotheses from the decoder were used to search for text continuations, followed by contextual biasing for generating the transcription. However, the key encoder lacked phonetic context making the retrieved token accuracy low for the initial tokens and at entity start positions. In contrast, we will demonstrate that this limitation can be overcome by using a multi-modal LM instead of a pre-trained text-only LM, by incorporating the audio information into the retrieval context.

Chan et al. (2023) built key-value databases from semantic text and its corresponding text-to-speech (TTS) audio embeddings. The text and TTS embeddings were independently created using two different models. The retrieval database was used with approximate k-nearest neighbour search to bias the ASR model using attention. Meanwhile, Wang et al. (2023b) experimented with a multi-modal LM and Speech2Entity retriever. The retriever, however, was not strictly multi-modal because keys were acoustic encoding of speech from a CTC model, later used to retrieve a set of textual candidate entities as values.

2 Proposed Approach

In the following sections, we describe the multimodal speech-LLM and the modelling approaches to incorporate the retrieved context.

2.1 Multi-Modal Language Models

Speech based multi-modal LMs model speech using quantized discrete audio tokens in addition to text tokens (Rubenstein et al., 2023). The discrete audio tokens are extracted from pre-trained HuBERT embeddings (Hsu et al., 2021) followed by k-means clustering. This is illustrated in Fig. 1.

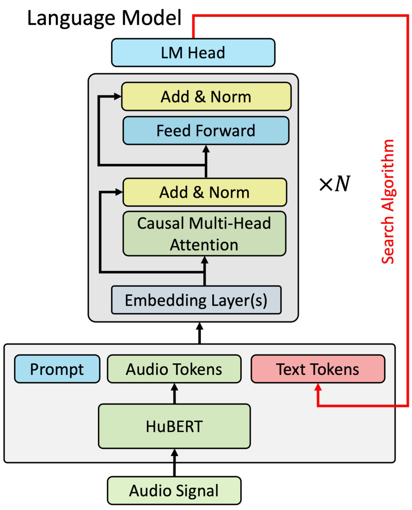

Figure 1: Illustration of a speech multi-modal LM. Inputs to the LM consist of three parts: a prompt specifying the task, audio tokens from an audio tokenizer, and text tokens.

For speech recognition, the multi-modal LM is decoded by concatenating the audio tokens x with a prompt p and generated text tokens y to form model inputs $z_{:i} = [p_0, \cdots, x_0, \cdots, y_0, \cdots, y_{i-1}]$. Next token probability is predicted by:

$$P_{SLM}(y_i) = \text{softmax}(E_o(f(z_{:i}))), \quad (1)$$

where $P_{SLM}(y_i)$ is the posterior distribution of the next token y_i; $f(z_{:i})$ is the multi-modal LM's last hidden state for token y_{i-1}; and E_o is the output embedding matrix that projects the hidden state to the vocabulary dimension. The first text token - y_0 - is a special start-of-sentence token and the generation is continued until an end-of-sentence token is obtained or maximum sequence length is reached. In this work, the maximum sequence length is 2048 for all models.

2.2 Retrieval for Multi-Modal Language Models

We consider two retrieval augmented models in this study: a kNN-LM (Khandelwal et al., 2019), and a novel neural cross-attention based model. Both models aim to augment the posterior token distribution via dynamic information retrieved based on the prior tokens and follow the same retrieval search process. However, they differ in the way retrieved values are constructed and used.

2.2.1 kNN-LM

kNN-LM can be directly applied for speech-recognition with an exception of formatting the inputs as described in the prior section. The principal idea of kNN-LM is to directly modify the token softmax probabilities by interpolating them with a multi-modal distribution constructed from the retrieved neighbors:

$$P_{kNN-LM} = \alpha P_{SLM}(y_i) + (1-\alpha) P_{kNN}, \quad (2)$$

where α is a scalar and P_{kNN} is the probability distribution predicted from the retrieved neighbors given by:

$$P_{kNN} \propto \sum_{j \in D(f(z_{:i}))} \mathbf{1}_{y_i = t_j} e^{-\beta \|\mathbf{k}_j - f(z_{:i})\|}. \quad (3)$$

Here t_j is the retrieved next token; $\mathbf{k}_j$ is its key embedding; $D(f(z_{:i}))$ is a set containing retrieved indices; β is a constant used to normalize the Euclidian norm; and $\mathbf{1}_{y_i=t_j}$ is the indicator function. Retrieved neighbors are obtained by finding the closest keys (in Euclidian sense) $\mathbf{k}_j$ to the models last hidden state $f(z_{:i})$. In this work, we used $\beta = 10^{-3}$ for all kNN-LM experiments and α is found by minimizing cross-entropy in a tuning data set.

For the prompt p and the audio tokens x we do not perform any retrieval, and retrieve the first neighbors starting with the first generated token y_0 - a start-of-sentence token. After this token, retrieval is performed for all tokens until end-of-sentence token or maximum sequence length is reached.

2.2.2 Cross-Attention Based Retrieval

The cross-attention model consists of three submodules: 1) f_r is a retrieval augmented decoder model; 2) g is a key encoder model; and 3) h is a value encoder model. The interplay between these models is illustrated in Fig. 2a. The retrieval augmented decoder model is used to decode the transcription from the prior tokens (including a prompt and audio tokens). The key encoder model is used to encode all retrieval keys similar to kNN-LM while the value encoder model is used to encode the context documents.

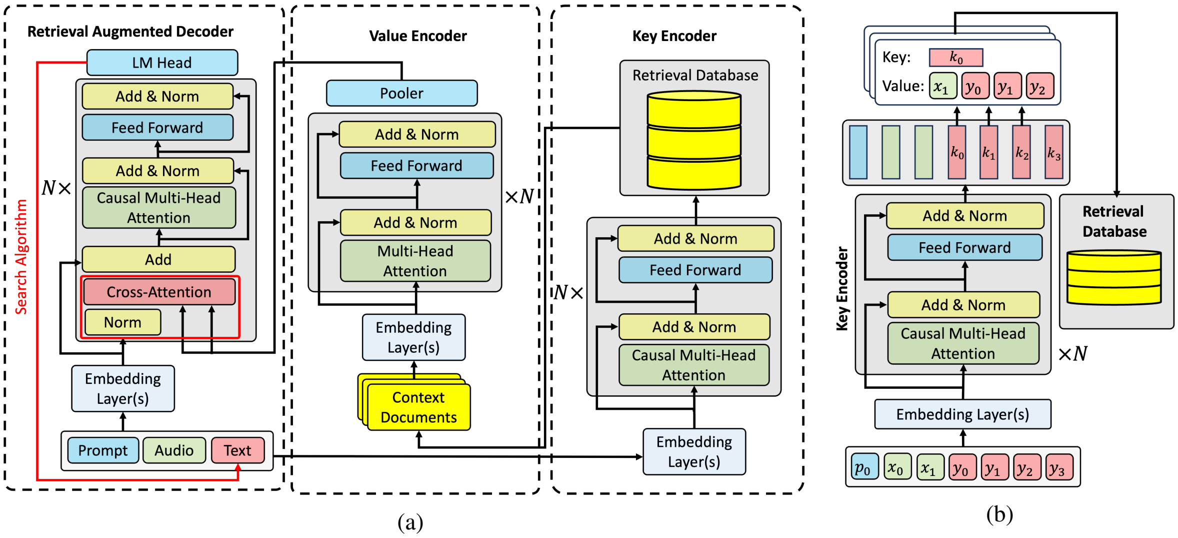

Figure 2: (a): Illustration of the cross-attention based retrieval model. Input tokens are used as inputs both to the decoder model (shown on left-hand side) and key encoder (shown on the right-hand side). The encoded value are used as inputs to the decoder as the key and query for multi-head cross-attention (shown with the red-block). The depicted transformer architecture (normalization layers, etc.) is for illustration purposes and may vary slightly between different models. (b) Illustration of retrieval database creation. Text tokens are encoded and used as keys for the database. Values are surrounding tokens of the key.

For each token y_i we encode the sequence $z_{:i}$ using the key encoder model g to obtain a key embedding $\mathbf{k}_i$. The encoder model is a multi-modal LM that uses the same tokens and audio tokenizer as the decoder model f_r. The last hidden state of the multi-modal LM is used as the key embedding k_i for retrieval lookup in the same way as in the kNN-LM model.

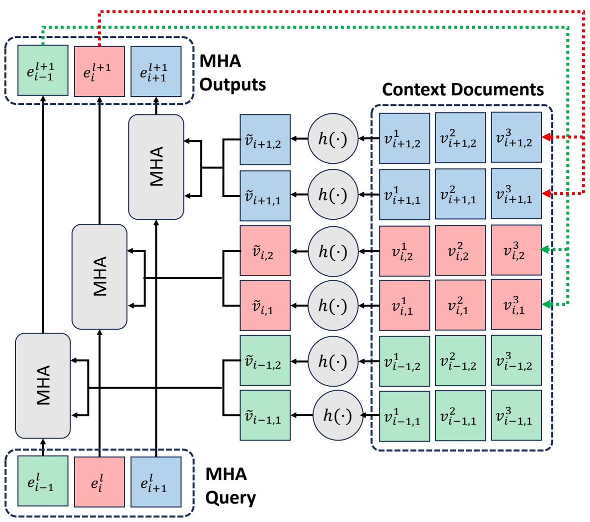

Figure 3: Illustration of the token level multi-head cross-attention. Here e_i^l is the i^{th} token of the l^{th} layer. Color highlights the interactions between the context and the query tokens. MHA stands for standard multi-head cross-attention. The dashed arrow lines from the MHA outputs illustrate the causal dependencies.

The retrieved values $v_{i,j}$ corresponding to the key $\mathbf{k}_i$ are contiguous token sequences of fixed length extracted from a window around the key token from the corpus. The value tokens may include both speech and text tokens as opposed to individual tokens used in the kNN-LM. The retrieval database construction is illustrated in Fig. 2b. Each value sequence $v_{i,j}$ is encoded by the value encoder model h to obtain an encoded embedding $\tilde{v}_{i,j} = h(v_{i,j})$.

In this study, we employ a small BERT (Devlin et al., 2018) model with two transformer layers as the value encoder model (other hyper-parameters matched key encoder model g). Pooling the BERT embeddings is done by selecting the token embedding that follows directly after the token used for computing the corresponding key. In this study, the context document length is fixed, hence this translates to selecting a token with a predefined index.

The encoded value vectors $\tilde{v}_{i,j}$ are stacked to form a context value matrix $\tilde{v}_i = [\tilde{v}_{i,0}, \cdots, \tilde{v}_{i,n}]$. The context value matrix is used as an input to the decoder model, f_r, token level cross-attention layers as the key and value matrices as illustrated in Fig. 3. More formally, for token i we have:

$$P(y_i) = \text{softmax}(E_o(f_r(z_{:i}, \tilde{v}_{:i})) \qquad (4)$$

The retrieval augmented decoder f_r is constructed from a multi-modal LM by adding a token level cross-attention block with a normalization layer to selected transformer layers before the self-attention as illustrated in Fig. 2a. In this work, we used the four topmost layers for the cross-attention blocks. During training, the original transformer weights are frozen and only the new parameters are updated. In this regard, the cross-attention layer can be considered as an adapter for new functionality.

We used pre-layer norm (Xiong et al., 2020) for the query inputs of the cross-attention block (see Fig. 2a) and applied an additional mask on the outputs that is constructed from the pooled context documents. If all context documents for a given token are omitted, we would zero out the output vector. This procedure in combination with the pre-layer norm and the parameter freezing guarantees that when there is no context documents provided the model predictions for all tokens matches the underlying LM.

3 Experiments

In the following sub-sections, we describe datasets used for experiments, model adaptation specifics, and retrieval data construction.

3.1 Datasets

We investigate retrieval augmentation using Spoken-Squad (Lee et al., 2018) and Spoken Language Understanding Evaluation (SLUE) Voxpopuli (Shon et al., 2022). These datasets differ in utterance lengths and topics widely, allowing us to gauge the models in wide range of applications. For, Spoken-Squad, we used Amazon Polly TTS service to synthesize speech for questions and answers.

We applied a simple text normalization for all datasets: lower-casing and punctuation removal. In the case of Spoken-Squad, since the test partition was not available, we used validation partition for testing.

3.2 Models

We use two base multi-modal LMs in this work: a small model based on the public OPT model with 330 million parameters (Zhang et al., 2022b); and an internal larger model using the Llama architecture (Touvron et al., 2023) with 6.8 billion parameters. Model attributes are listed in Table 1.

Table 1: Summary of model hyper-parameters.

Attribute	Small	Large
Parameters	$\sim 330M$	$\sim 6.8B$
Text tokens	50266	50001
Speech tokens	2000	2000
Embedding Dimension	512	4096
Hidden Size	1024	4096
Number of Layers	24	32
Attention Heads	16	32
Intermediate Dimension	4096	11008

Speech was encoded into continuous vectors with a pre-trained HuBERT model (Hsu et al., 2021) with ~ 1 billion parameters and further discretized using 2000 k-means centroids. Text was tokenized using the corresponding sentence piece model for all models. We used greedy search for the small model and beam search with beam width of two for the large model.

3.3 Model Training and Fine-Tuning

The multi-modal LMs were first pre-trained with a large text corpus. Pre-training setup for the small model is identical to Zhang et al. (2022b). The large model was pre-trained using RedPajama (Computer, 2023) with an exception that the books subset was replaced by internal text corpus. Training hyper-parameters are obtained from (Touvron et al., 2023).

After text pre-training, LM vocabularies were extended with the speech tokens followed by multi-task training. The small model was trained using multi-lingual Libri-Speech (Pratap et al., 2020) while the large model was trained using: multi-lingual Libri-Speech, Libri-Light (Kahn et al., 2020), People-Speech (Galvez et al., 2021), a large-scale multilingual speech-to-text translation corpus (CoVOST2) (Wang et al., 2020), Tedlium (Hernandez et al., 2018), and internal audio data. For all multi-tasking models, the training tasks included speech continuation, text continuation, speech recognition (ASR), and speech generation from text (TTS) with equal weights assigned. The training setup is similar to VoxtLM (Maiti et al., 2023). Task specific prompts are listed in Table 2.

For the cross-attention models, the underlying speech multi-modal LM parameters were frozen, with only the value encoder transformer layer and cross-attention adapter blocks trained. The value encoder embedding layer was initialized with

Task	Format
Text	$\langle st \rangle$ text $\langle st \rangle$
Speech	$\langle sa \rangle$ audio $\langle st \rangle$
ASR	$\langle st \rangle$ [ASR] $\langle sa \rangle$ audio $\langle st \rangle$ text $\langle st \rangle$
TTS	$\langle st \rangle$ [TTS] $\langle st \rangle$ text $\langle sa \rangle$ audio $\langle st \rangle$

Table 2: Token formats for pre-training tasks. Here $\langle \cdot \rangle$ are special tokens, [ASR] and [TTS] are text prompts.

weights from the corresponding multi-modal LM input embedding layer; for the small model we used the input projection layer to up-project the embedding to the hidden size whereas for the large model input embeddings were used directly.

The cross-attention model training was divided in two stages: 1) Starting with randomized weights, the model was trained with pre-training tasks and cross-entropy loss along with context extracted from a random document. For half of the tokens, one of the random context documents contained the correct next token while the rest were incorrect. This training approach helped the model distinguish between relevant and irrelevant context. 2) Next, the model was fine-tuned using retrieved context from SLUE Voxpopuli training partition and Spoken-Squad context paragraphs. For SLUE Voxpopuli, the training partition was split in two parts: one part used for the training samples and the other part to construct the retrieval corpus (along with all the context documents from Spoken-Squad).

3.4 Retrieval Data Construction

Retrieval data is constructed from the corresponding train partition of the datasets with the exception of Spoken-Squad where the context paragraphs are used. Audio-transcription pairs are encoded in ASR format shown in Table 2 for multi-modal memory. For text-only memory, transcriptions and contexts are encoded in text format. For both text-only and multi-modal retrieval memory, keys are encoded using the same model. Note that in both cases, the number of retrieval keys remains fixed because they correspond to the text tokens, but with the difference being whether audio was used as prompt or not. FAISS library (Johnson et al., 2019) is used for nearest neighbor search with a Voronoi based index.

For kNN-LM, the retrieved values (neighbours) contained the corresponding next token and their Euclidian distance from the retrieval query. For the cross-attention model, the values included the tokens within a fixed widow, namely, seven tokens preceding the token used for creating the retrieval key and also the following eight tokens.

4 Results and Discussion

4.1 Effect of Corpus Modality on Retrieval

We compare the effect of corpus modality using Spoken-Squad validation partition and quantify the recall of the transcription tokens in the retrieved values.

Retrieval recall statistics are shown in Table 3. The percentages show the fraction of the retrieved tokens matching the next token (relative to retrieval key) in the transcription. The next token is predicted with a high degree using a multi-modal corpus (69 % and 85 %) as opposed to the text-only corpus. We believe this is due to the corpus having both acoustic and semantic information. Higher number of retrieved documents increase the recall slightly for the multi-modal corpus as opposed to the text-only corpus, suggesting that lesser number of retrieved documents (and consequently compute) can be effectively used in applications, when using multi-modal memory. The subsequent tokens are predicted with significantly lower accuracy and recall than the first token, which can be attributed to the fact that LMs are trained to predict the next token but not the subsequent ones.

This result has two implications: (1) the retrieved documents likely work well on a token level model predicting the next such as the kNN-LM; (2) models relying on chunks such as chunked cross-attention used in RETRO (Borgeaud et al., 2022) are likely to have performance reduction when compared with token level models when the retrieved documents are obtained from a multi-modal autoregressive LM.

We demonstrate the impact of the retrieval modality on ASR with the large multi-modal kNN-LM. For Spoken-Squad validation partition, using text-only memory for decoding results in a higher WER of 17.9% when compared to the multi-modal memory case, which achieves a WER of 16.5%. Corresponding 1-best retrieval accuracy is 34 % and 85 % respectively as shown in 3. Hence, acoustic information is principally important for retrieval key construction in speech recognition applications using speech-text LLMs.

Table 3 also shows the retrieval statistics with the bert-large-uncased model (Devlin et al., 2018), which has similar number of parameters (336 mil-

Table 3: Token recall in Spoken-Squad validation partition. Index column shows the token used for recall computation relative to the token used for the retrieval key (in inputs). Document columns show the recall of subsequent retrieved tokens.

	Small Model				Large Model				BERT	
	Speech and Text		Text		Speech and Text		Text		Text	
Index	1st Doc.	8 Docs.	1st Doc.	8 Docs.	1st Doc.	8 Docs.	1st Doc.	8 Docs.	1st Doc.	8 Docs.
1	69 %	83 %	16 %	35 %	85 %	92 %	34 %	54 %	10 %	25 %
2	31%	50 %	6.8 %	19 %	47 %	65 %	16 %	35 %	8%	20 %
3	17 %	34 %	4.2 %	15 %	30 %	49 %	11 %	27 %	7 %	19 %
4	12 %	29 %	3.9 %	15 %	22 %	39 %	9 %	25 %	7 %	19 %

lion) as the small speech-text LM. For text-only retrieval corpus, the speech-text LM has higher recall for the next token (Index 1) for 1-best and 8-best neighbors compared to the BERT model . For subsequent tokens, BERT show a less steep decline in recall and fares better in terms of absolute recall values. When both speech and text are used for the retrieval corpus, the speech-text LM has higher 1-best and 8-best recall for all considered tokens (indices one to four).

4.2 Speech Recognition

We investigate our retrieval approaches in speech recognition setting. For these experiments, we only consider a multi-modal corpus for retrieval as it produced superior recall compared to the text-only corpus.

Table 4 shows the WER evaluated on all datasets, which are grouped by their relationship to the multi-modal LMs training. The speech adaptation datasets were not used for retrieval fine-tuning. Training partition of the datasets was used to construct the retrieval corpus. For Libri-Speech, Tedlium and SLUE-Voxpopuli, this corresponds to audio and transcription pairs and for Spoken-Squad, it is the context paragraphs' audio and text data over all Spoken-Squad titles. In the variant, Spoken-Squad (paragraph), we limit the retrieval corpus to each question/answer's corresponding context paragraph.

The small kNN-LM model show consistent improvement over the baseline with an exception of the Libri-Speech other dataset. We observe similar trend also for the large model with an exception that both Libri-Speech datasets degrade slightly. This can be explained by the strong in-domain baselines - in particular for the large model. The WER Reduction (WERR) from kNN-LM ranges from single digits to $\sim 40\,\%$ depending on the dataset and the baseline model.

For dynamic context in case of Spoken-Squad (paragraph), kNN-LM showed mixed results. For the small model we observe relative degradation of $13\,\%$, as opposed to improvement of of $23\,\%$ for the large model. This discrepancy can be attributed to underlying LMs retrieval recall (see Table 4) and the interpolation weight. In particular, re-tuning interpolation weight for the paragraph level Spoken-Squad would guarantee that the model does not degrade the baseline.

The cross-attention (CA) model requires training and the performance is impacted by the fine-tuning data and the retrieval corpus. This aspect is magnified in the large model that has a large number of trainable parameters from the value encoder and cross-attention layers. In the case of fine-tuning data and retrieval corpus overlapping (Spoken-Squad, SLUE-Voxpopuli), the CA model performs better than the kNN-LM. For Spoken-Squad, this is most prominent where the large CA model out performs all the other models by a wide margin.

The larger improvement from CA model compared with kNN-LM can be attributed to two main factors: 1) the model has more trainable parameters (about $\sim 40M$ parameters for the small model and $\sim 400M$ for the larger model) from the value-encoder and the additional CA blocks in the upper decoder layers; 2) kNN-LM is unable to discriminate incorrect context based on the context tokens and relies solely on the key distance while the cross-attention and value encoder allow more complex interactions between the context documents and the input tokens. With this context, one can conclude that the cross-attention approach tends to be a better candidate than kNN-LM when the dataset that we are domain adapting to is covered to some extent in the pre-training data of the model.

For dynamic context (paragraph Spoken-Squad) we see consistent improvement over the baselines

Table 4: WER on ASR datasets. Bold numbers indicate the best result obtained for the dataset. Numbers in the round brackets show WER Reduction (WERR) compared to the baseline model. Training column depicts the stage at which the datasets were used viz. speech-text LM adaptation or retrieval fine-tuning. CA stands for Cross-Attention

Training	Dataset	Small	kNN-LM (S)	CA (S)	Large	kNN-LM (L)	CA (L)
Adaptation	Libri-Speech (clean)	6.2	3.7 (40)	**3.4** (45)	3.5	3.7 (-5.7)	3.6 (-2.9)
	Libri-Speech (other)	8.1	10.6 (-30)	7.9 (2.5)	**6.5**	6.8 (-4.6)	7.5 (-15)
	Tedlium	16.8	12.2 (27)	10.0 (40)	**6.1**	7.5(-23)	8.6(-41)
Fine-tuning	SLUE Voxpopuli	21.4	20.1 (6.0)	15.9 (25)	12.5	12.4 (1)	11.3 (9.6)
	Spoken-Squad	27.2	20.6 (24)	15.5 (43)	18.4	16.5 (10)	**8.4** (54)
	Spoken-Squad (paragraph)	27.2	30.9 (-13)	16.9 (38)	18.4	14.2 (23)	**8.9** (52)

from both the small and large model. Interestingly, the improvement we see in the dynamic context case is less than we observe for the whole corpus case. This finding could be attributed to two effects: 1) The CA model is fine-tuned using the whole corpus and hence might perform better using that corpus. 2) The paragraph based retrieval corpus may be missing some tokens that would be present in the whole corpus. Overall, the CA model seem a better choice for dynamic context than the kNN-LM model.

Both models demonstrate also a high tolerance for unrelated keys in the retrieval corpus. Adding unrelated data to the retrieval corpus has minimal impact or can even improve the results (see Spoken-Squad results for the whole retrieval corpus compared to paragraph level retrieval corpus) on retrieval precision or recall of the next token. This property allows domain adaptation of the kNN-LM or the cross-attention model to multiple domains by combining multiple retrieval corpus.

Finally, in Table 5 we compare our results against the large whisper v2 model (Radford et al., 2023) for entity heavy SLUE and Spoken-Squad datasets. As can be seen, we reach parity on SLUE Voxpopuli dataset (used in training data of whisper), and improve the WER of whisper by $\sim 40\,\%$ for Spoken-Squad, achieving state-of-the-art results.

Table 5: WER Comparison of Whisper and Large Cross-Attention (CA) model on Fine-tuning datasets. WERR is shown in parentheses.

Dataset	Whisper	CA (L)
SLUE Voxpopuli	**11.2**	11.3 (-0.8%)
Spoken-Squad	14.3	**8.4** (41%)
Spoken-Squad (paragraph)	14.3	**8.9** (38%)

5 Conclusion

We investigate use of retrieval augmented multi-modal LMs for ingesting dynamic context and domain adaptation in speech recognition. We showed that a multi-modal LM can be effectively used for contextualizing retrieval database with audio, leading to an improvement of 10-50% (absolute) in retrieved token accuracy and recall compared to the text-only counterpart. Furthermore, we compared masked-LM (BERT) with a similar sized multi-modal LM as key encoder for constructing a text based retrieval database. Overall, the multi-modal LM fared better for next token retrieval by 6-10% (absolute), whereas the BERT model had better recall for subsequent tokens by 3-4% (absolute).

We considered two different approaches for domain adaptation: a kNN-LM and a cross-attention based neural approach. Both approaches are studied using two different model sizes: a small model with $\sim 300M$ parameters and a larger model with $\sim 7B$ parameters. In all cases we used a multi-modal LM to encode the keys used for retrieval. Domain adaptation using cross-attention outperformed the kNN-LM for both small and large models by 30-100% relative. The large multi-modal LM with cross-attention outperformed whisper model by 40% relative for the entity heavy Spoken-Squad QA dataset, achieving state-of-the-art results.

Additionally, the multi-modal LM's parameters can be shared between the key encoder and decoder. Activation from the early layers can be computed once, used for retrieval and re-used with the retrieved context in the modified top layers for retrieval augmented generation. This reduces memory footprint of the model, which can be an important consideration for embedded applications.

6 Limitations

We believe that the limitations of our work mainly stem from the limitations of Retrieval Augmented Generation (RAG), namely hallucinations and size of the retrieval database. With regards to hallucinations, a mismatch between the retrieval database and the task (in this case ASR) can lead to incorrect transcriptions. Homonyms and rare proper-nouns are especially prone to this. Additionally, size of the retrieval database is also a concern from two standpoints: 1. Having a very large database (billions of tokens) can negatively affect retrieval statistics like recall. Modeling better key encoder is one way to alleviate this. 2. Very large databases have higher memory requirements and can be costly to maintain.

In this work, we showed that using paired audio and text for creating retrieval store is better than using text alone for speech recognition. This places a dependency on having paired audio data for improved performance, which might not always be readily available or accessible. In such cases, using a Text-to-Speech system can be relied upon for generating paired data but it comes with its own pitfalls viz noisy speech, inaccurate pronunciation of rare words etc. These can lead to hallucinations in the RAG process.

References

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

David M. Chan, Shalini Ghosh, Ariya Rastrow, and Björn Hoffmeister. 2023. Domain adaptation with external off-policy acoustic catalogs for scalable contextual end-to-end automated speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Feng-Ju Chang, Jing Liu, Martin Radfar, Athanasios Mouchtaris, Maurizio Omologo, Ariya Rastrow, and Siegfried Kunzmann. 2021. Context-aware transformer transducer for speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 503–510. IEEE.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *CoRR*, abs/2111.09344.

Aditya Gourav, Linda Liu, Ankur Gandhe, Yile Gu, Guitang Lan, Xiangyang Huang, Shashank Kalmane, Gautam Tiwari, Denis Filimonov, Ariya Rastrow, et al. 2021. Personalization strategies for end-to-end speech recognition systems. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7348–7352. IEEE.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Librilight: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. `https://github.com/facebookresearch/libri-light`.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L. Seltzer. 2021. Deep shallow fusion for rnn-t personalization. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 251–257.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Proc. Interspeech 2018*, pages 3459–3463.

Linda Liu, Yile Gu, Aditya Gourav, Ankur Gandhe, Shashank Kalmane, Denis Filimonov, Ariya Rastrow, and Ivan Bulyko. 2021. Domain-aware neural language models for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7373–7377. IEEE.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2023. Voxtlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks. *arXiv preprint arXiv:2309.07937*.

Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. 2019. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

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.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, Jinru Su, Grant P Strimel, Athanasios Mouchtaris, and Siegfried Kunzmann. 2022. Contextual adapters for personalized speech recognition in neural transducers. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8537–8541. IEEE.

Suwon Shon, Ankita Pasad, Felix Wu, Pablo Brusco, Yoav Artzi, Karen Livescu, and Kyu J Han. 2022. Slue: New benchmark tasks for spoken language understanding evaluation on natural speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7927–7931. IEEE.

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*.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, et al. 2023a. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. *arXiv preprint arXiv:2304.06762*.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2023b. Speech-to-text adapter and speech-to-entity retriever augmented llms for speech understanding.

Qingyang Wu, Zhenzhong Lan, Kun Qian, Jing Gu, Alborz Geramifard, and Zhou Yu. 2020. Memformer: A memory-augmented transformer for sequence modeling. *arXiv preprint arXiv:2010.06891*.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR.

Ruiqing Xue, Yanqing Liu, Lei He, Xu Tan, Linquan Liu, Edward Lin, and Sheng Zhao. 2023. Foundationtts: Text-to-speech for asr custmization with generative language model. *arXiv preprint arXiv:2303.02939*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Bolaji Yusuf, Aditya Gourav, Ankur Gandhe, and Ivan Bulyko. 2023. On-the-fly text retrieval for end-to-end asr adaptation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Peitian Zhang et al. 2022a. Gateformer: Speeding up news feed recommendation with input gated transformers. *arXiv preprint arXiv:2201.04406*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-Fusion End-to-End Contextual Biasing. In *Proc. Interspeech 2019*, pages 1418–1422.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. *arXiv preprint arXiv:2205.12674*.

Jiaming Zhou, Shiwan Zhao, Yaqi Liu, Wenjia Zeng, Yong Chen, and Yong Qin. 2023. knn-ctc: Enhancing asr via retrieval of ctc pseudo labels. *arXiv preprint arXiv:2312.13560*.