# MASKED AUDIO TEXT ENCODERS ARE EFFECTIVE FEW-SHOT RESCORERS

**Jinglun Cai, Monica Sunkara, Xilai Li, Anshu Bhatia, Xiao Pan, Sravan Bodapati**
AWS AI Labs, Seattle, Washington, USA
{cjinglun, sunkaral, xilaili, anshubha, panxx, sravanb}@amazon.com

## ABSTRACT

Masked Language Models (MLMs) have proven to be effective for second-pass rescoring in Automatic Speech Recognition (ASR) systems. In this work, we propose **M**asked **A**udio **T**ext **E**ncoder (**MATE**), a multi-modal masked language model rescorer which incorporates acoustic representations into the input space of MLM. We adopt contrastive learning for effectively aligning the modalities by learning shared representations. We show that using a multi-modal rescorer is beneficial for domain generalization of the ASR system when target domain data is unavailable. MATE reduces word error rate (WER) by 4%-16% on in-domain, and 3%-7% on out-of-domain datasets, over the text-only baseline. Moreover, with very limited amount of training data (0.8 hours) MATE achieves a WER reduction of 8%-23% over the first-pass baseline.

## 1 INTRODUCTION

Performance of Automatic Speech Recognition (ASR) systems has been traditionally improved during inference time via second-pass rescoring (i.e., re-ranking ASR hypotheses) using language models (Xia et al., 2017; Hu et al., 2020). In recent studies, Transformer-based pre-trained Large Language Models (LLMs) have shown promising results when used as second-pass rescorers. Previous works (Xu et al., 2022; Salazar et al., 2020; Udagawa et al., 2022) have shown that deep bidirectional Transformers (Devlin et al., 2019) perform better than their unidirectional counterparts such as GPT-2 (Radford et al., 2019).

While LLMs are trained on giant text corpora, they may not be representative of the specific domain of interest, in this case, speech transcriptions. This may result in limited generalization ability without domain-specific fine-tuning. Further, ASR applications warrant robustness to noise and other distortions, which text-only LLMs are incapable of handling on their own at rescoring time.

A potential solution to mitigate these limitations is to incorporate the speech input into LLM rescorers. Recent studies have demonstrated the effectiveness of leveraging audio information during second-pass rescoring (Sainath et al., 2019; Gandhe & Rastrow, 2020; Hu et al., 2020; 2022) to improve performance. However, a tight integration of rescorer, attending to a shared speech encoder used in the first-pass, relies on ASR architecture, training mechanism and internal features, limiting the flexibility of being applied to other ASR systems.

Inspired by recent multi-modal LLM works (Tsimpoukelli et al., 2021; Gao et al., 2022), we propose MATE, a multi-modal MLM rescorer, which is compatible with encapsulated ASR systems. To the best of our knowledge, this is the first work to integrate a pre-trained self-supervised learning (SSL) speech representation model (Baevski et al., 2019; 2020; Hsu et al., 2021; Chen et al., 2021) into the second-pass rescoring. One key challenge of incorporating acoustic information into LLMs is to transform the speech into a form that can be accepted by the language model. We overcome this by using a cross-modal adaptation module consisting of Convolutional Neural Network (CNN) (LeCun et al., 1989) and adapter network (Houlsby et al., 2019). We experiment with different auxiliary alignment losses for audio-text alignment, to effectively learn shared representations across the two modalities, and adopt contrastive learning which significantly improves the model performance. Empirically, we show that MATE transfers well to new domains in zero-shot and few-shot settings, outperforming text-only baselines.

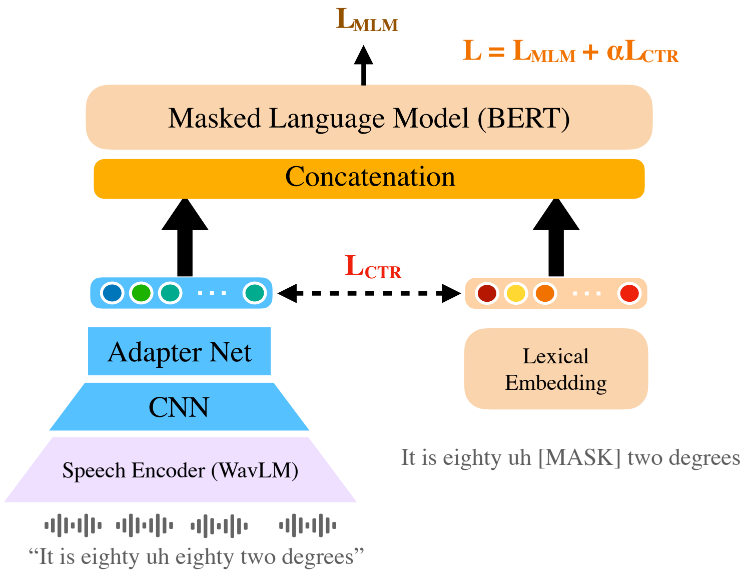
L_MLM
L = L_MLM + αL_CTR
Masked Language Model (BERT)
Concatenation
L_CTR
Adapter Net
CNN
Speech Encoder (WavLM)
"It is eighty uh eighty two degrees"
Lexical Embedding
It is eighty uh [MASK] two degrees

Figure 1: MATE is trained with two losses: (1) The MLM which takes concatenated cross-modal representation as input and computes $\mathcal{L}_{\text{MLM}}$ on masked text tokens. (2)$\mathcal{L}_{\text{CTR}}$ to align the audio and text latent representations.

## 2 Approach

MATE consists of a pre-trained masked language model BERT, an self-supervised learning (SSL) based speech encoder WavLM (Chen et al., 2021) and a modality matching module (CNN and adapter network), as illustrated in Figure 1.

### 2.1 System Architecture

**Masked Language Model** We use BERT, a pre-trained bidirectional MLM, as the primary component of our rescorer. In this work, we extend BERT to incorporate speech data along with text. The pre-trained embedding layers of BERT serve as the text embedding module, while the intermediate encoder layers take both acoustic and lexical representations as input.

**Pre-trained Speech Encoder** To extract the acoustic representation, we use WavLM model, pre-trained on masked speech prediction and speech denoising tasks, achieving state-of-the-art performance on various speech processing tasks and outperforming other models like Wav2Vec2(Baevski et al., 2020) and HuBERT(Hsu et al., 2021) on SUPERB (Yang et al., 2021) benchmark.

**Cross-modal Adaptation** To align the acoustic and lexical representations in the same feature space, we design a cross-modal adaptation module. It is composed of two sub-modules: (i) Convolutional Neural Network (CNN) based subsampling component, to balance the sequence length between the modalities, and (ii) A bottleneck adapter network to project the acoustic representations to the BERT encoder input space. The outputs from the adapter network $\boldsymbol{A}$ and lexical embedding $\boldsymbol{L}$ are concatenated[1] horizontally $\boldsymbol{A}^\frown\boldsymbol{L}$, and passed through the BERT encoder layers to fuse the information from the two modalities.

### 2.2 Alignment Loss

Pre-trained Masked Language Models are trained on text corpora (Devlin et al., 2019). To explicitly align audio and text modalities, we propose introducing an explicit alignment loss function, thereby further enhancing the quality of cross-modal learning.

[1] We have also experimented with a cross-attention based merging mechanism, which leads to inferior performance.

Table 1: Performance measured by WER ↓ and CWER ↓. All models except (2-3) are multi-modal. *(2) GPT2-text:* Full fine-tuning of GPT2 on training corpora transcriptions. *(3) BERT-text:* Full fine-tuning of BERT on training corpora transcriptions, also denoted as ”text-only baseline”. *(4) Multi-modal-GPT2:* A multi-modal uni-directional baseline with GPT2, accepting acoustic information from WavLM. *(5) MATE-NA*: MATE without additional alignment loss; *(6) MATE-MSE*: MATE trained with MSE loss instead of contrastive loss. *(8) Frozen-ME (Masked Encoder)*: Fine-tune all parameters in multi-modal system except masked encoder (BERT) layers with only MLM objective. *(9) WavLM-adapter*: add bottleneck adapter to speech encoder (WavLM) and do adapter-tuning on WavLM, all other parameters are trainable. *(10) ME (Masked Encoder)-adapter*: do adapter-tuning on masked encoder (BERT), all other parameters are trainable.

| | | **In-domain** | | | | | | | | **Out-of-domain** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **MTDialogue** | | **LS test-clean** | | **LS test-other** | | **Voxpopuli** | | **WSJ** | | **ConvAI** | | **SLURP** | |
| | | **WER** | **CWER** | **WER** | **CWER** | **WER** | **CWER** | **WER** | **CWER** | **WER** | **CWER** | **WER** | **CWER** | **WER** | **CWER** |
| (1) | No rescoring | 9.47 | 14.63 | 6.75 | 8.07 | 11.98 | 15.61 | 11.06 | 10.33 | 8.16 | 8.75 | 5.89 | 9.00 | 24.91 | 29.53 |
| (2) | GPT2-text | 9.32 | 14.37 | 6.45 | 7.78 | 11.70 | 15.11 | 10.72 | 9.94 | 7.64 | 8.40 | 5.76 | 8.66 | 24.91 | 29.53 |
| (3) | BERT-text | 9.05 | 13.88 | 5.50 | 7.20 | 10.70 | 14.45 | 10.33 | 9.96 | 6.46 | 8.20 | 5.38 | 8.37 | 24.48 | 29.27 |
| (4) | Multi-modal-GPT2 | 9.24 | 14.17 | 6.35 | 7.69 | 11.54 | 14.93 | 10.56 | 9.83 | 7.55 | 8.20 | 5.69 | 8.59 | 24.89 | 29.40 |
| (5) | MATE-NA | 9.05 | 13.90 | 5.55 | 7.29 | 10.75 | 14.51 | 10.34 | 9.92 | 6.49 | 8.10 | 5.40 | 8.36 | 24.46 | 29.24 |
| (6) | MATE-MSE | **7.49** | **11.41** | 5.22 | 6.95 | 10.31 | 13.97 | 10.10 | 9.62 | 6.10 | 7.65 | **5.07** | 7.92 | 23.84 | 28.24 |
| (7) | MATE (*ours*) | 7.64 | 11.70 | **5.16** | **6.84** | **10.30** | **13.81** | **9.91** | **9.47** | **6.01** | **7.46** | 5.10 | **7.91** | **23.77** | **28.14** |
| *Parameter-Efficient Tuning* | | | | | | | | | | | | | | | |
| (8) | Frozen-ME | 9.21 | 14.22 | 5.57 | 7.34 | 10.82 | 14.65 | 10.37 | 9.80 | 6.55 | 8.15 | 5.42 | 8.34 | 24.39 | 29.13 |
| (9) | WavLM-adapter | 9.15 | 14.02 | 5.58 | 7.41 | 10.81 | 14.69 | 10.23 | 9.86 | 6.52 | 8.05 | 5.47 | 8.39 | 24.56 | 29.27 |
| (10) | ME-adapter | 9.19 | 14.12 | 5.56 | 7.43 | 10.79 | 14.63 | 10.09 | 9.60 | 6.43 | 8.20 | 5.42 | 8.35 | 24.34 | 29.08 |

We adopt a contrastive loss function to enforce the mapping of acoustic representations $\boldsymbol{A}$ and lexical representations $\boldsymbol{L}$ to a shared feature space. We denote the average-pooled vectors by $(\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{l}}_j)$, from the acoustic or lexical representation $\boldsymbol{A}_i$ and $\boldsymbol{L}_i$ respectively. Given acoustic-lexical representations $(\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{l}}_i)_{1 \leq i \leq N}$ where $N$ is the batch size, we use the paired vectors $(\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{l}}_i)$ as positive samples and the unpaired vectors $(\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{l}}_j)_{i \neq j}$ in the same mini-batch as negative samples. The training objective is to minimize the following contrastive loss $\mathcal{L}_{\text{CTR}}$ with Negative Log-Likelihood (NLL) function:

$$\mathcal{L}_{\text{CTR}} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{l}}_i))}{\sum_{j=1}^{N} \exp(\text{sim}(\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{l}}_j))} \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ is a similarity metric, implemented as dot product in our experiments. Contrastive loss promotes a higher level of similarity between paired acoustic and lexical representations, as compared to unpaired representations, thus enhancing the alignment between the two modalities.

### 2.3 Training and Inference

**Training** MATE is trained jointly on the MLM objective $\mathcal{L}_{\text{MLM}}$, similar to that employed in the pre-training of BERT, and the contrastive loss $\mathcal{L}_{\text{CTR}}$.

$$\mathcal{L} = \mathcal{L}_{\text{MLM}} + \alpha \cdot \mathcal{L}_{\text{CTR}} \tag{2}$$

Following BERT pre-training, a portion of tokens in the text sequence are randomly selected for prediction, and are replaced by the [MASK] token, a random token or left unchanged. In order to optimize the model's performance, the model is trained end-to-end and all the parameters are updated during the training process.

**Inference** We use pseudo-log-likelihood (PLL) scoring (Wang & Cho, 2019; Salazar et al., 2020) to compute sequence level scores. Given an acoustic sequence $\boldsymbol{A} = (s_1, ..., s_R)$ and a lexical sequence $\boldsymbol{L} = (t_1, ..., t_T)$, let $\boldsymbol{L}_{\setminus k} = (t_1, ..., t_{k-1}, \text{ [MASK]}, t_{k+1}, ..., t_T)$, PLL score is computed by summing conditional log probabilities $\log P_{\text{MLM}}(\boldsymbol{L}_i | \boldsymbol{A}, \boldsymbol{L}_{\setminus i})$ of each masked lexical token:

$$\text{PLL}(\boldsymbol{A}; \boldsymbol{L}) = \sum_{i=1}^{T} \log P_{\text{MLM}}(\boldsymbol{L}_i | \boldsymbol{A}, \boldsymbol{L}_{\setminus i}) \tag{3}$$

The final score of an utterance is computed as a linear interpolation of the first-pass and second-pass PLL scores, leveraging the complementary information to improve performance while allowing a trade-off between them.

# 3 Experiments

## 3.1 Datasets

**Training Set** The training corpora consist of 10K+ hours of paired audio-text data, sampled from both public and in-house datasets. This data regime is representative of a variety of ASR systems used for various speech applications, with a mix of accents, speakers, sampling rates, and background noise. Less than 5% of the data are synthetic audios generated using AWS Polly Text-to-Speech (TTS) [2] neural backend.

**Evaluation Set** We evaluate MATE on six datasets: *MTDialogue* (movie-twitter) [3], *LibriSpeech (LS)* (Panayotov et al., 2015) and *VoxPopuli* (Wang et al., 2021) are in-domain sets, as the training set includes their corresponding train data splits. *Wall Street Journal (WSJ)* (Garofolo et al., 1993), *ConvAI* (in-house), *SLURP* (Bastianelli et al., 2020) datasets are out-of-domain (OOD) datasets for zero-shot evaluation. The details are described in Appendix A.

## 3.2 Evaluation Metrics

We use word error rate (**WER**) and content word error rate (**CWER**) as the evaluation metrics. CWER is computed on content words only (e.g., “pizza”, “parliament”, “airline”), where we apply rule based method to filter out function words. Furthermore, we evaluate Spoken Language Understanding (**SLU**) performance on SLURP dataset using standard SLU metrics (accuracy and F1 score); SLU predictions (scenario, action and entity) are generated by a bi-directional Long Short-Term Memory (BiLSTM) NLU module (Appendix B).

# 4 Results and Analysis

We summarize the observations and analysis of the results from our experiments [4] as follows:

**MATE excels at both in-domain and out-of-domain generalization:** Table 1 summarizes the performance of the proposed MATE and multiple baseline models, under various settings, across in-domain and OOD datasets. Overall, we observe that our proposed approach (row 7) significantly outperforms text-only baseline (row 3) on in-domain datasets indicating that audio information helps even when we have sufficient target domain corpus for fine-tuning. Furthermore, results on OOD datasets indicate that MATE generalizes much better to new domains in the complete absence of domain data (zero-shot setting), when compared to the text-only baseline, by utilizing the rich information from audio.

**MLMs are more effective multi-modal rescorers than uni-directional LMs:** Rows 2-4 indicate a significant performance gap between BERT and GPT-2 rescorers. BERT-Text, which is a text-only baseline, outperforms even the multi-modal GPT2 indicating the root cause of the gap is the lack of bi-directional (left and right) context in GPT2 which is necessary for reliable and effective LLM scoring, hence validating the choice of MLM in MATE.

**Alignment loss gives significant performance boost:** To study the effect of alignment loss, we train the multi-modal rescorer with two loss functions: Mean squared error (MSE) loss and contrastive loss. Significant performance gains (row 5 vs. row 6-7) in Table 1 indicate that explicit alignment techniques greatly improve learning of multi-modal representations. Specifically, contrastive loss not only aligns relevant pairs like MSE loss, but also promotes distancing irrelevant samples, leading to improved generalization on OOD sets.

**Parameter-efficient fine-tuning results in limited gains:** Rows 8-10 study the performance of a multi-modal rescorer under different parameter efficient fine-tuning settings. We observe that

[2] https://aws.amazon.com/polly/

[3] https://github.com/Phylliida/Dialogue-Datasets

[4] Appendix C contains experimental setup details, including hyperparameters and infrastructure setting.

Table 2: Zero-shot evaluation on SLURP SLU: Accuracy for Scenario/Action. F1 score for Entity.

| | **Scenario** | **Action** | **Entity** |
|---|---|---|---|
| No rescoring | 78.01 | 72.53 | 53.23 |
| GPT2-text | 78.01 | 72.53 | 53.23 |
| Multi-modal-GPT2 | 78.07 | 72.65 | 53.26 |
| BERT-text | 77.72 | 72.45 | 53.37 |
| MATE | **78.76** | **73.70** | **54.26** |

performance degrades as we move from full fine-tuning to adapter-tuning and freezing the full BERT encoder layers, indicating that fine-tuning BERT encoder is the most beneficial in terms of performance improvement. As expected, in comparison to model with full fine-tuning (row 5), rows 8-10 exhibit lower performance. This suggests that frozen or parameter-efficient training methods may lack the model capacity to fully leverage the acoustic information present in the multi-modal data.

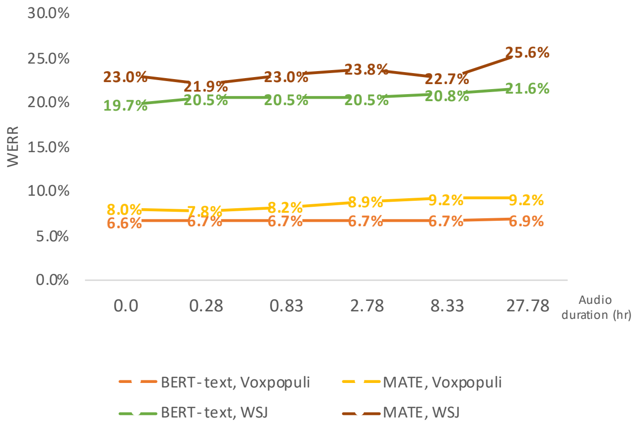

Figure 2: Relative WER reduction (over first-pass) versus domain specific training data size.

**MATE is the most effective few-shot learner:** To study the effect of few-shot learning, we plot the relative WER reduction (WERR) on Voxpopuli and WSJ datasets across different resource conditions as shown in Figure 2. We observe that MATE transfers well to the new domains in the zero-shot setting with no training or domain data at all. Few-shot performance clearly improves with more examples and goes a reasonable way towards closing the gap from zero-shot performance to full fine-tuning performance. We also observe that MATE consistently has superior performance to text-only baseline across both datasets, confirming the ability to rapidly adapt to new domains by leveraging additional information from the audio modality.

**MATE achieves best zero-shot performance improvement on downstream SLU tasks:** To evaluate the effectiveness of the proposed approach on the end goals in a dialog system, we compare it with other baselines using metrics such as scenario/action accuracy and entity F1 score in a zero-shot setting on SLURP dataset. From results in Table 2, we observe that MATE consistently outperforms[5] the other baselines on end-to-end goals indicating that the improvements are mainly on recognition of content words and slot entities. [6]

## 5 Conclusions

We propose a novel multi-modal rescorer, MATE, which achieves significant WER, CWER reduction on in-domain and OOD datasets. In zero-shot and few-shot settings, MATE performs well on unseen domains and adapts rapidly with limited data. The domain generalization capability of MATE makes it an effective choice as a second-pass rescorer for scaling ASR systems to new domains.

[5] The SLURP is a challenging corpus, which mimics the noisy use cases of smart home assistants. Hence, by improving rescoring method alone, we achieve less than 2% absolute improvement in WER and SLU metrics.

[6] Qualitative examples are presented in Appendix F.

## REFERENCES

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *ArXiv*, abs/1910.05453, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf`.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7252–7262, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.588. URL `https://aclanthology.org/2020.emnlp-main.588`.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Micheal Zeng, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1505–1518, 2021.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *NAACL 2022: the 2022 Conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, pp. 1693–1706, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Ankur Gandhe and Ariya Rastrow. Audio-attention discriminative language model for asr rescoring. In *ICASSP 2020*, 2020. URL `https://www.amazon.science/publications/audio-attention-discriminative-language-model-for-asr-rescoring`.

Heting Gao, Junrui Ni, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. WavPrompt: Towards Few-Shot Spoken Language Understanding with Frozen Language Models. In *Proc. Interspeech 2022*, pp. 2738–2742, 2022. doi: 10.21437/Interspeech.2022-11031.

John S. Garofolo, David Graff, Doug Paul, and David Pallett. CSR-I (WSJ0) Complete LDC93S6A. *Linguistic Data Consortium*, 1993. URL `https://catalog.ldc.upenn.edu/LDC93S6A`.

Anmol Gulati, James Qin, Chung Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020. doi: 10.21437/Interspeech.2020-3015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL `http://dblp.uni-trier.de/db/conf/icml/icml2019.html#HoulsbyGJMLGAG19`.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Ke Hu, Tara N Sainath, Yanzhang He, Rohit Prabhavalkar, Trevor Strohman, Sepand Mavandadi, and Weiran Wang. Improving deliberation by text-only and semi-supervised training. *arXiv preprint arXiv:2206.14716*, 2022.

Kevin Hu, Rohit Prabhavalkar, Ruoming Pang, and Tara Sainath (eds.). *Deliberation Model Based Two-Pass End-to-End Speech Recognition*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015. ISBN 9781467369978. doi: 10.1109/ICASSP.2015.7178964.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of bert rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 306–316, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462949. URL https://doi.org/10.1145/3404835.3462949.

Tara N. Sainath, Ruoming Pang, David Rybach, Yanzhang He, Rohit Prabhavalkar, Wei Li, Mirkó Visontai, Qiao Liang, Trevor Strohman, Yonghui Wu, Ian McGraw, and Chung-Cheng Chiu. Two-Pass End-to-End Speech Recognition. In *Proc. Interspeech 2019*, pp. 2773–2777, 2019. doi: 10.21437/Interspeech.2019-1341.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.240. URL https://aclanthology.org/2020.acl-main.240.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 200–212. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/01b7575c38dac42f3cfb7d500438b875-Paper.pdf.

Takuma Udagawa, Masayuki Suzuki, Gakuto Kurata, Nobuyasu Itoh, and George Saon. Effect and Analysis of Large-scale Language Model Rescoring on Competitive ASR Systems. In *Proc. Interspeech 2022*, pp. 3919–3923, 2022. doi: 10.21437/Interspeech.2022-11123.

Alex Wang and Kyunghyun Cho. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 30–36, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2304. URL https://aclanthology.org/W19-2304.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021. doi: 10.18653/V1/2021.ACL-LONG.80.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/c6036a69be21cb660499b75718a3ef24-Paper.pdf.

Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. Rescorebert: Discriminative speech recognition rescoring with bert. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6117–6121, 2022. doi: 10.1109/ICASSP43922.2022.9747118.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pp. 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

# APPENDIX

## A DATASET DETAILS

**In-domain Evaluation Set** We evaluate the proposed MATE approach on both synthetic and real datasets from various domains. **MTDialogue** (movie-twitter) is based on a public lexical dialogue corpus [7] which consists of movie subtitles and twitter user interactions. The audios are generated from TTS system. MTDialogue dataset is a seen dataset for open-book evaluation; i.e., all its data samples are covered in training data. An subset of 1.2 hour is sampled for evaluation. **LibriSpeech(LS)** (Panayotov et al., 2015) is a read English speech corpus based on LibriVox audiobooks. We consider the two official evaluation sets: *test-clean* and *test-other*, each with 5.0 hours of test audios. **VoxPopuli** (Wang et al., 2021) consists of public political speech, sampled from 2009-2020 European Parliament event recordings. For our evaluation purpose, we utilize a 5-hour subset of VoxPopuli English data.

**Out-of-Domain Evaluation Set** We also evaluate MATE on OOD evaluation sets: ConvAI, WSJ, and SLURP. The **Wall Street Journal (WSJ)** (Garofolo et al., 1993) corpus contains conventional and spontaneous dictation by journalists. The *test_eval93* split of 0.4 hour is selected for our evaluation. **ConvAI** is based on in-house user utterances of a task-oriented conversational AI system. The typical usage scenarios include booking flights, ordering food and querying health insurance information, etc. The 2.0 hours of audios are generated from TTS system. **SLURP** (Bastianelli et al., 2020) is a public dataset for smart home virtual assistant development. Top usage scenarios include checking calendar, playing music, and asking about time, etc. We utilized the 10 hr test set for evaluation.

**Ethical Considerations** We have reviewed all licenses of public datasets, which allow the usage for research and paper publication. The in-house dataset ConvAI is internally approved for research purposes. All datasets are sets are de-identified to ensure anonymity. We also make sure the datasets cover various English accents, speakers and backgrounds.

## B SLURP SLU SEMANTICS AND NLU MODULE

SLURP dataset consists of user interactions with smart home virtual assistants. The semantics are annotated with three levels of semantics: Scenario, Action and Entity. For example, ASR transcript "how do I make a turkey" is the annotated with semantics "scenario: cooking | action: recipe | entities: [(type: food | filler: turkey)]". The SLU semantics spans over 18 different scenarios, 46 defined actions and 55 different entity types (Bastianelli et al., 2020).

In the NLU module, we treat semantics prediction as a sequence-to-sequence problem. Specifically, given an ASR transcript after rescoring "how do I make a turkey", the goal is to predict: "scenario: cooking | action: recipe | entities: [(type: food | filler: turkey)]". The NLU module has an encoder-decoder structure based on bi-directional Long Short-Term Memory (Bi-LSTM). Both the encoder and the decoder have hidden dimention 256. The encoder has 2 layers while the decoder has 3 layers. We use Negative Log-Likelihood (NLL) loss for as training objective for sequence prediction. We train the model on ground truth ¡transcript, NLU semantics¿ paris from SLURP training dataset. The learning rate is set to 3e-4 and the training is conducted for 20 epochs with batch size 16.

## C EXPERIMENTAL SETUP

MATE has 217M parameters in total. For both masked language model and speech encoder, we utilize base size models for efficiency (BERT-Base 110M and WavLM-Base+ 95M respectively). The convolutional network contains 3 layers with 768 channels with strides $(2, 1, 2)$ and kernel widths $(3, 1, 1)$. The bottleneck adapter layer has compression factor $0.5$.

The training experiment for MATE is conducted end-to-end: we train all modules simultaneously. We use Adam optimizer (Kingma & Ba, 2014) with linear decay of learning rate. We set initial learning rate to $5e-5$ and batch size to 32. We searched the hyperparameter $\alpha$ in Eq.2 with $(1.0, 3.0, 10.0)$,

[7] https://github.com/Phylliida/Dialogue-Datasets

and the final value is set to $1.0$. The training was conducted for $88K$ steps. All the experiments are performed with NVIDIA Tesla V100 GPUs in a single run. The training for MATE takes 39.7 hours on a Tesla V100 8-GPU machine.

Our first-pass ASR model has a conformer-CTC (Gulati et al., 2020) architecture. which is trained on $50$K+ hours audio-transcript paired data. The conformer encoder consists of $20$ layers of conformer blocks with hidden dimension $2048$; while the shallow decoder is a single Transformer-based layer with the same hidden dimension of $2048$. The conformer-CTC model has approximately $140$M parameters,

We use SCTK[8] package for WER and CWER evaluation. CWER has the same logic as WER computation except that we filter out function words. We use SLURP toolkit[9] for SLU semantics evaluation.

## D Attention Visualization

We visualize the learned self-attention plots extracted from the proposed MATE model in Figure 3. The model has 12 Transformer layers and with 12 heads in each multi-head self-attention. We selected 6 representative plot from the 144 total attention plots with a sample utterance from wsj_eval93 test set. The input utterance has 33 tokens and 77 frames for the acoustic feature, the acoustic features are appended to the lexical embedding before fed into the BERT model. Our observations are listed as follows:

- **(a) (b) (c) and (d)** The plots highlight the border of the text input and audio input (the vertical straight line on position 32). We can conclude that even without feeding any modality border information to MATE, it can learn the border of two modalities itself.
- **(a), (d), (e) and (f)** The monotonic audio-to-text position alignment is clearly shown in the plots. This indicates that the acoustic and lexical representations are successfully mapped to one unified feature space. Interestingly, plots (a), (e) and (f) show that text-to-audio position alignment can also be learned by MATE.

[8] https://github.com/chinshr/sctk

[9] https://github.com/pswietojanski/slurp

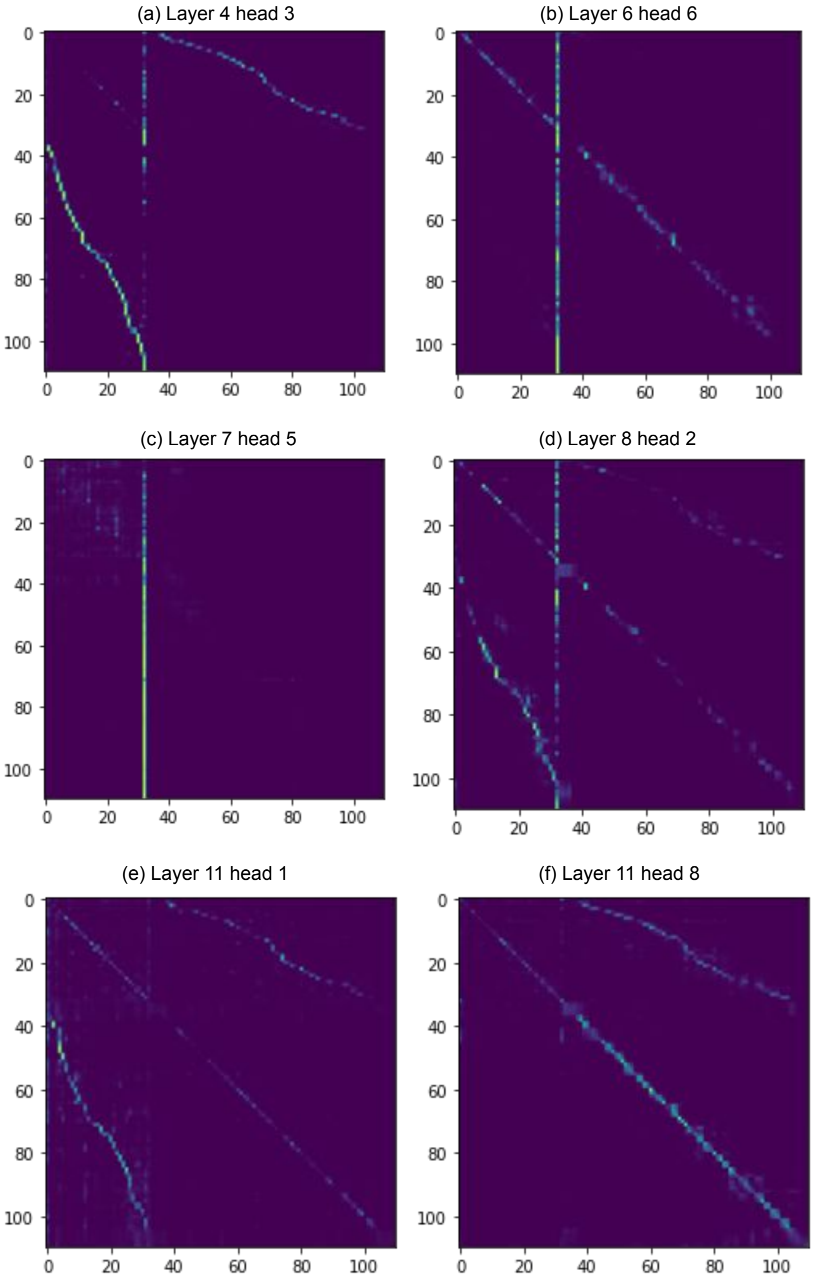

Figure 3: Selected attention plots from the self-attention layers of the 12-layer BERT encoder The sample utterance (from wsj_eval93) contains 110 total frames: the first 33 frames are lexical embedding, followed by 77 acoustic embedding frames. The utterance is: ”last year new hampshire enacted legislation enabling banks from outside the state to acquire new hampshire banks but restrictions in the bill discouraged potential buyers”

## E Limitations and Risks

One limitation of our approach is that incorporating acoustic features from an SSL speech encoder, in our case WavLM, introduces extra latency overhead, as we use a standalone ASR model for first-pass.

| **Dataset** | | **Utterance** |
|---|---|---|
| SLURP | Ground Truth | remove tuesday alarm of nine a m |
| | Rescored 1-best by BERT-text | move to alarm of nine a m |
| | Rescored 1-best by MATE | remove tuesday alarm at nine a m |
| | Ground Truth | hoover the hallway |
| | Rescored 1-best by BERT-text | who in the hallway |
| | Rescored 1-best by MATE | hoover the hallway |
| | Ground Truth | cancel business meeting on wednesday |
| | Rescored 1-best by BERT-text | council business meeting on wednesday |
| | Rescored 1-best by MATE | cancel business meeting on wednesday |
| | Ground Truth | can you let delta know i am never using them again |
| | Rescored 1-best by BERT-text | can you let doctor know i am never using them again |
| | Rescored 1-best by MATE | can you let delta know i am never using them again |
| | Ground Truth | i want to play fifa seventeen |
| | Rescored 1-best by BERT-text | i want to leave for seventeen |
| | Rescored 1-best by MATE | i want to play fifa seventeen |
| | Ground Truth | what do you know about fringe in edinburgh next year |
| | Rescored 1-best by BERT-text | what do you know about french in edinburgh next year |
| | Rescored 1-best by MATE | what do you know about fringe in edinburgh next year |
| Voxpopuli | Ground Truth | for example the report talks about the rule of law and corruption |
| | Rescored 1-best by BERT-text | for example the report talks about the rule of law on corruption |
| | Rescored 1-best by MATE | for example the report talks about the rule of law and corruption |
| | Ground Truth | i have met them they are young capable and visionary |
| | Rescored 1-best by BERT-text | i have met them they are young capable and missionary |
| | Rescored 1-best by MATE | i have met them they are young capable and visionary |
| MTDialogue | Ground Truth | it's muffled |
| | Rescored 1-best by BERT-text | it's muff |
| | Rescored 1-best by MATE | it's muffled |
| | Ground Truth | how much she got to pay |
| | Rescored 1-best by BERT-text | how much he got to pay |
| | Rescored 1-best by MATE | how much she got to pay |
| ConvAI | Ground Truth | why did the noodle box in greensborough fail its health inspection |
| | Rescored 1-best by BERT-text | why did the noodle box in greensboro fail its health inspection |
| | Rescored 1-best by MATE | why did the noodle box in greensborough fail its health inspection |
| | Ground Truth | tell me about duty free shopping |
| | Rescored 1-best by BERT-text | tell me about duty free shop |
| | Rescored 1-best by MATE | tell me about duty free shopping |

Table 3: Qualitative examples: We contrast the 1-best outputs of BERT-text model and MATE in reference to ground truth. We can observe that MATE improves recognition of content words and slot entities.

Therefore, our approach may not be appropriate for certain applications that have exceptionally low latency constraints.

Another limitation is that while multi-modal LLMs have the potential to improve ASR performance, they can be more complex and harder to interpret than text-only LLMs. This makes it more challenging to understand the model's decision making process or debug any potential errors.

The proposed system, MATE, incorporates both pre-trained language model (BERT) and speech model (WavLM) into its design. Such pre-trained models can contain biases and stereotypes against certain religion, race and gender groups (Rekabsaz et al., 2021; Delobelle et al., 2022).

## F Qualitative Examples

To better understand why MATE yields better WER score, we selected several representative cases from the evaluation sets. Table 3 clearly shows that MATE can correct more vocabulary or grammar errors present in the n-best list. We observe MATE is able to correct many ASR errors which are not resolvable by text information alone. In the example from SLURP, both "who in the hallway" and "hoover the hallway" are plausible utterances in an informal style of daily speech. With the aid of acoustic information, MATE is able to assign higher score to the correct utterance "hoover the hallway".