INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Dual Acoustic Linguistic Self-supervised Representation Learning for Cross-Domain Speech Recognition

Zhao Yang[1,3], *Dianwen Ng*[2,3], *Chong Zhang*[2], *Xiao Fu*[1], *Rui Jiang*[1], *Wei Xi*[1], *Yukun Ma*[2], *Chongjia Ni*[2], *Eng Siong Chng*[3], *Bin Ma*[2], *Jizhong Zhao*[1]

[1]Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, China
[2]Speech Lab of DAMO Academy, Alibaba Group [3]Nanyang Technological University, Singapore

zhaoyang9425@gmail.com, dianwen.ng@alibaba-inc.com, {xiwei,zjz}@xjtu.edu.cn, aseschng@ntu.edu.sg

Abstract

The integration of well-pre-trained acoustic and linguistic representations boosts the performance of speech-to-text cross-modality tasks. However, the potential of fine-tuning cross-modality integrated model on accented and noisy corpus is still under-explored. To address this gap, we propose an end-to-end acoustic and linguistic integrated representation learning model, namely Dual-w2v-BART. Our model incorporates acoustic representations from wav2vec2.0 and linguistic information from BART model by utilizing the cross-attention mechanism in the decoder, with paired speech-text dual inputs. To enhance model robustness on accent and noise, we propose a text-centric representation consistency component that helps to gain the similarity between different modality inputs while representing the same content. The results on accented and noisy speech recognition tasks demonstrate the effectiveness of the proposed model for reducing error rates compared to baseline and other competitive models.

Index Terms: Speech Recognition, Dual Acoustic Linguistic, Self-supervised Learning, Accented and Noisy

1. Introduction

The paradigm of self-supervised pre-training and fine-tuning has proven to be effective in transferring high-quality universal and contextual representations from high-resource unpaired data to various downstream tasks, resulting in significant performance gains. This approach has garnered considerable attention and has been extensively researched in the areas of speech, text, and computer vision. Methods like BERT [1], BART [2], wav2vec2.0 [3], and HuBERT [4] have emerged as the backbone of many speech and natural language processing tasks.

In self-supervised learning, the model is trained on unlabeled speech data using a pretext task, such as predicting masked, distorted speech signals, predicting the context of a given speech segment and so on. This pre-training stage helps the model to learn useful and discriminative representations of speech, which can then be fine-tuned on a downstream task, such as speech recognition or speaker identification, using a smaller amount of labeled data. By leveraging self-supervised learning, speech recognition models can achieve state-of-the-art performance with less dependence on human-annotated data, making them more scalable and cost-effective. Fine-tuning from self-supervised pre-trained acoustic models directly, *i.e.*, wav2vec2.0 [3], HuBERT [4], WavLM [5] has been well exhibited in downstream ASR tasks.

However, it suffers some performance drops on downstream ASR tasks in cross-domain scenarios. For example, the self-supervised pre-training data may be recorded in a studio environment with high-quality microphones, while the downstream task data may be recorded in a noisy environment with low-quality microphones. This domain mismatch can result in the model not generalizing well to the downstream task data. Consequently, additional fine-tuning or adaptation procedures are necessary to improve self-supervised models' performance in cross-domain ASR scenarios.

Recent end-to-end approaches leverage the complementary strengths of speech and text modalities to improve performance in various speech-related tasks, which can better handle variations in acoustic conditions, reduce transcription errors, and capture higher-level semantic features. Unlike previous work [6, 7, 8, 9, 10] that focused on pre-training using large-scale speech and text data, our work emphasize more on fine-tuning ASR tasks on small or moderate amounts of speech-text data.

The most recent approaches have been devoted to fusing pre-trained acoustic and linguistic models into a single end-to-end model for downstream ASR tasks, thereby fully exploiting the acoustic and linguistic information in the low-resource corpus. Additionally, existing end-to-end approaches can be categorized into the two: *a) Stacked Acoustic-and-BERT-style Models*. These methods [11, 12, 13, 14] cascade the pre-trained acoustic encoder and BERT-style linguistic encoder. *b) Acoustic-and-Autoregressive Models*. These methods [15, 16] are straightforward to design encoder-decoder architecture that incorporates pre-trained acoustic encoder and autoregressive linguistic decoder, *i.e.*, DistilGPT2 [17], mBART [18].

Inspired by mentioned studies and other joint speech-text training work [19, 20, 21], we aim to design an end-to-end encoder-decoder based architecture utilizing self-supervised wav2vec2.0 and BART models with paired speech-text dual inputs in this paper. The proposed model follows the design principle of effectively combining acoustic and linguistic models while modifying each model at the minimum cost to avoid the catastrophic forgetting problem. To produce robust acoustic and textual representations, the proposed model encodes speech and noisy texts via pre-trained wav2vec2.0 and BART, respectively. We introduce a dual-attention decoder in replacement of the vanilla decoder to capture the useful speech-text and text-text dependencies, then generate final predicted outputs. In addition, a sampling training strategy is adopted to bridge the gap between differences between acoustic and linguistic pre-trained models in the early stage of fine-tuning. We evaluate Dual-w2v-BART in accented and noisy speech conditions and demonstrate its effectiveness. To sum up, we propose a Dual-w2v-BART model and the contributions can be summarized as follows:

- A dual-attention decoder is proposed to effectively utilize the knowledge embedded in the decoder of the pre-trained language/seq2seq model.
- A text-centric representation consistency component is proposed to mitigate the domain mismatch between speech and

10.21437/Interspeech.2023-387

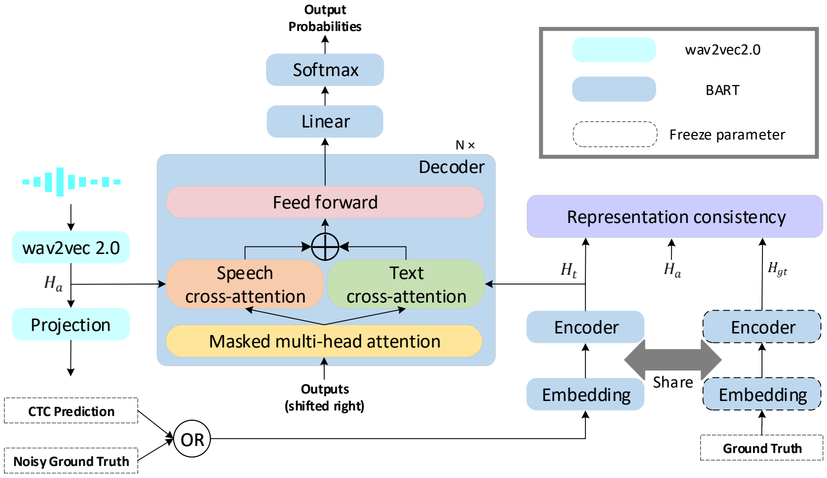

Figure 1: *The model architecture of Dual-w2v-BART consists of wav2vec2.0 as acoustic encoder, BART encoder as text encoder, and a decoder with two parallel cross-attention modules. The representation consistency unit is used to increase the similarity of representations from different modalities via pivots from the ground truth. The sampling training strategy is utilized to alleviate the inaccurate intermediate CTC predictions due to domain shift.*

text representations learned from the encoder.

- We conduct extensive experiments on cross-domain ASR datasets to demonstrate the effectiveness of our proposed approach, which outperforms many strong baselines.

2. Dual Acoustic and Linguistic Model

2.1. Model Description

The architecture of Dual-w2v-BART is illustrated in Figure 1, where the acoustic encoder is a pre-trained speech model wav2vec2.0 [3], and the linguistic part is an encoder-decoder based language model BART [22]. The acoustic encoder takes raw waveform X_a as input and outputs acoustic representation $\mathbf{H}_a$. The output of the last encoder layer is first mapped to a posterior probability distribution over vocabularies using a linear projection layer and softmax function with connectionist temporal classification (CTC) loss criterion $\mathcal{L}_{CTC}$. At the same time, intermediate CTC predictions without language dependencies are calculated which can be refined to further improve accuracy.

$$\mathbf{H}_a = Accoustic(X_a) \tag{1}$$

$$X_t^{CTC} = CTC(\mathbf{H}_a) \tag{2}$$

The input X_t of the linguistic model is noisy text, either from CTC predictions or from corrupted ground truth during training. In the case of corrupted ground truth, various text corruption schemes involving token deletion, text infilling, and token replacement inspired by BART pre-training are employed. The purpose of applying these operations is to simulate ASR predictions and stabilize the training process of the linguistic model. The noisy sequence X_t passes through the corresponding text encoder to produce contextualized linguistic representation $\mathbf{H}_t$. The text decoder consumes both acoustic and linguistic features for sequence generation, which enables better modeling of the correlations between the two modalities. Specifically, we insert a speech cross-attention module to the transformer decoder layer consisting of masked multi-head attention, text cross-attention and feed-forward network. The speech cross-attention module and text cross-attention module focus on pronunciation and semantics aspects respectively and both are deployed in parallel. After that, acoustic and linguistic information are fused and aligned into a unified semantic space via a fully connected network.

$$\mathbf{E}_t = LinguisticEmbedding(X_t) \tag{3}$$

$$\mathbf{H}_t = LinguisticEncoder(\mathbf{E}_t) \tag{4}$$

Besides, a novel text-centric representation consistency regularization unit is proposed to reduce the distance between encoder outputs from different input modalities. The overall training objective is the combination of CTC loss, attention loss, and text-centric representation consistency loss, formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CTC} + \lambda_2 \mathcal{L}_{ATT} + \lambda_3 \mathcal{L}_{RCR} \tag{5}$$

where the hyper-parameters λ_1, λ_2 and λ_3 are set to be 0.3, 0.7, and 0.2, respectively. The ASR task loss consists of both CTC loss and attention loss. Similar with [23], we ensure the ASR loss weight sum to 1 and adjust the auxiliary task loss weight to hold the relative importance of the auxiliary task.

2.2. Representation Consistency Regularization

Representation Consistency Regularization (RCR) aims to align the acoustic and linguistic contextual latent spaces by increasing the representation similarity between the output of the acoustic encoder and linguistic encoder when paired with speech and text inputs. This alignment is important for effectively leveraging multimodal inputs. Specifically, RCR facilitates the alignment between representations from different modalities and pivot embeddings from ground truth by implicitly pushing input representations from different modalities closer together.

The representation consistency regularization is a cross-attention based structure. Specifically, to obtain the pivot embedding $\mathbf{H}_{gt}$, we feed the ground truth text $T_{gt} = (t_1, t_2, ..., t_n)$ into BART encoder. For the acoustic representation $\mathbf{H}_a$, we feed the speech into the wav2vec2.0 model, and for the linguistic representation $\mathbf{H}_t$, we feed the noisy text into the BART encoder. To prevent label leakage during training, we freeze the network for pivot embedding extraction and share its parameters with the one for noisy text. We then use $\mathbf{H}_{gt}$ as a query, and $\mathbf{H}_a$ as the key and value for acoustic representation.

Similarly, we use $\mathbf{H}_{gt}$ as a query, and $\mathbf{H}_t$ as key and value for linguistic representation. This can be formulated as:

$$\mathbf{Y}_{inter} = FFN(Attention(\mathbf{H}_{gt}, \mathbf{H}_a, \mathbf{H}_a)) \tag{6}$$

$$\mathbf{Y}_{corr} = FFN(Attention(\mathbf{H}_{gt}, \mathbf{H}_t, \mathbf{H}_t)) \tag{7}$$

Finally, the cross-entropy criterion is used to align $\mathbf{Y}_{inter}$ and $\mathbf{Y}_{corr}$ to the ground truth transcript respectively. The representation consistency loss is defined as the weighted sum of interaction loss and correction loss:

$$\begin{aligned}\mathcal{L}_{RCR} = &-\frac{\alpha}{n}\sum_{i=1}^{n} logp_i^{inter}(t_i|sg[\mathbf{H}_{gt}], \mathbf{H}_a) \\ &-\frac{(1-\alpha)}{n}\sum_{i=1}^{n} logp_i^{corr}(t_i|sg[\mathbf{H}_{gt}], \mathbf{H}_t)\end{aligned} \tag{8}$$

where $sg[\cdot]$ is the stop-gradient operator and p_i^{inter}, p_i^{corr} represents the predicted probability of the output being t_i at time step i. The hyper-parameter α is set to 0.5.

2.3. Sampling Training Strategy

The intermediate CTC predictions produced by the acoustic model are inaccurate due to the domain shift problem between pre-training and fine-tuning. To alleviate this problem, this paper adopts a very simple yet effective strategy. Either the noisy ground truth transcript or the predicted CTC result is fed into the linguistic model with a certain probability. The probability $p\%$ linearly decreases to 0 as the training step ratio increases to $q\%$. In all experiments, we set p, q to 100 and 10, respectively.

Furthermore, since the vocabulary of BART is very large, only a very small subset of vocabulary is used in ASR tasks. We find that eliminating redundant tokens from the vocabulary when fine-tuning is necessary for stable training, and the trick is used throughout all our experiments.

3. Experiments

3.1. Dataset

The proposed model is evaluated on two challenge tasks: Accented ASR task and Noisy ASR task. For the accented ASR task, we evaluate using the AESRC2020 [24] corpus, which consists of 164 hours of accented English speech recordings from non-native speakers. Since no labeled test set is publicly released, we create a test set by splitting approximately 10% of the training speech. The speakers and accents for both training and test sets are overlapped. As for the noisy ASR task, we construct a noisy speech corpus by randomly sampling noise clips from the MUSAN [1] noise dataset and adding them to a clean subset of 360 hours sourced from the LIBRISPEECH[2] dataset. The Signal-to-Noise Ratio (SNR) levels are sampled from a uniform distribution in 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. Upon adding noise to the dev-clean and test-clean sets of LIBRISPEECH in the same manner, we use them as the validation and test sets for the noisy dataset. All transcriptions are normalized by removing special punctuations and casing for both tasks.

3.2. Experimental Setup

The Dual-w2v-BART model is trained on a single A100 40GB GPU using the Adam optimizer with a tri-state learning scheduler where the learning rate is warmed up for the first 10% steps, hold as a constant for the following 40% steps, and is decayed linearly for the rest steps. We use an average batch size of 1.6m samples. For accented and noisy speech datasets, we train 80k updates in total. We average the models' parameters at the last 5 epochs to avoid overfitting. In decoding, we use beam-search with a beam size of 10.

Table 1: *The WER (%) results of Dual-w2v-BART on accented and noisy English speech datasets in comparison with other competitive models. The proposed Dual-w2v-BART outperforms other compared models under accented and noisy scenarios.*

Model	**Accented**		**Noisy**	
	Dev	**Test**	**Dev**	**Test**
wav2vec2.0-base	-	8.31	10.37	9.29
wav2vec2.0-large	-	7.55	8.77	7.82
w2v-BART [16]	7.24	7.94	8.63	7.76
w2v-cif-bert [12]	-	9.18	9.58	8.75
Fairseq S2T (Scratch)	9.14	9.93	10.14	9.23
Supervised Pre-train & FT	7.54	8.36	9.01	8.33
Dual-w2v-BART	**6.53**	**7.17**	**7.93**	**7.12**

3.3. Experimental Results

We compare our proposed model with baseline and other mainstream models, including two types of models: the first being self-supervised acoustic and linguistic integrated models and the second being supervised models. For self-supervised setting, the w2v-BART [16] model is an end-to-end encoder-decoder model that stacks the pre-trained wav2vec2.0-base as an acoustic encoder and BART decoder as the linguistic decoder. The w2v-cif-bert [12] model fuses pre-trained wav2vec2.0 and BERT into a single end-to-end ASR model. In a supervised setting, one model is trained from scratch, while the other is fine-tuned based on supervised pre-training with LIBRISPEECH.

The detailed results are shown in Table 1. We measure the performance of ASR by the word error rate (WER). In the comparison, the fixed pre-trained wav2vec2.0 base model is used as a baseline. On the accented and noisy test sets, Dual-w2v-BART outperformed wav2vec2.0 and achieved the WER of 7.17% and 7.12%, respectively. This indicates that the linguistic decoder improves the performance of the acoustic encoder. We re-implemented w2v-BART [16] and used w2v-cif-bert [12] source codes to reproduce results on the AESRC2020 dataset. We use Fairseq[3] S2T recipe to train from scratch and the traditional supervised pre-train & finetuning pipeline as baselines. Based on WER results from Table 1, the proposed Dual-w2v-BART model achieves significant improvements in comparison with other joint acoustic and linguistic models and baselines in accented and noisy speech settings.

3.4. Alleviation of Length Inconsistency

In contrast to encoder-based BERT-type acoustic models, the encoder-decoder based pre-trained acoustic model has been shown to effectively alleviate the length inconsistency problem in ASR. Our experiments demonstrate that the use of a linguistic model can improve the overall ASR performance by reducing the WER. Additionally, we report the correction length ra-

[1] https://www.openslr.org/17

[2] https://www.openslr.org/12

[3] https://github.com/facebookresearch/fairseq

tio, which measures the ratio of deletion and insertion errors in the ASR-recognized transcription text compared to the ground truth. This ratio serves as an important metric for evaluating the ability of linguistic models to mitigate the length inconsistency problem in ASR.

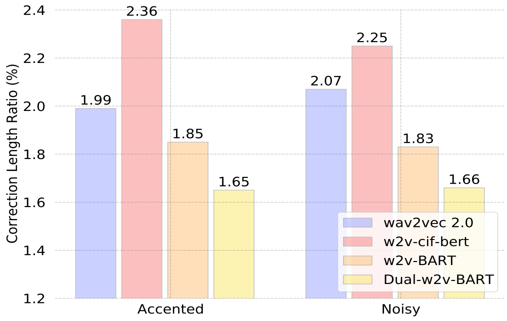

Figure 2: *Performance comparison with different models in terms of correction length ratio. The smaller the correction length ratio, the better ASR performance. Dual-w2v-BART outperforms the compared mainstream models with the lowest correction length ratio.*

The results of our experiments on accented and noisy English datasets are presented in Figure 2. It is observed that models employing a decoder structure exhibit better performance, suggesting that a well-trained decoder can effectively correct over-recognized and under-recognized text, leading to a reduction in WER. However, it is important to note that the use of a decoder may increase the decoding time during the recognition phase, which may be a potential drawback.

3.5. Impact of Sampling Training Strategy

To demonstrate the impact of the sampling training strategy on the learning process, we conduct experiments to compare the performance of models trained with and without this strategy.

Table 2: *The WER(%) results of with/without sampling training strategy on the accented and noisy test sets.*

Sampling Training	Accented	Noisy
Yes	**7.17**	**7.12**
No	7.21	7.14

Our experiments on accented and noisy speech datasets are summarized in Table 2. We observed that the use of the sampling training strategy resulted in a slight improvement in the ASR performance in terms of lower WER. This finding is both intuitive and experimentally valid, as the strategy helps to stabilize the training process. Moreover, the results also show that the strategy effectively compensated for the inconsistency between the acoustic and linguistic models at different stages during fine-tuning.

3.6. Ablation Study

Ablation studies are conducted to verify the effectiveness by removing each main component from the proposed model. WER results on AESRC2020 dataset are shown in Table 3.

The ablation study reveals that all components of the Dual-w2v-BART model are essential for achieving optimal performance, as removing any component results in degraded performance. However, the contributions of each component can be distinguished. When the speech cross-attention module is removed from the Dual-w2v-BART model (Row 2), the integrated model degenerates into a cascaded model with ASR followed by error correction, and the performance drops significantly. This emphasizes the importance of joint speech-text dependency for decoding. We also examined the impact of adding a decoder to the wav2vec2.0 model. Fine-tuning with a randomly initialized decoder (Row 5) resulted in improved performance compared to the wav2vec2.0 model, but worse than the Dual-w2v-BART model. However, pre-training the decoder (Row 4) resulted in further performance improvements, even when the pre-trained parameters were from an independent model. This highlights the importance of a pre-trained decoder for improving cross-domain ASR performance.

Table 3: *Ablation study of Dual-w2v-BART with/without different model components on AESRC2020 test set in terms of WER(%).*

Number	Model	Test
1	Dual-w2v-BART	**7.17**
2	*w/o* speech cross-attention	8.24
3	*w/o* BART encoder	
4	*w/* Pre-trained decoder	7.85
5	*w/* Random Init. decoder	7.98
6	*w/o* BART (only wav2vec2.0)	8.31
7	*w/o* Representation consistency	7.40
8	*w/* Interaction	7.22
9	*w/* Correction	7.32

Furthermore, the experimental results demonstrate that the recognition performance is the worst when using only a pre-trained acoustic model (Row 6), highlighting the crucial role of the linguistic model in the cross-domain ASR task. Moreover, when any part of the representation consistency regularization is excluded (Rows 8 and 9), the performance decreases, underscoring the importance of the representation loss in improving speech and text representations, and thus benefiting downstream cross-domain ASR tasks.

4. Conclusion

In this work, we proposed a self-supervised dual acoustic and linguistic representation learning framework for cross-domain speech recognition, leveraging the effective modeling capabilities of self-supervised models. To address the challenges of heterogeneous modality inputs and single textual outputs, we developed a speech-aware cross-attention module and a text-aware cross-attention module to improve model dependency and facilitate the cooperation of wav2vec2.0 and BART. Additionally, we introduced a representation consistency regularization to reduce domain mismatches between speech and text representations. Experimental results demonstrate that our proposed Dual-w2v-BART model significantly improves cross-domain ASR performance on accented and noisy English speech datasets.

5. Acknowledgements

This work was supported in part by National Key R&D Program of China 2020YFB2104000, NSFC Grant No. 61832008, 62002284, Alibaba Group through Alibaba Innovative Research (AIR) Program and the China Scholarship Council.

6. References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. 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)*, 2019, pp. 4171–4186.

[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[6] R. Ye, M. Wang, and L. Li, "End-to-end speech translation via cross-modal progressive training," in *Proc. of INTERSPEECH*, Aug. 2021.

[7] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5723–5738.

[8] Z. Zhang, S. Chen, L. Zhou, Y. Wu, S. Ren, S. Liu, Z. Yao, X. Gong, L. Dai, J. Li, and F. Wei, "Speechlm: Enhanced speech pre-training with unpaired textual data," *arXiv preprint arXiv:2209.15329*, 2022.

[9] J. Ao, Z. Zhang, L. Zhou, S. Liu, H. Li, T. Ko, L. Dai, J. Li, Y. Qian, and F. Wei, "Pre-Training Transformer Decoder for End-to-End ASR Model with Unpaired Speech Data," in *Proc. Interspeech 2022*, 2022, pp. 2658–2662.

[10] D. Ng, R. Zhang, J. Q. Yip, Z. Yang, J. Ni, C. Zhang, Y. Ma, C. Ni, E. S. Chng, and B. Ma, "De'hubert: Disentangling noise in a self-supervised model for robust speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] G. Zheng, Y. Xiao, K. Gong, P. Zhou, X. Liang, and L. Lin, "Wav-bert: Cooperative acoustic and linguistic representation learning for low-resource speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 2765–2777.

[12] C. Yi, S. Zhou, and B. Xu, "Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 788–792, 2021.

[13] K. Deng, S. Cao, Y. Zhang, L. Ma, G. Cheng, J. Xu, and P. Zhang, "Improving ctc-based speech recognition via knowledge transferring from pre-trained language models," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8517–8521.

[14] C. Xu, B. Hu, Y. Li, Y. Zhang, S. Huang, Q. Ju, T. Xiao, and J. Zhu, "Stacked acoustic-and-textual encoding: Integrating the pre-trained models into speech translation encoders," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2619–2630.

[15] K. Deng, S. Cao, Y. Zhang, and L. Ma, "Improving hybrid ctc/attention end-to-end speech recognition with pretrained acoustic and language models," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 76–82.

[16] X. Li, C. Wang, Y. Tang, C. Tran, Y. Tang, J. Pino, A. Baevski, A. Conneau, and M. Auli, "Multilingual speech translation from efficient finetuning of pretrained models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 827–838.

[17] T. Li, Y. E. Mesbahi, I. Kobyzev, A. Rashid, A. Mahmud, N. Anchuri, H. Hajimolahoseini, Y. Liu, and M. Rezagholizadeh, "A short study on compressing decoder-based language models," in *ENLSP Workshop, 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

[18] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[19] Y. Tang, H. Gong, N. Dong, C. Wang, W.-N. Hsu, J. Gu, A. Baevski, X. Li, A. Mohamed, M. Auli, and J. Pino, "Unified speech-text pre-training for speech translation and recognition," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1488–1499.

[20] Z. Chen, Y. Zhang, A. Rosenberg, B. Ramabhadran, P. J. Moreno, A. Bapna, and H. Zen, "MAESTRO: Matched Speech Text Representations through Modality Matching," in *Proc. Interspeech 2022*, 2022, pp. 4093–4097.

[21] Y. Du, J. Zhang, Q. shi Zhu, L. Dai, M. Wu, X. Fang, and Z. Yang, "A Complementary Joint Training Approach Using Unpaired Speech and Text A Complementary Joint Training Approach Using Unpaired Speech and Text," in *Proc. Interspeech 2022*, 2022, pp. 2613–2617.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[23] Y. Kubo, S. Karita, and M. Bacchiani, "Knowledge transfer from large-scale pretrained language models to end-to-end speech recognizers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8512–8516.

[24] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie, "The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6918–6922.