Regarding the Existence of the Internal Language Model in CTC-Based E2E ASR

Zeyu Zhao
The Centre for Speech Technology Research
University of Edinburgh
Edinburgh, United Kingdom
zeyu.zhao@ed.ac.uk

Peter Bell
The Centre for Speech Technology Research
University of Edinburgh
Edinburgh, United Kingdom
peter.bell@ed.ac.uk

***Abstract*—Some End-to-End (E2E) Automatic Speech Recognition (ASR) models, such as Attention-based Encoder-Decoder (AED) and Recurrent Neural Network Transducer (RNN-T) are known to have components that effectively act as internal language models (ILM), implicitly modelling the prior probability of the output sequence. However, the existence of an ILM in pure Connectionist Temporal Classification (CTC) ASR systems remains debated. In this paper, we investigate the existence and strength of an ILM in CTC systems. Since CTC posterior probabilities cannot be analytically factorised, we propose a novel empirical method to probe the ILM. After validating our method on a hybrid DNN model with various external language models, we apply it to CTC models trained under different conditions, examining the effects of training data, modelling units, and training or pre-training methods. Our results show no strong evidence of an ILM in CTC-based ASR systems, even with the largest training dataset in our experiments. However, we make the surprising finding that when a CTC encoder is jointly trained with an AED loss, an ILM emerges, even when only the CTC component is used in decoding.**

***Index Terms*—Automatic Speech Recognition, Connectionist Temporal Classification, Internal Language Model**

I. Introduction

Since its introduction in 2006, Connectionist Temporal Classification (CTC) [1] has gained widespread use in End-to-End (E2E) Automatic Speech Recognition (ASR) due to its simplicity and efficient combination with self-supervised pre-training [2], [3]. Although newer architectures like Attention-based Encoder-Decoder (AED) [4], [5] and Recurrent Neural Network Transducer (RNN-T) [6] have emerged, sometimes offering better Word Error Rate (WER) performance, CTC remains popular for its straightforward approach and training efficiency [7]–[9].

E2E systems, directly modelling the posterior probability of the output sequence $P(Y|X)$, do not require the independent language modelling component used in traditional HMM systems. However, unlike CTC, both AED and RNN-T architectures do incorporate a language model-like component – the decoder in AED and the prediction network in RNN-T – that plays a strong role in capturing language-related information from training data. It is widely accepted and has been demonstrated that these systems contain a strong internal language model as a consequence [10]. The existence of an ILM can pose a problem when too closely tied to the language of a specific training dataset, which can limit their generalisability across different domains [11]. This limitation has driven significant research into ILM adaptation for AED and RNN-T systems, underscoring the need for more adaptable models [12]–[14]. In contrast, the existence and importance of an ILM in CTC-based ASR systems remains a topic of debate. While some researchers argue that in practical terms, CTC functions purely as an acoustic model [15], others suggest that an ILM exists and must be estimated and subtracted when the model is applied to a domain that differs significantly from the training data [16]. However, neither view has been definitively proven or disproven.

This paper aims to empirically investigate the existence and impact of an ILM in CTC systems. As we cannot analytically factorise CTC's posterior into acoustic and language model parts, we employ a masking strategy that allows us to probe a model's sensitivity to inter-word dependencies. We first verify the effectiveness of this method on a hybrid acoustic model with a range of external language models. Then, we apply this method to CTC systems under different conditions, exploring how factors, such as training data, modelling units, and training or pre-training methods, affect the ILM's presence and strength.

II. Method

A. CTC

Given an input speech $\mathbf{x}$, represented either as a sequence of feature vectors or as a raw waveform, we first process it through an acoustic encoder. This encoder transforms $\mathbf{x}$ into a sequence of hidden features or high-level representations. Subsequently, one or more linear layers, culminating in a softmax activation function, are applied. This process yields a probability distribution over each token c in the token set $\mathbb{C}$ (which includes the blank label) at each time step t, denoted as $p_t(c|\mathbf{x})$, where $t = 1, 2, \ldots, T$. With the conditional independence assumption, CTC models the posterior probability $p(\mathbf{y}|\mathbf{x})$ by marginalising all possible alignments between X and the target tokenised sequence Y (e.g., graphemes, phonemes, Byte Pair Encodings (BPE) [17]),

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\pi \in B^{-1}(\mathbf{y})} \prod_{t=1}^{T} p_t(\pi_t|\mathbf{x}). \tag{1}$$

where $B(\pi)$ is the operation of compressing repeated neighbouring tokens and then removing blank labels from an alignment sequence π. Note that despite the conditional independence assumption, the potential for implicit language modelling theoretically arises from the conditioning of each output token on the complete sequence $\mathbf{x}$. We aim to investigate whether such long-span dependencies are learned in practice from CTC training.

B. Internal LM

Most existing E2E ASR methods, including CTC, AED, and RNN-T, directly model the posterior probability $p(\mathbf{y}|\mathbf{x})$, which, according to Bayes' theorem, can be decomposed as

$$p(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{y})p(\mathbf{y})}{p(\mathbf{x})}, \tag{2}$$

where $p(\mathbf{x}|\mathbf{y})$ is the acoustic model, $p(\mathbf{y})$ is the language model, and $p(\mathbf{x})$ is the prior probability of the input sequence. The decomposition

implies that an E2E ASR system which directly models the posterior probability inherently includes an ILM, regardless of the specific method applied. However, the strength of the ILM might vary significantly across different E2E ASR approaches. Since there is no analytical method to explicitly separate the components in the CTC posterior, we empirically investigate the characteristics of the ILM using a masking strategy, as described below.

C. ILM Detection – Masking Strategy

To detect the potential existence of an internal language model (ILM) within an E2E ASR system, we propose a novel evaluation method. First, we introduce disturbances by introducing masking to the utterances in the evaluation set. Whilst inspired by the work of [16], our masking is applied at the word level, based on the alignments generated by a HMM-based ASR system. Each word in an evaluation utterance has a probability p_m of being masked, with its corresponding segment in the original waveform silenced, guided by the word's start time and duration from the alignment.

After applying the masking process, we decode the masked evaluation set using a CTC-based E2E ASR model and measure the WER for the unmasked words only. If the model functions mainly as an acoustic model with minimal ILM influence, we would expect the WER on the unmasked words to remain relatively stable, regardless of the p_m value. This expectation is based on the assumption that the overall speech quality is preserved, even with the introduction of silent segments from the masking.

In practice, of course, pure acoustic models also benefit from both local cross-word context information – allowing the modelling of effects such as co-articulation – as well as utterance-level information, so can never be completely unaffected by increasing values of p_m. We therefore do expect the WER for unmasked words to rise as p_m increases. However, if the model has a strong ILM, we expect that the degradation caused by masking will be greater because the masked segments disrupt the language patterns learned during training. The stronger the ILM, the more sensitive the ASR model will be to changes in p_m, leading to confusion and errors for the unmasked words. Therefore, if model A has an acoustic model that is as strong as, or stronger than, model B's, but is more sensitive to masking, we can conclude that model A has a stronger ILM.

It is important to clarify that when calculating the WER on a masked evaluation set, the masked words are excluded from the calculation. Further, we do not penalise the model for outputting words during the silent region. We take this approach because the masking process may not be perfect, and some portions of the masked words might still be exposed to the model, potentially causing substitution errors. Of course, if the masking is executed perfectly and the model makes no predictions at the masked word's position, it is not penalised as a deletion error. Note that the masked words are excluded from both the numerator and denominator in the WER calculation. Further details on the methodology for calculating WER in the context of masked utterances will be provided in section III-B.

III. Experiments

A. Settings

In our study, we utilise the LibriSpeech dataset for experiments. This dataset comprises three training subsets: train-clean-100, train-clean-360, and train-other-500. To ensure a uniform acoustic condition across our data, we first merge these three subsets into a single dataset, train960. We then derive different subsets from this merged dataset, including train10, train100, train320, train640, and train960, each containing 10, 100, 320, 640, and 960 hours of data, respectively. Importantly, a small dataset is always a subset of a larger one, ensuring that the acoustic model trained on a smaller dataset is always no better than the model trained on a larger dataset

We manually apply masking to 10% of the training data in each training set to make the acoustic model robust to the edge effects introduced by the masking process. Incorporating masked data helps the acoustic model better handle masked inputs and also smoothes the language patterns, as the masking is applied randomly at the word level. It acts as gentle regularisation that we expect will not significantly impact the learning of the internal language model (ILM) if one does exist.

All E2E models in our experiments are Transformer-based, but they differ in modelling units, training or pre-training methods, and the amount of data used. For neural network training, we use two optimisers: the AdamW optimiser with a learning rate of 10^{-4} for all Transformer and convolution layers (if applicable). For the output linear layers, we use the Adadelta optimiser with a learning rate of 0.9. Training is stopped when no further reduction in loss is observed on the development set (specifically the dev-clean subset) for two consecutive epochs.

As for decoding, for all CTC-based ASR models, we construct decoding graphs and apply Viterbi decoding [18], using Kaldi's `decode-faster` [19] with a beam of 32 and a maximum activate state of 5000, and with no external language models, to focus solely on ILMs.

For word masking, we generate word-level alignments using a Kaldi HMM-based model (tri6) trained on the LibriSpeech dataset. To analyse how WER changes with different masking probabilities (p_m), we experiment with a range of p_m values: $0.0, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35$, and 0.40. To minimise acoustic differences between dataset partitions, we merge the test-clean and test-other subsets into a single test set. Masking is then applied to the larger test set across the different p_m values, resulting in seven additional test sets. In total, we create eight test sets, including the original unmasked set.

B. Metrics

For unmasked datasets, we use the standard WER calculation. However, for masked evaluation sets, we adjust the calculation by treating masked words as optional. When calculating WER for these sets, we ignore one substitution or deletion error for each masked word, as shown in fig. 1. In this example, the reference contains

```
Ref: The (quick) brown fox jumps over (the) **** lazy dog
Hyp: The ******* brown fox jump  over t     lazy lazy ***
Res:                        S               I         D
```

Fig. 1. Example of word error rate computation with optional masked words, where "Ref:", "Hyp", and "Res" denote the reference sequence, the hypothesis and the WER alignment results respectively. The words in parenthesis are masked and optional in the reference. Capital letters S, I, and D denote substitution, insertion, and deletion errors, respectively.

7 unmasked words and 2 masked words (shown in parentheses). There is one substitution error, one insertion error, and one deletion error. However, the substitution for the masked word "(the)" and the deletion for the masked word "(quick)" are ignored. As a result, the WER is calculated as $3/7$.

When the model recognises parts of a masked word, we do not count these as substitution errors, as seen with the masked word "(the)" in fig. 1. Similarly, when a word is completely masked and the model produces no output, we do not classify it as a deletion error, as shown with the word "(quick)" in fig. 1.

C. Results and Analysis

1) Results on Hybrid Models: To validate our proposed masking evaluation strategy, we first examine how the masking probability affects the WER with a hybrid model [20], [21]. We use Kaldi's standard E2E chain model recipe to train a TDNN-F acoustic model [22]–[24]. The model is then evaluated on the test set with different masking probabilities, using various external language models. This approach helps us determine whether the choice of language model affects the WER sensitivity to p_m while keeping the acoustic model constant.

The results are shown in fig. 2. The LibriSpeech dataset provides

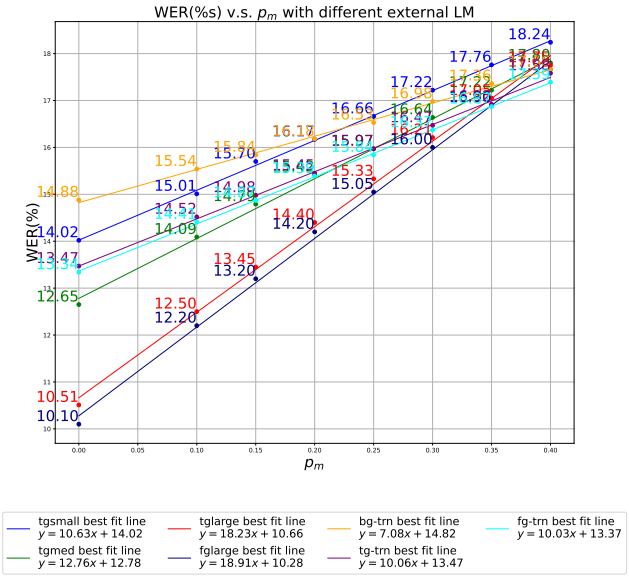

Fig. 2. The WER(%) results of the hybrid models with different external language models, where the x-axis represents the masking probability p_m and the y-axis denotes WER(%).

four external LMs: `tgsmall`, `tgmed`, `tglarge`, and `fglarge`, all trained on the LibriSpeech LM training data. Since this text dataset is much larger than the train960 transcription data, we also introduce three additional external LMs, trained on the transcriptions using KenLM [25]. These are `bg-trn`, `tg-trn`, and `fg-trn`, representing 2-gram, 3-gram, and 4-gram LMs, respectively. Each point in the figure represents the WER of the model with a specific external LM and masking probability. To compare the effect of different LMs, we perform linear regression, where the slope of the regression line shows how sensitive the model is to masking.

As expected, the largest external LM, `fglarge`, has the steepest slope, indicating that the model is most sensitive to masking when using this LM. This occurs because the masking disrupts the language model's predictions, leading to more errors. In contrast, the model is least sensitive to masking when using the weakest LM, `bg-trn`, a 2-gram model trained on the train960 transcription data. Since this LM is not strong enough to significantly influence the predictions, masking has less effect.

These findings demonstrate that our masking evaluation strategy effectively measures the influence of external LMs in hybrid models. Assuming that the external LM in hybrid models behaves similarly to the ILM in E2E ASR models, this approach can be applied to explore ILMs in E2E ASR models as well.

2) Results on Data Quantity: First, we examine whether the amount of training data affects the existence or strength of the ILM in CTC ASR models. To do this, we fine-tune a pre-trained wav2vec 2.0 model [2], specifically `WAV2VEC2_LARGE_LV60K`, which was trained on 60,000 hours of unlabelled data. We use five different training data sizes: train10, train100, train320, train640, and train960, with characters as the modelling units. The results are presented in fig. 3. It's reasonable to assume that a larger training dataset would

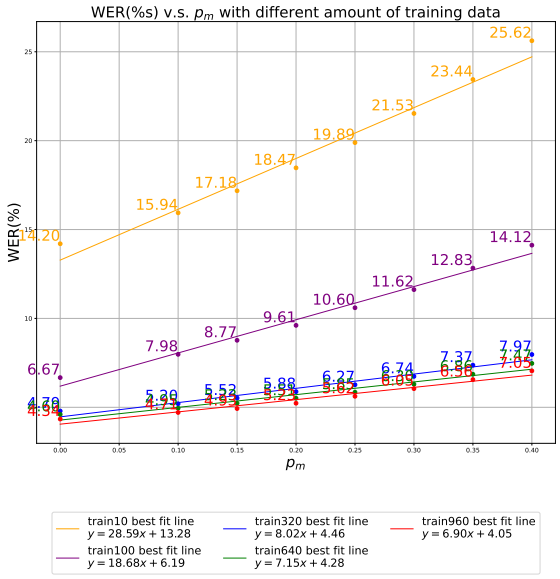

Fig. 3. The WER(%) results of the wav2vec 2.0 models trained with different amount of training data.

lead to a stronger ILM. If the ILM were strong enough, we would expect to see a pattern similar to that in hybrid models—where larger datasets make the model more sensitive to masking. However, our results show the opposite: models trained on smaller datasets are more sensitive to masking.

It's important to remember that masking challenges both the ILM and the acoustic model. Therefore, we can conclude that either the ILM in CTC models is too weak to be significantly impacted by masking, or the acoustic model plays a larger role in recognition than the ILM. Another possibility is that even 960 hours of training data may not be enough to develop a strong ILM in CTC models. More data might be needed to observe the trend seen in fig. 2, even though we have already shown that the training transcription data is sufficient to create a strong external LM sensitive to masking.

3) Modelling Units: Next, we investigate whether the type of modelling units affects the existence of an ILM in CTC models. To explore this, we continue using the wav2vec2.0 model, but this time we train it on the largest available dataset, train960, only. We experiment with four different types of modelling units: characters, 500 BPE units, 1000 BPE units, and 5000 BPE units [17].

We chose the train960 dataset for two main reasons. First, using the largest BPE set (5000 units) with smaller datasets could lead to sparsity issues, which would negatively affect performance. Second, we aim to observe the maximum possible difference in ILM strength

across different modelling units. By training on the largest dataset, we can assess whether changing the modelling units has a significant impact on the ILM. The results of this experiment are shown in fig. 4. Although one might expect different modelling units to affect

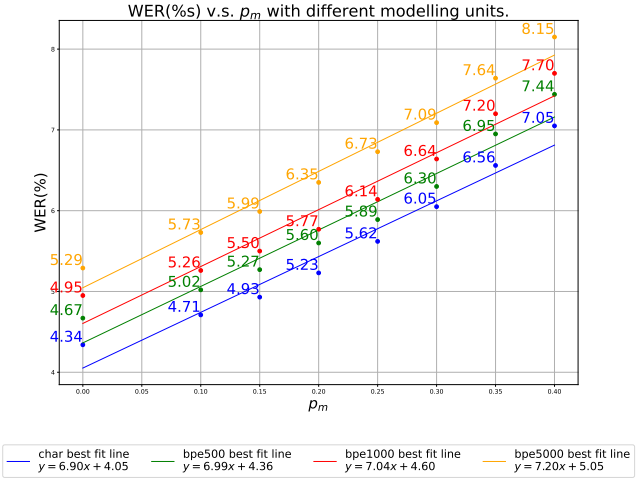

Fig. 4. The WER(%) results of the wav2vec 2.0 models with different modelling units.

the ILM in CTC systems, our findings suggest otherwise. Regardless of the type of modelling units used, the increase in WER remains consistent as the masking probability changes. This indicates that for CTC ASR models, the ILM's ability to learn from the training data is not significantly influenced by the choice of modelling units. In summary, while the selection of modelling units is important for ASR system design, it does not appear to have a significant impact on the ILM in CTC models.

4) Results on Pre-training and Training Methods: The final aspect we examine is the impact of pre-training and training methods on the ILM in E2E ASR systems. We investigate four different models, each trained or pre-trained under different conditions:

- The wav2vec 2.0 model (W2V2-CTC)[1] fine-tuned with CTC
- The Whisper model (WSP-CTC) [5], pre-trained on a large corpus of transcribed data[2]. For our study, we use only the encoder part of this model to fine-tune our CTC models.
- A Transformer model (TRSF-CTC) trained from scratch, without any pre-training, using the same hyper-parameters as the Whisper model encoder, trained with CTC.
- A Transformer model trained with ESPnet, using the standard CTC-AED joint training recipe [26], [27], with the same hyper-parameters as the Whisper model. During decoding, we set the CTC weight to 1.00 to make it function as a CTC model, which we denote as TRSF-JCTC.
- The same Transformer model as above, but during decoding, we assign the CTC weight to 0.00 to make it function as an AED model, which we denote as TRST-JAED.

For fine-tuning the wav2vec 2.0 model, we adjust all 24 Transformer layers along with the two linear output layers. Similarly, for both the Whisper and Transformer models, we train all 24 Transformer layers, along with two convolutional layers at the beginning and two linear layers at the end. Each model contains approximately 300 million trainable parameters. Our goal is to determine whether pre-training affects the ILM in CTC models. Additionally, by comparing the CTC-AED joint-trained model with different CTC weights during decoding, we can assess how the ILM responds to our masking evaluation strategy.

For this analysis, we use the train960 dataset to train each of the four models. The first three models use characters as the modelling units, while the CTC-AED joint-trained model uses 5000 BPE units. The results are shown in fig. 5. We observe that models trained purely

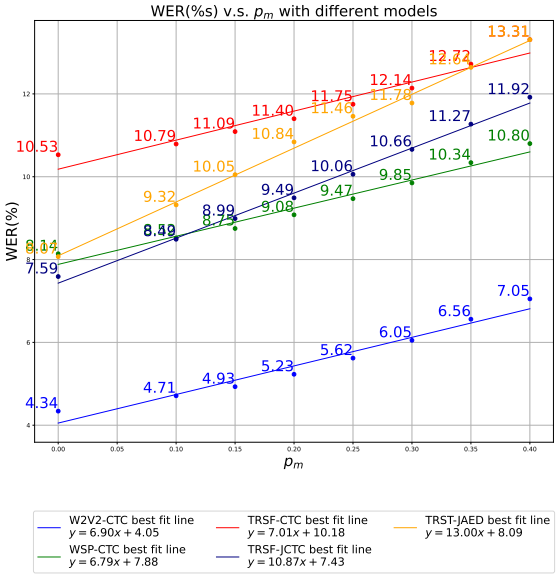

Fig. 5. The WER(%) results with different training/pre-training methods.

with CTC, W2V2, WSP and TRSF-CTC show similar sensitivity to masking, indicating that pre-training does not significantly affect ILM. However, when examining the joint-trained model with a CTC weight of 1.00, TRSF-JCTC, the model becomes more sensitive to masking, suggesting that joint training is more likely to encourage the development of a stronger ILM, even in the encoder. Additionally, when the CTC weight is set to 0.00, resulting in an AED model, TRSF-JAED, the sensitivity to masking increases even further. This suggests that joint training can indeed lead to a strong ILM, with some ILM components residing in the decoder of the AED model. We assume this is because the independence assumption in CTC limits the development of a strong ILM, while joint training with AED, which lacks this assumption, encourages a stronger ILM.

IV. Conclusion

In this paper, we introduced a novel masking evaluation strategy to investigate the existence of the ILM in E2E ASR systems. Our analysis of CTC-based models showed that the ILM is relatively weak, with minimal impact from masking, regardless of training data quantity, modelling units, or pre-training methods. However, we cannot deny that using larger datasets beyond 960 hours might reveal a stronger ILM. Further investigation with the CTC-AED joint-trained model revealed the interesting finding that joint training with an AED objective strengthens the ILM of the model, even when it is used as a pure encoder with a CTC output layer.

[1] `WAV2VEC2_LARGE_LV60K`
[2] `medium.en`

References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 369–376.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[3] 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.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 4960–4964.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023, pp. 28 492–28 518.

[6] A. Graves, "Sequence Transduction with Recurrent Neural Networks," *arXiv:1211.3711 [cs, stat]*, Nov. 2012.

[7] Y. Fujita, S. Watanabe, X. Chang, and T. Maekaku, "LV-CTC: Non-Autoregressive ASR With CTC and Latent Variable Models," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2023, pp. 1–6.

[8] Y. Yang, X. Yang, L. Guo, Z. Yao, W. Kang, F. Kuang, L. Lin, X. Chen, and D. Povey, "Blank-regularized CTC for Frame Skipping in Neural Transducer," in *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 4409–4413.

[9] Z. Yao, W. Kang, F. Kuang, L. Guo, X. Yang, Y. Yang, L. Lin, and D. Povey, "Delay-penalized CTC Implemented Based on Finite State Transducer," in *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 1329–1333.

[10] Z. Meng, N. Kanda, Y. Gaur, S. Parthasarathy, E. Sun, L. Lu, X. Chen, J. Li, and Y. Gong, "Internal Language Model Training for Domain-Adaptive End-To-End Speech Recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2021, pp. 7338–7342.

[11] Z. Meng, S. Parthasarathy, E. Sun, Y. Gaur, N. Kanda, L. Lu, X. Chen, R. Zhao, J. Li, and Y. Gong, "Internal Language Model Estimation for Domain-Adaptive End-to-End Speech Recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 243–250.

[12] Z. Meng, Y. Gaur, N. Kanda, J. Li, X. Chen, Y. Wu, and Y. Gong, "Internal Language Model Adaptation with Text-Only Data for End-to-End Speech Recognition," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 2608–2612.

[13] Z. Meng, W. Wang, R. Prabhavalkar, T. N. Sainath, T. Chen, E. Variani, Y. Zhang, B. Li, A. Rosenberg, and B. Ramabhadran, "JEIT: Joint End-to-End Model and Internal Language Model Training for Speech Recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5.

[14] Z. Meng, T. Chen, R. Prabhavalkar, Y. Zhang, G. Wang, K. Audhkhasi, J. Emond, T. Strohman, B. Ramabhadran, W. R. Huang, E. Variani, Y. Huang, and P. J. Moreno, "Modular Hybrid Autoregressive Transducer," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 197–204.

[15] T. Zenkel, R. Sanabria, F. Metze, J. Niehues, M. Sperber, S. Stüker, and A. Waibel, "Comparison of Decoding Strategies for CTC Acoustic Models," in *Proc. Interspeech 2017*, 2017, pp. 513–517.

[16] N. Das, M. Sunkara, S. Bodapati, J. Cai, D. Kulshreshtha, J. Farris, and K. Kirchhoff, "Mask the Bias: Improving Domain-Adaptive Generalization of CTC-Based ASR with Internal Language Model Estimation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.

[17] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, E. Blanco and W. Lu, Eds. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[18] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015, pp. 167–174.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[20] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec. 2013, pp. 273–278.

[21] W. Zhou, R. Schlüter, and H. Ney, "Full-Sum Decoding for Hybrid Hmm Based Speech Recognition Using LSTM Language Model," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7834–7838.

[22] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 3743–3747.

[23] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech 2015*. ISCA, Sep. 2015, pp. 3214–3218.

[24] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "Flat-Start Single-Stage Discriminatively Trained HMM-Based Models for ASR," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1949–1961, 2018.

[25] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, Jul. 2011, pp. 187–197.

[26] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Interspeech 2018*. ISCA, Sep. 2018, pp. 2207–2211.

[27] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.