Abstract—In automatic speech recognition (ASR) research, discriminative criteria have achieved superior performance in DNN-HMM systems. Given this success, the adoption of discriminative criteria is promising to boost the performance of end-to-end (E2E) ASR systems. With this motivation, previous works have introduced the minimum Bayesian risk (MBR, one of the discriminative criteria) into E2E ASR systems. However, the effectiveness and efficiency of the MBR-based methods are compromised: the MBR criterion is only used in system training, which creates a mismatch between training and decoding; the on-the-fly decoding process in MBR-based methods results in the need for pre-trained models and slow training speeds. To this end, novel algorithms are proposed in this work to integrate another widely used discriminative criterion, lattice-free maximum mutual information (LF-MMI), into E2E ASR systems not only in the training stage but also in the decoding process. The proposed LF-MMI training and decoding methods show their effectiveness on two widely used E2E frameworks: Attention-Based Encoder-Decoder (AEDs) and Neural Transducers (NTs). Compared with MBR-based methods, the proposed LF-MMI method: maintains the consistency between training and decoding; eschews the on-the-fly decoding process; trains from randomly initialized models with superior training efficiency. Experiments suggest that the LF-MMI method outperforms its MBR counterparts and consistently leads to statistically significant performance improvements on various frameworks and datasets from 30 hours to 14.3 k hours. The proposed method achieves state-of-the-art (SOTA) results on Aishell-1 (CER 4.10%) and Aishell-2 (CER 5.02%) datasets. Code is released\(^1\).

Index Terms—Automatic speech recognition, discriminative training, end-to-end, maximum mutual information, minimum Bayesian risk, sequential training.

I. INTRODUCTION

IN RECENT research of automatic speech recognition (ASR)\(^1\), great progress has been made due to the advances in neural network architecture design [1], [2], [3], [4], [5] and end-to-end (E2E) frameworks [6], [7], [8], [9], [10], [11], [12]. Without the compulsory forced-alignment and external language model integration, the end-to-end systems are also becoming increasingly popular due to its compact working pipeline. Today, E2E ASR systems have achieved state-of-the-art results on a wide range of ASR tasks [13], [14]. Currently, attention-based encoder-decoders (AEDs) [6], [7] and neural transducers (NTs) [8] are two of the most popular frameworks in E2E ASR. In general practice, training criteria like cross-entropy (CE), connectionist temporal classification (CTC) [15] and transducer loss [8] are adopted in AED and NT systems. However, all of the three E2E criteria try to directly maximize the posterior of the transcription given acoustic features but never attempt to consider the competitive hypotheses and optimize the model discriminatively.

Given the success of discriminative training criteria (e.g., MPE [16], [17], [18], [19], sMBR [20], [21] and MMI [22], [23], [24], [25], [26], [27], [28]) in DNN-HMM systems, integrating these criteria into E2E ASR systems is promising to further advance the performance of E2E ASR systems. Although several efforts [29], [30], [31], [32], [33], [34], [35], [36], [37], [38] have been spent on this field, most of these works [30], [31], [32], [33], [34], [35], [36], [37], [38] focus on applying minimum Bayesian risk (MBR) discriminative training criterion to E2E ASR systems. However, the effectiveness and efficiency of these MBR-based methods are compromised by several issues. Firstly, all of these methods adopt the MBR criterion in system training but still use the maximize-a-posterior (MAP) paradigm during decoding. The mismatch between system training and decoding can lead to sub-optimal recognition performance. Secondly, the Bayesian risk objective is usually approximated based on the N-best hypothesis list and the corresponding posteriors, which is generated by the on-the-fly decoding process. This process, however, requires a pre-trained model for initialization and makes these methods in the two-stage style. Thirdly, the on-the-fly decoding process is much more time-consuming, which results in the slow training speed of MBR-based methods [29].

To this end, we propose to integrate lattice-free maximum mutual information [22], [23] (LF-MMI, another discriminative

\(^{1}\)https://github.com/jctian98/e2e_lfmmi
This paper is among the first works that propose to adopt M
is a token in a
and
and
A systemic analysis between the MBR-based methods and
∈V
<

(1)
W
ˆ
is a
U
. Finally, we define
w
(3)
=R
is usually adopted as the evaluation metric of
o
The proposed method achieves consistent error reduction
Three novel decoding algorithms with LF-MMI criterion
E2E ASR
U
=end-of-sentence
W
as the training
E
and
This paper proposes an efficient way to incorporate dis-
P
(2)
O
¯
and the
5.02% on Aishell-1 test set and Aishell-2 test-ios set respec-
tively, which, to the best of our knowledge, are the state-of-the-
performance. Moreover, integrating LF-MMI scores in the decoding
hypothesis list generated without LF-MMI is further rescored by
the LF-MMI scores. Compared with the MBR-based methods,
our method (1) maintains consistent training and decoding under
MAP paradigm; (2) eschews the sampling process, approximate
the objective through efficient finite-state automaton algorithms
and work from scratch; and (3) maintains a much faster training
speed than its MBR counterparts.

Experimental results suggest that adding LF-MMI as an ad-
ditional criterion in training can improve the recognition perform-
ance. Moreover, integrating LF-MMI scores in the decoding
stage can further improve the performance of both AED and NT
systems. The best of our models achieves CER of 4.10% and
5.02% on Aishell-1 test set and Aishell-2 test-ios set respect-
ively, which, to the best of our knowledge, are the state-of-the-
art results on these two datasets. The proposed method is also
applicable to tiny-scale and large-scale ASR tasks: up to 19.6%
and 9.9% relative CER reductions are obtained on a 30 hours
corpus and a 14.3k-hour corpus respectively.

To conclude, we propose a novel approach to integrate the dis-
criminative training criteria, LF-MMI, into E2E ASR systems.
The main contributions of this work are summarized as follow:

- This paper is among the first works that propose to adopt
the LF-MMI criterion in both training and decoding stages of
AED and NT frameworks while previous works [30], [31], [32], [33] only consider the training process and only
address a single E2E ASR framework.
- This paper proposes an efficient way to incorporate dis-
criminative training criteria into E2E system training. Compared with previous work [30], [31], [32], [33], the
proposed approach is free from the pre-trained model and
the on-the-fly decoding process.
- Three novel decoding algorithms with LF-MMI criterion
are presented, covering the first-pass decoding and the
second-pass rescoring of AED and NT systems.
- A systemic analysis between the MBR-based methods and
proposed LF-MMI method is delivered to provide more
insight into discriminative training of E2E ASR.
- The proposed method achieves consistent error reduction
on various data volume and achieve state-of-the-art (SOTA)
results on two widely used Mandarin datasets (Aishell-1 and Aishell-2).

This journal paper is an extension of our conference pa-
paper [39]. The extension of this paper mainly falls in three ways:
(1) the decoding algorithms, including the efficient computation
of MMI Prefix Score and the look-ahead mechanism in MMI
Alignment Score, are updated from the original version to
achieve better performance; (2) the MBR-based methods are
carefully investigated and compared with the proposed method,
which is also untouched in the conference paper; (3) more
detailed experimental analysis on various hyper-parameters and
corpus scale (range from 30 hours to 14.3 k hours) is provided
while only preliminary results are reported in the conference
paper. The rest of this paper is organized as follows. The MBR-
based training methods in E2E ASR are analyzed in Section II.
The proposed LF-MMI training and decoding methods are pre-
sented in Sections III and IV respectively. The experimental
setup and the results are reported in Sections V and VI. We
discuss and conclude in Section VII.

II. MBR-BASED METHODS IN E2E ASR

In the rest of this paper, O = [o1, . . . , oT] and W = [w1, . . . , wU]
represent the input feature sequence and the token
sequence with length T and U respectively. Here oi ∈ Rd is a
dimensional speech feature vector while wi ∈ V is a token in a
known vocabulary V. In addition, we note O1 = [o1, . . . , o1] and
W1 = [w1, . . . , wu] as the first T frames and the first u tokens of
O and W respectively. Subsequently, H(W) is defined as the
set of token sequences that start with W i. Finally, we define <
<sos> and <eos> as the start-of-sentence and end-of-sentence
respectively while <blk> stands for the blank symbol used in
NT systems.

In this section, we briefly introduce the MBR-based methods
and their deficiencies.

A. MBR-Based Methods in E2E ASR

The motivation of MBR-based methods in E2E ASR is to
solve the mismatch between the training objective and the eval-
uation metric. Conventional E2E ASR models are optimized to
maximize the posterior of the transcription given the acoustic
features:

JMAP(W, O) = P(W|O) (1)

where W is the transcription text sequence. In decoding, the
most probable hypothesis is considered as the recognition result.

W = arg max P(W|O) (2)
W∈H([<sos>])

However, as the edit-distance (a.k.a., Word Error Rate, WER,
for languages like English; Character Error Rate, CER, for
languages like Chinese) between the hypothesis W and the
transcription W is usually adopted as the evaluation metric of
ASR task [40], there is no guarantee that the hypothesis with
the highest posterior is exactly the hypothesis with smallest
edit-distance, which is a mismatch problem and may result in
sub-optimal system performance.

To alleviate this mismatch, the MBR-based methods have
been proposed to directly take the edit-distance as the training
objective. Formally, the Bayesian risk objective to minimize is
formulated as:

JMBR(W, O) = EJ(W|O) | u(W, W)

= ∑ W∈H([<sos>]) P(W|O) · u(W, W) (3)
where the $u(\cdot)$ is the risk function (also known as utility function) while $W$ is any possible hypothesis. In most of the MBR-based methods in E2E ASR, edit-distance($\cdot$) is adopted as the risk function so the objective to minimize is exactly the expected edit-distance. If so, the MBR methods are also known as minimum word error rate (MWER) methods.

In real practice, however, the computation of the Bayesian risk is prohibitive, as the elements in the distribution $P(W|O)$ cannot be fully explored. Therefore, the objective of MBR-based methods is approximated by sampling over the set of all hypothesis:

$$J_{\text{MBR}}(W, O) = \frac{\sum_{W \in \mathcal{W}} P(W|O) \cdot u(W, W)}{\sum_{W \in \mathcal{W}} P(W|O) + \epsilon}$$  \hspace{1cm} (4)$$

where $\mathcal{W}$ is a subset of $H([<\text{sos}>])$ sampled from the posterior distribution $P(W|O)$ while the probability of each hypothesis in $\mathcal{V}$ is normalized by the summed probability over $\mathcal{W}$, $\epsilon$ is a smooth constant to avoid numerical errors. In existing MBR-based E2E methods, $\mathcal{W}$ is the N-best hypotheses list so the accumulated probability mass over $\mathcal{W}$ is maximized and a better approximation of $H([<\text{sos}>])$ is achieved. The N-best hypothesis list is generated by the decoding process, which is usually implemented on-the-fly during training (a.k.a, on-the-fly decoding). The workflow of MBR-based training methods is described in Fig. 1.

Although all methods [30], [31], [32], [33], [34], [35], [36], [37], [38] roughly follow the paradigm above, there are still some differences among these methods. Firstly, [30], [31] work on AED systems while [32], [33], [34], [35], [36], [37], [38] serve NT systems and their derivatives [34], [36]. Secondly, [31] also tries to generate $\mathcal{V}$ by random sampling rather than the N-best hypothesis list but performance degradation is observed. Thirdly, [32] proposes to replace the on-the-fly decoding with offline decoding and update the decoding results partially every after an epoch of data is exhausted for faster training speed. Fourthly, [33], [34], [35], [36], [37] propose to integrate external language models during the MBR training to better leverage the linguistic information [33], [34], [35] or emphasize the rare words [36], [37]. Fifthly, [38] discusses the impact of the input length in the MBR training.

Besides, there are also some attempts to introduce the MBR criterion during the decoding stage of DNN-HMM system [41], [42] or in machine translation community [43], which is out of the scope of this work.

### B. Deficiencies of MBR-Based Methods

Although MBR-based methods are found effective in improving recognition accuracy, we claim there are still several deficiencies in these methods.

1) **Mismatch Between Training and Decoding:** As shown in (1) and (2), E2E ASR systems that adopt non-discriminative training criteria usually achieve the consistency between training and decoding: the posterior probability of the transcription is maximized during training while the goal of decoding is to find the hypothesis with maximum posterior probability. This consistency, however, fails in MBR-based methods, as the training objective is shifted to minimize the Bayesian risk (a.k.a, expected WER) but the searching target is still to find the most probable hypothesis, rather than the hypothesis with the smallest Bayesian risk or its approximation.

2) **The Need for a Pre-Trained Model for Initialization:** As mentioned above, the approximation of the Bayesian risk objective in existing MBR-based methods requires the on-the-fly decoding process. In addition, the hypotheses generated on-the-fly should be roughly correct and occupy considerable probability mass to make the objective optimizeable. For this reason, the MBR-based methods need a seed model pre-trained from non-discriminative training criteria for initialization to ensure the success of the on-the-fly decoding, which is in the two-stage style and leads to complex training workflow.

3) **Slow Training Speed:** Even though the on-the-fly decoding is successfully implemented, it would significantly slow down the training speed of the MBR-based methods multiple times. This is also reported in [29].

Besides the deficiencies aforementioned, we experimentally find the Bayesian risk objective is hard to be optimized and the performance ceiling of these methods is limited. As the training data is well-fitted after several epochs of training, it is much unlikely that an erroneous hypothesis with high probability mass would be generated during the on-the-fly decoding after the training data. Instead, we experimentally find that the transcription $W$ with zero Bayesian risk usually occupies a dominant share of probability mass so the Bayesian risk is too small to optimize. We further discuss this observation in Section VI-G.
where $W$ represents any possible hypothesis. Similar to MBR-based methods, directly enumerating $W$ is almost impossible in practice. However, instead of using the N-best hypothesis list like MBR methods, Lattice-Free MMI [22], [23] is proposed to approximate the numerator and denominator in (5) by forward-backward algorithm on two Finite-State Acceptors (FSAs). The log-posterior of $W$ is then converted into the ratio of likelihood given the FSAs and acoustic features as follow [24]:

$$J_{LF-MMI} = \log P_{LF-MMI}(W|O) \approx \log \frac{P(O|G_{num})}{P(O|G_{den})}$$ (6)

where $G_{num}$ and $G_{den}$ denotes the FSA numerator graph and denominator graph respectively.

Unlike the lattice-based MMI method [25], the construction of the denominator graph in LF-MMI is identical to all utterances, which avoids the pre-decoding process before training and allows the system to be built from scratch. The mono-phone modeling units are adopted in the LF-MMI criterion, as a large number of modeling units (e.g. Chinese characters) makes the denominator graph computationally expensive and memory-consuming.

**B. AED Training With LF-MMI**

Attention-based Encoder-Decoder (AEDs) are a series of frameworks that adopt the encoder-decoder architecture to directly learn the mapping from the acoustic features to the transcriptions. In this work, we take the widely used hybrid CTC-attention framework [7] as the baseline model of AEDs. As shown in Fig. 2, the acoustic encoder is used to encode the acoustic features $O$. Given the embeddings of the token sequence $W$ and the hidden output of the encoder, the attention decoder tries to predict the next tokens in auto-regressive style and is supervised by the cross-entropy criterion. In [7], the acoustic encoder is additionally supervised by the token-level CTC criterion [15]. Thus, the training objective of this system is the interpolated value between the cross-entropy (CE) criterion and token-level CTC criterion:

$$J_{AED} = \alpha_{AED} \cdot J_{CE} + (1 - \alpha_{AED}) \cdot J_{CTC}$$ (7)

where $0 < \alpha_{AED} < 1$ is a adjustable hyper-parameter. We refer [7] for more details.

Similarly to [7], the LF-MMI criterion is added as an auxiliary criterion to optimize the acoustic encoder in AED systems in this work. As shown in (8), the LF-MMI criterion is used to replace or cooperate with the CTC criterion.

$$J_{AED} = \begin{cases} \alpha_{AED} \cdot J_{CE} + (1 - \alpha_{AED}) \cdot J_{CTC} \\ + (1 - \alpha_{AED}) \cdot J_{LF-MMI} \\ \alpha_{AED} \cdot J_{CE} + (1 - \alpha_{AED}) \cdot J_{LF-MMI} \end{cases}$$ (8)

**C. NT Training With LF-MMI**

NT is another framework that is widely used in E2E ASR. As shown in Fig. 3, a typical NT system consists of an acoustic encoder, a prediction network and a joint network. Unlike the AED system that directly maximizes the posterior of $W$ (a.k.a., $P(W|O)$), the NT system tries to maximize the summed probability of a set of expanded hypotheses:

$$J_{NT} = \log P(W|O) = \log \sum_{W' \in B^{-1}(W)} P(W'|O)$$ (9)

where each element in $B^{-1}(W)$, a.k.a., $W = [\hat{w}_1, \ldots, \hat{w}_{T+U}]$, is an expanded hypothesis generated from the transcription $W$. The length of $W$ is $T+U$ and each element in $W$ is either a token in vocabulary $V$ or $<blk>$. We refer [8] for more details.

Similar to the AED system, the LF-MMI criterion is added on the acoustic encoder and the token-level CTC criterion is optionally used. The training objective is then revised as:

$$J_{NT} = \begin{cases} J_{NT} + \alpha_{NT} \cdot J_{CTC} + \alpha_{NT} \cdot J_{LF-MMI} \\ J_{NT} + \alpha_{NT} \cdot J_{LF-MMI} \end{cases}$$ (10)

**IV. LF-MMI DECODING FOR E2E ASR**

To consistently use the LF-MMI criterion in both system training and decoding stages, three different decoding algorithms are proposed in this section. Specifically, MMI Prefix Score ($J_{MMI}^{pref}$) and MMI Alignment Score ($J_{MMI}^{ali}$) are proposed in Sections IV-A and IV-B respectively to integrate LF-MMI scores into beam search of AEDs and NTs consistently. In addition, we also propose a rescoring method using LF-MMI in Section IV-C. We finally compare our LF-MMI training and decoding method with MBR-based methods in Section IV-D.

**A. AED Decoding With LF-MMI**

Following (2), given the acoustic feature sequence $O$ and $W|_{<sos>} = [\text{<sos>}]$, the goal of MAP decoding for AED systems is to search the most probable hypothesis $W \in \mathcal{H}([\text{<sos>}])$. Normally, this searching process is approximated by the beam search algorithm. Assume $\Omega_n$ is the set of active partial hypotheses with length $n$ during decoding. The set $\Omega_n$ is recursively
generated by expanding each partial hypothesis in $\Omega_{a-1}$ and pruning those expanded partial hypotheses with lower scores. This iterative process would continue until a certain stopping condition is met. Typically, we set $\Omega_0 = \{\{< sos>\}\}$ while all hypotheses in any $\Omega_a$ that end with $<eos>$ would be moved to a finished hypothesis set $\Omega_F$ for final decision.

The computation of partial scores is the core of beam search. Partial score $\alpha(W^u_1, O)$ of a partial hypothesis $W^u_1$ is recursively computed as:

$$\alpha(W^u_1, O) = \alpha(W^{u-1}_1, O) + \log p(w_u|W^{u-1}_1, O)$$ (11)

where $\log p(w_u|W^{u-1}_1, O)$ is the weighted sum of different log-probabilities possibly delivered by the attention decoder, the acoustic encoder and the language models. In this work, log-probability distribution provided by LF-MMI, namely log $p_{LF-MMI}(w_u|W^{u-1}_1, O)$, is additionally considered as a component of log $p(w_u|W^{u-1}_1, O)$ with a pre-defined weight $\beta_{\text{LF}}$. The LF-MMI posterior $\log p_{LF-MMI}(w_u|W^{u-1}_1, O)$ can be derived from the first-order difference of $S_{\text{pref}}^{\text{num}}$:

$$\log p_{LF-MMI}(w_u|W^{u-1}_1, O) = S_{\text{pref}}^{\text{num}}(W^{u}_1, O) - S_{\text{pref}}^{\text{num}}(W^{u-1}_1, O)$$ (12)

where MMI Prefix Score $S_{\text{pref}}^{\text{num}}$ is defined as the summed probability of all hypotheses that start with the known prefix $W^u_1$. Given the fully known $W$ and $O$ (typically non-streaming scenarios), the MMI Prefix Score in (13) is firstly formulated. Secondly, $W$ is split into two part with the known $u$. Next, the $O$ is further split into two part. Since any $t$ is valid, the enumeration along t-axis is needed. In the approximation, we consider the conditional independence: for any known $t$, $O^t_1$ is independent to $O^{t+1}_1$ and $W^{t+1}_1$, while $W^{t}_1$ is independent to $O^{t}_1$ and $W^{t+1}_1$ (also see (2.17) and (2.18) in [44]). In the fifth line, the probability sum over the set $H(W^t_1)$ is equal to 1 and then discarded. In the final equation, each element $P_{\text{HMT}}(W^u_1|O^t_1)$ is approximated by (6), where $G_{\text{num}}(W^u_1)$ is the numerator graph built from $W^u_1$.

$$S_{\text{HMT}}^{\text{num}}(W^u_1, O) = \log \sum_{W \in H(W^t_1)} P_{\text{HMT}}(W|O)$$

$$= \log \sum_{W \in H(W^t_1)} P_{\text{HMT}}(W^U_{u+1}|O^T_1)P_{\text{HMT}}(W^U_1W^U_{u+1}, O^T_1)$$

$$= \log \sum_{W \in H(W^t_1)} P_{\text{HMT}}(W^U_{u+1}|O^t_1, O^{t+1}_1)$$

$$\approx \log \sum_{t=1}^{T} \sum_{W \in H(W^t_1)} P_{\text{HMT}}(W^U_1|O^t_1)P_{\text{HMT}}(W^U_{u+1}|O^{t+1}_1)$$

$$\approx \log \sum_{t=1}^{T} \sum_{W \in H(W^t_1)} P_{\text{HMT}}(W^U_1|O^t_1)P_{\text{HMT}}(W^U_{u+1}|O^{t+1}_1)$$

In (13), the accumulation of probability along the t-axis seems computationally expensive. However, several properties of it can be considered to greatly alleviate this problem. Firstly, unlike in the training stage, only the forward part of the forward-backward algorithm is needed to calculate all terms in (13). Secondly, the computation on the denominator graph is independent of the partial hypothesis $W^t_1$, which can be done before the searching process and reused for any partial hypothesis proposed during beam search. Thirdly, all items in series $P(O^t_1|G^s)$ can be obtained through the computation of the last term $P(O^t_1|G^s)$ by recording the intermediate variables of the forward process. As a result, the forward computation is conducted only once for each partial hypothesis. A more detailed description of this process is provided in Appendix A.

Details of the AED beam search process with MMI Prefix Score integrated are shown in Algorithm 1. In each decoding step, the log-posteriors from attention decoder, CTC and LF-MMI are updated into the score of each hypothesis $s(W^t_1)$ and the ended hypotheses will be moved into the finished hypothesis set. After the scores of all newly proposed hypotheses are updated, only $b$ hypotheses with highest scores can be kept in BeamPrune() process before entering the next decoding step. The iteration will continue until the StopCondition() is satisfied (see (50) in [7]). After the loop is ended, the hypotheses in the finished hypothesis set are sorted by their scores and returned.

### B. NT Decoding With LF-MMI

For NTs, MMI Alignment Score $S_{\text{all}}^{\text{num}}$ is proposed to cooperate with the decoding algorithm ALSD [45]. Note tuple $(W^u_1, s_t(W^u_1))$ as a hypothesis where $W^u_1$ is the output sequence (including no $<blk>$) with length $u$, $s_t(W^u_1)$ is the hypothesis score. The subscript $t$ in $s_t(W^u_1)$ means the hypothesis is aligned to first $t$ frames $O^t_1$.

As hypotheses in NT decoding obtain explicit alignments $(t, u)$, the proposed $S_{\text{all}}^{\text{num}}$ should also depend on this $t$-$u$ relationship. Unlike the MMI Prefix Score that only considers the $u$-axis, we define the MMI Alignment Score as:

$$S_{\text{all}}^{\text{num}}(W^u_1, O^t_1) = \max_{0 \leq i \leq \tau} \log P_{\text{HMT}}(W^t_1|O^{t+i}_1)$$ (14)

where $\tau$ is a hyper-parameter named look-ahead steps. Thus, it is natural to conduct the beam search process of ALSD based on the interpolated value of the NT log-posterior and the MMI Alignment Score (also a log-posterior). In our implementation, once a new hypothesis is proposed (a $<blk>$ or a new non-blank token is added), its score $s_{t+1}(W^t_1)$ or $s_t(W^t_1)$ is computed recursively using (15) and (16):

$$s_{t+1}(W^t_1) = s_t(W^t_1) + \log P_{\text{HMT}}(W^t_1|O^{t+i}|<blk>) + \beta_{\text{all}} * (S_{\text{all}}^{\text{num}}(W^t_1, O^{t+i}_1) - S_{\text{all}}^{\text{num}}(W^t_1, O^{t+i}_1))$$

$$s_t(W^{t+1}_1) = s_t(W^t_1) + \log P_{\text{HMT}}(W^{t+1}_1|O^t_1) + \beta_{\text{all}} * (S_{\text{all}}^{\text{num}}(W^{t+1}_1, O^t_1) - S_{\text{all}}^{\text{num}}(W^t_1, O^t_1))$$ (15)
Algorithm 1: AED Beam Search With MMI Prefix Score.

Require: acoustic feature sequence \( O \), vocabulary \( V \)
Require: beam size \( b \)
Require: decoding weight \( \beta_{\text{ATT}}, \beta_{\text{CTC}}, \beta_{\text{MMI}} \)

1: Initial Hypothesis Set \( \Omega_0 \leftarrow \{ \langle \text{eos}, 0 \rangle \} \)
2: Finished Hypothesis Set \( \Omega_F \leftarrow \emptyset \)
3: Decoding Step \( u \leftarrow 1 \)
4: while not StopCondition(\( \Omega_F, u \)) do
5: \( \Omega_u \leftarrow \emptyset \)
6: for \( (W_1^{u-1}, s(W_1^{u-1})) \in \Omega_{u-1} \) do
7: for \( w_u \in V \) do
8: \( W_1^{u} \leftarrow W_1^{u-1} + w_u \)
9: \( s(W_1^{u}) \leftarrow s(W_1^{u-1}) \)
10: \( s(W_1^{u}) \leftarrow s(W_1^{u}) + \beta_{\text{ATT}} \cdot \log P_{\text{ATT}}(w_u|W_1^{u-1}, O) \)
11: \( s(W_1^{u}) \leftarrow s(W_1^{u}) + \beta_{\text{CTC}} \cdot \log P_{\text{CTC}}(w_u|W_1^{u-1}, O) \)
12: \( s(W_1^{u}) \leftarrow s(W_1^{u}) + \beta_{\text{MMI}} \cdot \log P_{\text{MMI}}(w_u|W_1^{u-1}, O) \)
13: end for
14: if \( w_u = <\text{eos}> \) then
15: \( \Omega_F \leftarrow \Omega_F \cup (W_1^{u}, s(W_1^{u})) \)
16: else
17: \( \Omega_u \leftarrow \Omega_u \cup (W_1^{u}, s(W_1^{u})) \)
18: end if
19: end for
20: \( \Omega_u \leftarrow \text{BeamPrune}(\Omega_u, b) \)
21: \( u \leftarrow u + 1 \)
22: end while
23: return Sorted(\( \Omega_F \))

where the score provided by MMI criterion is obtained from the first-order difference over \( S_{\text{MMI}}^1(W_1^{u}, O_1^{i}) \) along t-axis or u-axis respectively. \( \beta_{\text{MMI}} \) is an adjustable hyper-parameter in decoding stage. Similar to \( S_{\text{pref}}^1 \), we approximate \( S_{\text{MMI}}^1 \) by (6). The denominator scores can also be reused in all hypotheses and the series \( P_{\text{MMI}}(W_1^{t}, O_1^{i+t}) \), \( i = 0, \ldots, t \) can also be computed in one go.

Compared with the peak scores provided by the NT system, we observe that there is usually a time-dimensional delay in LF-MMI peak scores (see dashed lines in Fig. 4). An explanation of this delay is: the LF-MMI criterion needs more frames to reach the peak score since more arcs for the newly proposed token are added in the numerator graph. By contrast, when a non-blank token is proposed by the NT system, its score can reach the peak value even without increasing the \( t \) in the t-u alignment. To align the scores provided by NT and LF-MMI, a look-ahead mechanism is needed: it encourages the MMI criterion to provide an early validation against the newly proposed tokens during decoding.

We also note, in each step when all proposed hypotheses are evaluated, scores of hypotheses that have identical \( W_1^{t} \) but different alignment paths should be merged. But \( S_{\text{MMI}}^1 \) should not participate in this process, since \( S_{\text{MMI}}^1 \) directly assesses the validity of the aligned sequence pair \( (W_1^{t}, O_1^{i}) \) and is the summed posterior of all alignment paths.

We finally provide an overview of the NT beam search process with the MMI Alignment Score integrated into Algorithm 2. The log-posteriors provided by the LF-MMI criterion are adopted whenever a <\text{eos}> or a non-blank token is newly proposed. After the scores \( s_i(W_1^{u}) \) of each hypothesis is updated, the hypotheses with the identical blank-removed sequence but different alignment paths should be merged [45] using the function CombineHypothesis(). Then, only \( b \) hypotheses with highest scores can be kept in the BeamPrune() process before entering the next decoding step. After the decoding loop is ended, the hypotheses in the finished hypothesis set are sorted by their scores and returned.

C. LF-MMI Rescoring

We further propose a unified rescoring method called LF-MMI Rescoring for both AEDs and NTs that are jointly optimized with the LF-MMI criterion. Compared with using \( S_{\text{MMI}}^1 \) and \( S_{\text{pref}}^1 \) in beam search, rescoring method is more computationally efficient.

Assume the AED or NT system has been optimized by the LF-MMI criterion before decoding. As illustrated in Fig. 5, the N-best hypothesis list is firstly generated by beam search without the LF-MMI criterion. Along this process, the log-posterior of each hypothesis \( W \), namely \( \log P_{\text{AED/NT}}(W|O) \), is also calculated. Next, another log-posterior for each hypothesis in the N-best hypothesis list, \( \log P_{\text{LF-MMI}}(W|O) \), is computed.

Fig. 4. Log-posteriors \( \log P(W_1^{t}, O_1^{i}) \) provided by LF-MMI and transducer with various frame index \( t \). \( W_1^{t} \) is a correct partial hypothesis with \( u = U - 2 \).

Fig. 5. Diagram for MMI Rescoring. Interpolation from original posteriors and MMI posteriors are used for the final decision.
Algorithm 2: NT Beam Search With MMI Alignment Score.

Require: acoustic feature sequence O, vocabulary V
Require: beam size b, maximum hypothesis length U_{\text{max}}
Require: decoding weight $\beta_{\text{mmi}}$

1: Initial Hypothesis Set $\Omega_1 \leftarrow \{\langle \text{sos} \rangle, \text{0}\}$
2: Finished Hypothesis Set $\Omega_F \leftarrow \emptyset$
3: Decoding Step $t \leftarrow 1$
4: for $l = 1, \ldots, T + U_{\text{max}}$ do
5: \hspace{1em} $\Omega_{l+1} \leftarrow \emptyset$
6: \hspace{1em} for $(W_1^l, s_t(W_1^l)) \in \Omega_l$ do
7: \hspace{2em} if $t > T$ then
8: \hspace{3em} continue
9: \hspace{2em} end if
10: \hspace{2em} $s_{t+1}(W_1^{l+1}) \leftarrow s_t(W_1^l) + \log P_{\text{MMI}}^{\text{WF}}(\langle \text{blank} \rangle | W_1^l, O)$
11: \hspace{2em} $s_{\text{mmi}} \leftarrow S_{\text{mmi}}^{\text{WF}}(W_1^l, O_1^{l+1}) - S_{\text{mmi}}^{\text{WF}}(W_1^l, O_1^l)$
12: \hspace{2em} $s_{t+1}(W_1^{l+1}) \leftarrow s_{t+1}(W_1^l) + \beta_{\text{mmi}} \cdot s_{\text{mmi}}$
13: \hspace{2em} $\Omega_{l+1} \leftarrow \Omega_{l+1} \cup (W_1^{l+1}, s_{t+1}(W_1^{l+1}))$
14: \hspace{2em} if $t = T$ then
15: \hspace{3em} $\Omega_F \leftarrow \Omega_F \cup (W_1^{l+1}, s_{t+1}(W_1^{l+1}))$
16: \hspace{2em} end if
17: \hspace{2em} end for
18: \hspace{2em} end for
19: \hspace{2em} CombineHypothesis($\Omega_{l+1}$)
20: \hspace{2em} BeamPrune($\Omega_{l+1}, b$)
21: \hspace{2em} end for
22: \hspace{2em} return Sorted($\Omega_F$)

according to the LF-MMI criterion. Finally, the interpolation of the two log-posteriors are calculated as follows:

$$\log P(W|O) = \log P_{\text{MMI}}^{\text{WF}}(W|O) + \lambda_{\text{mmi}} \cdot \log P_{\text{LF-MMI}}(W|O)$$

(17)

where $\lambda_{\text{mmi}}$ is the interpolation weight used in LF-MMI Rescoring. As the LF-MMI criterion is applied to the acoustic encoder, MMI Rescoring can better emphasize the validity of hypotheses from the perspective of acoustics. Moreover, since the denominator score $P(O|\mathcal{G}_{\text{den}})$ is independent of the hypotheses, it can be considered as a constant for different hypotheses of a given utterance. Thus, only the numerator scores need to be calculated during LF-MMI Rescoring:

$$\log P_{\text{MMI}}(W|O) = \log P(O|\mathcal{G}_{\text{num}}(W)) - \text{constant}$$

D. Compare With MBR-Based Methods

Here the proposed LF-MMI training and decoding method is compared with its MBR counterparts. We claim that the three deficiencies of MBR-based methods discussed in section II-B are solved or eschewed in our method.

1) Consistent Training and Decoding: The LF-MMI method achieves the consistency between training and decoding since the LF-MMI criterion is applied in both stages. By contrast, the MBR methods only consider the training process and apply no discriminative criterion during decoding.

2) Train From Randomly Initialized Model: To compute the Bayesian risk, the MBR-based methods require a pre-trained model for on-the-fly decoding. However, this is not necessary for the LF-MMI training method since the on-the-fly decoding process is no longer needed. Instead, the LF-MMI criterion, along with the non-discriminative criteria, is capable to train from a randomly initialized model.

3) Training Efficiency: The MBR-based methods are slow in training speed due to the adoption of the sampling process (usually the on-the-fly decoding process) over the hypothesis space. The LF-MMI training method, however, replaces the sampling process with the forward-backward algorithm on FSAs, which can be effectively implemented by matrix operators and is much faster.

V. EXPERIMENTAL SETUP

Experimental setup is described in this section. The datasets are described in Section V-A. The model, optimization and decoding configurations are described in Section V-B. The implementation of LF-MMI and MBR methods is presented in Section V-C.

A. Datasets

We evaluate the effectiveness of the proposed method on three Mandarin datasets: Aishell-1 [46], Aishell-2 [47] and our internal dataset Oteam. All utterances with non-Mandarin tokens are excluded. To examine the proposed method on low-resource ASR tasks, we additionally split a subset of the Aishell-1 training corpus, called Aishell-s, in which only 70 utterances for each speaker are kept. The scales of these datasets range from 30 hours to 14.3 k hours. Details of these datasets are listed in Table I.

| Dataset       | Train hrs | Num. utts | Vocabulary Size |
|---------------|-----------|-----------|-----------------|
| Aishell-1     | 30        | 23.8k     | 4231            |
| Aishell-1     | 178       | 120k      | 4231            |
| Aishell-2     | 1k        | 963k      | 5214            |
| Oteam         | 14.3k     | 14.6M     | 6267            |

B. Model, Optimization and Evaluation

The proposed method is evaluated on both AED and NT systems. For all experiments, we adopt a unified model architecture, optimization and decoding settings as described below.

1) AED: For the AED system, a 12-layer Conformer [2] encoder and a 6-layer transformer decoder are adopted. For all attention modules in encoder and decoder, the feed-forward dimension, the attention dimension and the number of attention
heads are 2048, 256 and 4 respectively. The CNN module in each Conformer layer has 31 kernels. All batch-normalization layers are replaced by group-normalization [49] layers with group number of 2 for training efficiency. The acoustic input is 80-dim Fbank features plus 3-dim pitch features. The input features are down-sampled by a 2-layer CNN with the factor of 4 before being fed into the Conformer encoder. The dimension of token embedding is 256. This architecture consumes around 46 M parameters in total.

2) NT: For the NT system, a 12-layer Conformer encoder, a single-layer LSTM prediction network and an MLP joint network are adopted. The encoder has the same architecture as that in AED except the down-sampling factor is 6. The dimension of token embedding and prediction network is 1024 and 512 respectively. The joint network is also in the size of 512. This architecture consumes around 90 M parameters.

3) Optimization: All experiments adopt the Adam optimizer [50] with warm-up steps of 25 k and inverse square root decay schedule [1]. The global batch-size, peak learning rate, number of GPUs and number of epochs for different datasets are listed in Table II. SpecAugment [51] is consistently adopted with two time masks and two frequency masks. We use the speed perturbation with factors 0.9, 1.0 and 1.1 for all Aishell datasets so their data scales are expanded by 3 times. All models are trained on Tesla P40 GPUs.

4) Evaluation: We average the checkpoints from the last 10 epochs for evaluation. The decoding algorithms for AED and NT systems are presented in [7] and [45] respectively. The beam size for all decoding stages is fixed to 10. During decoding, Word-level N-gram language models [52] trained from the training transcriptions are optionally integrated with a default weight of 0.4. Our implementation is mainly revised from Espnet [53], [54].

C. LF-MMI and MBR Implementation

1) LF-MMI Method Implementation: Our LF-MMI implementation mainly follows [23]. In terms of topology, the CTC-topology [55] is adopted in all experiments. As discussed in Section III-A, to reduce the memory and computational cost, phoneme is adopted as the modeling unit to compute the LF-MMI loss in this work, which means a phone-level lexicon is used. All lexicons of open-accessible datasets are from standard Kaldi recipes, and the lexicon in Aishell-2 is also used in Oteam dataset. All phones are used in context-independent format. When compiling the lexicon into FST, optional silence with the probability of 0.5 is also added. The order of the phone language model used in the FSA compilation is fixed to 2. No alignment information in any form is used. The generation of numerator and denominator graphs still follows [23] and the gradients are computed by standard forward-backward algorithm [56]. By default, the training hyper-parameters $\alpha_{\text{AED}}$ and $\alpha_{\text{NT}}$ are empirically set to 0.3 and 0.5 respectively. The decoding hyper-parameters $\beta_{\text{AED}}$ and the rescoring parameter $\lambda_{\text{NT}}$ are set to 0.2 consistently. The look-ahead step $\tau$ in MMBL Alignment Score computation is set to 3. The LF-MMI criterion is based on differentiable FST algorithms [57] and is implemented by k2.4

2) MBR-Based Method Implementation: To implement the MBR training, we initialize our model from the last checkpoint of the baseline model (AED or NT models without LF-MMI). The training lasts 10% epochs described in Table II, which is found necessary to achieve full convergence and obtain the benefit of model average. Due to the GPU memory limitation, we consistently adopt the beam size of 4 in all on-the-fly decoding processes. As MBR methods consume more GPU memory, we halve the mini-batch size and double the gradient accumulation number to ensure the same batch size. The smooth constant $\epsilon$ in (4) is set to $1e-10$. Following the implementation of [30], [33], the non-discriminative training criteria of the AED and NT system are also adopted for regularization and are interpolated with the MBR criterion equally. Note we use the raw input features for on-the-fly decoding, which is not corrupted by SpecAugment. As the number of epochs is greatly reduced in MBR training, we average the checkpoints from all epochs for evaluation.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

This section is organized as follows: To begin with, an overview of the proposed method on two open-source datasets is provided in Section VI-A to show the strength of the proposed method against previous methods. To provide a detailed analysis of each component in the proposed method, the ablation study and the hyper-parameter investigation are conducted in Sections VI-B and VI-C respectively. To verify the generalization of the proposed method on both small and large scale speech corpus, experiments on a 30-hour dataset and a 14.3k-hour dataset are then provided in Sections VI-D and VI-E. Finally, to show the performance difference between the proposed LF-MMI method and previous MBR-based methods on E2E ASR, a comparison between these two methods is given in Sections VI-F and VI-G.

A. Overall Results on Aishell-1 and Aishell-2

In Table IV, we present our best results obtained by the proposed LF-MMI method on the open-sourced Aishell-1 and Aishell-2 datasets for both AED and NT systems. To show the strength of the proposed method, competitive results in the previous publication are also provided.

As suggested in Table III, the proposed LF-MMI training and decoding methods achieve significant performance improvement on both AED and NT systems and on all evaluation sets. For the NT system, up to 15.3% relative error reduction is achieved

| Dataset | Batch size | Peak LR | Num. GPUs | Num. epochs |
|---------|------------|---------|-----------|-------------|
| Aishell-1 | 64 | 3e-4 | 8 | 100 |
| Aishell-2 | 512 | 3e-4 | 8 | 50 |
| Oteam | 1024 | 3e-3 | 32 | 30 |

4https://github.com/k2-fsa/k2
between the baseline systems and the LF-MMI works.

Similarly, the token-level CTC criterion can be considered as another regularization for LF-MMI in E2E frameworks. In previous literature working on DNN-HMM systems [22], [23], the LF-MMI criterion is sensitive to over-fitting and requires regularization to alleviate this problem. In previous literature, significant CER reduction is achieved by word N-gram LMs no matter LF-MMI criterion is adopted or not. On the other hand, the adoption of word N-gram LMs provides larger performance improvement on NT system trained by the LF-MMI criterion.

B. Ablation Study

To systematically analyze the effectiveness of each component in the proposed LF-MMI training and decoding method, an ablation study is conducted from various perspectives on Aishell-1 and Aishell-2 datasets. The experimental results are reported in Table IV. Our main observations are listed as follows.

1) LF-MMI Training Method and CTC Regularization: For all experiments conducted on Aishell-1 and Aishell-2, token-level CTC criterion is found as a necessary regularization during training. As suggested in Table IV, if the CTC regularization is not adopted, the impact of using LF-MMI as an auxiliary criterion is marginal on the NT system (exp.8 vs. exp.9,14), which indicates that the potential benefit of additional phone-level information is limited and the improvement achieved by LF-MMI is from its discriminative nature.

2) LF-MMI Decoding Methods: The results in table IV also suggest that consistently using the LF-MMI criterion in both training and decoding stages can further improve the recognition performance (exp.5 vs. exp.6,7; exp.10 vs. exp.11,12; exp.15 vs. exp.16,17). In addition, the on-the-fly decoding methods slightly outperform the rescoring methods (exp.6 vs. exp.7; exp.11 vs. exp.12; exp.16 vs. exp.17).

3) Impact of Phone-Level Information: As our implementation of the LF-MMI criterion still adopts phone-level information (the external lexicon), one may challenge that the improvement of LF-MMI training should be attributed to this additional information. To address this concern, the LF-MMM criterion is replaced by the phone-level CTC criterion with the same lexicon for comparison. Compared with the baseline systems, the adoption of phone-level CTC criterion during training provides no performance improvement (exp.1 vs. exp.4; exp.8 vs. exp.9,14), which indicates that the potential benefit of additional phone-level information is limited and the improvement achieved by LF-MMI is from its discriminative nature.

4) Impact of Language Models: Word N-gram LMs are optionally adopted in all experiments and achieve consistent performance improvement. However, the improvement provided by the LMs varies for different E2E ASR frameworks. In AED systems, significant CER reduction is achieved by word N-gram LMs no matter LF-MMI criterion is adopted or not. On the other hand, the adoption of word N-gram LMs provides larger performance improvement on NT system trained by the LF-MMI criterion.

C. Impact of Hyper-Parameters

In this part, we investigate the impact of the training and decoding hyper-parameters on the Aishell-1 dataset and present the results in Table V. Based on these results, we claim the proposed method maintains superior robustness against hyper-parameters. During the training stage of both AED and NT systems, adopting LF-MMI as an auxiliary criterion is consistently helpful with various weights in {0.1, 0.3, 0.5, 0.8} (exp.1 vs. exp.2-5; exp.11 vs. exp.12-15). Additionally, the adoption of LF-MMI decoding methods, including both the beam search methods and the rescoring method, also provides consistent CER reduction with all weights in {0.05, 0.1, 0.2, 0.3} (exp.6 vs. exp.7-10; exp.16 vs. exp.17-20). These results suggest that the proposed LF-MMI training and decoding methods may lead to improvement without much engineering effort.

Also note the training weights of 0.3 (for AED), 0.5 (for NT) and decoding weights of 0.2 (for both AED and NT) achieve much promising results, which are the default settings in the experiments aforementioned. Finally, exp.21 suggests that the adoption of the look-ahead mechanism in the MMI Alignment Score is beneficial, even though the improvement is marginal.

D. Results on the Low-Resource Dataset

To further show the generalization ability of the LF-MMI-based methods on the low-resource dataset, experiments conducted on the 30-hour Aishell-s dataset are shown in Table VI. Note that the original dev and test sets of Aishell-1 are still used for evaluation in this experiment.
TABLE IV
DETAILED RESULTS ON AISHELL-1 AND AISHELL-2 DATASETS

| No. | System | \( \text{AISHELL-1 (17.8 hrs)} \) | \( \text{AISHELL-2 (1k hrs)} \) |
|-----|--------|-------------------------------|-------------------------------|
|     |        | w/o LM | w LM | w/o LM | w LM | w/o LM | w LM |
|     |        | dev | test | dev | test | dev | test | dev | test |
| 1   | Attention + char. CTC (baseline) | 4.60 | 5.07 | 4.33 | 4.71 | 6.60 | 5.72 | 6.38 | 6.07 |
| 2   | Attention + LF-MMI | - | - | - | - | - | - | - | - |
| 3   | + MMI Prefix Score | 4.83 | 5.36 | 4.32 | 4.83 | 7.42 | 6.08 | 6.91 | 5.58 |
| 4   | Attention + char. CTC + phone CTC | 4.69 | 5.24 | 4.40 | 4.85 | 6.77 | 5.83 | 6.67 | 6.27 |
| 5   | Attention + char. CTC + LF-MMI | 4.55 | 5.05 | 4.22 | 4.59 | 6.69 | 5.80 | 6.44 | 5.98 |
| 6   | + MMI Prefix Score | 4.55 | 5.10 | 4.08 | 4.45 | 6.75 | 5.80 | 6.46 | 5.88 |
| 7   | + LF-MMI Resoring | 4.55 | 5.10 | 4.08 | 4.49 | 6.75 | 5.80 | 6.46 | 5.92 |

Neural Transducer (NT)

| No. | System | \( \text{AISHELL-1 (17.8 hrs)} \) | \( \text{AISHELL-2 (1k hrs)} \) |
|-----|--------|-------------------------------|-------------------------------|
|     |        | w/o LM | w LM | w/o LM | w LM | w/o LM | w LM |
|     |        | dev | test | dev | test | dev | test | dev | test |
| 8   | Transducer (baseline) | 4.41 | 4.82 | 4.47 | 4.84 | 6.52 | 5.81 | 6.52 | 6.33 |
| 9   | Transducer + Phone CTC | 4.81 | 5.24 | 4.59 | 4.96 | 7.08 | 6.01 | 6.86 | 6.94 |
| 10  | Transducer + LF-MMI | 4.37 | 4.86 | 4.27 | 4.73 | 6.64 | 5.80 | 6.57 | 6.33 |
| 11  | + MMI Alignment Score | 4.29 | 4.81 | 3.92 | 4.38 | 6.58 | 5.70 | 6.56 | 5.97 |
| 12  | + LF-MMI Resoring | 4.28 | 4.78 | 3.98 | 4.46 | 6.58 | 5.73 | 6.57 | 6.05 |
| 13  | Transducer + char. CTC | 4.91 | 5.02 | 4.88 | 4.88 | 6.45 | 5.47 | 6.26 | 6.23 |
| 14  | Transducer + char. CTC + phone CTC | 4.54 | 5.04 | 4.39 | 4.82 | 6.62 | 5.58 | 6.45 | 6.49 |
| 15  | Transducer + char. CTC + LF-MMI | 4.25 | 4.64 | 4.08 | 4.37 | 6.47 | 5.51 | 6.29 | 6.14 |
| 16  | + MMI Alignment Score | 4.18 | 4.54 | 3.79 | 4.10 | 6.41 | 5.40 | 6.18 | 5.85 |
| 17  | + LF-MMI Resoring | 4.17 | 4.57 | 3.84 | 4.21 | 6.41 | 5.42 | 6.17 | 5.95 |

Cer % is reported. Word n-gram lms are optionally used.

TABLE V
CER% RESULTS ON AISHELL-1 WITH DIFFERENT TRAINING WEIGHTS \( (\alpha_{AED}, \alpha_{NT}) \), DECODING WEIGHTS \( (\beta_{MMI}, \gamma_{MMI}) \) AND LOOK-AHEAD STEPS \( \tau \) OF THE PROPOSED LF-MMI METHOD

| No. | System | Train Weight | Decoding Weight | look-ahead step | Beam Search | LF-MMI Resoring |
|-----|--------|---------------|-----------------|-----------------|-------------|-----------------|
|     |        | dev | test | dev | test | dev | test | dev | test |
| 1   | Attention + CTC (ABD, baseline) | - | - | - | - | 4.33 | 4.71 | 4.33 | 4.71 |
| 2   | AED + LF-MMI | 0.1 | - | - | - | 4.11 | 4.37 | 4.13 | 4.60 |
| 3   | + MMI Alignment Score | 0.3 | - | - | - | 4.08 | 4.45 | 4.08 | 4.49 |
| 4   | + LF-MMI Resoring | 0.5 | - | - | - | 4.09 | 4.51 | 4.10 | 4.54 |
| 5   | Transducer (NT, baseline) | 0.8 | - | - | - | 4.11 | 4.56 | 4.12 | 4.58 |
| 6   | NT + LF-MMI | 0.05 | - | - | - | 4.22 | 4.59 | 4.22 | 4.59 |
| 7   | + MMI Alignment Score | 0.3 | - | - | - | 4.16 | 4.51 | 4.17 | 4.53 |
| 8   | + LF-MMI Resoring | 0.1 | - | - | - | 4.13 | 4.49 | 4.13 | 4.51 |
| 9   | Transducer + char. CTC | 0.2 | - | - | - | 4.08 | 4.45 | 4.08 | 4.49 |
| 10  | + LF-MMI Resoring | 0.3 | - | - | - | 4.06 | 4.46 | 4.07 | 4.48 |
| 11  | NT + LF-MMI | 0.0 | 0.2 | 3 | - | 3.82 | 4.23 | 3.91 | 4.36 |
| 12  | + MMI Alignment Score | 0.3 | 3 | - | - | 3.77 | 4.18 | 3.86 | 4.29 |
| 13  | + LF-MMI Resoring | 0.5 | 0.2 | 3 | - | 3.79 | 4.10 | 3.86 | 4.21 |
| 14  | Transducer + char. CTC | 0.8 | 0.05 | 3 | - | 4.01 | 4.22 | 3.90 | 4.27 |
| 15  | + LF-MMI Resoring | 0.5 | 0.1 | 3 | - | 3.85 | 4.17 | 3.93 | 4.27 |
| 16  | + LF-MMI Resoring | 0.5 | 0.2 | 3 | - | 3.79 | 4.10 | 3.86 | 4.21 |
| 17  | Transducer + char. CTC | 0.3 | 0.3 | - | - | 3.86 | 4.14 | 3.84 | 4.17 |

All results are with word n-gram lms.

E. Results on the Large-Scale Dataset

To further show the effectiveness of the proposed method on the large-scale dataset, experiment conducted on the...
internal 14.3k-hour Oteam dataset is further presented in this section. In this experiment, five test sets are provided to assess the model performance in various real applications: speech translation (st), television (tv), music (mu), education (ed) and reading (re). We change the decoding coefficient $\beta_{\text{MMI}}$ and rescoring coefficient $\gamma_{\text{MMI}}$ to 0.05 while keeping the other hyper-parameters unchanged. All results are presented in Table VII. Our main observations on this dataset are reported as follows.

1) **Overall Improvement:** The proposed LF-MMI training and decoding method is still beneficial in large-scale scenario. For the AED system, our method provides an absolute CER reduction of 7.9% on average (exp.1 vs. exp.3) while this number for the NT system is 9.9% (exp.7 vs. exp.12). The maximum improvement is observed on the music (mu) test set, where the CER is reduced from 15.23% to 11.36%.

2) **Impact of CTC Regularization:** In the large-scale experiment, it is no longer that crucial to adopt CTC regularization during training. For the AED system, the model performance is consistently compromised by the adoption of CTC regularization (exp.3 vs. exp.4-6). For the NT system, the averaged CERs achieved by the systems with or without CTC regularization show limited difference. We suppose the LF-MMI criterion can work independently when the data scale is comparatively large as the over-fitting problem is alleviated.

### F. Compare With MBR-Based Method

To compare the proposed LF-MMI method with the MBR-based methods, we conduct MBR training methods on both AED and NT systems over Aishell-1 and Aishell-2 datasets. All results are presented in Table VIII.

In these experiments, the proposed LF-MMI method outperforms the MBR-based methods on both accuracy and speed consistently. Compared with the baseline systems, the MBR-based methods achieve marginal improvement on Aishell-1 AED experiment and encounter degradation in all other experiments. By contrast, the proposed LF-MMI method achieves consistent and considerable performance improvement on both datasets and both frameworks.

### G. Further Analysis on MBR-Based Methods

Although improvements are consistently observed in previous literature [30], [31], [32], [33], we can hardly achieve better CER results on either AED or NT systems with these MBR-based methods. To find the reasons, further investigation is conducted on the MBR-trained models with Aishell-1 dataset. Specifically, the results of on-the-fly decoding over the training set are analyzed and the statistics are presented in Fig. 6. Our main observations are reported as below.

1) **CER Over the Training Set:** As shown in the left column of Fig. 6, for all models trained by MBR criterion, the CER over the training set is very small (all below 0.5%). This small CER suggests the MBR-trained models have fit the training data to a satisfactory level. Note the motivation of MBR-based methods is to reduce the expected CER, this observation also verifies the MBR-based methods have fulfilled their original intention and reduced the CER over the training set effectively.

2) **Probability of Reference Token Sequence:** As presented in the middle column of Fig. 6, the accumulated probability of the N-best hypothesis list and the probability of the reference token sequence are much close, which means the probability occupied

### Table VII

| No. | System | w/o LM | w LM |
|-----|--------|-------|------|
|     |        | st    | tv   | mu   | ed   | re   | Mean | st    | tv   | mu   | ed   | re   | Mean |
| 1   | Attention + CTC (baseline) | 17.20 | 9.24 | 17.24 | 11.43 | 5.44 | 12.11 | 16.30 | 7.71 | 14.66 | 11.42 | 4.98 | 11.01 |
| 2   | Attention + LF-MMI | - | - | - | - | - | - | - | - | - | - | - |
| 3   | + MMI Prefix Score | 16.85 | 8.32 | 14.75 | 10.87 | 4.98 | 11.15 | 16.00 | 7.10 | 14.16 | 11.02 | 4.88 | 10.63 |
| 4   | Attention + CTC + LF-MMI | 17.13 | 8.69 | 15.48 | 11.26 | 5.38 | 11.59 | 16.14 | 7.38 | 13.25 | 11.28 | 4.92 | 10.59 |
| 5   | + MMI Prefix Score Decoding | 17.20 | 8.72 | 15.51 | 11.25 | 5.37 | 11.61 | 16.20 | 7.44 | 13.32 | 11.15 | 4.94 | 10.61 |
| 6   | + LF-MMI Rescoring | 17.16 | 8.73 | 15.46 | 11.27 | 5.34 | 11.59 | 16.17 | 7.43 | 13.29 | 11.28 | 4.91 | 10.62 |

### Table VIII

| System | Aishell-1 | Aishell-2 | relative train time |
|--------|-----------|-----------|---------------------|
|        | dev | test | android | ios | mic | 100 |
| AED    | 4.33 | 4.71 | 6.07 | 5.29 | 5.88 | 1.00 |
| + MBR  | 4.30 | 4.70 | 6.36 | 5.33 | 6.02 | 5.81 |
| + LF-MMI | 4.08 | 4.45 | 5.92 | 5.15 | 5.77 | 1.62 |
| NT     | 4.47 | 4.84 | 6.33 | 5.58 | 6.25 | 1.00 |
| + MBR  | 4.48 | 4.88 | 6.52 | 5.72 | 6.37 | 2.78 |
| + LF-MMI | 3.79 | 4.10 | 5.85 | 5.02 | 5.66 | 1.77 |
by the reference token sequence is dominant. This observation suggests the posterior distribution output by the MBR-trained models is very sharp and these MBR-trained models are much confident in the reference token sequence.

3) Approximated Bayesian Risk: With the two observations above, the approximated Bayesian risk in (4) tends to be very small, as the accumulated probability in the denominator is considerably large but the risk of the erroneous hypotheses with small probability is marginal in the numerator. This is also experimentally verified in the right column of Fig. 6: the mean value of the approximated Bayesian risk is less than 0.07 for the AED system and 0.02 for the NT system.

Given these observations, the MBR-trained models are believed to have fit the training data successfully: the approximated Bayesian risk objective becomes very small after this training. In addition, the adoption of larger beam size during the on-the-fly decoding provides limited increment on the Bayesian risk objective. However, given this promising results over the training set, generalizing these MBR-trained models to the evaluation sets still achieves limited improvement. This ineffectiveness may be attributed to the over-fitting problem: take the NT system as an example, the CER over test set is more than 10 times larger than that of the training set. By contrast, we observe this ratio for LF-MMI objective in both AED and NT systems is below 3.

VII. DISCUSSION AND CONCLUSION

This work proposes to integrate the discriminative training criterion, LF-MMI, into two of the most popular E2E ASR systems: AED and NT. Unlike the previous MBR-based methods that only consider the training stage, the proposed method can be consistently applied to training and decoding for better performance. Specifically, the LF-MMI criterion acts as an auxiliary criterion during training. In decoding, two algorithms are proposed for the on-the-fly decoding process of AED and NT systems respectively and a unified rescoring method using LF-MMI criterion is also provided.

In experiments, we show that the proposed methods lead to consistent improvement. The training and decoding methods are carefully examined on 4 datasets ranging from tens hours to more than 10 k hours. The impact of regularization, language model fusion, hyper-parameters and multi-scale data volume are analyzed in depth. The impact of phone-level supervision is experimentally excluded so the improvement obtained in the training stage should be attributed to the discriminative nature of the LF-MMI criterion. For results, state-of-the-art performances are achieved on two of the most popular Mandarin datasets: Aishell-1 and Aishell-2. In addition, considerable CER reduction is also achieved on a 30-hour low-resource ASR dataset and a 14.3k-hour industrial ASR dataset.

This work also compares the proposed method with the MBR-based method. Experimentally we find the proposed LF-MMI method outperforms the MBR-based methods on the two E2E ASR frameworks. Also, the proposed training method achieves a much faster training speed than the MBR counterparts as the on-the-fly decoding process during training is reasonably eschewed. Finally, the proposed training method requires no pre-trained model for initialization but the MBR methods do. This sets us free from the complex training pipeline so the model can be trained in one go.

For all, this work provides a new way to adopt LF-MMI, the criterion that was mainly adopted in the DNN-HMM system in previous research, into the E2E ASR systems. We believe it is feasible to use the proposed training and decoding method not only in the ASR research but also in practical systems. In future work, we may explore the way to make the proposed method totally end-to-end so the adoption of phone lexicon can be eschewed. We release all source code in the hope to make our work helpful to other researchers.
APPENDIX A

All items in series $P(O_i^t | G^+)$ can be obtained through the computation of the last term $P(O_i^t | G^+)$ by recording the intermediate variables of the forward process. The forward computation is implemented frame-by-frame in logarithmic domain. Each state $i$ in graph (numerator or denominator) will have a state score $\alpha_i(t)$ (initialized by $\alpha_i(t) = -\infty$) that will be updated in every frame step. The state scores are updated recursively until the last frame using the log-semiring:

$$\alpha_{t+1}(i) = \log \sum_{a_{j,i} \in A(i)} \exp(\alpha_t(j) + w_t(a_{j,i})) \tag{19}$$

where $A(i)$ is the set of all arcs entering state $i$; $a_{j,i}$ is an arc from state $j$ to state $i$; $w_t(a_{j,i})$ is the predicted log-posterior for the input label on arc $a_{j,i}$ in $t$-th frame. Next, if we consider $t$-th frame as the last frame (a.k.a., to compute $\log P(O_i^t | G^+)$), we should only consider the scores on the end states and add the state weights on those states. So, for any $1 \leq t \leq T$, we have:

$$\log P(O_i^t | G^+) = \log \sum_i \exp(s'_i(i)) \tag{20}$$

and

$$s'_i(i) = \begin{cases} s_i(i) + \pi_i & \text{if state } i \text{ is a final state} \\ -\infty & \text{otherwise} \end{cases} \tag{21}$$

where $\pi_i$ is the weight on the final state $i$.

REFERENCES

[1] A. Vaswani et al., “Attention is all you need,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., 2017, pp. 6000–6010.

[2] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2020, pp. 5036–5040.

[3] Y. Shi et al., “Enformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2021, pp. 6783–6787.

[4] S. Zhang, M. Lei, Z. Yan, and L. Dai, “Deep-FSMN for large vocabulary continuous speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2018, pp. 5869–5873.

[5] C.-C. Chiu* and C. Raffel*, “Monotonic chunkwise attention,” in Proc. Int. Conf. Learn. Representations, 2018. [Online]. Available: https://openreview.net/forum?id=Hk05pCtW

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2016, pp. 4960–4964.

[7] S. Watanabe, T. Hori, S. Kim, and T. Hayashi, “Hybrid CT/attention architecture for end-to-end speech recognition,” IEEE J. Sel. Topics Signal Process., vol. 11, no. 8, pp. 1240–1253, Dec. 2017.

[8] A. Graves, “Sequence transduction with recurrent neural networks,” 2012, arXiv:1211.3711.

[9] H. Sak, M. Shannon, K. Rao, and F. Beaufays, “Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping,” in Proc. Annu. Int. Conf. Speech Commun. Assoc., 2017, pp. 1298–1302.

[10] M. Malik, M. K. Malik, K. Mehmoov, and I. Makhdoom, “Automatic speech recognition: A survey,” Multimedia Tools Appl., vol. 80, no. 6, pp. 9411–9457, 2021.

[11] J. Padmanabhan and M. J. J. Premkumar, “Machine learning in automatic speech recognition: A survey,” IEEE Tech. Rev., vol. 32, no. 4, pp. 240–251, 2015.

[12] N. M. Markovnikov and I. S. Kipiatkova, “An analytic survey of end-to-end speech recognition systems,” Informal Automat., vol. 58, pp. 77–110, 2018.

[13] P. Guo et al., “Recent developments on ESPnet toolkit boosted by conformer,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2021, pp. 3874–3878.

[14] Z. Yao et al., “WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., Brno, Czech Republic, 2021.

[15] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 369–376.

[16] D. Povey and P. C. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2002, pp. I-105–I-108.

[17] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2013, pp. 2345–2349.

[18] D. Povey and B. Kingsbury, “Evaluation of proposed modifications to MPE for large scale discriminative training,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2007, pp. IV-321–IV-324.

[19] H. Xu, D. Povey, J. Zhu, and G. Wu, “Minimum hypothesis phone error as a decoding method for speech recognition,” in Proc. 10th Annu. Conf. Int. Speech Commun. Assoc., 2009.

[20] M. Gibson and T. Hain, “Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2006, pp. 2406–2409.

[21] B. Kingsbury, “Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2009, pp. 3761–3764.

[22] D. Povey et al., “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2016, pp. 2751–2755. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-595

[23] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, “End-to-end speech recognition using lattice-free MMI,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2018, pp. 12–16. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1423

[24] Y. Shao, Y. Wang, D. Povey, and S. Khudanpur, “PyChain: A fully parallelized PyTorch implementation of LF-MMI for end-to-end ASR,” 2020, arXiv:2005.09824.

[25] D. Povey, D. Kanevsky, B. Kingsbury, R. Manabhadran, G. Saon, and K. Visweswaraih, “Boosted MMI for model and feature-space discriminative training,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2008, pp. 4057–4060.

[26] H. Hadian, D. Povey, H. Sameti, J. Trmal, and S. Khudanpur, “Improving LF-MMI using unconstrained supervisions for ASR,” in Proc. IEEE Spoken Lang. Technol. Workshop, 2018, pp. 43–47.

[27] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, “Semi-supervised training of acoustic models using lattice-free MMI,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2019, pp. 4844–4848.

[28] P. Gahrehmani, V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, “Investigation of transfer learning for ASR using LF-MMI trained neural networks,” in Proc. IEEE Autom. Speech Recognit. Understanding Workshop, 2017, pp. 279–286.

[29] W. Michel, R. Schlüter, and H. Ney, “Early stage LMs integration using local and global log-linear combination,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2020.

[30] C. Weng et al., “Improving attention based sequence-to-sequence models for end-to-end English conversational speech recognition,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2018, pp. 761–765.

[31] R. Prabhavalkar et al., “Minimum word error rate training for attention-based sequence-to-sequence models,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2018, pp. 4839–4843.

[32] J. Guo, G. Tiwari, J. Droppo, M. Van Segoeloeck, C.-W. Huang, and A. Stolcke, “Effective minimum word error rate training of RNN-transducer for end-to-end speech recognition,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2020, pp. 2807–2811.

[33] C. Weng, C. Yu, I. Cui, G. Wang, and D. Yu, “Minimum bayes risk training of RNN-transducer for end-to-end speech recognition,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2020, pp. 966–970.

[34] L. Lu, Z. Meng, N. Kanda, J. Li, and Y. Gong, “On minimum word error rate training of the hybrid autoregressive transducer,” 2020, arXiv:2010.12673.

[35] Z. Meng et al., “Minimum word error rate training with language model fusion for end-to-end speech recognition,” 2021, arXiv:2106.02302.

[36] W. Wang et al., “Improving rare word recognition with LM-aware MWER training,” 2022, arXiv:2204.07553.
[37] C. Peyser, T. N. Sainath, and G. Pundak, “Improving proper noun recognition in end-to-end ASR by customization of the MWER loss criterion,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 7789–7793.

[38] Z. Lu et al., “Input length matters: An empirical study of RNN-T and MWER training for long-form telephony speech recognition,” 2021, arXiv:2110.03841.

[39] J. Tian et al., “Consistent training and decoding for end-to-end speech recognition using lattice-free MML,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2022, pp. 7782–7786.

[40] T. Hori and A. Nakamura, “Speech recognition algorithms using weighted finite-state transducers,” Synth. Lectures Speech Audio Process., vol. 9, no. 1, pp. 1–162, 2013.

[41] H. Xu, D. Povey, L. Mangu, and J. Zhu, “Minimum bayer risk decoding and system combination based on a recursion for edit distance,” Comput. Speech Lang., vol. 25, no. 4, pp. 802–828, 2011.

[42] H. Xu, D. Povey, L. Mangu, and J. Zhu, “An improved consensus-like method for minimum bayer risk decoding and lattice combination,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2010, pp. 4938–4941.

[43] B. Eikema and W. Aziz, “Is MAP decoding all you need? The inadequacy of the mode in neural machine translation,” in Proc. 28th Int. Conf. Comput. Linguistics, 2020, Barcelona, Spain, pp. 4506–4520.

[44] Y. Wang, “Model-based approaches to robust speech recognition in diverse environments,” Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 2015.

[45] G. Saon, Z. Tüske, and K. Audhkhasi, “Alignment-length synchronous decoding for RNN transducer,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 7804–7808.

[46] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline,” in Proc. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment, 2017, pp. 1–5.

[47] J. Du, X. Na, X. Liu, and H. Bu, “AISHELL-2: Transforming mandarin ASR research into industrial scale,” 2018, arXiv:1808.10583.

[48] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 448–456.

[49] Y. Wu and K. He, “Group normalization,” in Proc. Eur. Conf. Comput. Vix., 2018, pp. 3–19.

[50] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[51] D. S. Park et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[52] J. Tian, J. Yu, C. Weng, Y. Zou, and D. Yu, “Improving mandarin end-to-end speech recognition with word n-gram language model,” IEEE Signal Process. Lett., vol. 29, pp. 812–816, 2022.

[53] S. Watanabe et al., “ESPnet: End-to-end speech processing toolkit,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[54] S. Watanabe et al., “The 2020 ESPnet update: New features, broadened applications, performance improvements, and future plans,” in Proc. IEEE Data Sci. Learn. Workshop, 2021, pp. 1–6.

[55] X. Zhang et al., “On lattice-free boosted MMI training of HMM and CTC-based full-context ASR models,” in Proc. IEEE Autom. Speech Recognit. Understanding Workshop, 2021, pp. 1026–1033.

[56] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” Proc. IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[57] A. Hannun, V. Pratap, J. Kahn, and W.-N. Hsu, “Differentiable weighted finite-state transducers,” 2020, arXiv:2010.01003.

[58] M. Ravanelli et al., “SpeechBrain: A general-purpose speech toolkit,” 2021, arXiv:2106.04624.