

Minimising Biasing Word Errors for Contextual ASR With the Tree-Constrained Pointer Generator

Guangzhi Sun , *Student Member, IEEE*, Chao Zhang , *Member, IEEE*, and Philip C. Woodland , *Fellow, IEEE*

Abstract—Contextual knowledge is essential for reducing speech recognition errors on high-valued long-tail words. This paper proposes a novel tree-constrained pointer generator (TCPGen) component that enables end-to-end ASR models to bias towards a list of long-tail words obtained using external contextual information. With only a small overhead in memory use and computation cost, TCPGen can structure thousands of biasing words efficiently into a symbolic prefix-tree, and creates a neural shortcut between the tree and the final ASR output to facilitate the recognition of the biasing words. To enhance TCPGen, we further propose a novel minimum biasing word error (MBWE) loss that directly optimises biasing word errors during training, along with a biasing-word-driven language model discounting (BLMD) method during the test. All contextual ASR systems were evaluated on the public Librispeech audiobook corpus and the data from the dialogue state tracking challenges (DSTC) with the biasing lists extracted from the dialogue-system ontology. Consistent word error rate (WER) reductions were achieved with TCPGen, which were particularly significant on the biasing words with around 40% relative reductions in the recognition error rates. MBWE and BLMD further improved the effectiveness of TCPGen, and achieved more significant WER reductions on the biasing words. TCPGen also achieved zero-shot learning of words not in the audio training set with large WER reductions on the out-of-vocabulary words in the biasing list.

Index Terms—Contextual speech recognition, end-to-end, language model discounting, minimum Bayes' risk, Pointer generator.

I. INTRODUCTION

END-TO-END ASR systems often suffer from high recognition errors on long-tailed words that are rare or not included in the training set. Contextual biasing integrates external contextual knowledge into ASR systems at test-time, which plays an increasingly important role in addressing the long-tail word problem in many applications [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. Contextual knowledge

is often represented as a list (referred to as a *biasing list*) of words or phrases (referred to as *biasing words*) that are likely to appear in a given context. There exist a variety of resources from which biasing lists can be extracted or organised, such as a user's contact book or playlist, recently visited websites and the ontology of a dialogue system *etc.* Although biasing words occur infrequently and hence may only have a small impact on the overall word error rate (WER), they are mostly content words, such as nouns or proper nouns, and are thus particularly important to downstream tasks and highly valuable. A word is more likely to be recognised if it is incorporated in the biasing list, which makes contextual biasing critical to the correct recognition of those rare content words in an end-to-end ASR system.

As end-to-end ASR systems [18], [26] are often designed to incorporate all of the required knowledge into a single static model, it is particularly challenging for such systems to integrate contextual knowledge specific to a dynamic test-time context. Therefore, dedicated contextual biasing approaches have been proposed, including shallow fusion (SF) with a special weighted finite-state transducer (WFST) or a language model (LM) adapted for the contextual knowledge [1], [2], [3], [11], [12], [14], attention-based deep context approaches [4], [5], [6], [7], [8], as well as deep biasing (DB) with a prefix tree for improved efficiency when dealing with large biasing lists [6], [9], [10]. More recently, contextual biasing components with a neural shortcut that directly modifies the output distribution has been proposed [63], [64], which can be jointly optimised with ASR systems.

In this paper, a tree-constrained pointer generator (TCPGen) component is proposed and developed for end-to-end contextual speech recognition. This paper extends our work in [63]. TCPGen directly interpolates the original model distribution with an extra distribution (the TCPGen distribution) estimated from contextual knowledge, based on a dynamic interpolation weight predicted by the TCPGen component. TCPGen creates a neural shortcut between biasing lists and the final ASR output distribution to allow end-to-end training of a single neural network model. In contrast to the original work on pointer generators [15], [16], [17] which rely on the attention-mechanism to attend to all biasing words, TCPGen represents biasing lists as wordpiece-level symbolic prefix trees and only attends to the valid subset of the biasing words at each time step in decoding, and can thus handle large biasing lists with high efficiency. Furthermore, TCPGen also uses wordpieces instead of whole words for pointer generators, which allows the entire system to

Manuscript received 18 May 2022; revised 19 October 2022; accepted 28 October 2022. Date of publication 23 November 2022; date of current version 2 December 2022. The work of Guangzhi Sun was supported by International Scholarship offered by Cambridge Trust. This work has been performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service (www.hpc.cam.ac.uk) funded by EPSRC Tier-2 under Grant EP/T022159/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rohit Prabhavalkar. (*Corresponding author: Guangzhi Sun.*)

Guangzhi Sun and Philip C. Woodland are with the Department of Engineering, University of Cambridge, CB2 1TN Cambridge, U.K. (e-mail: gs534@cam.ac.uk; pcw@eng.cam.ac.uk).

Chao Zhang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100190, China (e-mail: cz277@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TASLP.2022.3224286

use wordpieces to address the out of vocabulary (OOV) issue. Therefore, as not only a few frequent words but also OOV words exist in our biasing lists, TCPGen can be viewed as a structure that achieves zero-shot learning of previously unseen words without changing model parameters. As a result, TCPGen combines the advantages of both neural and symbolic methods and improves pointer generators for ASR applications with large biasing lists.

To further improve the effectiveness of contextual biasing with TCPGen, a minimum biasing word error (MBWE) loss is proposed to directly optimise the model performance on the biasing words. By changing the risk function of the widely used minimum word error (MWE) loss [40], [41], [42], [43], [44], the proposed MBWE loss has a greater focus on minimising the expected errors of the rare and OOV words in the biasing list associated with that utterance. An efficient beam search algorithm is proposed for MBWE training in RNN-T by limiting the number of output wordpieces at each encoder step to one [65], [66]. Moreover, to address the issue that end-to-end models often suffer from the internal LM effect that biases towards common words, a biasing-driven LM discounting (BLMD) method is proposed. The BLMD method extends the density ratio-based LM discounting method [49] to incorporate an additional discounting factor for the TCPGen distribution before interpolation.

In this paper, TCPGen, a generic component for end-to-end ASR systems, is integrated into both attention-based encoder-decoder (AED) [18], [19], [20], [21], [22], [38] and recurrent neural network transducer (RNN-T) [26], [27], [28], [29], [30]. MBWE and BLMD methods are also applied to both types of end-to-end models in conjunction with TCPGen. Experiments were performed on two different types of data, including the Librispeech audiobook data and goal-oriented dialogue data. Improvements in word error rate (WER) were achieved by using TCPGen in both AED and RNN-T configurations across all test sets compared to the baseline and the DB method, with a significant WER reduction on the biasing words.

The remainder of this paper is organised as follows. Section II reviews related work. Section III introduces the TCPGen component. Sections IV and V describe MBWE training and BLMD for TCPGen respectively. Sections VI and VII present the experimental setup and results. Section VIII gives the conclusions.

II. RELATED WORK

A. End-to-end Contextual Speech Recognition

Various contextual biasing algorithms have been recently developed for end-to-end ASR. One of the major streams of research in this area focuses on representing biasing lists as external WFST which is incorporated into a class-based LM via SF [1], [2], [3], [11]. Such methods usually rely on special context prefixes like “call” or “play,” which limit their flexibility in handling more diverse grammars in natural speech. Neural network-based deep context approaches using attention mechanisms have also been proposed. These approaches encode a biasing list into a vector to use as a part of the input to the end-to-end ASR models [4], [5], [6], [7], [8]. Although the deep context approaches address the dependency issue on syntactic

prefixes in the SF methods, they are more memory intensive and less effective for handling large biasing lists.

Work in [9] jointly adopted the use of deep context and the SF of a WFST together in an RNN-T. It also improved the efficiency by extracting the biasing vector from only a subset of wordpieces constrained by a prefix tree representing the biasing list, which is referred to as deep biasing (DB) in this paper. Work in [10] extended the prefix-tree-based method to RNN LMs which are used for SF to achieve further improvements in biasing words. Moreover, while previous studies only focused on industry datasets, researchers in [10] proposed and justified a simulation of the contextual biasing task on open-source data by adding a large number of distractors to the list of biasing words in an utterance. More recently, [63], [64] have simultaneously proposed creating a neural shortcut between the biasing list and the final model output distribution.

B. MWE Training for end-to-end ASR Models

The MWE loss that directly minimises the expected WER [40], has become increasingly popular in training end-to-end ASR models [41], [42], [43], [44], [45], [46], [47]. MWE training using the N-best hypotheses to approximate the expected word errors has been proposed in [41] for the AED model, which was then improved in [44]. More recently, work in [46] applied LM fusion and internal LM estimation during MWE training of an AED model to improve the N-best approximation. Work in [45] exploited a lattice structure in place of the N-best list to calculate the expected word errors. For RNN-T models, [42] applied the same N-best approximation as in AED to calculate the expected errors. Work in [43] further improved MWE training by relating the gradient calculation for each alignment of a hypothesis in the N-best list to the original RNN-T loss function, which enabled offline decoding of N-best lists when abundant CPU resource was available. Work in [67] proposed the customisation of MWE loss for proper nouns. Moreover, work in [48] first discussed the necessity of LM discounting in sequence discriminative training for HMM-based large-vocabulary continuous speech recognition, and [47] applied MWE to the hybrid auto-regressive transducer (HAT) [53].

C. LM Discounting

Various LM discounting algorithms have recently been proposed to minimise the internal LM effect of end-to-end ASR models, in particular when applying the model to a test set in a different domain from the training data. In [49], a density ratio method was introduced to estimate the score from a source-domain external LM that is to be subtracted from the target-domain LM score. Later, the HAT was proposed to preserve the modularity of a hybrid system and allowed internal LM scores to be estimated and discounted [53]. An internal LM estimation method was proposed in [54] to estimate the source-domain LM score directly from the end-to-end ASR model. This method was further improved by performing internal LM training [55] in order to better estimate the internal LM score. Moreover, [56] proposed regularising the internal LM in RNN-T training to avoid over-fitting to text priors.

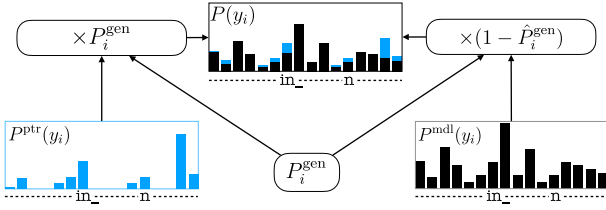


Fig. 1. Illustration of interpolation in TCPGen with corresponding terms in (3). $P^{\text{ptr}}(y_i)$ is the TCPGen distribution. $P^{\text{mdl}}(y_i)$ is the distribution from a standard end-to-end model. $P(y_i)$ is the final output distribution. $\hat{P}_i^{\text{gen}}$ and P_i^{gen} are the scaled and unscaled generation probabilities.

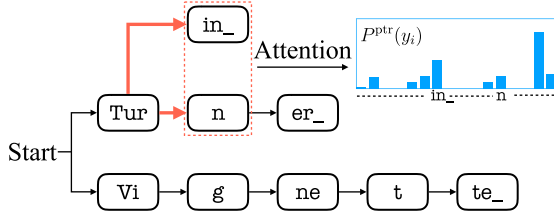


Fig. 2. An example of prefix tree search and attention in TCPGen. With previous output *Tur*, *in_* and *n* are two valid wordpieces on which attention will be performed. A word end unit is denoted by *_*.

III. TREE-CONSTRAINED POINTER GENERATOR

TCPGen is a neural network-based component combining symbolic prefix-tree search with a neural pointer generator for contextual biasing, which also enables end-to-end optimisation. TCPGen represents the biasing list as a wordpiece-level prefix tree. At each output step, TCPGen calculates a distribution over all valid wordpieces constrained by the prefix tree. TCPGen also predicts a generation probability which indicates how much contextual biasing is needed at a specific output step. The final output distribution is the weighted sum of the TCPGen distribution and the original AED or RNN-T output distribution (Fig. 1).

The key symbolic representation of the external contextual knowledge in TCPGen is the prefix tree. For simplicity, examples and equations in this section are presented for a specific search path, which can be generalised easily to beam-search with multiple paths. In the example prefix tree with biasing words (turner, vignette and turin) shown in Fig. 2, if the previously decoded wordpiece is *Tur*, wordpieces *in_* and *n* form the set of valid wordpieces $\mathcal{Y}_i^{\text{tree}}$. Denoting $\mathbf{x}_{1:T}$ and y_i as input acoustic features and output wordpiece, $\mathbf{q}_i$ as the query vector carrying the decoding history and acoustic information, $\mathbf{K} = [\dots, \mathbf{k}_j, \dots]$ as the key vectors, a scaled dot-product attention is performed between $\mathbf{q}_i$ and $\mathbf{K}$ to compute the TCPGen distribution P^{ptr} , and Section III-A and Section III-B introduces how keys and queries are computed for AED and RNN-T. An output vector $\mathbf{h}_i^{\text{ptr}}$ as shown in Eqns. (1) and (2).

$$P^{\text{ptr}}(y_i | y_{1:i-1}, \mathbf{x}_{1:T}) = \text{Softmax}(\text{Mask}(\mathbf{q}_i \mathbf{K}^T / \sqrt{d})), \quad (1)$$

$$\mathbf{h}_i^{\text{ptr}} = \sum_j P^{\text{ptr}}(y_i = j | y_{1:i-1}, \mathbf{x}_{1:T}) \mathbf{v}_j^T, \quad (2)$$

where d is the size of $\mathbf{q}_i$ (see [62]), $\text{Mask}(\cdot)$ sets the probabilities of wordpieces that are not in $\mathcal{Y}_i^{\text{tree}}$ to zero, and $\mathbf{v}_j$ is the value

vector relevant to j . For more flexibility, an *out-of-list* (OOL) token is included in $\mathcal{Y}_i^{\text{tree}}$ indicating that no suitable wordpiece can be found in the set of valid wordpieces. To ensure that the final distribution sums up to 1, the generation probability, P_i^{gen} , is scaled as $\hat{P}_i^{\text{gen}} = P_i^{\text{gen}}(1 - P^{\text{ptr}}(\text{OOL}))$, and the final output can be calculated as shown in (3).

$$P(y_i) = P^{\text{mdl}}(y_i)(1 - \hat{P}_i^{\text{gen}}) + P^{\text{ptr}}(y_i)\hat{P}_i^{\text{gen}}, \quad (3)$$

where conditions, $y_{1:i-1}$, $\mathbf{x}_{1:T}$, are omitted for clarity. $P^{\text{mdl}}(y_i)$ represents the output distribution from the standard end-to-end model.

A. TCPGen in AED

A standard AED contains three components: an *encoder*, a *decoder* and an *attention network*. The encoder encodes the input, $\mathbf{x}_{1:T}$, into a sequence of high-level features, $\mathbf{h}_{1:T}^{\text{enc}}$. At each decoding step i , an attention mechanism is first used to combine the encoder output sequence into a single context vector, $\mathbf{c}_i$. The decoder computation is thus as follows:

$$\mathbf{h}_i^{\text{dec}} = \text{Decoder}(\mathbf{y}_{i-1}, \mathbf{h}_{i-1}^{\text{dec}}, \mathbf{c}_i), \quad (4)$$

where $\text{Decoder}(\cdot)$ denotes the decoder network and $\mathbf{y}_{i-1}$ is the embedding of the preceding wordpiece. The model output can be calculated using a Softmax layer taking $\mathbf{h}_i^{\text{dec}}$ as input.

To calculate the TCPGen distribution in AED, the query combines the context vector $\mathbf{c}_i$ and the previously decoded token embedding $\mathbf{y}_{i-1}$ as shown in (5).

$$\mathbf{q}_i = \mathbf{W}_c^Q \mathbf{c}_i + \mathbf{W}_y^Q \mathbf{y}_{i-1}, \quad (5)$$

where $\mathbf{W}_c^Q$ and $\mathbf{W}_y^Q$ are parameter matrices. The keys and values are computed from the decoder wordpiece embedding matrix as shown in (6).

$$\mathbf{k}_j = \mathbf{W}^K \mathbf{y}_j \quad \mathbf{v}_j = \mathbf{W}^V \mathbf{y}_j, \quad (6)$$

where $\mathbf{y}_j$ denotes the j -th row of the embedding matrix. $\mathbf{W}^K$ and $\mathbf{W}^V$ are key and value parameter matrices which are shared throughout this paper. The TCPGen distribution and the TCPGen output can be computed using Eqns. (1) and (2) respectively. The generation probability are calculated from the decoder hidden state $\mathbf{h}_i^{\text{dec}}$ and the TCPGen output $\mathbf{h}_i^{\text{ptr}}$, as shown in (7).

$$P_i^{\text{gen}} = \sigma \left(\mathbf{W}^{\text{gen}} \left[\mathbf{h}_i^{\text{dec}}, \mathbf{h}_i^{\text{ptr}} \right] \right), \quad (7)$$

where $\mathbf{W}^{\text{gen}}$ is a parameter matrix. The distribution of y_i can be calculated using (3). In AED, deep biasing can be applied as shown in (8).

$$P^{\text{mdl}}(y_i) = \text{Softmax} \left(\mathbf{W}^O \left[\mathbf{h}_i^{\text{dec}}, \mathbf{c}_i \right] + \mathbf{W}^{\text{db}} \mathbf{h}_i^{\text{db}} \right), \quad (8)$$

where the biasing vector $\mathbf{h}_i^{\text{db}}$ is obtained from the sum of embeddings of all wordpieces in $\mathcal{Y}_i^{\text{tree}}$, similar to [9].

B. TCPGen in RNN-T

A standard RNN-T consists of an encoder, a predictor and a joint network. The encoder in RNN-T is similar to that in AED which outputs $\mathbf{h}_{1:T}^{\text{enc}}$. The predictor encodes all wordpieces in the history into a vector, $\mathbf{h}_{1:T}^{\text{pred}}$, analogous to an LM. Given K encoder outputs and N predictor outputs, the joint network

determines the output distribution $P(z_{i,t}|y_{1:i}, \mathbf{h}_t^{\text{enc}})$ for each combination of i and t as shown in Eqns. (9) and (10).

$$\mathbf{h}_{i,t}^{\text{joint}} = \tanh(\mathbf{W}^{\text{joint}}[\mathbf{h}_i^{\text{pred}}; \mathbf{h}_t^{\text{enc}}]), \quad (9)$$

$$P(z_{i,t}|y_{1:i-1}, \mathbf{h}_t^{\text{enc}}) = \text{Softmax}(\mathbf{W}_2^{\text{joint}} \mathbf{h}_{i,t}^{\text{joint}}), \quad (10)$$

where the $\mathbf{W}^{\text{joint}}$ matrices are the parameter matrices of the joint network, and $z_{i,t} \in \mathcal{Y} \cup \{\emptyset\}$ where $\mathcal{Y}$ represents the set of all wordpieces.

In RNN-T, the query for the TCPGen distribution is computed for each combination of i, t as shown in (11).

$$\mathbf{q}_{i,t} = \mathbf{W}_c^Q \mathbf{h}_t^{\text{enc}} + \mathbf{W}_y^Q \mathbf{y}_{i-1}, \quad (11)$$

where $\mathbf{y}_{i-1}$ is the wordpiece embedding from the predictor. Key and value vectors are derived from the predictor embedding matrix. The generation probability for each i, t -pair is computed using the penultimate layer output of the joint network and the TCPGen output vector.

$$P_{i,t}^{\text{gen}} = \sigma(\mathbf{W}^{\text{gen}}[\mathbf{h}_{i,t}^{\text{joint}}; \mathbf{h}_{i,t}^{\text{ptr}}]) \quad (12)$$

As $\emptyset$ only exists in P^{mdl} , its value is directly copied to the final output distribution as shown in (13)

$$\hat{P}(z_{i,t}) = \begin{cases} P^{\text{mdl}}(\emptyset), & \text{if } z_{i,t} = \emptyset \\ P(z_{i,t}), & \text{otherwise} \end{cases}, \quad (13)$$

where $P(z_{i,t})$ is the interpolated probability for the output token $z_{i,t}$ in (3), except that $P^{\text{ptr}}(z_{i,t})$ is scaled by a factor of $1 - P^{\text{mdl}}(z_{i,t} \neq \emptyset)$ to ensure all probabilities sum up to 1. Moreover, whenever TCPGen is used in RNN-T, the biasing vector, $\mathbf{h}_{i,t}^{\text{ptr}}$ is always sent to the input of the joint network as shown in (14) which yielded the best results as discussed in [63].

$$\mathbf{h}_{i,t}^{\text{joint}} = \tanh(\mathbf{W}^{\text{joint}}[\mathbf{h}_i^{\text{pred}}; \mathbf{h}_t^{\text{enc}}; \mathbf{h}_{i,t}^{\text{ptr}}]), \quad (14)$$

As for AED, DB can be applied as described in [6], [9].

IV. MBWE TRAINING FOR TCPGEN

This section introduces minimum biasing word error (BMWE) training. The MWE loss in end-to-end ASR systems minimises the expected value of word errors across all possible output sequences of a given input. Denoting the output sequence $Y = \{y_1, \dots, y_L\}$ and input sequence $X = \mathbf{x}_{1:T}$ for convenience, the MWE loss can be written as (15).

$$\mathcal{L}^{\text{mwe}} = \sum_Y P(Y|X) \mathcal{W}^{\text{mwe}}(Y, Y^*), \quad (15)$$

where Y^* is the ground-truth output sequence and $\mathcal{W}^{\text{mwe}}(\cdot)$ is the risk function representing the number of word errors calculated using an edit-distance between each possible sequence Y and the ground-truth sequence Y^* . The sum is performed over all possible sequences and $P(Y|X)$ is the probability of a specific sequence calculated from the end-to-end ASR model output. As it is intractable to enumerate over all possible sequences and calculate their probabilities, a common practice which has been widely adopted in MWE training for end-to-end ASR systems [40], [41], [42], [43], [44], [45] is to use the N-best hypotheses to approximate the expected word errors, as shown

in (16).

$$\begin{aligned} \mathcal{L}^{\text{mwe}} &\approx \frac{\sum_{Y_i \in \text{Beam}_N(X)} P(Y_i|X) \mathcal{W}^{\text{mwe}}(Y_i, Y^*)}{\sum_{Y_i \in \text{Beam}_N(X)} P(Y_i|X)} \\ &= \sum_{Y_i \in \text{Beam}_N(X)} \hat{P}(Y_i|X) \mathcal{W}^{\text{mwe}}(Y_i, Y^*), \end{aligned} \quad (16)$$

where $\text{Beam}_N(Y) = \{Y_1, \dots, Y_N\}$ is the N-best hypotheses obtained via beam search, and $\mathcal{W}^{\text{mwe}}(Y_i, Y^*)$ represents the edit-distance function. The probabilities of the N-best hypotheses are normalised to form a valid distribution, where each normalised probability is represented as $\hat{P}(Y_i|X)$. Moreover, the average number of word errors over the N-best hypotheses is often subtracted from the number of word errors in each hypothesis as a form of variance reduction.

To apply the MWE loss to contextual ASR with TCPGen which focuses on the correct recognition of biasing words, we propose a new risk function that includes an additional biasing word error term to the word error term as shown in (17).

$$\mathcal{W}^{\text{mbwe}}(Y, Y^*) = \mu_1 \mathcal{W}^{\text{mwe}}(Y_i, Y^*) + \mu_2 \mathcal{W}^{\text{bias}}(Y_i, Y^*), \quad (17)$$

where $\mathcal{W}^{\text{bias}}$ is the additional biasing word error term which is the edit-distance between the sequence of biasing words in Y and the sequence of biasing words in Y^* . Scaling factors μ_1 and μ_2 control the importance of the word error term and the new biasing word error term. As a result, if $\mu_1 = 1$ and $\mu_2 = 0$, $\mathcal{W}^{\text{mbwe}}$ becomes $\mathcal{W}^{\text{mwe}}$. If $\mu_1 = 1$ and $\mu_2 = 1$, it is equivalent to giving a weight of 2 to any rare word errors in the original word error rate. A new MBWE loss function is proposed to use $\mathcal{W}^{\text{mbwe}}$ instead of $\mathcal{W}^{\text{mwe}}$. That is,

$$\begin{aligned} \mathcal{L}^{\text{mbwe}} &= \sum_Y P(Y|X) \mathcal{W}^{\text{mbwe}}(Y, Y^*) \\ &\approx \sum_{Y_i \in \text{Beam}_N(X)} \hat{P}(Y_i|X) \mathcal{W}^{\text{mbwe}}(Y_i, Y^*). \end{aligned} \quad (18)$$

As a generic loss for end-to-end ASR systems, MBWE can be applied to the standard AED and RNN-T models, as well as other deep context models.

A. MBWE Training in AED

The MBWE loss can be applied to the AED model following a similar MWE training scheme proposed in [41] which also interpolated the MBWE loss with cross-entropy (CE) loss for better training stability, as shown in (19).

$$\mathcal{L} = \mathcal{L}^{\text{mbwe}} + \mathcal{L}^{\text{ce}}, \quad (19)$$

where $\mathcal{L}$ is the total loss function to be minimised, $\mathcal{L}^{\text{mbwe}}$ is the proposed MBWE loss in (18) and $\mathcal{L}^{\text{ce}}$ is the CE loss. As it is hard to train a randomly initialised model with the MBWE loss, which is similar to MWE, the MBWE loss is applied from the epoch when the model is reasonably trained with CE loss, which depends on optimisation algorithms. Moreover, to boost the efficiency of beam search which is the bottleneck in the time taken for training, batched beam search is implemented by parallelising the model forward computation of all beams of all utterances in the same mini-batch on a GPU.

TABLE I
STATISTICS OF THE NUMBER OF OUTPUT WORDPIECE TOKENS AT EACH
ENCODER STEP FOR RNN-T 1-BEST HYPOTHESIS ON LIBRISPEECH DEV SET.
PERCENTAGE IS OF THE TOTAL NUMBER OF ENCODER STEPS

# Wordpiece tokens	0	1	2+
Count	80%	18%	2%

B. MBWE Training in RNN-T

The MBWE loss can also be applied to the RNN-T model following a similar MWE training scheme proposed in [42], except that the original RNN-T loss is also included for stable training, as shown in (20).

$$\mathcal{L} = \mathcal{L}^{\text{mbwe}} + \mathcal{L}^{\text{rnn-t}}, \quad (20)$$

where $\mathcal{L}^{\text{rnn-t}}$ represents the original RNN-T loss [26]. Although previous work [42], [43] tried to obtain N-best lists using standard beam search for RNN-T, it is infeasible to perform such a training scheme on a single GPU with a limited number of CPUs, even with batched decoding. The major obstacle that restricts the level of parallel computation is the unknown number of wordpiece tokens to output at a given encoder step, as beams requiring two or more output tokens have to be handled separately. However, as the encoder output sequence is usually longer than the number of output wordpiece tokens, cases where two or more tokens are output at a specific encoder step, should be rare. To verify this conjecture, taking a standard RNN-T model trained on the Librispeech training set and decoded on its dev set as an example, the path taken and the number of output tokens at each encoder step for each 1-best hypothesis were recorded, as shown in Table I.

As shown in the table, the vast majority of encoder steps output 0 or 1 wordpiece token, where 0 means a $\emptyset$ token is output. Although in the standard beam search, it is always required to compare the score with a second output token, it is often unnecessary for generating reasonably good N-best hypotheses, especially for MBWE training where a strong approximation using the N-best list has already been made. Therefore, a restricted beam search which only allows 0 or 1 output token at each encoder step is proposed for efficient MBWE training in RNN-T, which is similar to the one-step constrained beam search algorithm in [65] but with all neural network computations parallelised across all samples in the mini-batch. With restricted beam search, the model forward computation can be efficiently parallelised for all beams of all utterances in the same mini-batch on a single GPU.

V. BLMD FOR TCPGEN

This section introduces the biasing language model discounting (BLMD) method for TCPGen. To incorporate an external LM, SF is often performed for both AED and RNN-T models via log-linear interpolation. Define the source domain data as the text of the training data for the end-to-end model, and the target domain data as the data used to train an external LM such that it generates better probability estimates for the test data. Then,

the recognised sequence for LM SF can be written as (21).

$$Y^* = \arg \max_Y \log P^{\text{mdl}}(Y|X) + \alpha \log P^{\text{tgt}}(Y), \quad (21)$$

where $P^{\text{mdl}}(Y)$ is the output of the end-to-end system and $P^{\text{tgt}}(Y)$ is the probability of the output sequence predicted by an LM trained on the target domain. The interpolation factor α is a hyper-parameter to be determined. After decomposing the probability of each possible sequence Y into a token-level sequence, $P^{\text{sf}}(y_i)$, the probability of each output token y_i after SF, can be written as

$$P^{\text{sf}}(y_i) = P^{\text{mdl}}(y_i)P^{\text{tgt}}(y_i)^\alpha, \quad (22)$$

where the conditions on acoustic and history information were omitted for clarity. The density ratio method provides a Bayes' rule-grounded way to reduce the effect of the internal LM in the end-to-end system especially when there is a difference between the source and target domain data. That is,

$$P^{\text{sf}}(y_i) = P^{\text{mdl}}(y_i) \frac{P^{\text{tgt}}(y_i)^\alpha}{P^{\text{src}}(y_i)^\beta}, \quad (23)$$

where $P^{\text{src}}(Y)$ refers to the probability of the output sequence predicted by an LM trained on the source domain. The factors α and β are hyper-parameters. In this way, the probabilities of commonly seen text patterns in the source domain are penalised, whereas those of text patterns specific to the target domain are boosted. Therefore, the density ratio LM discounting method can also be applied to the TCPGen component to further improve its performance on the biasing words that are rare in the source domain. As the final distribution comes from both the model and the TCPGen distribution which use different parameters and history information, density ratio SF is separately performed as

$$P^{\text{sf}}(y_i) = (1 - P^{\text{gen}})P^{\text{mdl}}(y_i) \frac{P^{\text{tgt}}(y_i)^{\alpha_1}}{P^{\text{src}}(y_i)^{\beta_1}} + P^{\text{gen}}P^{\text{ptr}}(y_i) \frac{P^{\text{tgt}}(y_i)^{\alpha_2}}{P^{\text{src}}(y_i)^{\beta_2}}, \quad (24)$$

where P^{ptr} is the TCPGen distribution, and the same source and target LMs are used for both distributions, but with different sets of hyper-parameters α_1, β_1 and α_2, β_2 . To avoid a complicated hyper-parameter search, the best set of α, β obtained from the standard end-to-end system can be directly applied to α_1, β_1 , and only α_2, β_2 tuned to find the best values for the TCPGen distribution.

VI. EXPERIMENTAL SETUP

A. Data

Experiments were performed on two different data sets, including the Librispeech audiobook corpus and the dialogue state-tracking challenge (DSTC) data. The Librispeech corpus [60] contains 960 hours of read English from audiobooks. The dev-clean and dev-other sets were held out for validation, and the test-clean and test-other sets were used for evaluation. Models trained on the Librispeech data were finetuned and tested on the DSTC data.

The DSTC data was included as a realistic application with a limited amount of audio training resources of TCPGen where the

ontology was used to extract contextual knowledge. The DSTC data contains human-machine task-oriented dialogues where user-side input audio was used for recognition. The DSTC track2 train and dev sets were used as the training and validation sets, and the DSTC track3 test set was used for evaluation. The 80-dim FBANK features at a 10 ms frame rate concatenated with 3-dim pitch features were used as the model input. SpecAugment [57] with the setting $(W, F, m_F, T, p, m_T) = (40, 27, 2, 40, 1.0, 2)$ was used without any other data augmentation or speaker adaptation.

B. Biasing List Selection

For Librispeech simulation of real tasks, the full rare word list containing 200 k distinct words proposed in [10] was used as the collection of all biasing words, in which more than 60% were OOV words that did not appear in the Librispeech training set. Following the scheme in [10], biasing lists were organised by finding words that belong to the full rare word list from the reference transcription of each utterance and adding a certain number of distractors to it. There were 10.3% word tokens in the test sets that belong to the full rare word list and hence were covered by the biasing list during testing.

On the DSTC data, the biasing list arrangement by adding distractors was used for training only, where the full rare word list was augmented with words that occurred less than 100 times in the DSTC training data. During the evaluation, the ontology of DSTC3 which contains lists of named entities corresponding to different types (e.g. restaurant name, restaurant food etc.) was used to form the biasing list by extracting distinct words from those named entities and removing common words with 200 or more occurrences in the training data. This biasing list contained 243 distinct words, and 4.7% word tokens in the test set belong to this biasing list.

C. Model Specification

Systems were built using the ESPnet toolkit [58]. A unigram wordpiece model with 600 distinct wordpieces was built on the Librispeech data and was directly applied to the DSTC data. For both the AED and RNN-T models, a Conformer [39] encoder was used. The Conformer encoder contained 16 conformer blocks consisting of 4 512-d attention heads. The AED additionally contained a single-layer 1024-d LSTM decoder and a 4-head 1024-d location-sensitive attention. The RNN-T additionally had a 1024-d predictor with a 1024-d single fully-connected layer joint network.

For BLMD, a 2-layer 2048-d LSTM-LM trained on the Librispeech 800 million-word text training corpus was used as the target domain LM for Librispeech experiments. Each source domain LM trained on the text of the audio training data used a single-layer 1024-d LSTM. Each LM had the same wordpieces as the corresponding ASR system.

D. Training Specifications

During training, biasing lists with 1000 distractors were used for the Librispeech experiments and 100 distractors for the DSTC data. These lists were organised by finding biasing

words from the reference and adding distractors. The dropping technique described in [10] was applied for AED model training, where biasing words that were found in the reference transcription had a 30% probability to be removed from the biasing list. This was to prevent the model from being over-confident about TCPGen outputs. The Noam [62] optimiser was used for the Conformer. The MBWE loss was applied after 16 epochs. The beam size for MBWE training was 5 for all experiments and 30 for decoding. A coverage penalty [61] of 0.01 was applied to AED models.

E. Evaluation Metrics

In addition to WER, a rare word error rate (R-WER) was used to evaluate the system performance on biasing words that were “rare” in the training data for that system. R-WER is the total number of *error* word tokens that belong to the biasing list divided by the total number of word tokens in the test set that belong to the biasing list. Insertion errors were counted in R-WER if the inserted word belonged to the biasing list [10]. As WERs on Librispeech test sets were all small, for the rest of this paper, 2 decimal places were included for WER whereas one decimal place was used for other results. Moreover, for small WER and R-WER reductions, a project-by-project sign test was performed for Librispeech experiments based on the “project ID” of each utterance. The same sign test was performed dialogue-by-dialogue for the DSTC data.

VII. RESULTS

A. Conformer AED Experiments on Librispeech

The TCPGen component together with proposed MBWE and BLMD algorithms was first applied to the conformer AED model. With the Noam optimiser, the learning rate was a smooth function and preliminary experiments found that adjusting the weight of the cross-entropy loss yielded significantly worse results as it effectively introduced an abrupt change to the learning rate. Therefore, to adjust the contribution of the MBWE loss, different values of μ_1 and μ_2 in (17) were used while keeping the coefficient of the cross-entropy loss to 1. The effect of using small and large values of μ_1 and μ_2 on the training and dev set WER are shown in Fig. 3. As shown in Fig. 3, applying MBWE with both small and large values had a similar effect on the training set WER, whereas using smaller values produced better results on the dev set and hence was adopted for the experiments.

Then, the best set of BLMD parameters was searched for and applied to the trained models during decoding. The search procedure is illustrated in Fig. 4. The left part of Fig. 4 shows the dev set WER of different sets of BLMD parameters α_1, β_1 for the baseline standard AED system. The best values found here were $\alpha_1 = 0.5, \beta_1 = 0.3$, which were then fixed for the search of α_2, β_2 , as shown on the right part of Fig. 4 for the system with TCPGen. As a result, $\alpha_2 = 0.3, \beta_2 = 0.3$ were used which indicates that a stronger LM discounting effect was needed for the TCPGen distribution.

The results for the Conformer AED model are summarised in Table II. As shown in Table II, for the baseline standard Conformer AED system, using the MBWE loss benefits the

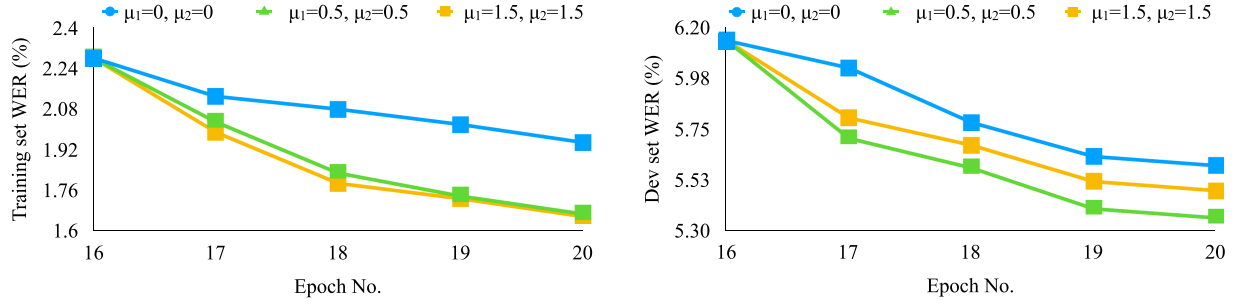


Fig. 3. Plots of training (left) and dev (right) set WERs across 4 training epochs. Training set WER was calculated on 5% randomly sampled utterances from the full 960-hour training set. Dev-set combines both dev-clean and dev-other sets. MBWE parameters μ_1, μ_2 were defined in (17).

TABLE II
WER AND R-WER ON LIBRISPEECH TEST-CLEAN AND TEST-OTHER SETS FOR CONFORMER-BASED SYSTEMS TRAINED ON LIBRISPEECH FULL 960-HOUR TRAINING SET. MBWE PARAMS. INCLUDE μ_1 AND μ_2 IN (17), AND BLMD PARAMS INCLUDE α_1, β_1 AND α_2, β_2 IN (24). THE BASELINE HERE REFERS TO THE STANDARD CONFORMER AED MODEL

System	MBWE params.	BLMD params.	test-clean		test-other	
			%WER	%R-WER	%WER	%R-WER
Baseline	$\mu_1 = 0.0, \mu_2 = 0.0$	-	3.71	13.2	9.36	29.5
Baseline	$\mu_1 = 0.5, \mu_2 = 0.0$	-	3.65	13.0	9.02	28.9
Baseline	$\mu_1 = 0.5, \mu_2 = 0.5$	-	3.62	12.8	9.08	28.6
TCPGen	$\mu_1 = 0.0, \mu_2 = 0.0$	-	3.23	8.6	8.43	21.3
TCPGen	$\mu_1 = 0.5, \mu_2 = 0.5$	-	2.96	7.6	7.88	19.5
Baseline	$\mu_1 = 0.0, \mu_2 = 0.0$	$\alpha_1 = 0.5, \beta_1 = 0.3$	3.33	12.3	8.04	27.6
Baseline	$\mu_1 = 0.5, \mu_2 = 0.0$	$\alpha_1 = 0.5, \beta_1 = 0.3$	3.19	12.2	7.95	27.1
Baseline	$\mu_1 = 0.5, \mu_2 = 0.5$	$\alpha_1 = 0.5, \beta_1 = 0.3$	3.17	11.7	7.92	27.3
TCPGen	$\mu_1 = 0.0, \mu_2 = 0.0$	$\alpha_1 = 0.5, \beta_1 = 0.3, \alpha_2 = 0.3, \beta_2 = 0.3$	2.79	6.9	7.40	19.5
TCPGen	$\mu_1 = 0.5, \mu_2 = 0.5$	$\alpha_1 = 0.5, \beta_1 = 0.3, \alpha_2 = 0.3, \beta_2 = 0.3$	2.59	6.4	7.13	18.2

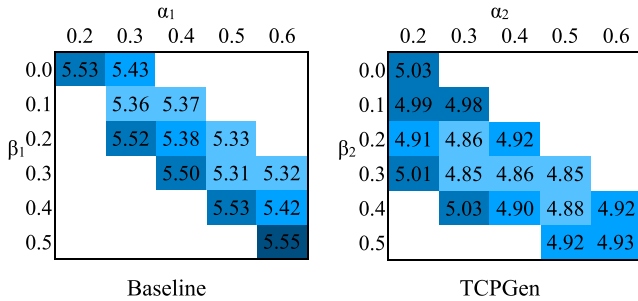


Fig. 4. Illustration of tuning BLMD hyper-parameters for the baseline standard Conformer AED model and the Conformer AED model with TCPGen. Numbers in each grid are dev set WER in percentage. Left: Tuning α_1, β_1 on the baseline model. Right: Tuning α_2, β_2 on the TCPGen model with the best set of α_1, β_1 found from the baseline on the left.

R-WER. When MBWE is applied to TCPGen, there was a 12% relative reduction in R-WER on the test-clean set and 9% on the test-other set compared to the TCPGen system without MBWE training. This increased the relative R-WER improvement by using TCPGen from 33% to 41% on the test-clean set, and from 28% to 32% on the test-other set compared to the baseline with the same training loss (i.e. comparing row 1 with row 4 and row 3 with row 5 in Table II).

After applying BLMD, obvious reductions in R-WER were observed for both the baseline and TCPGen systems on both test sets. In particular, large R-WER reductions were found

when different discounting factors were applied to the TCPGen distribution, which further increased the relative R-WER improvement by using TCPGen from 41% to 46% on test-clean and from 32% to 33% on test-other (comparing row 5 and row 8 with row 10 in Table II).

B. Conformer RNN-T Experiments on Librispeech

Experiments were then performed on Librispeech full 960-hour training data as shown in Table III. Preliminary experiments on Librispeech found that μ_2 for the MBWE loss should be set larger than μ_1 for better performance when using TCPGen in RNN-T. The baseline here is a standard Conformer-based RNN-T model. The MBWE and BLMD hyper-parameters were found in the same way as for the AED experiments. In addition to the standard baseline system, the DB method proposed in [10] was used as a biasing method for comparison. In general, consistent and significant WER and R-WER reductions were achieved using TCPGen compared to both the baseline and the DB system, with a p-value less than .001. MBWE with a restricted beam search achieved WER and R-WER reductions for both baseline and TCPGen systems. In particular, the relative R-WER improvement increased from 37% to 41% on the test-clean set and from 33% to 37% on the test-other set compared to the baseline system (comparing row 4 to row 1 and row 6 to row 3 in Table III).

Applying BLMD achieved further WER and R-WER reductions for all systems. In particular, BLMD increased the relative

TABLE III

WER AND R-WER ON LIBRISPEECH TEST-CLEAN AND TEST-OTHER SETS FOR CONFORMER-BASED RNN-T MODELS TRAINED ON LIBRISPEECH FULL 960-HOUR TRAINING SET. MBWE PARAMS. INCLUDE μ_1 AND μ_2 IN (17), AND BLMD PARAMS INCLUDE α_1, β_1 AND α_2, β_2 IN (24). THE BASELINE HERE REFERS TO THE STANDARD CONFORMER RNN-T MODEL

System	MBWE params.	BLMD params.	test-clean		test-other	
			%WER	%R-WER	%WER	%R-WER
Baseline	$\mu_1 = 0.0, \mu_2 = 0.0$	-	4.02	14.1	10.12	33.1
Baseline	$\mu_1 = 0.5, \mu_2 = 0.0$	-	4.01	14.0	9.96	32.5
Baseline	$\mu_1 = 0.5, \mu_2 = 0.5$	-	3.87	13.8	9.80	31.8
DB	$\mu_1 = 0.0, \mu_2 = 0.0$	-	3.57	10.4	9.45	25.0
TCPGen	$\mu_1 = 0.0, \mu_2 = 0.0$	-	3.40	8.9	8.79	22.2
TCPGen	$\mu_1 = 0.5, \mu_2 = 1.0$	-	3.12	8.1	8.64	20.7
Baseline	$\mu_1 = 0.0, \mu_2 = 0.0$	$\alpha_1 = 0.4, \beta_1 = 0.1$	3.55	12.5	8.90	30.4
Baseline	$\mu_1 = 0.5, \mu_2 = 0.0$	$\alpha_1 = 0.4, \beta_1 = 0.1$	3.38	12.6	8.59	29.5
Baseline	$\mu_1 = 0.5, \mu_2 = 0.5$	$\alpha_1 = 0.4, \beta_1 = 0.1$	3.28	12.2	8.50	28.8
TCPGen	$\mu_1 = 0.0, \mu_2 = 0.0$	$\alpha_1 = 0.4, \beta_1 = 0.1, \alpha_2 = 0.2, \beta_2 = 0.1$	3.02	8.0	7.49	18.6
TCPGen	$\mu_1 = 0.5, \mu_2 = 1.0$	$\alpha_1 = 0.4, \beta_1 = 0.1, \alpha_2 = 0.1, \beta_2 = 0.0$	2.79	7.0	7.44	18.2

TABLE IV

ZERO-SHOT WERS ON OOV WORDS ON THE COMBINED LIBRISPEECH TEST-CLEAN AND TEST-OTHER SET USING THE BASELINE AND TCPGen SYSTEMS WITH MBWE AND BLMD. SAME BIASING LISTS WERE USED AS THOSE IN TABLE II AND TABLE III

Systems	AED	RNN-T
Baseline + MBWE + BLMD	75.6%	78.0%
TCPGen + MBWE + BLMD	37.8%	41.7%

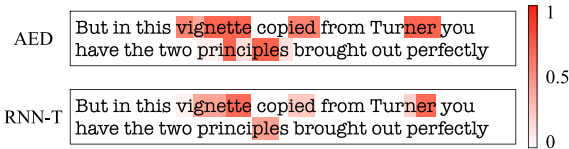


Fig. 5. Heat map of the generation probability for each wordpiece in an utterance taken from recognition results to show how each system spots where to use contextual biasing. Biasing words are vignette and Turner.

R-WER reduction from 41% to 43% on test-clean and from 37% to 38% on test-other (comparing row 6 to row 3 and row 11 to row 9 in Table III). The sign test was used to compare TCPGen with and without MBWE training, as the WER and R-WER were smaller than those observed in AED. All R-WER improvements after applying MBWE, including those when BLMD was applied, were significant at $p < .05$.

Compared to the results for the AED model, the R-WER improvement was smaller in general, and to investigate this discrepancy, the generation probabilities P^{gen} for TCPGen in AED and RNN-T models are plotted in Fig. 5 for an example utterance from the test-clean set.

As shown in Fig. 5, the shade on biasing words was lighter for RNN-T than AED, indicating that RNN-T had a smaller dependency on TCPGen than AED. This resulted in a smaller R-WER reduction using TCPGen in RNN-T. During RNN-T training, the probability of each alignment to be maximised contained only a tiny portion of the factorised probabilities that correspond to a new wordpiece output with a large portion of the blank symbol, $\emptyset$, output, and TCPGen has an effect on only a small portion of those that output a new wordpiece token. Therefore, RNN-T used TCPGen much less frequently than AED during training. Moreover, when TCPGen was used, the generation probability was scaled by $1 - P(\emptyset)$ to ensure a valid

output distribution. To encourage RNN-T to use TCPGen, the scaling factor for the rare word error loss for TCPGen was set to 5.0 during MBWE training.

TCPGen also achieved *zero-shot learning* of OOV words incorporated in the biasing list. There were 468 OOV word tokens in the combined test-clean and test-other set that were covered by the biasing list. The OOV WER which was measured in the same way as R-WER but for OOV words on the combined test sets was separately reported in Table IV for the baseline and for TCPGen systems using MBWE and BLMD. Note that these systems used exactly the same biasing lists with 1000 distractors as those in Table II and Table III, so they had the same WER and R-WER as those corresponding systems. As a result, TCPGen achieves a large OOV WER reduction compared to the best baseline system, and the majority of OOV words could be correctly recognised once incorporated in the biasing list.

C. DSTC Experiments

Finally, TCPGen and the proposed MBWE and BLMD methods were evaluated on the DSTC data where the biasing list was extracted from the ontology. Models trained on the Librispeech 960-hour data were finetuned on the DSTC track2 training set. For the baseline and TCPGen systems without MBWE training, finetuning was performed only with the CE loss, whereas for systems trained with MBWE, the MBWE loss was also applied during finetuning. The hyper-parameters for MBWE were found in the same way as before, with $\mu_1 = 0.5$ and $\mu_2 = 5.0$. Moreover, as it is difficult to obtain large external task-oriented dialogue data for LM training, an LM was trained only on the DSTC2 training data to perform either shallow fusion or LM discounting. For the baseline system, this DSTC LM was found to be more effective as an SF LM with $\alpha_1 = 0.1, \beta_1 = 0.0$, which achieved a limited WER improvement. For TCPGen, BLMD was applied with the same α_1 and β_1 as the baseline, and $\alpha_2 = 0.0, \beta_2 = 0.1$, which, in addition to the SF, the internal LM effect was discounted in the TCPGen distribution. The WER and R-WER are reported in Table V where the R-WER was measured for biasing words that appeared in the ontology.

Progressive WER and R-WER reductions were achieved by applying MBWE and BLMD successively for both AED and

TABLE V

WER AND R-WER (IN BRACKETS) ON THE DSTC3 TEST SET FOR CONFORMER AED AND RNN-T MODELS TRAINED ON LIBRISPEECH FULL 960-HOUR TRAINING SET AND FINETUNED ON DSTC2 TRAIN AND DEV SETS

System	AED(%)	RNN-T(%)
Baseline	21.31 (61.5)	21.26 (64.2)
Baseline + MBWE	20.81 (60.4)	21.15 (64.1)
Baseline + MBWE + SF	20.73 (60.4)	20.63 (64.1)
TCPGen	20.38 (45.2)	20.05 (49.2)
TCPGen + MBWE	20.00 (43.6)	19.87 (47.4)
TCPGen + MBWE + BLMD	19.74 (40.3)	19.13 (40.9)

RNN-T using TCPGen. For AED, using TCPGen achieved a 26% relative R-WER reduction compared to the baseline, which then increased to 33% with BLMD for AED. For RNN-T, TCPGen alone achieved a 23% relative R-WER reduction compared to the baseline, which increased to 36% relative R-WER reduction compared to the corresponding baseline with SF. Moreover, a dialogue-by-dialogue sign test was performed between TCPGen and TCPGen with MBWE loss. For both AED and RNN-T, R-WER improvements were significant at a p-value less than 0.05.

VIII. CONCLUSION

This paper proposed the TCPGen component for contextual ASR. TCPGen combines a neural pointer generator with a symbolic prefix-tree search. Meanwhile, the minimum biasing word error (MBWE) loss was proposed to improve the training of TCPGen with an emphasis on biasing words, and the biasing-word-driven LM discounting (BLMD) method was proposed for decoding to account for the internal LM effect. Experiments were performed on Librispeech and DSTC dialogue data. Consistent and significant WER improvements were found using TCPGen, especially on biasing words. Applying MBWE and BLMD achieved further significant R-WER reductions compared to the original TCPGen reductions.

REFERENCES

- [1] I. Williams, A. Kannan, P. Aleksic, D. Rybach, and T. Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search," in *Proc. Interspeech*, Hyderabad, India, 2018, pp. 2227–2231.
- [2] Z. Chen, M. Jain, Y. Wang, M. L. Seltzer, and C. Fuegen, "End-to-end contextual speech recognition using class language models and a token passing decoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6186–6190.
- [3] D. Zhao et al., "Shallow-fusion end-to-end contextual biasing," in *Proc. Interspeech*, 2019, pp. 1418–1422.
- [4] G. Pundak, T. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: End-to-end contextual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 418–425.
- [5] Z. Chen, M. Jain, Y. Wang, M. L. Seltzer, and C. Fuegen, "Joint grapheme and phoneme embeddings for contextual end-to-end ASR," in *Proc. Interspeech*, 2019, pp. 3490–3494.
- [6] M. Jain, G. Keren, J. Mahadeokar, G. Zweig, F. Metze, and Y. Saraf, "Contextual RNN-T for open domain ASR," in *Proc. Interspeech*, 2020, pp. 11–15.
- [7] U. Alon, G. Pundak, and T. Sainath, "Contextual speech recognition with difficult negative training examples," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6440–6444.
- [8] Z. Chen, M. Jain, Y. Wang, M. Seltzer, and C. Fuegen, "End-to-end contextual speech recognition using class language models and a token passing decoder," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6186–6190.
- [9] D. Le, G. Keren, J. Chan, J. Mahadeokar, C. Fuegen, and M. L. Seltzer, "Deep shallow fusion for RNN-T personalization," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 251–257.
- [10] D. Le et al., "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," 2021, *arXiv: 2104.02194*.
- [11] R. Huang, O. Abdel-hamid, X. Li, and G. Evermann, "Class LM and word mapping for contextual biasing in end-to-end ASR," in *Proc. Interspeech*, 2020, pp. 4348–4351.
- [12] Y. M. Kang and Y. Zhou, "Fast and robust unsupervised contextual biasing for speech recognition," 2020, *arXiv: 2005.01677*.
- [13] A. Garg, A. Gupta, D. Gowda, S. Singh, and C. Kim, "Hierarchical multi-stage word-to-grapheme named entity corrector for automatic speech recognition," in *Proc. Interspeech*, 2020, pp. 1793–1797.
- [14] D. Liu, C. Liu, F. Zhang, G. Synnaeve, Y. Saraf, and G. Zweig, "Contextualizing ASR lattice rescoring with hybrid pointer network language model," in *Proc. Interspeech*, 2020, pp. 3650–3654.
- [15] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 1073–1083.
- [16] Z. Liu, A. Ng, S. Lee, A. T. Aw, and N. F. Chen, "Topic-aware pointer-generator networks for summarizing spoken conversations," in *Proc. Autom. Speech Recognit. Understanding*, 2019, pp. 814–821.
- [17] W. Li, R. Peng, Y. Wang, and Z. Yan, "Knowledge graph based natural language generation with adapted pointer-generator networks," *Neuro-computing*, vol. 328, pp. 174–187, 2020.
- [18] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.
- [19] L. Lu, X. Zhang, K. Cho, and S. Renals, "A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition," in *Proc. Interspeech*, 2015, pp. 3249–3253.
- [20] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4945–4949.
- [21] C. C. Chiu et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4774–4778.
- [22] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," in *Proc. Interspeech*, 2018.
- [23] Y. Zhang et al., "Towards end-to-end speech recognition with deep convolutional neural networks," in *Proc. Interspeech*, 2016, pp. 410–414.
- [24] T. Hayashi et al., "Back-translation-style data augmentation for end-to-end ASR," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 426–433.
- [25] S. Karita, A. Ogawa, M. Delcroix, and T. Nakatani, "Sequence training of encoder-decoder model using policy gradient for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5839–5843.
- [26] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6645–6649.
- [27] R. Prabhavalkar, K. Rao, T. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition," in *Proc. Interspeech*, 2017, pp. 939–943.
- [28] E. Battenberg et al., "Exploring neural transducers for end-to-end speech recognition," in *Proc. Autom. Speech Recognit. Understanding*, 2017, pp. 206–213.
- [29] J. Li, R. Zhao, H. Hu, and Y. Gong, "Improving RNN transducer modeling for end-to-end speech recognition," in *Proc. Autom. Speech Recognit. Understanding*, Singapore, 2019, pp. 114–121.
- [30] Y. He et al., "Streaming end-to-end speech recognition for mobile devices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6381–6385.
- [31] M. Ghodsi, X. Liu, J. Apfel, R. Cabrera, and E. Weinstein, "Rnn-transducer with stateless prediction network," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 7049–7053.
- [32] Q. Li, C. Zhang, and P. C. Woodland, "Integrating source-channel and attention-based sequence-to-sequence models for speech recognition," in *Proc. Autom. Speech Recognit. Understanding*, 2019, pp. 39–46.
- [33] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 369–376.
- [34] A. Zeyer, E. Beck, R. Schlüter, and H. Ney, "CTC in the context of generalized full-sum HMM training," in *Proc. Interspeech*, 2017, pp. 944–948.
- [35] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4835–4839.

- [36] Q. Liu, Z. Chen, H. Li, M. Huang, Y. Lu, and K. Yu, "Modular end-to-end automatic speech recognition framework for acoustic-to-word model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2174–2183, 2020.
- [37] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end speech recognition using lattice-free MMI," in *Proc. Interspeech*, 2018, pp. 12–16.
- [38] W. Chan, N. Jaitly, Q. V. 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.
- [39] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.
- [40] 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.
- [41] 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.
- [42] C. Weng, C. Yu, J. Cui, C. Zhang, and D. Yu, "Minimum bayes risk training of RNN-Transducer for end-to-end speech recognition," in *Proc. Interspeech*, 2020, pp. 966–970.
- [43] J. Guo et al., "Efficient minimum word error rate training of RNN-transducer for end-to-end speech recognition," in *Proc. Interspeech*, 2020, pp. 2807–2811.
- [44] C. Weng et al., "Improving attention based sequence-to-sequence models for end-to-end english conversational speech recognition," in *Proc. Interspeech*, 2018, pp. 761–765.
- [45] N. P. Wnand, W. Michel, J. Rosendahl, R. Schlüter, and H. Ney, "Efficient sequence training of attention models using approximative recombination," 2021, *arXiv: 2110.09245*.
- [46] Z. Meng et al., "Minimum word error rate training with language model fusion for end-to-end speech recognition," in *Proc. Interspeech*, 2021.
- [47] L. Lu, Z. Meng, N. Kanda, J. Li, and Y. Gong, "On minimum word error rate training of the hybrid autoregressive transducer," in *Proc. Interspeech*, 2021.
- [48] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden markov models for speech recognition," *Comput. Speech Lang.*, Vol. 16, pp. 25–47, 2002.
- [49] E. McDermott, H. Sak, and E. Variani, "A density ratio approach to language model fusion in end-to-end automatic speech recognition," in *Proc. Autom. Speech Recognit. Understanding*, 2019, pp. 434–441.
- [50] A. Kannan, Y. Wu, P. Nguyen, T. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5828.
- [51] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 389–396.
- [52] S. Kim et al., "Improved neural language model fusion for streaming recurrent neural network transducer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7333–7337.
- [53] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid autoregressive transducer (HAT)," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020, pp. 6139–6143.
- [54] Z. Meng et al., "Internal language model estimation for domain-adaptive end-to-end speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 243–250.
- [55] Z. Meng et al., "Internal language model training for domain-adaptive end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 243–250.
- [56] C. Zhang, B. Li, Z. Lu, T. N. Sainath, and S. Chang, "Improving the fusion of acoustic and text representations in RNN-T," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 8117–8121.
- [57] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.
- [58] S. Watanabe et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.
- [59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2014.
- [60] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
- [61] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," in *Proc. Interspeech*, 2017, pp. 523–527.
- [62] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017.
- [63] G. Sun, C. Zhang, and P. C. Woodland, "Tree-constrained pointer generator for end-to-end contextual speech recognition," in *Proc. Autom. Speech Recognit. Understanding*, 2021, pp. 780–787.
- [64] C. Huber, J. Hussain, S. Stüker, and A. Waibel, "Instant one-shot word-learning for context-specific neural sequence-to-sequence speech recognition," in *Proc. Autom. Speech Recognit. Understanding*, 2021, pp. 1–7.
- [65] J. Kim and Y. Lee, "Accelerating RNN transducer inference via one-step constrained beam search," *IEEE Signal Process. Lett.*, vol. 27, pp. 2019–2023, 2020.
- [66] A. Tripathi, H. Lu, H. Sak, and H. Soltau, "Monotonic recurrent neural network transducer and decoding strategies," in *Proc. Autom. Speech Recognit. Understanding*, 2019, pp. 944–948.
- [67] 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.



Guangzhi Sun (Student Member, IEEE) received the B.A. and M.Eng. degrees in information engineering in 2019 from the Department of Engineering, University of Cambridge, Cambridge, U.K., where he is currently working toward the Ph.D. degree in information engineering. His research interests include contextual speech recognition, spoken dialogue and language understanding, speaker diarisation, and speech synthesis. He was awarded the Best Student Paper in Interspeech 2022.



Chao Zhang (Member, IEEE) received the B.E. and M.Sc. degrees from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2009 and 2012, respectively, and the Ph.D. degree from Cambridge University Engineering Department (CUED), Cambridge, U.K. He is currently an Assistant Professor with the Department of Electronic Engineering, Tsinghua University. Before that, he was a Senior Research Scientist with Google, a Research Associate with CUED, and an Advisor and Speech Team Co-Leader of JD.com. He has authored or coauthored 70 peer-reviewed speech and language processing papers. His research interests include spoken language processing, machine learning, and cognitive neuroscience. He is also a Visiting Fellow with CUED, and an Associate Member of the IEEE Speech and Language Processing Technical Committee. He was the recipient of multiple paper awards.



Philip C. Woodland is currently a Professor of information engineering with the Engineering Department, University of Cambridge, Cambridge, U.K., where he is the Head of the Machine Intelligence Laboratory and a Professorial Fellow of Peterhouse. After working with British Telecom Research Labs for three years, he returned to a Lectureship with Cambridge in 1989 and became a Full Professor in 2002. He has authored or coauthored more than 250 papers in the area of speech and language technology with a main focus on speech recognition systems. He was the recipient of number of Best Paper awards including for work on speaker adaptation and discriminative training. He is one of the original coauthors of the HTK toolkit and has continued to play a major role in its development. He was a Member of the Editorial Board of Computer Speech and Language (1994–2009) and is currently a Member of the Editorial Board Member of Speech Communication. He was a Member of the Speech Technical Committee of the IEEE Signal Processing Society from 1999 to 2003. He is a Fellow of the International Speech Communication Association and the Royal Academy of Engineering.