Post-decoder Biasing for End-to-End Speech Recognition of Multi-turn Medical Interview

Heyang Liu [1], **Yu Wang**[1,2]*, **Yanfeng Wang**[1,2]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]Shanghai Artificial Intelligence Laboratory
{liuheyang, yuwangsjtu,wangyanfeng622}@sjtu.edu.cn

Abstract

End-to-end (E2E) approach is gradually replacing hybrid models for automatic speech recognition (ASR) tasks. However, the optimization of E2E models lacks an intuitive method for handling decoding shifts, especially in scenarios with a large number of domain-specific rare words that hold specific important meanings. Furthermore, the absence of knowledge-intensive speech datasets in academia has been a significant limiting factor, and the commonly used speech corpora exhibit significant disparities with realistic conversation. To address these challenges, we present Medical Interview (MED-IT), a multi-turn consultation speech dataset that contains a substantial number of knowledge-intensive named entities. We also explore methods to enhance the recognition performance of rare words for E2E models. We propose a novel approach, post-decoder biasing, which constructs a transform probability matrix based on the distribution of training transcriptions. This guides the model to prioritize recognizing words in the biasing list. In our experiments, for subsets of rare words appearing in the training speech between 10 and 20 times, and between 1 and 5 times, the proposed method achieves a relative improvement of 9.3% and 5.1%, respectively.

Keywords: automatic speech recognition, end-to-end, contextual speech recognition, knowledge-intensive

1. Introduction

Automatic speech recognition (ASR) is a fundamental task that converts speech signals into corresponding textual formats. The hybrid model (Dahl et al., 2011; Hinton et al., 2012) consists of distinct components: an acoustic model, a language model, and a lexicon. Subsequently, the sequence-to-sequence paradigm (Graves, 2012; Sutskever et al., 2014; Chan et al., 2016; Prabhavalkar et al., 2017) has provided an E2E pattern, gradually revealing its potential to disrupt the research field. Although this paradigm has shown great superiority, it often declines in the scenarios of many rare words with low frequency in the training corpus. These rare words often contain important meanings with a significant impact on downstream tasks such as question answering. Contextual automatic speech recognition (CASR) (Aleksic et al., 2015; Michaely et al., 2017; Pundak et al., 2018; Alon et al., 2019; Zhao et al., 2019; Le et al., 2021b) aims to improve the recognition accuracy of hot words, with the most challenging aspect being rare words. It finds wide-ranging applications, for instance, in medical consultation or company meeting scenarios. The common characteristic is that specialized nouns of a knowledgeable nature hold higher significance, and they may have lower word frequencies during training. We refer to such scenarios as knowledge-intensive contexts, and we concentrate on enhancing the performance of E2E models in recognizing rare words within this setting.

The scarcity of speech data in knowledge-intensive scenarios, especially multi-turn medical consultation, has become one of the limiting factors in the academic community. In medical settings, the discourse is replete with medical terminology that is not commonly found in daily language. The significant roles played by these terms are beyond the scope of generic data. The industry possesses the capability to acquire large-scale speech data from specific settings. Google has recorded medical dialogues covering 151 different diseases, totaling 14,000 hours (Chiu et al., 2017). However, due to concerns regarding internal corporate information and speaker privacy, such data is not made public. Alternatives adopted by the academic community include training on general datasets or pre-collecting audio from relevant domains, which will lead to severe domain mismatch or great human workload. In an authentic open-source setting, a unified speech dataset rich in named entities serves as a crucial prerequisite for academic methodology comparison and system optimization.

In light of the current shortage of relevant speech corpus, this work focuses on two aspects: we construct an English consultation dataset MED-IT. Subsequently, we conduct research on rare word recognition and propose corresponding algorithms. The main contributions of this paper are as follows:

1) Dataset: We are dedicated to the field of medical consultations and have segmented authen-

*Corresponding author

This work is supported by National Key R&D Program of China (No. 2022ZD0162101), STCSM (No. 21511101100, No. 22DZ2229005), and State Key Laboratory of UHD Video and Audio Production and Presentation.

LREC-COLING 2024, pages 12917–12926

tic speech data from four departments to build a knowledge-intensive speech corpus called MED-IT. This corpus consists of multi-turn conversational medical dialogues and includes a substantial number of named entities such as medication names, disease symptoms, and treatment plans. It offers robust data support for CASR experiments and other related research.

2) Algorithm: We propose a novel lightweight and portable scheme called post-decoder biasing to enhance the recognition of rare words. Compared to previous approaches, our method has a minimal impact on the recognition of non-rare words. Furthermore, it can be easily integrated into different E2E models without significant computational cost or decoding latency.

2. Related Works

2.1. Knowledge-Intensive Corpus

Constructing a speech corpus in the knowledge-intensive scenario presents several challenges. Records in the real world are generally characterized by noisy conditions, which makes it difficult to guarantee speech quality. The introduction of specialized vocabulary implies that the data annotation process requires the involvement of professionals. There are also ethical concerns, such as privacy protection. Prior to data being made public, it necessitates scrutiny and protective measures, especially in contexts like medical consultations. Due to the limited availability of knowledge-intensive speech corpora, most studies on CASR have been conducted using general datasets. Many experiments are conducted on LibriSpeech (Panayotov et al.; Le et al., 2021b; Han et al., 2022), while some researchers prefer GigaSpeech (Chen et al., 2021; Fox and Delworth, 2022).

Knowledge-intensive datasets align closely with real-world applications, providing excellent material for CASR. To our knowledge, only Earnings-21 (Del Rio et al., 2021) meets the requirements. It collects English speech from nine different financial sectors, totaling 39 hours, which includes specialized domain-specific factual terms such as organization, company names, and financial vocabulary. However, there has consistently been a shortage of knowledge-intensive speech data in medical consultation scenarios and many other settings.

2.2. E2E Contextual Speech Recognition

As the superiority of E2E speech recognition systems gradually becomes evident, research in CASR has also shifted towards this paradigm. Previous efforts primarily encompass three approaches: shallow fusion based on language models (Aleksic et al., 2015; Williams et al., 2018; Kannan et al., 2018; Zhao et al., 2019), deep context utilizing attention mechanisms (Pundak et al., 2018; Han et al., 2022), and deep biasing based on word pieces (Le et al., 2021b,a; Zhang and Zhou, 2022). Shallow fusion enhances the prediction probability of rare words by refining the language model for specific vocabulary. This approach has been extensively studied in both hybrid models and E2E systems. The deep context approach explicitly incorporates information from the biasing list into the network architecture. A bias encoder is used to generate embeddings for rare words. Deep biasing takes advantage of the powerful representation and modeling capabilities of neural networks. It constructs a prefix tree at the word pieces level, capturing the concatenation patterns of word segments in the biasing list. This process occurs simultaneously with decoding, facilitating the retrieval of potential subsequent context concatenation patterns.

2.3. Multimodal Knowledge Fusion

For multimodal processing, the lexicon of text modality provides an intuitional method for knowledge fusion. Taking the field of multimodal emotion detection as an example, emotion lexicons compass the importance of tokens in the specific task. Words imbued with significant emotional valence are assigned higher weights, consequently affording them greater prominence in subsequent recognition and detection processes. The lexicon has been widely used by concept and knowledge retrieval for enhancing meaningful words (Zhong et al., 2019). Other works directly fuse the information from the word list with the feature vectors of the textual modality (Zhao et al., 2023). In the field of speech recognition, a similar approach is proposed by Das et al. (Das et al., 2022), which introduces a knowledge graph as an external knowledge base. After the first decoding is completed and one of the hot words is correctly recognized, they guide the model to recognize another real word located at a neighboring node by adding additional language model scores. However, such a knowledge-based lexicon is not always available for CASR. Instead, the biasing list provides the importance rate, which can be fused directly into the decoding process alongside the connection distribution of recognition units obtained from the training transcriptions.

3. MED-IT Dataset

3.1. Creation Pipeline

A previous study has collected real-life simulated clinical consultation speech recordings in a hospital (Fareez et al., 2022). Each dialogue involves one doctor and one patient, both portrayed by medical students, ensuring no potential issues related to

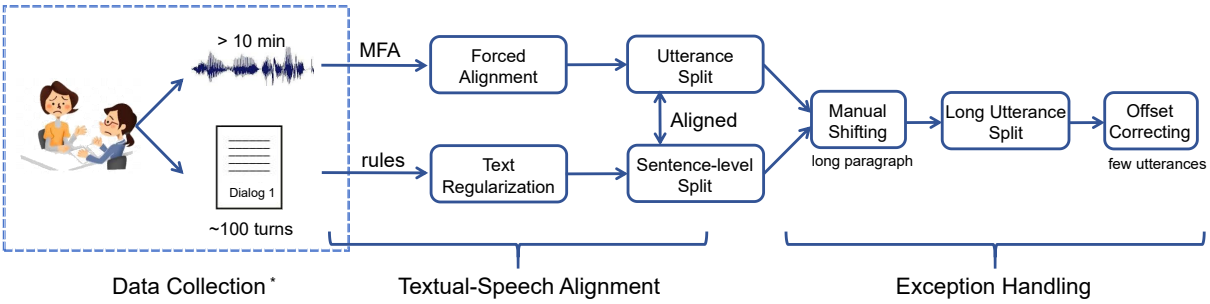


Figure 1: Creation pipeline of MED-IT. Data collection was done in published research. Cleaning and segmentation on both modalities have been performed serialized for textual-speech alignment, and then manual examination for exception handling.

privacy disclosure. However, accurate time annotation has not been released, and transcription quality is not trustful. Building upon this open-source corpus, we establish the MED-IT speech dataset, which can be applied to recognition tasks. Our data processing procedure is present in Figure 1:

1) Data Cleaning: We employ a standardized annotation approach for cleaning textual annotation, including using a special token to represent colloquial pauses, converting special characters based on their pronunciation, and systematically rectifying annotation format errors. Speech wav with great noise or overlap is discarded.

2) Data Segmentation: The original audio segments typically have durations of over ten minutes. Obtaining precise utterance-level alignment is a necessary step in speech recognition training. The Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) is initially applied for forced alignment. We acquire the pronunciation time annotations of each word in a long conversation as a basis for segmenting speech. Sentence-level split on text is performed to ensure the correspondence between speech wav and transcription. Practical experience has shown that the performance of the MFA is closely tied to the acoustic environment. For the few instances where significant alignment errors occur, we utilize the phonetic software Praat (Boersma and Van Heuven, 2001) for re-segmentation and annotation. To ensure computational efficiency, we limit the duration of each individual utterance to within 24 seconds. Any statements exceeding this duration are further split.

3) Data Examination: After the initial segmentation, we conduct regular interval sampling checks on the speech data to prevent any potential audio offsets in shorter passages. Additionally, a small portion of multi-channel audio was converted into single-channel through linear interpolation. The inspected speech data exhibits appropriate durations and precise alignment with the text annotations.

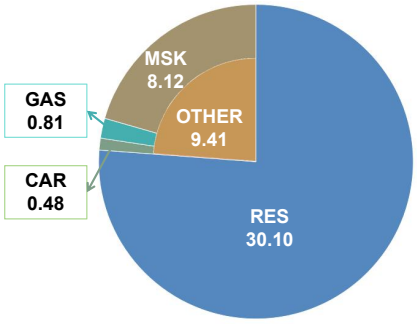
3.2. Dataset Details and Application

MED-IT is a medical consultation speech corpus recorded in real-world scenarios. The doctor-patient dialogues are structured according to the Objective Structured Clinical Examination (OSCE) (Zayyan, 2011). The main process includes the patient's introduction of symptoms, the doctor's inquiries about the disease condition, and finally, diagnostic and treatment recommendations. Our speech dataset encompasses diagnostic and treatment consultant recordings from four departments: Respiratory (RES), Musculoskeletal (MSK), Gastrointestinal (GAS), and Cardiovascular (CAR). Among these, the majority of the recordings pertain to RES, with the composition ratios and other details as illustrated in Figure 2.

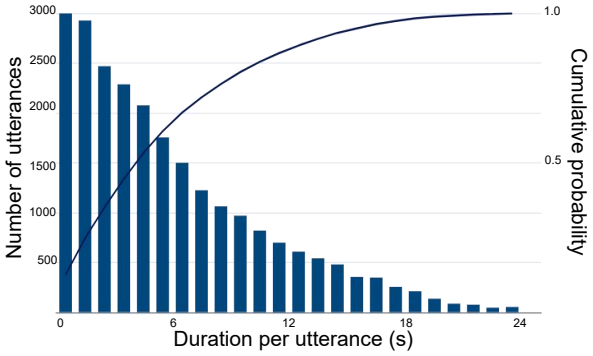
MED-IT is a standard speech dataset that can be used to evaluate speech recognition models. It contains a significant number of medical-specific terms that are not commonly found in general corpora. This presents a greater and unique challenge for ASR and can be used as evaluation data to assess the generalization capability. The main objective of constructing this dataset is CASR. In addition, MED-IT can also be utilized for research in the field of natural language processing. For example, the manually annotated textual data can be used to train named entity recognition (NER) models with additional annotations. It can also be used to develop semantic understanding and dialogue-generation models for doctor-patient interactions based on the standard OSCE diagnostic process.

Previous work has shown significant interest in this corpus for ASR tasks, although there is still a performance gap (Liu et al., 2023; Whetten et al., 2023). The entire MED-IT dataset and also the annotations have been uploaded to Hugging Face[1].

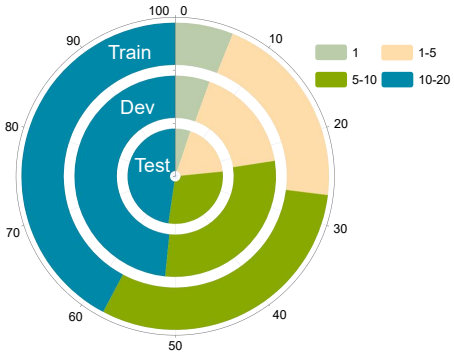
[1]The dataset with transcriptions will be made available at https://huggingface.co/datasets/SandO114/Medical_Interview



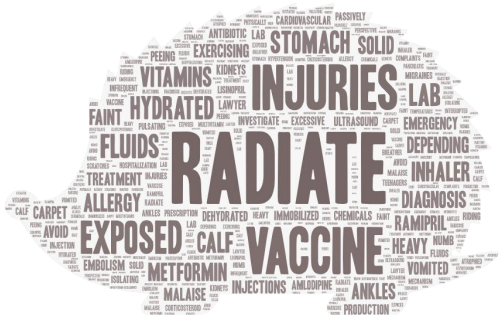
(a) Distribution of speech duration



(b) Number of seconds per utterance



(c) Partition of different rare words subset



(d) Rare words slices with different frequencies

Figure 2: Statistics of our dataset. (a) shows the speech duration of each department with RES making the most. (b) shows the number of seconds per utterance, with most lasting for less than 10 seconds. (c) indicates the partition portion of each biasing list, and (d) shows word slices from them.

3.3. Biasing List Selection for CASR

As for the CASR task, we select words with training set frequencies falling into the ranges of (10,20], (5,10], (1,5], and 1 to form different subsets of rare words. Figure 2(c) exhibits the statics of the partition of rare word subsets with different frequencies, with the outer ring indicating the portion of unique words in each rare word set (i.e. word "vaccine" accounts for 1 even if it occurs 20 times), and other rings indicating the total frequency in each evaluation set(i.e. "vaccine" accounts for 5 if it occurs 5 times in corresponding set). Figure 2(d) shows part of the biasing rare words, with the word size proportional to the relative word frequencies. It can be observed that, although the construction of rare word subsets is based on word frequency rather than semantics, it contains a large number of knowledge-intensive named entities, such as human organs, disease symptoms, and drug names. Rare words carry rich semantic information, and accurately recognizing these vocabulary items will have a profound impact.

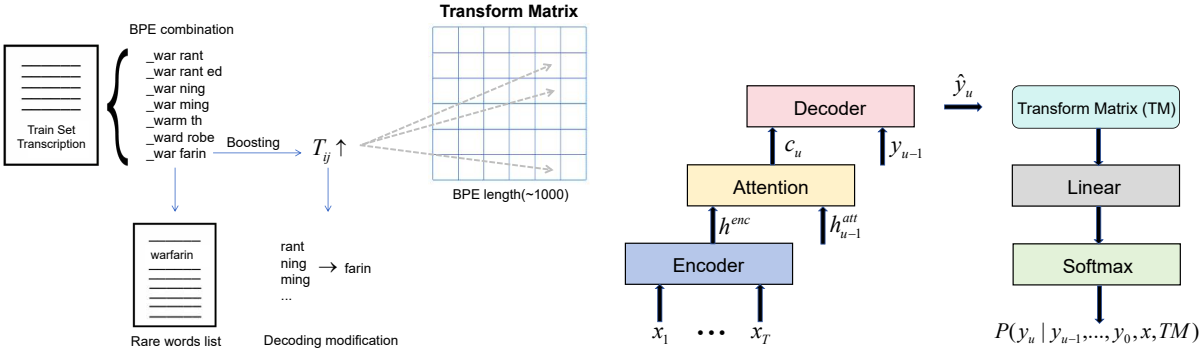
4. Post-Decoder Biasing

Rare words can be considered as hot words that appear less frequently in the training set. While the textual transcription in speech recognition inference is unknown, the decoded hypothesis provides an unbiased recognition probability, along with an approximate confidence score. In addition, training transcription provides pivotal connection distribution in-domain, which serves as a steady knowledge base for decoded hypothesis biasing. As illustrated in Figure 3, the post-decoder fuses the information from the transcription distribution and conducts biased recognition results based on unbiased decoding outputs. This is achieved by the replacement of recognition units that favor the biasing words. The replacement rules are related to the word frequency in the training set. By boosting the word frequencies of rare words, post-decoder biasing guides the model to consider the rare words with a suboptimal recognition probability.

Compared to previous approaches, our method leverages information from the initial decoding and does not require the introduction of additional knowledge. Furthermore, it avoids the computational cost and latency associated with architectural improvements or secondary decoding fusion.

4.1. Post-Decoder Architecture

The post-decoder introduces a transform probability matrix, which facilitates biasing subword replace-



(a) Caculation of the transform matrix (b) LAS model with post-decoder biasing

Figure 3: Overview of the post-decoder biasing for attention-based encoder-decoder. The transcription of the train set contains the biasing word "warfarin". BPE "rant" might be transformed to "farin" if the corresponding probability is abnormally high, for "warrant" and "warfarin" are both valid words. The biasing decoding results are determined by both neural architecture and transform matrix obtained from the in-domain sub-word combination distribution.

ment. Generally, common words appear more frequently in the training set, leading the model to output higher probabilities for corresponding subword units during inference. On the other hand, rare words tend to have lower probabilities. If the model outputs several subword units with the highest probabilities that are listed as rare words, it suggests the possibility of the model incorrectly identifying a rare word as a common one. Specifically, the decoder outputs token probability with a shape of $[batch, max\_len, bpe\_size]$, where $batch$ denotes the batch size, max_len represents the maximum number of subwords in a single sentence, and bpe_size is the number of tokens chosen by the model. When the subwords contained in the rare word list exhibit higher probabilities, the replacement probability matrix executes substitutes from non-rare words to rare words. We incorporate a linear layer connection and use the output probabilities as the basis for recognizing the E2E model. The added network in the post-decoder biasing architecture is a single connected layer, which results in minimal time delay and computation cost.

4.2. Transform Matrix Calculation

The transform matrix models the replacement probabilities of byte pair encoding (BPE) units. We do not employ any criteria related to pronunciation but rather consider the concatenations and word combinations in the training transcriptions to obtain an approximate statistical representation. Assuming a BPE partition $B = (b_1, b_2, ..., b_k)$, the replacement probability matrix T ideally represents the probability of replacing b_i with b_j as T_{ij}. Constructing the transform matrix involves the computation of the following two steps:

1) Computing BPE connection probabilities. We divide BPEs into two sets: word prefixes or standalone tokens P with each element containing the special marker "_" and word suffixes or mid-segments S without special symbols. For the former, we calculate the subsequent subword unit distribution for this BPE connection as follows:

$$p_i^j = \frac{n_{ij}}{\sum_k n_{ik}} \quad (b_i \in P) \tag{1}$$

p_i^j represents the probability of subword unit b_j immediately following word prefix (or standalone token) b_i within the corpus domain and without any prior conditions, and n_{ij} denotes the frequency of subword unit b_j immediately following word prefix b_i in the training set. Similarly, for word suffixes and mid-segments, we calculate the distribution of word pieces connected before them:

$$p_i^j = \frac{n_{ji}}{\sum_k n_{ki}} \quad (b_i \in S) \tag{2}$$

2) Calculating BPE replacement probabilities. For subword unit b_i, assuming the replacement probability is p_i, the probability of not being replaced is $1 - p_i$, i.e., $T_{ii} = 1 - p_i$. We model the replacement probability of BPE as the substitutability of subword units in the training text, following the calculation method as follows:

$$T_{ij} = p_i \sum_k p_i^k p_k^j \quad (i \neq j, i \neq k, j \neq k) \tag{3}$$

The calculation of the transform matrix is done independently before the training of the neural network, which means it does not affect the efficiency of the training and inferring process. Additionally, we have not introduced any additional prior knowledge or relied on pronunciation-based similarity. In practical applications, this approach exhibits strong scalability, for providing the text annotations of the

training set (the in-domain BPE combination distribution) is always sufficient.

5. Experiments Setup

5.1. Dataset and Evaluation

Experiments are conducted on MED-IT, which has been explained in Section 3. In the following work, we implement E2E models for the CASR task. Most experiments are conducted for biasing lists with frequencies of (10,20] and (1,5] to represent rare words with relatively higher frequencies and extremely low occurrence in the training corpus.

As for the evaluation metrics, WER (word error rate) is the most popular choice for assessing the overall performance of a speech recognition system. We introduce a similar concept into rare word recognition, defining RWER (rare word error rate) as the evaluation metric for recognizing rare words. It encompasses deletion errors and substitution errors for rare words, as well as insertion errors where rare words wrongly appear in the decoding hypothesis.

5.2. Model Specification

One noteworthy exemplar of E2E speech recognition is the Attention-based Encoder-Decoder (AED) (Chan et al., 2016), which utilizes an attention mechanism to enhance the importance of specific information. Connectionist temporal classification (CTC) (Graves et al., 2006) uses intermediate representation allowing repetitions of labels and blank tokens. The model widely used to leverage both advantages is CTC-Attention (Watanabe et al., 2017), which we use as our recognition backbone. The implementation is carried out following the instructions of the well-known ESPnet (Watanabe et al., 2018) toolkit. We use Conformer (Gulati et al., 2020) as our encoder, which is improved from Transformer (Vaswani et al., 2017) by capturing both long-distance and local representations. Our Conformer encoder contains 6 blocks, with hidden linear units $d^{ff} = 1024$ and attention output size $d^{att} = 256$. Attention head H is set to 4 and the front CNN kernel size is 31. As for the decoder, we use the Transformer with 6 blocks ($d^{ff} = 1024$,$H = 8$). We use 1k BPE as our recognition unit obtained by SentencePiece (Kudo and Richardson, 2018). In addition to the attention training objectives, We use a certain degree of CTC loss function. CTC weight λ is 0.3 during both training and infering. SpecAugment (Park et al., 2019) is applied with time mask width $T = 40$ and frequency mask with $F = 30$. We train our model on two 24GB 3090 RTX GPUs for 80 epochs. The top 10 checkpoints are preserved for model averaging. It is worth mentioning that our model does not employ the common batching method used in ESPnet. Instead, we strive to ensure that the same conversation is assigned to the same worker in data parallelization, which provides a certain optimization for the recognition of rare words (Kim and Metze, 2018).

6. Results

6.1. CTC-Attention Results

The most common batch processing modes are numel (num element) and unsorted. Both involve random iterations, with the distinction lying in whether the sampling within each batch is sequential or random. We implement dialog batch, which aims to ensure that utterances in the same dialogue are assigned to the same worker in distributed training. Any discrepancies caused by varying utterance quantities are compensated for by padding at the end. The optimal parameters we use and the experimental results obtained with three different batch processing modes are shown in Table 1. RWER(20) stands for the word error rate of the subset with frequencies between 10 and 20 in the training set, and similarly, RWER(10) quantifies the recognition performance of words with frequencies between 5 and 10. Although the unsorted batch achieves the best overall performance on the whole dataset, the dialog batch shows superiority in rare word recognition, which will be used in the following experimental settings. It is worth mentioning that the recognition performance is much better than previous studies with self-trained or commercial systems on the origin corpus(over 21% WER) (Liu et al., 2023; Whetten et al., 2023), thanks to our decent dataset preprocess.

6.2. Post-Decoder Biasing Results

To evaluate the performance of post-decoding biasing for CASR, we increase the frequency of each word in the rare word subset of (10,20] and (5,10] by 100 and calculate the transform probability matrix T. Before that, it is necessary to quantify p_i. Here, we treat p_i as a hyperparameter based on empirical knowledge. In the next subsection, we will use an automatic and quantified method to specify this value. To some extent, p_i affects the confidence threshold of the recognition probability from the decoder: the larger p_i, the higher the confidence threshold, and only the optimal outputs that exceed this confidence threshold will be retained, while outputs with confidence below this threshold will be replaced with biasing tokens towards rare words. In general, recognition units with lower word frequencies in the training set tend to have fewer learned features, resulting in smaller corresponding probabilities. Therefore, we choose larger values of p_i

Table 1: Experiment results on MED-IT(%)

Method/Batch	biasing set	p_i	WER	RWER(20)	RWER(10)	RWER(5)	RWER(1)
Azure(Whetten et al., 2023)	-	-	21.0	-	-	-	-
IFNT(Liu et al., 2023)	-	-	22.3	-	-	-	-
CTC-Attention							
Numel	-	-	17.4	40.4	55.9	70.0	88.1
Unsorted	-	-	**15.5**	38.0	48.1	**66.8**	84.1
Dialog	-	-	16.0	**37.8**	**47.1**	67.3	**82.8**
Dialog	(1,5]	0.3	16.4	38.3	50.4	66.3	82.1
	(1,5]	0.7	16.2	35.6	46.6	**64.1**	80.1
	(10,20]	0.3	16.4	**34.3**	46.3	67.1	80.1
	(10,20]	0.7	16.2	36.1	49.6	68.6	83.4
	(1,5]	Auto	16.3	35.3	48.4	**63.9**	82.1
	(10,20]	Auto	15.9	**34.5**	49.1	66.1	82.7

for rare words with lower frequencies and smaller values for rare words with relatively higher frequencies. The experimental results with post-decoder biasing are shown in the third part of Table 1.

The baseline experiment using dialog batch without rare words biasing achieves rare word error rates of 37.8% and 67.3% for the two subsets respectively. When using the post-decoder and increasing the word frequency with occurrences in the range of (1,5] by 100 times, the corresponding rare word error rate can be reduced to 34.3% with a relative improvement of 9.3%. Similarly, increasing the word frequency with occurrences in the range of (10,20] by 100 times leads to a rare word error rate of 64.1% with 4.8% relative improvement.

6.3. Automatic Tuning Results

For a subword unit BPE b_i, if it occurs frequently enough in the training set, the decoding probability tends to be stable for many speech features learned by the neural model. In this case, we choose a higher p_i which means a high threshold to keep the output unchanged. Let n_i denote the frequency of b_i in the training transcriptions. The auto-selection process is present in Equation 4. We follow the linear interpolation scheme and avoid the need to adjust hyperparameters for different frequency ranges of rare word subsets. The experimental results, shown in the third part of Table 1, demonstrate that the automatic adjustment of replacement probabilities leads to similar performance. It achieves a relative reduction of 5.1% in RWER on the subset of extremely rare words with occurrences in the range of (1, 5]. Furthermore, it shows minimal degradation in performance on the other subset. It is worth noting that various interpolation schemes have minimal influence on the efficacy of rare word recognition. Therefore, the primary consideration should be the associated computational cost.

$$p_i = \begin{cases} 0.9 & n_i \geq 1000 \\ 0.9\dfrac{n_i}{1000} & 100 < n_i < 1000 \\ 0.09 & n_i \leq 100 \end{cases} \quad (4)$$

6.4. Ablation Experiment

In order to demonstrate the joint effectiveness of each component of the post-decoder, we conduct ablation experiments, constructing the same structure that solely uses the replacement probability matrix, or adds a linear layer. The same architecture without increasing the training set word frequency is also conducted. The experimental results are shown in Table 2.

Compared to the direct decoding results, using only a linear layer as the post-decoder leads to a slight decline in overall recognition performance (from 16.0% to 16.5%), while the recognition performance of rare word subsets remains approximately unchanged. Using only the probability replacement matrix without enhancing the probabilities of rare words also results in a degradation of the overall recognition performance. At the same time, no stable gain in rare word recognition is observed. Enhancing specific subsets and changing the transform matrix alters the model's output distribution, leading to a noticeable decline in both the overall accuracy and rare word recognition performance. When only enhancing the (10,20] rare word subset with the replacement probability matrix, the recognition performance of this subset remains stable (first and sixth row), while the recognition abilities of other subsets decline dramatically. We believe that the enhancement of the probability replacement matrix changes the distribution tendency of the model's output, but its quantification is achieved through the linear layer. We implement the complete post-decoder but do not enhance the training set frequencies (seventh and eighth row). The ex-

Table 2: Ablation Experiment for Post-Decoder(%)

linear	TM	biasing set	p_i	WER	RWER(20)	RWER(10)	RWER(5)	RWER(1)
-	-	-	-	16.0	37.8	47.1	67.3	82.8
✓	-	-	-	16.5	37.1	50.1	67.3	82.8
-	✓	-	0.3	16.4	37.6	47.9	67.3	85.4
-	✓	-	0.7	16.3	37.3	48.1	67.3	83.4
-	✓	(1,5]	-	16.6	38.4	49.1	71.0	85.4
-	✓	(10,20]	-	16.8	37.8	52.6	70.3	86.1
✓	✓	-	0.3	16.1	37.3	46.6	67.6	82.8
✓	✓	-	0.7	16.3	37.3	48.6	66.6	86.8
✓	✓	(1,5]	0.7	16.2	35.6	46.6	**64.1**	80.1
✓	✓	(10,20]	0.3	16.4	**34.3**	46.3	67.1	80.1

periment shows a relatively small impact on the overall recognition performance of the model, decreasing from the baseline of 16.0% to 16.1% and 16.3%. However, there is no significant gain in rare word recognition (less than 1%). The probability replacement matrix without rare word enhancement has a minor impact on the output distribution. Although it provides the possibility of substituting suboptimal paths, it does not guide the model in a specific direction during inference.

6.5. Rare Words Enhancing Frequency

In the previous experiments, we increase the word frequency of the biasing subset in the training set by 100 times when constructing the transform matrix. In this subsection, we investigate the effects of the increased rare word frequency on the recognition performance. We use six different frequencies to construct the replacement probability matrix. The performance of post-decoder biasing with different enhancing frequencies is illustrated in Figure 4.

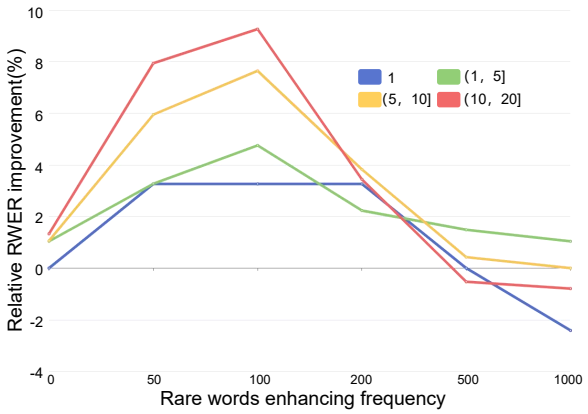


Figure 4: Relative rare words recognition improvement with different enhancing frequencies. Negative values indicate a certain degree of decline compared to the baseline.

The experimental results indicate that the frequency enhancement affects the performance of post-decoder biasing. For rare words with extremely low frequencies in the training set, even with an enhancement of 1000 times, there is still a performance gain. However, for other rare word subsets, a certain decrease in recognition performance is observed. As for relatively higher-frequency rare words, when the frequency enhancement exceeds 500 or 1000 times, the post-decoder shows a negative impact on this rare word subset. In general, with higher frequency enhancements, the model's replacement probability increases, leading to a decline in rare word deletion errors and an increase in insertion errors. The increase in insertion errors is more noticeable when the rare word subset itself has a relatively high decoding score. Among the enhancing frequencies, 100 shows the best performance for all the biasing lists. It is encouraging because when the experiment performs well on a specific subset of rare words, this hyperparameter should also be applicable to other subsets.

7. Conclusion

In knowledge-intensive scenarios, rare words often have extremely important meanings. However, the scarcity of speech datasets in this context has limited academic research. In this study, we reconstruct a speech corpus focused on medical inquiries, which contains a wealth of specialized named entities. To enhance the E2E model's ability to recognize rare words, we propose a lightweight and easily transferable post-decoder biasing method. The experiments show that post-decoder has a positive effect on CASR. By simply increasing the word frequency, the model achieves relative performance improvements of 9.3% and 5.1% on two subsets of rare words with frequency ranges of (10,20] and (1,5], respectively. In future work, we will explore rare word replacements based on pronunciation similarities and word-level rules.

8. Bibliographical References

Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, and Pedro Moreno. 2015. Bringing contextual information to google speech recognition.

Uri Alon, Golan Pundak, and Tara N Sainath. 2019. Contextual speech recognition with difficult negative training examples. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6440–6444. IEEE.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Chung-Cheng Chiu, Anshuman Tripathi, Katherine Chou, Chris Co, Navdeep Jaitly, Diana Jaunzeikare, Anjuli Kannan, Patrick Nguyen, Hasim Sak, Ananth Sankar, et al. 2017. Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.

George E Dahl, Dong Yu, Li Deng, and Alex Acero. 2011. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, 20(1):30–42.

Nilaksh Das, Monica Sunkara, Dhanush Bekal, Duen Horng Chau, Sravan Bodapati, and Katrin Kirchhoff. 2022. Listen, know and spell: Knowledge-infused subword modeling for improving asr performance of oov named entities. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7887–7891. IEEE.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Jennifer Drexler Fox and Natalie Delworth. 2022. Improving contextual recognition of rare words with an alternate spelling prediction model. *arXiv preprint arXiv:2209.01250*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *Interspeech 2020*.

Minglun Han, Linhao Dong, Zhenlin Liang, Meng Cai, Shiyu Zhou, Zejun Ma, and Bo Xu. 2022. Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8532–8536. IEEE.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Anjuli Kannan, Yonghui Wu, Patrick Nguyen, Tara N Sainath, Zhijeng Chen, and Rohit Prabhavalkar. 2018. An analysis of incorporating an external language model into a sequence-to-sequence model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5828. IEEE.

Suyoun Kim and Florian Metze. 2018. Dialog-context aware end-to-end speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 434–440. IEEE.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent

subword tokenizer and detokenizer for neural text processing. *EMNLP 2018*, page 66.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, Yangyang Shi, Jay Mahadeokar, Julian Chan, Yuan Shangguan, Christian Fuegen, Ozlem Kalinli, et al. 2021a. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. *arXiv preprint arXiv:2104.02194*.

Duc Le, Gil Keren, Julian Chan, Jay Mahadeokar, Christian Fuegen, and Michael L Seltzer. 2021b. Deep shallow fusion for rnn-t personalization. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 251–257. IEEE.

Junzhe Liu, Jianwei Yu, and Xie Chen. 2023. Improved factorized neural transducer model for text-only domain adaptation. *arXiv preprint arXiv:2309.09524*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Assaf Hurwitz Michaely, Xuedong Zhang, Gabor Simko, Carolina Parada, and Petar Aleksic. 2017. Keyword spotting for google assistant using contextual speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 272–278. IEEE.

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

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*.

Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. 2017. A comparison of sequence-to-sequence models for speech recognition.

Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. Deep context: end-to-end contextual speech recognition. In *2018 IEEE spoken language technology workshop (SLT)*, pages 418–425. IEEE.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, et al. 2018. Espnet: End-to-end speech processing toolkit. *INTERSPEECH 2018, Hyderabad, India*.

Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. 2017. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Ryan Whetten, Mir Tahsin Imtiaz, and Casey Kennington. 2023. Evaluating automatic speech recognition in an incremental setting. *arXiv preprint arXiv:2302.12049*.

Ian Williams, Anjuli Kannan, Petar S Aleksic, David Rybach, and Tara N Sainath. 2018. Contextual speech recognition in end-to-end neural network systems using beam search. In *Interspeech*, pages 2227–2231.

Marliyya Zayyan. 2011. Objective structured clinical examination: the assessment of choice. *Oman medical journal*, 26(4):219.

Zhengyi Zhang and Pan Zhou. 2022. End-to-end contextual asr based on posterior distribution adaptation for hybrid ctc/attention system. *arXiv preprint arXiv:2202.09003*.

Ding Zhao, Tara N Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang. 2019. Shallow-fusion end-to-end contextual biasing. In *Interspeech*, pages 1418–1422.

Zihan Zhao, Yu Wang, and Yanfeng Wang. 2023. Knowledge-aware bayesian co-attention for multimodal emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.