Interspeech 2024
1-5 September 2024, Kos, Greece

Contextual Biasing with Confidence-based Homophone Detector for Mandarin End-to-End Speech Recognition

Chengxu Yang[1,2], *Lin Zheng*[1,2], *Sanli Tian*[1,2], *Gaofeng Cheng*[1], *Sujie Xiao*[1,2], *Ta Li*[1,2]

[1]Key Lab of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS, China
[2]University of Chinese Academy of Sciences, China

{yangchengxu,zhenglin,tiansanli,chenggaofeng,xiaosujie,lita}@hccl.ioa.ac.cn

Abstract

Deep biasing methods and shallow fusion methods have been demonstrated to improve the performance of end-to-end ASR effectively. However, accurate recognition often becomes challenging when specific words within the contextual phrases occur too infrequently in the training corpus or are out-of-vocabulary. To address this issue, we introduce a confidence-based homophone detector and syllable bias model to correct context phrases that may have been recognized incorrectly. The detector utilizes confidence distribution peaks resulting from homophone substitutions in ASR decoding outputs and employs their coefficient of variation for discrimination to avoid loss of general performance. Experiments on the biased word subset of Aishell-1 show that our proposed method obtains a 31.2% relative CER improvement over the baseline and a relative decrease of 52.0% for context phrases. When cascaded with the deep fusion and shallow fusion methods, the improvements become 13.7% and 33.5% respectively.

Index Terms: End-to-end Speech Recognition, Contextual Biasing, Confidence, Coefficient of Variation

1. Introduction

In recent years, with the continuous development of deep learning, end-to-end automatic speech recognition (ASR) methods have become mainstream in the field of speech recognition. Currently, many end-to-end ASR methods are integrated into specific tasks such as voice assistants, including connectionist temporal classification (CTC) [1, 2], recurrent neural network transducer (RNN-T) [3, 4], and attention-based encoder-decoder (AED) [5, 6, 7]. However, in practice, the phrases in the training data may not cover all the phrases appearing in the audio. Some rare proper nouns, specialized terms in specific domains, and uncommon characters are difficult to include in the training data. In specific application scenarios, the recognition accuracy of these words will receive additional attention from users. For example, in the scenario of voice assistants, ASR systems need to recognize names in the user's contact list or specific geographic locations accurately. Fine-tuning models with relevant data can be very unfriendly to user experience. Therefore, integrating contextual phrases into ASR systems is becoming increasingly important.

In previous studies, contextual biasing methods for end-to-end models have been broadly categorized into two types: shallow fusion-based contextual biasing[8, 9, 10, 11, 12], such as methods based on weighted finite state transducers (WFST) [8, 9, 10], and deep biasing based on neural attention-based contextual biasing (deep biasing) [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. In WFST-based methods[8], contextual phrases form an additional decoding graph. During the decoding process, when a specific contextual phrase is decoded, scores are added to the path, thereby making the recognition result more inclined towards that contextual phrase. However, its performance improvement on rare word biasing is limited and requires constant adjustments to determine the optimal fusion weights. In contrast, deep biasing methods incorporate an additional biasing module into end-to-end models to model contextual phrases. Compared to shallow fusion methods, deep biasing methods can adapt to contextual phrase lists in different scenarios and demonstrate superior performance. Some prior works have already modified the basic end-to-end speech recognition models to achieve contextualized models, such as CLAS[17], CATT[18], and CPPN[19].

The aforementioned deep biasing methods generally integrate contextual information implicitly within the bias model, making it challenging to control the degree of biasing for specific contexts. In contrast, shallow fusion typically modifies the posterior of the model output during decoding to achieve biasing explicitly. To combine the advantages of both approaches, a collaborative decoding method based on the continuous integrate-and-fire (CIF) is proposed in [15]. The CIF module extracts token-level embeddings from frame-level information, allowing for explicit control over the influence of contextual biasing. As a result, the encoding precision of both the context decoder and ASR decoder is highly similar, enabling collaborative decoding to obtain the final result. Extending this method to AED models, beyond just CIF models, was discussed in [25]. They introduced a spike-triggered deep contextual biasing method, filtering emitting frames from the encoder output of CTC posterior as inputs to the bias model, while also supporting explicit and implicit biasing in AED models.

Nevertheless, even methods combining deep and shallow fusion struggle to bias these characters effectively, which appear too infrequently in the training data or are out-of-vocabulary (OOV). Therefore, additional information is needed to address this issue. Sainath et al. [26] proposed a method called JOIST, which is a modality-matching text injection method. During training, it injects text data that is representative of contextually relevant phrases that will be seen during inference. Sudo et al. [27] introduced a method that does not require retraining and is based on phoneme similarity estimation. It extracts phonemes of possible named entities and compares them with the phonemes of the target named entity. If the similarity exceeds the threshold, it will be replaced. For Mandarin, homophones, which are always overlooked in end-to-end models, are more suitable to utilize.

This work is partially supported by the National Science and Technology Major Project (No. 2022ZD0116103), and the Goal-Oriented Project Independently Deployed by Institute of Acoustics, Chinese Academy of Sciences (No. MBDX202106)

10.21437/Interspeech.2024-869

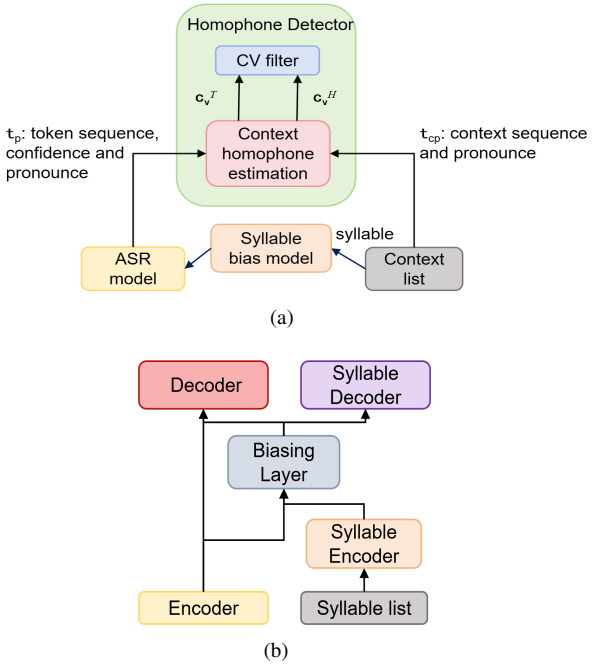

Figure 1: *Architecture of the proposed confidence-based homophone detector: (a) The input and output of the detector (b) Basic structure of syllable bias model.*

To address the aforementioned issues, we propose a confidence-based homophone detection method. Firstly, We propose a novel bias model based on syllable-to-syllable mapping. And we obtain decoding output results through the ASR model and syllable bias model. We assess whether a certain phrase is in the contextual list based on character-level pronunciation similarity. Then, we calculate the coefficient of variation using the posterior distribution of confidence for this potential contextual phrase. If it exceeds the threshold, it indicates a high likelihood of homophone substitution in the contextual phrase. At this point, correcting with phrases from the context list can improve the recognition accuracy of rare and OOV characters. Furthermore, for a larger context list, performing homophone detection at this stage can be time-consuming. Inspired by [19, 23], the context list of the biasing model undergoes two-stage filtering to ensure only relevant contextual phrases are used for biasing, reducing the computational burden and inference time. Therefore, we apply the same filtering method to homophone detection.

2. Preliminary

This section briefly introduces why the confidence distribution of decoding results can serve as a basis for homophone correction. In the results obtained through decoding by the ASR model, errors involving homophone substitutions in the contextual phrase section are quite common. This is mainly because the replaced characters appear too infrequently or are absent altogether in the training data, making it difficult for the biasing model to learn their features. As a result, the biasing model tends to output other homophones. Given this situation, it is reasonable to infer that the confidence of tokens in the contextual phrase results is relatively low due to the rarity of character combinations. To enhance the utilization of homophones in contextual biasing by end-to-end speech recognition models, we introduce phonetic information of contextual phrases. It allows us to replace potentially misrecognized contextual phrases in the decoding results based on the distribution of their confidence levels, thereby improving the recognition accuracy of the contextual phrases.

Typically, in the recognized results, the confidence of correct characters is higher, while the confidence of incorrect characters tends to be lower, resulting in a peaked distribution of confidence levels. The confidence distribution reflects the system's certainty about the correctness of its recognition. By analyzing this distribution, we can identify instances where the system is less confident, indicating potential errors or ambiguous interpretations. Homophones often result in lower confidence scores due to the increased uncertainty in selecting the correct character. To describe this situation better, we use the coefficient of variation to ascertain whether there might be contextual phrase substitutions. The coefficient of variation (CV) is a statistical measure of the dispersion of data points around the mean in a data sequence. It represents the ratio of the standard σ deviation to the mean μ ,

$$c_v = \frac{\sigma}{\mu} \tag{1}$$

and is a useful statistic for comparing the variability between one data sequence and another. In the contextual phrase sections, it is often the case that a small number of characters are mistakenly recognized as homophones, leading to lower confidence scores for these characters compared to others. As a result, this section tends to have a higher CV.

To identify homophone substitution errors in a sentence, we can compare the CV of confidence distribution for tokens in the contextual phrase section with a predetermined threshold. If the CV for this section is larger than the CV for other parts of the sentence, it indicates that the distribution pattern aligns with such errors. In such cases, it is reasonable to replace the erroneous contextual phrases with the correct ones.

3. Proposed method

3.1. Context homophone estimation and CV filter

The overall architecture of the proposed method is depicted in Figure 1. Figure 1(a) illustrates the input and output of the confidence-based homophone detector, where the input consists of a sequence decoded by a conventional ASR model, the confidence for each token, and the sequence of phrases in the context list. The ASR model can be either augmented with a bias module or not. The input sequence $\mathbf{t}_p$ and $\mathbf{t}_{cp}$ can be represented as follows:

$$\mathbf{t}_p = \{\{y_1, \ldots, y_N\}, \{c_1, \ldots, c_N\}\} \tag{2}$$

$$\mathbf{t}_{cp} = \{z_1, \ldots, z_K\} \tag{3}$$

where $y_i (i = 1, \ldots, N)$ represents each token in the output sequence of ASR model, $c_i (i = 1, \ldots, N)$ represents the confidence of the corresponding token in the output sequence of ASR model, and $z_k (k = 1, \ldots, K)$ represents the k-th phrase in the context list. It should be pointed out that c_i are exp(log softmax). Then, in the context homophone estimation, the token sequences corresponding to the pinyin in $\mathbf{t}_p$ and $\mathbf{t}_{cp}$ will be extracted separately (lines 2-5 in Algorithm 1). Subsequently, the pronunciation of each phrase in the context list will be compared with the pronunciation of the current sequence (lines 6-7

in Algorithm 1). The comparison method involves calculating the degree of phonetic matching, using the formula described as follows:

$$s_n = \alpha(1 - \frac{M}{max(|\mathbf{t}_{cp_k}|, |\mathbf{t}_{p_n}|)}) \tag{4}$$

where M represents the edit distance, $|\mathbf{t}_{cp_k}|$ and $|\mathbf{t}_{p_n}|$ represent the lengths of the two strings respectively. We denote a scaling factor α specific to Mandarin, as

$$\alpha = \begin{cases} \alpha_{high}, & \text{if } only\ different\ tune \\ \alpha_{low}, & \text{if } different\ phoneme\ \&\ tune \end{cases} \tag{5}$$

Where $\alpha_{high} > \alpha_{low}$. Due to the existence of tones as a special variable in Mandarin, we need to include it in the comparison scope. Errors caused by tone mismatches in the output of the ASR model are quite common and reasonable. Therefore, we use α_{high} to reward cases where only tone mismatches exist. For cases where other mismatches occur, we use α_{low}. When $s_n = 1$, it means the pronunciation of the two sequences is identical, while $s_n = 0$ indicates they are completely different. We introduce a similarity threshold to make it more robust. When s_n exceeds the set similarity threshold, we consider this part of the tokens as candidate contexts.

After obtaining the candidate contexts filtered by homophones, the coefficient of variation $\mathbf{c}_v^H$ for their corresponding confidence distribution, as mentioned in Section 2, is calculated. And then a final decision is made through the CV filter (lines 8-10 in Algorithm 1). Here, CV_{th} can be represented as the maximum acceptable level of dispersion set by a predetermined threshold. More reasonably, the coefficient of variation $\mathbf{c}_v^T$ for the entire sequence's confidence distribution can be computed, representing the fluctuation degree of the candidate part compared to the entire sequence. If it is greater, it further indicates that the candidate context is the target context. Then, the phrase in the sequence $\mathbf{t}_p$ corresponding to the position is replaced with the target context phrase $\mathbf{t}_{cp}$, and the maximum acceptable similarity threshold for this match is updated (line 11 in Algorithm 1). Introducing SIM_{th} and CV_{th} effectively prevents non-context phrases from being replaced with context phrases.

Algorithm 1 Context homophone estimation and CV filter

1: $s_{max} = SIM_{th}$;
2: **for** $k = 1, ..., K$ **do**
3: $\quad \mathbf{t}_{cp_k} = ExtractPinyin(z_k)$;
4: $\quad$ **for** $n = 1, ..., N$ **do**
5: $\quad\quad \mathbf{t}_{p_n} = ExtractPinyin(y_n)$;
6: $\quad\quad s_n = CompareProunce(\mathbf{t}_{cp_k}, \mathbf{t}_{p_n})$;
7: $\quad\quad$ **if** $s_n > s_{max}$ **then**
8: $\quad\quad\quad \mathbf{c}_v^H = CalConfidenceCV(\mathbf{t}_{p_n})$;
9: $\quad\quad\quad$ **if** $\mathbf{c}_v^H > CV_{th}$ **then**
10: $\quad\quad\quad\quad \mathbf{t}_{new} = ReplaceWord(\mathbf{t}_p, \mathbf{t}_{cp})$;
11: $\quad\quad\quad\quad s_{max} = s_n$;
12: $\quad\quad\quad$ **end if**
13: $\quad\quad$ **end if**
14: $\quad$ **end for**
15: **end for**

3.2. Syllable bias model

In Mandarin, each character corresponds to one syllable, which inspired us to replace characters with syllables. We then calculate the syllable CTC loss during the training of the bias model, considering the same length of syllables as in the original context. Figure 1(b) illustrates the core structure of the syllable bias model, whose basic architecture closely resembles that of the deep bias model in CPPN [19]. The key difference lies in transforming the context encoder and decoder into syllable encoder and decoder. This operation enables the bias module to generate outputs that are more inclined towards homophones of the labels, rather than strictly adhering to the correct labels, thus enhancing the model's learning capability. The syllable bias model primarily aims to induce errors in ASR output that lean towards phonetically similar substitutions, thereby making it more conducive for the methods described in Section 3.1.

4. Experiments

4.1. Data

Our experiments were conducted on the publicly available Mandarin dataset based on Aishell-1[28], which comprises recordings totaling 178 hours and spans 11 domains, including smart home, autonomous driving, and industrial production. To test the performance improvement of the proposed method on bias lists specifically, we selected a subset of the Aishell-1 dataset referred to as *Test-Aishell1-Small*, which was used in [29] for evaluating the effectiveness of the Paraformer-large model on hotwords. Additionally, to demonstrate the enhancement of our method for OOV words, we followed the filtering method outlined in [24] to construct two additional test sets, *Test-Aishell1-Middle* and *Test-Aishell1-Large*, along with their respective hotword lists, from the open-source Aishell-1 dataset [28]. The final test sets encompassing various domains, are presented in Table 1.

Table 1: *Test sets for ASR*

	#utt	**#hotword**	**#OOV**
Test-Gerneral	7176	-	-
Test-Aishell1-Small	235	187	9
Test-Aishell1-Middle	808	400	10
Test-Aishell1-Large	1334	600	15

4.2. Experimental setup

We conducted experiments based on Wenet[30], an open-source automatic speech recognition framework. The baseline model selected is a pre-trained checkpoint of Wenet's conformer architecture on the Aishell-1 dataset, which includes 12 conformer layers in the encoder and 6 bi-transformer layers in the decoder, both with 256-dimensional inputs and 4 self-attention heads.[1]

We use a WFST-based decoding graph method as the shallow fusion baseline, where the context graph score is set to 6.0. Additionally, we employ the contextual phrase prediction network[19] as our baseline for deep biasing, comprising a context encoder, biasing layer, and context decoder. The context encoder consists of a BLSTM layer followed by a linear layer. The biasing layer includes 4-head attention layers and another linear layer. The context decoder contains a linear layer that maps the input dimension to the size of the vocabulary. The CTC loss weight is set to 0.3, the bias loss weight is set to 0.03, and the deep biasing score is set to 2.0. Furthermore, its

[1] https://github.com/wenet-e2e/wenet/blob/main/docs/pretrained_models.md

Table 2: *Experiment results with different methods and test sets*

	Test-Aishell1-Small	**Test-Aishell1-Middle**	**Test-Aishell1-Large**
	Total Set CER/B-CER/U-CER (%) // All Biased Words Recall Rate/Precision/F1-Score (%)		
Baseline	13.9/42.0/5.98 26.9/98.7/42.3	9.67/25.2/5.48 50.7/98.1/66.9	8.28/22.3/4.66 58.6/**99.1**/73.7
SF	10.9/28.5/6.01 53.3/97.5/68.9	7.79/16.4/5.46 71.8/97.6/82.7	6.50/13.5/4.69 77.3/98.3/86.6
CPPN	13.0/38.9/5.78 31.4/96.8/47.4	8.67/21.6/**5.17** 58.2/**99.3**/73.4	7.58/18.7/4.72 64.0/99.0/77.7
CBCB	**8.95/20.3/5.78 66.2/99.5/79.5**	**6.91/12.3**/5.46 **78.2**/97.5/**86.8**	**5.81/10.5/4.60 82.5**/97.6/**89.4**
SF + CPPN	9.18/20.8/5.94 66.2/96.0/78.4	7.27/14.5/**5.32** 76.4/**97.7**/85.8	6.25/11.8/4.80 80.8/**97.5**/88.3
SF + CBCB	**8.00/15.5/5.91 77.0/97.8/86.2**	**6.46/9.82**/5.56 **85.5**/95.7/**90.3**	**5.19/6.77/4.78 90.4**/95.6/**92.9**

training process is consistent in [19]. We named our proposed confidence-based context biasing method CBCB, which incorporates context homophone estimation, CV filter and syllable bias model. Unlike CPPN, in the syllable bias model, the vocabulary will be replaced with a Mandarin syllable table. The special parameters α_{high} and α_{low} for Mandarin, described in Section 3.1 were set to 0.9 and 0.75 respectively. And the similarity threshold SIM_{th} was set to 0.7. Moreover, we chose $\mathbf{c}_v^T$ as CV_{th} for general cases.

In addition to CER, we also employ a biased character error rate (B-CER) and an unbiased character error rate (U-CER) for evaluation[19]. B-CER measures the error rate for characters that exist in the bias list, while U-CER assesses the error rate for characters not in the bias list. Furthermore, we will consider the recall, precision, and F1 score (R/P/F) averaged over all biased words.

4.3. Results

The experimental results obtained by using different methods, varying sizes of test sets, and different sizes of context lists are presented in Table 2. When using a single biasing method, CBCB achieves the best performance across all three test sets. Compared to the baseline, CBCB results in an average relative reduction of 31.2% in CER, a 52.0% reduction in B-CER, an increase in average recall of biased words from 45.4% to 75.6%, and an increase in F1 score from 61.0% to 85.2%. These results suggest that homophone substitution in the context contributes significantly to overall errors in standard ASR models, and thus, detecting and correcting homophones can lead to improved performance.

The experimental results of the cascaded methods are also listed in Table 2. We cascade the shallow fusion method with CPPN[19] method as the baseline and introduce our proposed CBCB method additionally. The results show that our proposed method still achieves a 13.7% average relative reduction in CER and a 33.5% average relative reduction in B-CER almost without increasing U-CER. Additionally, there are improvements in average recall and F1-score of biased words. We believe this is due to the introduction of homophone information by the CBCB method, which corrects some errors introduced by the shallow fusion and deep biasing methods. The slight increase in U-CER may be attributed to lower confidence in some non-contextual content that shares pronunciation with contextual phrases, leading to false positives. However, this content can be ignored compared to the majority of biased words that have been recalled.

The influence of the CV threshold of confidence was also examined on the *Test-Aishell1-Large* test set, as shown in Table 3. The first column indicates whether our proposed method is cascaded with shallow fusion and CPPN[19].$\mathbf{c}_v^T$ always shows the best performance, indicating that the CV filter needs to consider the relationship between the CV of confidence between the contextual part and the whole sentence. Therefore, $\mathbf{c}_v^T$ should be chosen as CV_{th} in general.

Table 3: *Effect of* CV_{th}

Cas.	CV_{th}	**CER/B-CER/U-CER**	**R/P/F**
N	0	**5.90/10.6**/4.69	**86.6**/93.3/**89.8**
	0.2	6.10/11.6/4.67	80.0/97.7/87.9
	0.5	6.95/15.8/**4.66**	70.3/**98.1**/81.9
	$\mathbf{c}_v^T$	5.97/11.0/4.67	81.1/97.9/88.7
Y	0	5.54/8.18/4.85	**91.2**/92.6/91.9
	0.2	5.35/7.42/4.82	88.9/96.5/92.6
	0.5	5.62/8.74/**4.82**	85.9/**97.0**/91.1
	$\mathbf{c}_v^T$	**5.30/7.14**/4.83	89.3/96.5/**92.8**

Table 4: *Results of ablation study*

Method	**CER**	**B-CER**	**U-CER**
CBCB	**5.19**	**6.77**	**4.78**
- homophone detector	6.02	10.84	**4.78**
- syllable bias	5.30	7.14	4.83

4.4. Ablation study

Table 4 presents the results of ablation study. The removal of syllable bias refers to the utilization of a general bias model. It can be seen that the homophone detector significantly reduces B-CER, while syllable bias further enhances this improvement and also causes a decrease in U-CER.

5. Conclusion

In this paper, we present a contextual biasing method with a confidence-based homophone detector, composed of two modules: context homophone estimation and CV filter. And we have transformed the mapping relationship within the deep bias module from character-to-character to syllable-to-syllable. Compared to previous works, our approach achieves better effects in the Mandarin dataset. In the future, we will explore the generality of our method in other languages.

6. References

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

[2] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[3] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[4] Y. Zhang, S. Sun, and L. Ma, "Tiny transducer: A highly-efficient speech recognition model on edge devices," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6024–6028.

[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[6] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., 2014, pp. 1724–1734.

[7] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[8] 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 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6186–6190.

[9] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and R. Pang, "Shallow-fusion end-to-end contextual biasing." in *Interspeech*, 2019, pp. 1418–1422.

[10] 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*, 2020, pp. 4348–4351.

[11] I. Williams, A. Kannan, P. S. Aleksic, D. Rybach, and T. N. Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search." in *Interspeech*, 2018, pp. 2227–2231.

[12] S. Kim, Y. Shangguan, J. Mahadeokar, A. Bruguier, C. Fuegen, M. L. Seltzer, and D. Le, "Improved neural language model fusion for streaming recurrent neural network transducer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7333–7337.

[13] Z. Chen, M. Jain, Y. Wang, M. L. Seltzer, and C. Fuegen, "Joint grapheme and phoneme embeddings for contextual end-to-end asr." in *Interspeech*, 2019, pp. 3490–3494.

[14] D. Le, M. Jain, G. Keren, S. Kim, Y. Shi, J. Mahadeokar, J. Chan, Y. Shangguan, C. Fuegen, O. Kalinli, Y. Saraf, and M. L. Seltzer, "Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion," in *Proc. Interspeech 2021*, 2021, pp. 1772–1776.

[15] M. Han, L. Dong, S. Zhou, and B. Xu, "Cif-based collaborative decoding for end-to-end contextual speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6528–6532.

[16] S. Dingliwal, M. Sunkara, S. Ronanki, J. Farris, K. Kirchhoff, and S. Bodapati, "Personalization of ctc speech recognition models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 302–309.

[17] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: end-to-end contextual speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 418–425.

[18] F.-J. Chang, J. Liu, M. Radfar, A. Mouchtaris, M. Omologo, A. Rastrow, and S. Kunzmann, "Context-aware transformer transducer for speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 503–510.

[19] K. Huang, A. Zhang, Z. Yang, P. Guo, B. Mu, T. Xu, and L. Xie, "Contextualized end-to-end speech recognition with contextual phrase prediction network," in *Annual Conference of the International Speech Communication Association, INTERSPEECH 2023*, 2023, pp. 4933–4937.

[20] Y. Xu, B. Liu, Q. Huang, X. Song, Z. Wu, S. Kang, and H. Meng, "Cb-conformer: Contextual biasing conformer for biased word recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] T. Xu, Z. Yang, K. Huang, P. Guo, A. Zhang, B. Li, C. Chen, C. Li, and L. Xie, "Adaptive contextual biasing for transducer based streaming speech recognition," in *Annual Conference of the International Speech Communication Association, INTERSPEECH 2023*, 2023, pp. 1668–1672.

[22] X. Fu, K. M. Sathyendra, A. Gandhe, J. Liu, G. P. Strimel, R. McGowan, and A. Mouchtaris, "Robust acoustic and semantic contextual biasing in neural transducers for speech recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] Z. Yang, S. Sun, X. Wang, Y. Zhang, L. Ma, and L. Xie, "Two stage contextual word filtering for context bias in unified streaming and non-streaming transducer," in *Annual Conference of the International Speech Communication Association, INTERSPEECH 2023*, 2023, pp. 3257–3261.

[24] X. Shi, Y. Yang, Z. Li, and S. Zhang, "Seaco-paraformer: A non-autoregressive asr system with flexible and effective hotword customization ability," *arXiv preprint arXiv:2308.03266*, 2023.

[25] K. Huang, A. Zhang, B. Zhang, T. Xu, X. Song, and L. Xie, "Spike-triggered contextual biasing for end-to-end mandarin speech recognition," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[26] T. N. Sainath, R. Prabhavalkar, D. Caseiro, P. Rondon, and C. Allauzen, "Improving contextual biasing with text injection," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[27] Y. Sudo, K. Hata, and K. Nakadai, "Retraining-free Customized ASR for Enharmonic Words Based on a Named-Entity-Aware Model and Phoneme Similarity Estimation," in *Proc. INTERSPEECH 2023*, 2023, pp. 491–495.

[28] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[29] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and Accurate Parallel Transformer for Non-autoregressive End-to-End Speech Recognition," in *Proc. Interspeech 2022*, 2022, pp. 2063–2067.

[30] B. Zhang, D. Wu, Z. Peng, X. Song, Z. Yao, H. Lv, L. Xie, C. Yang, F. Pan, and J. Niu, "Wenet 2.0: More productive end-to-end speech recognition toolkit," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association*, 2022, pp. 1661–1665.