Enhancing the Robustness of Contextual ASR to Varying Biasing Information Volumes Through Purified Semantic Correlation Joint Modeling

Yue Gu, Zhihao Du, Ying Shi, Shiliang Zhang, Qian Chen, and Jiqing Han, *Member, IEEE*

Abstract—Recently, cross-attention-based contextual automatic speech recognition (ASR) models have made notable advancements in recognizing personalized biasing phrases. However, the effectiveness of cross-attention is affected by variations in biasing information volume, especially when the length of the biasing list increases significantly. We find that, regardless of the length of the biasing list, only a limited amount of biasing information is most relevant to a specific ASR intermediate representation. Therefore, by identifying and integrating the most relevant biasing information rather than the entire biasing list, we can alleviate the effects of variations in biasing information volume for contextual ASR. To this end, we propose a purified semantic correlation joint modeling (PSC-Joint) approach. In PSC-Joint, we define and calculate three semantic correlations between the ASR intermediate representations and biasing information from coarse to fine: list-level, phrase-level, and token-level. Then, the three correlations are jointly modeled to produce their intersection, so that the most relevant biasing information across various granularities is highlighted and integrated for contextual recognition. In addition, to reduce the computational cost introduced by the joint modeling of three semantic correlations, we also propose a purification mechanism based on a grouped-and-competitive strategy to filter out irrelevant biasing phrases. Compared with baselines, our PSC-Joint approach achieves average relative F1 score improvements of up to 21.34% on AISHELL-1 and 28.46% on KeSpeech, across biasing lists of varying lengths.

Index Terms—semantic correlation joint modeling, group competitive purification, varying-length biasing lists, contextual ASR.

I. INTRODUCTION

IN recent years, remarkable advancements have been made on end-to-end automatic speech recognition (E2E ASR), such as connectionist temporal classification [1], recurrent neural network transducer [2], [3], and attention encoder-decoder [4]–[7]. E2E ASR models have outperformed traditional hybrid speech recognition systems [8]. Despite these advancements, the performance of E2E ASR models degrades significantly when recognizing personalized contexts such as the names of contacts, songs, and locations. To alleviate the degradation, contextual ASR has been introduced, which incorporates flexible context and improves the recognition performance on biasing phrases. In contextual ASR, personalized phrases are typically considered as biasing information and are introduced into the model as a biasing list.

Contextual speech recognition approaches can be broadly divided into four categories: shallow fusion approaches, prompt-based contextual approaches, trie-based deep biasing techniques, and attention-based deep context approaches. Shallow fusion approaches [9]–[18] integrate biasing information through a weighted finite state transducer or language model during the inference stage, going against the benefits derived from the joint optimization in sequence-to-sequence models. Prompt-based contextual approaches leverage the acoustic-textual alignment capability of large-scale multimodal speech models by treating the biasing phrase as the prompt, improving the recognition of biasing phrases [19], [20]. Trie-based deep biasing approaches [21]–[26] focus on matching the prefix tree of biasing phrases using recognized partial tokens, where only some word pieces of the biasing phrase are considered. In contrast, attention-based deep context approaches [27]–[49] encode all word pieces or tokens of a biasing phrase into an embedding vector. In such approaches, the cross-attention module uses ASR intermediate representations as queries to attend to all biasing phrases simultaneously. Due to their simplicity and flexibility, cross-attention-based contextual ASR models have attracted significant attention.

In cross-attention-based approaches, the correlations between ASR intermediate representations and biasing phrases are learned by the cross-attention module, which identifies the relevant biasing information from the given biasing list. However, in real-world applications, the effectiveness of the attention module is affected by variations of biasing information volumes [33], [49], particularly when the number of biasing phrases is large, introducing irrelevant biasing information to the contextual ASR. Although injecting distractors into the biasing list during training allows the cross-attention module to handle a longer biasing list, it also introduces more confusion between the target phrase and its similar ones when the number of injected distractors is large [27], [33]. Furthermore, injecting more distractors improves performance on larger biasing lists but degrades performance on smaller lists [21], indicating that it cannot enhance the robustness of contextual ASR models to variations in biasing list length.

We find that, no matter how long the biasing list is, only a limited amount of biasing information is relevant to a specific ASR intermediate representation. Therefore, we can mitigate the impact of biasing information volume variations for contextual ASR by identifying and integrating the most relevant biasing information rather than the entire biasing list. To this end, we propose a novel purified semantic correlation joint modeling (PSC-Joint) approach, in which three semantic correlations are defined and calculated from coarse to fine: list-level, phrase-level, and token-level. Then, based on the cross-granularity consistency of the most relevant biasing information, the three types of semantic correlations are jointly
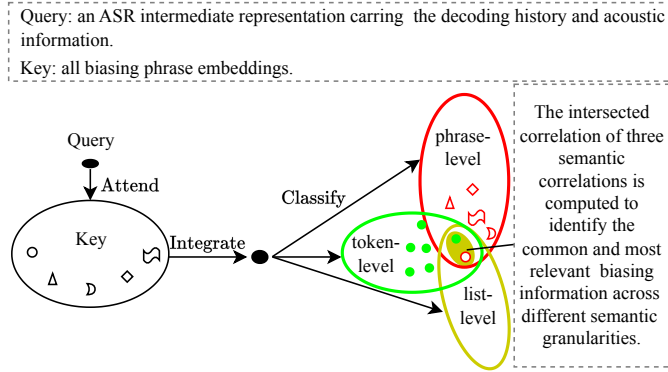
Fig. 1. The schematic diagram of semantic correlation joint modeling method.

modeled to highlight and integrate this information through their intersection. In addition, during the inference stage, a divide-and-conquer-based purification mechanism is also proposed to filter out the irrelevant phrases and refine the biasing list, thereby enhancing the computational efficiency of the multi-level semantic correlation joint modeling.

Fig. 1 shows more details about the semantic correlation joint modeling method, in which an ASR intermediate representation and all biasing phrases serve as "Query" and "Key", respectively. For the list-level correlation, the biasing list is treated as a single entity, and the correlations between "Query" and "Key" are modeled as a binary prediction task, representing the correlation score of the "Query" being relevant to the entire biasing list. Similarly, the phrase-level correlation is modeled as a phrase classification task by treating each biasing phrase as an independent entity, representing the correlation score of the "Query" belonging to a specific biasing-phrase category. For the token-level correlation, each token in the biasing list is considered as an entity. The correlations are modeled as a token classification task, representing the correlation score of the "Query" belonging to a specific ASR vocabulary token. Obviously, the most relevant biasing information are consistent across three types of semantic correlations when the "Query" is relevant to any "Key" and it can be identified by the intersected correlations. Thus, we compute the joint score of three semantic correlations and the containment relationship between phrases and tokens to produce the intersected correlations, in which the most relevant information is spotlighted. As a result, the proposed semantic correlation joint modeling method can significantly enhance the robustness of contextual ASR against changes in the biasing information volume.

To reduce the computational cost introduced by the joint modeling of three semantic correlations, we further propose a divide-and-conquer-based biasing information purification mechanism, which purifies biasing phrases in a grouped and competitive manner before semantic correlation joint modeling. Experiments conducted on the AISHELL-1, KeSpeech, and AISHELL-NER datasets show that PSC-Joint significantly improves the F1 score and reduces character error rate across varying-length biasing lists, outperforming the baseline models.

II. RELATED WORK

Various attention-based deep context approaches for autoregressive (AR) ASR and non-autoregressive (NAR) ASR [50]–[54] are developed recently. During inference, the AR decoder generates tokens one by one since each token is conditioned on all previous tokens. In contrast, the NAR decoder generates output sequences iteratively or simultaneously by removing temporal dependency [55]. The primary distinction between AR and NAR decoding is the time consumption of token generation, thus most attention-based approaches can be expanded for both AR and NAR ASR models in theory. The continuous Integrate-and-Fire (CIF) [51] based NAR ASR model is often employed as the ASR backbone considering its faster inference speed and comparable recognition performance.

Recently, several attention-based deep context algorithms have been developed for CIF-based ASR. Within the CIF-based collaborative decoding (ColDec) framework for contextual ASR [30], an attention-based context processing network (CPN) is employed to produce the context detection results, in which the CIF outputs are used as the queries and the biasing phrases are regarded as the keys and values. The target of CPN consists only of the biasing phrase and the "<no-bias>" token. For instance, the target of the CPN is set to "# peace_ dove_ # #", whereas the ASR backbone target is "the_ peace_ dove_ symbolizes_ peace_", where "peace_ dove_" represent a biasing phrase and "#" denotes the special "<no-bias>" token. At each output step in inference, the ASR decoder and CPN decoder conduct collaborative decoding with interpolated log-probability. To alleviate the token-level prediction uncertainty caused by the confusion between similar context-specific phrases, the fine-grained contextual knowledge selection (FineCoS) method employs a phrase selection (PS) strategy to narrow down phrase candidates and applies token-level attention to the selected phrases [33]. FineCoS absorbs the thought of collaborative decoding and further improves the ability to recognize the given biasing phrases. Similar to FineCoS, semantic-augmented contextual paraformer (SeACo) [49] also employs the cross-attention module and the collaborative decoding method to produce the contextualized results. In SeACo, not only the CIF outputs but also the ASR backbone recognition outputs are fed into the cross-attention module in which the context detection outputs of the cross-attention module are semantic-augmented. SeACo utilizes an attention score filtering (ASF) strategy to discard irrelevant biasing phrases, by summing the attention scores among the sentence for each biasing phrase and picking the most active candidates based on these scores for the second-pass decoding. Notably, the PS and ASF strategies filter out the irrelevant biasing phrases according to the attention score, which may demonstrate limited effectiveness as the length of biasing lists increases significantly.

III. PURIFIED SEMANTIC CORRELATION JOINT MODELING APPROACH

A. Overview and Symbol Definition

Fig. 2(a) illustrates that attention-based contextual NAR ASR models, such as ColDec [30], typically consist of four com-
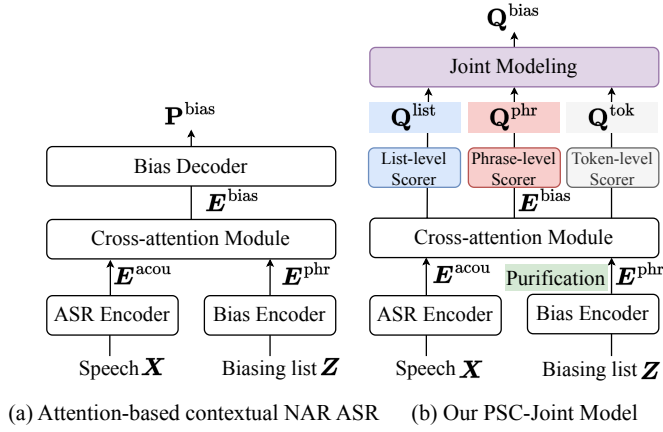
Fig. 2. Schematic diagram of (a) current attention-based contextual NAR ASR model and (b) our PSC-joint Model.

ponents to generate transcriptions biased toward the provided phrases. First, an ASR encoder encodes a sequence of T speech features $\boldsymbol{X} = \{\boldsymbol{x}_t \mid t = 1, \cdots, T\}$ into acoustic hidden representations $\boldsymbol{E}^{\mathrm{acou}} = \{\boldsymbol{e}_u^{\mathrm{acou}} \mid u = 1, \cdots, U\}$. Second, a biasing encoder transforms the biasing phrase list $\boldsymbol{Z} = \{\boldsymbol{z}_m \mid m = 1, \cdots, M\}$ into phrase embeddings $\boldsymbol{E}^{\mathrm{phr}} = \{\boldsymbol{e}_m^{\mathrm{phr}} \mid m = 1, \cdots, M\}$. U and M denote the transcription length of an utterance and the length of the biasing list, respectively. Third, a cross-attention module integrates all biasing phrases $\boldsymbol{E}^{\mathrm{phr}}$ to produce biased hidden representations $\boldsymbol{E}^{\mathrm{bias}} = \{\boldsymbol{e}_u^{\mathrm{bias}} \mid u = 1, \cdots, U\}$, according to the correlations between $\boldsymbol{E}^{\mathrm{acou}}$ and $\boldsymbol{E}^{\mathrm{phr}}$. Finally, a bias decoder transform the biased hidden representations $\boldsymbol{E}^{\mathrm{bias}}$ into the token-level probability distributions $\mathbf{P}^{\mathrm{bias}} = \{\mathbf{P}_u^{\mathrm{bias}} \in \mathbb{R}^V \mid u = 1, \cdots, U\}$ over a vocabulary of V tokens, yielding the biased output. Additionally, a vanilla ASR model is also involved as backbone, whose decoder transforms $\boldsymbol{E}^{\mathrm{acou}}$ into token probabilities $\mathbf{P}^{\mathrm{bb}} = \{\mathbf{P}_u^{\mathrm{bb}} \in \mathbb{R}^V \mid u = 1, \cdots, U\}$ over the same vocabulary. Then, $\mathbf{P}^{\mathrm{bb}}$ is interpolated with biased output $\mathbf{P}^{\mathrm{bias}}$ to yield the final recognition results.

At each decoding step u, the correlation $\boldsymbol{corr}_u$ between the acoustic hidden representation $\boldsymbol{e}_u^{\mathrm{acou}}$ and the set of phrase embeddings $\boldsymbol{E}^{\mathrm{phr}}$ is quantified by the attention scores that are typically computed using a scaled dot-product attention mechanism, formulated as $corr_{u,m} = \frac{1}{\sqrt{d}} < \boldsymbol{e}_u^{\mathrm{acou}}, \boldsymbol{e}_m^{\mathrm{phr}} >$. This correlation is then used to perform a weighted sum over embeddings in $\boldsymbol{E}^{\mathrm{phr}}$, resulting in a biased representation $\boldsymbol{e}_u^{\mathrm{bias}}$.

$$e_u^{\mathrm{bias}} = \sum_{m=1}^{M} corr_{u,m} e_m^{\mathrm{phr}} \qquad (1)$$

Subsequently, $\boldsymbol{e}_u^{\mathrm{bias}}$ is passed through the bias decoder to produce probabilities $\mathbf{P}_u^{\mathrm{bias}}$ of the biased output. Since the acoustic hidden representation $\boldsymbol{e}_u^{\mathrm{acou}}$ can be mapped to V vocabulary tokens, $\boldsymbol{corr}_u$ reflects the correlations between the M biasing phrases and the V vocabulary tokens, residing in a space of $\mathbb{R}^{M \times V}$. However, during inference, the dimension of the correlation space can fluctuate significantly with the varying list length M, which may impair the integration

of biasing phrases (Eq. (1)) and introduce irrelevant biasing information into the contextual ASR model.

By defining semantic correlations at multiple granularities and computing their intersection, the most relevant biasing information can be effectively identified. As this information is inherently limited regardless of the list length, leveraging it helps mitigate the impact of list length variations in contextual ASR models. Thus, we propose a semantic correlation joint modeling (SC-Joint) method, as illustrated in Fig.2(b). Specifically, three semantic correlation scores at various granularities, i.e., list-level $\mathbf{Q}^{\mathrm{list}} = \{\mathrm{Q}_u^{\mathrm{list}} \in [0,1] \mid u = 1, \cdots, U\}$, phrase-level $\mathbf{Q}^{\mathrm{phr}} = \{\mathbf{Q}_u^{\mathrm{phr}} \in \mathbb{R}^M \mid u = 1, \cdots, U\}$, and token-level $\mathbf{Q}^{\mathrm{tok}} = \{\mathbf{Q}_u^{\mathrm{tok}} \in \mathbb{R}^V \mid u = 1, \cdots, U\}$, are produced by three scorers, respectively. These scorers are trained by list-level, phrase-level, and token-level classification losses, respectively. Then, these correlations are jointly modeled by computing their intersection $\mathbf{Q}^{\mathrm{bias}} = \{\mathbf{Q}_u^{\mathrm{bias}} \in \mathbb{R}^V \mid u = 1, \cdots, U\}$, in order to identify the most relevant biasing information consistent across three levels:

$$\mathbf{Q}_u^{\mathrm{bias}} = \mathbb{N}(\mathrm{Q}_u^{\mathrm{list}}(\mathbf{Q}_u^{\mathrm{phr}})^\top \boldsymbol{\Phi} \mathbf{Q}_u^{\mathrm{tok}}) \qquad (2)$$

where $\mathbb{N}$ denotes a normalization function that ensures V sub-scores are positive and sum to 1, such as the softmax function. Note that the joint modeling takes into account the inclusion relationship between V ASR vocabulary tokens and M biasing phrases, represented by $\boldsymbol{\Phi} \in \mathbb{R}^{M \times V}$.

To reduce the computational cost of SC-Joint, we further propose a divide-and-conquer-based biasing information purification mechanism, which is performed before SC-Joint, composing the PSC-Joint approach. In each round of purification, the biasing phrase embedding $\boldsymbol{E}^{\mathrm{phr}}$ is divided into G groups, then the irrelevant phrases in each group $\boldsymbol{E}_g^{\mathrm{phr}}$ are competitively filtered out in parallel based on the list-level $\mathbf{Q}_g^{\mathrm{list}}$ and phrase-level $\mathbf{Q}_g^{\mathrm{phr}}$ correlations:

$$\boldsymbol{E}^{\mathrm{phr}} \in \mathbb{R}^M \xrightarrow{\mathrm{Group}} \{\boldsymbol{E}_g^{\mathrm{phr}} \in \mathbb{R}^{group\_size} \mid g = 1, \cdots, G\} \quad (3)$$

$$\left\{ \left(\boldsymbol{E}_g^{\mathrm{phr}}, \boldsymbol{E}^{\mathrm{acou}} \right) \xrightarrow{\mathrm{Attend}} \boldsymbol{E}_g^{\mathrm{bias}} \xrightarrow{\mathrm{Score}} \left(\mathbf{Q}_g^{\mathrm{list}}, \mathbf{Q}_g^{\mathrm{phr}} \right) \right\}_{g=1}^{G}$$
$$\xrightarrow{\mathrm{Topk\ (group\text{-}wise)\ and\ Merge}} \boldsymbol{E}^{\mathrm{phr}} \in \mathbb{R}^{M^{\mathrm{pur}}} \quad (4)$$

where M^{pur} represents the number of retained biasing phrases with $M^{\mathrm{pur}} \ll M$. In this way, many irrelevant phrases are discarded. Then, the multi-level semantic correlations can be predicted more accurately on the shorter biasing list. In addition to discarding irrelevant phrases, the purification mechanism also reduces the number of biased vocabulary tokens by eliminating tokens not included in the refined list. Thus, this mechanism enhances both computational efficiency and the identification of the most relevant biasing information.

Finally, the recognition results of ASR backbone $\mathbf{P}^{\mathrm{bb}}$ can be interpolated with the joint modeling results $\mathbf{Q}^{\mathrm{bias}}$ automatically according to the list-level correlation $\mathbf{Q}^{\mathrm{list}}$ to obtain the contextualized results $\mathbf{Q}^{\mathrm{casr}} = \{\mathbf{Q}_u^{\mathrm{casr}} \in \mathbb{R}^V \mid u = 1, \cdots, U\}$:

$$\mathbf{Q}^{\mathrm{casr}} = (1 - \mathbf{Q}^{\mathrm{list}})\mathbf{P}^{\mathrm{bb}} + \mathbf{Q}^{\mathrm{list}}\mathbf{Q}^{\mathrm{bias}} \qquad (5)$$
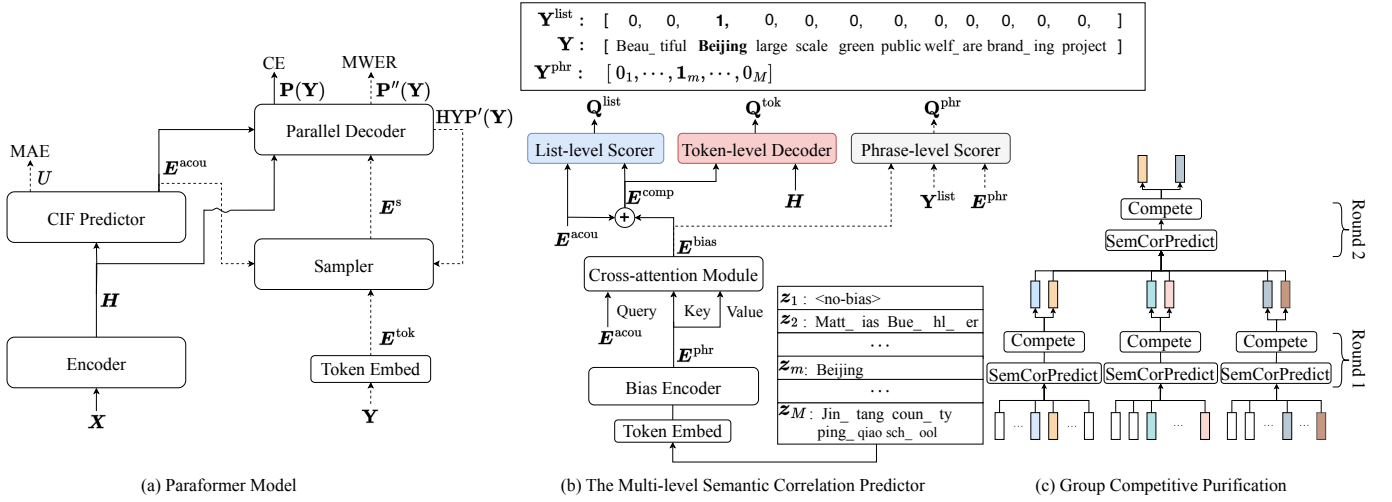
Fig. 3. Schematic diagram of the purified semantic correlation joint modeling approach. (a) Non-autoregressive ASR backbone model: Paraformer. (b) Multi-level semantic correlation predictor (SemCorPredict). (c) Purification flowchart. In Figures (a) and (b), solid lines denote inference paths, whereas dotted lines represent connections used exclusively during training.

B. ASR Backbone

The powerful NAR ASR model, Paraformer [55], is the backbone to obtain the basic recognition results. As shown in Fig. 3(a), the Paraformer consists of an encoder, a continuous integrate-and-fire (CIF) predictor, and a parallel decoder. Firstly, the speech feature X is fed into the encoder to extract the speech representation $H = \{h_t \mid t = 1, \cdots, T\}$. Then, the CIF predictor accumulates H by weighted summation and predicts the acoustic embeddings E^{acou} for each ASR output token. In the parallel decoder, E^{acou} integrates the acoustic information of H to predict the probabilities of a sequence of target tokens, $Y = \{y_u \mid u = 1, \cdots, U\}$.

During the training stage, a glance language model [56] based sampler module is utilized to strengthen the NAR decoder by modeling token inter-dependency. Specifically, the parallel decoder takes in E^{acou} and H to generate the first pass decoding result $P(Y)$. Then, the greedy search is performed to obtain the first pass hypothesis $\text{HYP}'(Y)$. The Hamming distance d between $\text{HYP}'(Y)$ and the ground-truth labels Y is calculated. Subsequently, Paraformer processes the tokens in Y through the embedding layer to obtain dense token representations $E^{\text{tok}} = \{e_u^{\text{tok}} \mid u = 1, \cdots, U\}$. Then, the sampler randomly selects the E^{tok} with a computed ratio, and the chosen token embeddings are used to replace the acoustic embeddings in E^{acou}, forming the sampled embedding sequence $E^{\text{s}} = \{e_u^{\text{s}} \mid u = 1, \cdots, U\}$. Finally, the speech embedding sequence H and E^{s} are fed into the parallel decoder again to produce the second pass decoding results $P''(Y)$. The cross-entropy (CE) loss, mean absolute error (MAE), and minimum word error rate (MWER) [57] losses are employed to train the Paraformer jointly. The second pass hypotheses are sampled to generate negative candidates for the MWER training, and MAE is computed between the predicted token count and its reference label U.

C. Multi-level Semantic Correlation Predictor

Fig. 3(b) describes the architecture of the multi-level semantic correlation predictor (SemCorPredict). During the training of SemCorPredict, the "Token Embed" layer encodes the l-th token $z_{m,l}$ in the m-th biasing phrase $z_m = \{z_{m,l} \mid l = 1, \cdots, L\}$, which has a length of L_m and is derived from a given biasing phrase list Z. Then, a long short-term memory (LSTM) [58] based bias encoder processes all token embeddings of one phrase and outputs the phrase-level embedding e_m^{phr}. Among the biasing list Z, a no-bias token "<no-bias>" is involved as an option to indicate the absence of context. Subsequently, a cross-attention module takes in the "Query" embeddings E^{acou}, "Key" and "Value" embeddings E^{phr} to produce the relevant biased embeddings E^{bias}:

$$E^{\text{bias}} = \text{CrossAttn}(E^{\text{acou}}, E^{\text{phr}}, E^{\text{phr}}) \tag{6}$$

Then, E^{bias} and E^{acou} are added to obtain the compound embeddings $E^{\text{comp}} = \{e_u^{\text{comp}} \mid u = 1, \cdots, U\}$.

The list-level correlation prediction is regarded as a binary classification task at each decoding step u, in which the list-level scorer takes in the acoustic embeddings e_u^{acou} and e_u^{comp} to predict the correlation Q_u^{list}, indicating whether e_u^{acou} is related to E^{phr}. In practice, biasing phrases may not appear in all sentences, and their length is relatively limited compared to the entire transcription, leading to label imbalance. To alleviate this issue, we apply the focal loss [59] on the prediction and its reference label $Y^{\text{list}} = \{y_u^{\text{list}} \in \{0, 1\} \mid u = 1, \cdots, U\}$:

$$\tau_u^{\text{list}} = Q_u^{\text{list}} y_u^{\text{list}} + (1 - Q_u^{\text{list}})(1 - y_u^{\text{list}}) \tag{7}$$

$$\theta_u = \alpha y_u^{\text{list}} + (1 - \alpha)(1 - y_u^{\text{list}}) \tag{8}$$

$$\mathcal{L}_{\text{list}} = \sum_{u=1}^{U} (-\theta_u (1 - \tau_u^{\text{list}})^\gamma \log(\tau_u^{\text{list}})) \tag{9}$$

where α, ranging from 0 to 1, is used to balance positive and negative labels. γ is used to modulate the factor $(1 - \tau_u^{\text{list}})$ and balance easy and hard labels.

In the phrase-level correlation prediction, each biasing phrase is treated as an entity for classification. To simplify this classification at the training stage, we assume that each sentence contains only one biasing phrase. In the phrase-level scorer, the element-wise product between the list-level reference label $\mathbf{Y}^{\text{list}}$ and $\boldsymbol{E}^{\text{bias}}$ is calculated first. Then, the product is summed over u to obtain a representation $\boldsymbol{e}^{\text{phr}'}$:

$$\boldsymbol{e}^{\text{phr}'} = \sum_{u=1}^{U} \text{y}_u^{\text{list}} \boldsymbol{e}_u^{\text{bias}} \tag{10}$$

The cosine similarities between $\boldsymbol{e}^{\text{phr}'}$ and all biasing phrase embeddings $\boldsymbol{E}^{\text{phr}}$ are finally calculated:

$$s_m^{\text{phr}'} = \frac{< \boldsymbol{e}^{\text{phr}'}, \boldsymbol{e}_m^{\text{phr}} >}{\|\boldsymbol{e}^{\text{phr}'}\|_2 \|\boldsymbol{e}_m^{\text{phr}}\|_2} \tag{11}$$

To optimize the SemCorPredict, contrastive loss [60] is utilized to maximize the similarity within the target biasing phrase and minimize its similarity with negative biasing phrases:

$$\mathcal{L}_{\text{phr}} = \sum_{m=1}^{M} \{ -s_m^{\text{phr}'} \text{y}_m^{\text{phr}} + s_m^{\text{phr}'} (1 - \text{y}_m^{\text{phr}}) \} \tag{12}$$

where y_m^{phr} denotes the reference label of the m-th phrase, and $\mathbf{Y}^{\text{phr}} = \{ \text{y}_m^{\text{phr}} \in \{0,1\} \mid m = 1, \cdots, M \}$ denotes the full label set. It is worth noting that the biasing-phrase classification serves as an auxiliary task to guide the attention module training. At the inference stage, due to the absence of the list-level label $\mathbf{Y}^{\text{list}}$, the phrase-level correlation at the u-th decoding step is extracted from the multi-head attention matrix $\boldsymbol{A} \in \mathbb{R}^{U \times M \times N}$ (N is the number of heads) of the cross-attention module:

$$\mathbf{Q}_u^{\text{phr}} = \max_{n \in [1,N]} (\boldsymbol{A}_{u,:,n}) \tag{13}$$

The token-level correlation prediction is similar to the decoding process of the ASR backbone. The token-level scorer shares the same architecture as the parallel decoder of the ASR backbone. Their differences are that the token-level scorer uses $\boldsymbol{E}^{\text{comp}}$ instead of $\boldsymbol{E}^{\text{acou}}$ and has an extra feed-forward output layer. The SemCorPredict can be optimized by the cross-entropy loss $\mathcal{L}_{\text{tok}}$ between the correlation distribution of token classification $\mathbf{Q}^{\text{tok}}$ and its reference label $\mathbf{Y}$.

The ASR backbone is trained first, then the SemCorPredict is jointly trained with three loss functions, i.e., focal loss $\mathcal{L}_{\text{list}}$, contrastive loss $\mathcal{L}_{\text{phr}}$ and cross-entropy loss $\mathcal{L}_{\text{tok}}$:

$$\mathcal{L}_{\text{SemCorPredict}} = \mathcal{L}_{\text{list}} + \mathcal{L}_{\text{phr}} + \mathcal{L}_{\text{tok}} \tag{14}$$

During the training of SemCorPredict, the open-source tool Spacy[1] is utilized to perform textual parsing on each transcription of the training set. The semantically coherent nouns are randomly selected to form the biasing list, with a probability of 90% for a sentence to be chosen with a biasing phrase. Fig. 3(b) shows a simple case that "Beijing" is the m-th biasing phrase in the given biasing list and it appears in the transcription $\mathbf{Y}$, thus the indices among u of "Beijing" in $\mathbf{Y}^{\text{list}}$ are set to 1 and the m-th value in $\mathbf{Y}^{\text{phr}}$ is set to 1.

[1]https://spacy.io/models/zh#zh_core_web_lg

D. Semantic Correlation Joint Modeling Method

The cross-granularity consistency of the most relevant biasing information is intuitive and intrinsic, making it possible to jointly model the intersection of the three semantic correlations. Specifically, at the u-th decoding step, if an ASR intermediate representation $\boldsymbol{e}_u^{\text{acou}}$ shares consistent information with any biasing phrase $\boldsymbol{e}_m^{\text{phr}}$, the list-level correlation $\mathbf{Q}_u^{\text{list}}$ becomes prominent, and there must be corresponding phrase and ASR vocabulary tokens that exhibit higher phrase-level correlation $\mathbf{Q}_u^{\text{phr}}$ and token-level correlation $\mathbf{Q}_u^{\text{tok}}$. Moreover, if certain biasing information shows higher correlations across all three semantic granularities simultaneously, it will achieve the highest correlation within the intersected correlations, making it the most relevant biasing information. An overview of the joint modeling process for multiple semantic correlations is presented in Eq. (2).

To improve the stability of the predicted $\mathbf{Q}^{\text{list}}$ and $\mathbf{Q}^{\text{phr}}$, two different sliding window smoothing methods, triangular window smoothing, and list-correlation-guided smoothing, are proposed. The smoothed list-level correlation $\mathbf{Q}^{\text{slist}}$ is calculated as below:

$$\mathbf{Q}^{\text{slist}} = \text{Conv1d}(\mathbf{Q}^{\text{list}}, \boldsymbol{\mathcal{W}}^{\text{tri}}) \tag{15}$$

where $\boldsymbol{\mathcal{W}}^{\text{tri}} = [\frac{1-\omega}{2}, \omega, \frac{1-\omega}{2}]$ and the hyperparameter ω belongs to $[0, 1]$. When designing the sliding window for the phrase-level correlation $\mathbf{Q}^{\text{phr}}$, the length of the possible biasing phrase should be considered, thus we propose the list-correlation-guided smoothing method. Specifically, $\mathbf{Q}^{\text{slist}}$ is summed among u to estimate the length L' of the possible biasing phrase, which is then used as the window size. At each output step u, the token could appear at any position within the potential biasing phrase. The position of the biasing phrase is determined by finding the sliding window with the highest sum of $\mathbf{Q}^{\text{slist}}$ that includes the token. The calculation process of the list-correlation-guided smoothing is as follows:

$$L' = \text{round}(\sum_{u=1}^{U} \mathbf{Q}_u^{\text{slist}}) \tag{16}$$

$$j_u = \arg\max_{j \in [u-L'+1, u+L'-1]} (\text{Conv1d}(\mathbf{Q}^{\text{list}}, \mathbb{1}^{L'})[j]) \tag{17}$$

where $\mathbb{1}^{L'}$ is a one-dimensional convolution kernel of length L' with weights of 1, j_u is the index of the most relevant location for the potential biasing phrase. This location determines the corresponding smoothed phrase-level correlation $\mathbf{Q}_u^{\text{sphr}}$ associated with the ASR intermediate representation, the index selection function, IdxSel, indexes the $\mathbf{Q}_u^{\text{sphr}}$ using j_u:

$$\mathbf{Q}^{\text{sphr}} = \text{Conv2d}(\mathbf{Q}^{\text{phr}}, \mathbb{1}^{L' \times 1}) \tag{18}$$

$$\mathbf{Q}_u^{\text{sphr}} = \mathbf{Q}^{\text{sphr}}[u - L' + 1 : u + L' - 1] \tag{19}$$

$$\mathbf{Q}_u^{\text{sphr}} = \tanh(\text{IdxSel}(\mathbf{Q}_u^{\text{sphr}}), j_u) \tag{20}$$

The smoothed phrase-level correlation $\mathbf{Q}^{\text{sphr}}$ becomes sharper after list-correlation-guided smoothing, alleviating the issue of flattened attention weight $\mathbf{Q}^{\text{phr}}$ caused by extremely long

---

Algorithm 1: Group Competitive Purification

Data: biasing phrase embeddings $\boldsymbol{E}^{\mathrm{phr}}$, acoustic embeddings $\boldsymbol{E}^{\mathrm{acou}}$, speech representations $\boldsymbol{H}$, purification group size $group\_size$, purification rounds n_r, list-level correlation threshold $thres\_list$, top number n_top

Result: purified biasing phrase embeddings $\boldsymbol{E}^{\mathrm{phr,pur}}$

1 # Using decoding of a sentence as an example.
2 $i \leftarrow 1$;
3 $G \leftarrow \lceil M/\ group\_size \rceil$
4 **while** $i \leq n\_r$ and $G > 1$ **do**
5     $\{\boldsymbol{E}_g^{\mathrm{phr}}\}_{g=1}^G \leftarrow \boldsymbol{E}^{\mathrm{phr}}.\mathrm{reshape}(G, group\_size, :)$
6     $\{\ \mathbf{Q}_g^{\mathrm{list}}, \mathbf{Q}_g^{\mathrm{phr}} = \mathrm{SemCorPredict}(\boldsymbol{E}^{\mathrm{acou}}, \boldsymbol{E}_g^{\mathrm{phr}}, \boldsymbol{H})\ \}_{g=1}^G$
7     $\{\ \mathbf{Q}_g^{\mathrm{list}} \in \{0,1\} \leftarrow \mathrm{bool}(\mathbf{Q}_g^{\mathrm{list}} > thres\_list)\ \}_{g=1}^G$
8     $\boldsymbol{I} \leftarrow \mathrm{shuffle}(\mathrm{unique}(\{(\mathbf{Q}_g^{\mathrm{list}} \odot \mathbf{Q}_g^{\mathrm{phr}}).\mathrm{topk}(n\_top)\}_{g=1}^G))$
9     $M \leftarrow \mathrm{len}(\boldsymbol{I})$
10     $\boldsymbol{E}^{\mathrm{phr}} \leftarrow \mathrm{IdxSel}(\boldsymbol{E}^{\mathrm{phr}}, \boldsymbol{I})$
11     $G \leftarrow \lceil M/\ group\_size \rceil$
12     $i \leftarrow i + 1$
13 **end**
14 $M^{\mathrm{pur}} \leftarrow M$
15 **return** $\boldsymbol{E}^{\mathrm{phr}}, M^{\mathrm{pur}}$

TABLE I
STATISTICS OF AISHELL-NER

Sets	Categ.	#Sent.	# Phraselist	Length of phrase min,max,avg	Inclusion Rate (%)	OOV Rate (%) AISHELL-1	OOV Rate (%) KeSpeech
Dev	NE	14326	2204	2,19,4	34.94	51.57	49.52
	PER		706	2,9,3	11.20	56.31	65.25
	LOC		612	2,12,3	15.41	43.86	32.57
	ORG		888	2,19,5	19.11	53.10	48.70
Test	NE	7176	1196	2,16,4	31.73	50.71	49.12
	PER		414	2,11,3	**10.30**	**62.22**	**65.38**
	LOC		343	2,11,3	15.25	38.60	30.99
	ORG		441	2,16,5	16.97	49.32	47.95

of purification is that only a few biasing phrases with higher list- and phrase-level correlations are selected in each group for constructing the purified short biasing list. The strategy of dividing biasing phrases into groups based on the principle of divide and conquer, combined with performing parallel purification in group competitive purification (GCP), can yield more accurate semantic correlation prediction, especially when M varies significantly compared to the training condition. Algorithm 1 provides the pseudo-code of the calculation procedure. In each round, the embeddings of biasing list $\boldsymbol{E}^{\mathrm{phr}}$ are grouped randomly into several short lists $\boldsymbol{E}_g^{\mathrm{phr}}$. The SemCorPredict uses $\boldsymbol{E}^{\mathrm{acou}}$, $\boldsymbol{E}_g^{\mathrm{phr}}$ and H to generate grouped list-level correlation $\mathbf{Q}_g^{\mathrm{list}}$ and grouped phrase-level correlation $\mathbf{Q}_g^{\mathrm{phr}}$ in a parallel way. At each decoding step, $\mathbf{Q}_{u,g}^{\mathrm{list}}$ is reset to **1** (True) or **0** (False) according to the relationship with list-level correlation threshold $thres\_list$. The top n_top active biasing phrases are the winners at each "True" decoding step, and then they are shuffled and re-grouped for the next round. Notably, the purification running **once** means that the $group\_size$ equals the length of the biasing list, which is termed once competitive purification (OCP).

biasing lists. After smoothing, the intersected correlation $\mathbf{Q}_u^{\mathrm{bias}}$ can be modeled:

$$\mathbf{Q}_u^{\mathrm{bias}} = \mathrm{Softmax}(\max_{m \in [1,M]} \mathbf{Q}_u^{\mathrm{slist}}\ \mathbf{Q}_{u,m}^{\mathrm{sphr}}\ \boldsymbol{\Phi}_{m,v}\ \mathbf{Q}_{u,v}^{\mathrm{tok}}) \quad (21)$$

$$\boldsymbol{\Phi}_{m,v} = \begin{cases} 1, & if\ \mathrm{y}_v \in \boldsymbol{z}_m \\ 0, & others \end{cases} \quad (22)$$

where $\boldsymbol{\Phi}_{m,v}$ is an indicator function representing whether the v-th ASR vocabulary token y_v appears in a biasing phrase $\boldsymbol{z}_m$.

Whether to introduce contexts can be automatically controlled by list-level correlation interpolating, which is used in the collaborative decoding of the joint modeling result $\mathbf{Q}^{\mathrm{bias}}$ and the backbone recognition result $\mathbf{P}^{\mathrm{bb}}$ to obtain the contextualized result $\mathbf{Q}^{\mathrm{casr}}$:

$$\mathbf{Q}_u^{\mathrm{casr}} = (1 - \mathbf{Q}_u^{\mathrm{slist}})\mathbf{P}_u^{\mathrm{bb}} + \mathbf{Q}_u^{\mathrm{slist}}\mathbf{Q}_u^{\mathrm{bias}} \quad (23)$$

In real applications, many sentences do not contain any specified biasing phrases in that the introduced biasing information may be harmful to the recognition results, which is termed as over-biasing problem. To alleviate the over-biasing issue, a simple and efficient post-processing strategy is utilized: the contextual hypothesis $\mathrm{HYP}^{\mathrm{casr}}$ is selected as the final output only when it detects more biasing phrases than the backbone hypothesis $\mathrm{HYP}^{\mathrm{bb}}$.

$$\mathrm{HYP}^{\mathrm{casr}} = \begin{cases} \mathrm{HYP}^{\mathrm{casr}}, if\ \mathcal{C}(\mathrm{HYP}^{\mathrm{casr}}) > \mathcal{C}(\mathrm{HYP}^{\mathrm{bb}}) \\ \\ \mathrm{HYP}^{\mathrm{bb}}, & others \end{cases} \quad (24)$$

where $\mathcal{C}$ denotes the counter function which counts the number of the detected biasing phrases in a hypothesis.

E. Group and Competitive Purification

The purification mechanism is proposed to reduce the computational cost introduced by the joint modeling across three semantic granularities. As shown in Fig. 3(c), the key idea

IV. EXPERIMENTS

A. Data and Metrics

In practice, the biasing list may be set arbitrarily for contextual ASR, thus the length of the biasing list and the number of tokens per phrase can vary significantly. To simulate this situation, we employ an open-source Chinese named entity data set called AISHELL-NER [61] to form the biasing list for evaluation, in which a named entity is viewed as a biasing phrase. A named entity phrase usually contains several words and the token number of a phrase is larger than that of a word. Thus, named entity phrases may be more suitable for the method evaluation than word-level biasing phrases [21].

Table I provides statistics of the biasing information on the development and test sets[2], including the number of sentences, the length of biasing lists, the length statistics of phrases, the percentages of sentences including the biasing phrase (Inclusion Rate), and the rate of out-of-vocabulary (OOV). The OOV rate measures the fraction of phrases appearing in the biasing lists but not in the training data.

[2]All biasing lists are present at https://github.com/shibeiing/Purified-Semantic-Correlation-Joint-Modeling-PSC-Joint.

The development and test sets of AISHELL-NER both consist of three categories of the named entity: person names (PER), location names (LOC), and organization names (ORG). We combine the three mentioned named entity categories to create the fourth biasing list called "NE". As indicated in the fourth and fifth columns of Table I, the "NE" lists contain thousands of biasing phrases, with the number of tokens per phrase (phrase length) ranging from 2 to 19. To evaluate the impact of varying biasing list lengths, we split the "NE" biasing list randomly into several lists with different lengths, e.g., "NE-101". The lower inclusion rate means that biasing phrases are rarer in the text, i.e., the biasing information is redundant for most sentences. Thus, the "PER-414" biasing list of the test set, which has a higher OOV rate and lower inclusion rate, is also employed to evaluate the performance degradation caused by the over-biasing issue.

To investigate the impact of training data scale for the PSC-Joint approach, experiments are mainly conducted on two open-source Chinese corpora, AISHELL-1 [62] and Kespeech [63]. The AISHELL-1 contains 178 hours of standard Mandarin audio, and the KeSpeech Phase-1 corpus contains 895 hours of transcribed audio in Mandarin and its eight subdialects. The former one is used to perform rapid experimental validation and analysis, and the latter one, which contains many more accented speech utterances, is suitable to verify the generalization of the proposed methods in reality.

In all experiments, we evaluate the performance using the following metrics: character error rate (CER), recall, precision and F1 score. The CER and F1 score, which reflect the trade-off between under- and over-biasing in contextual ASR, are mainly discussed in the results subsection. Notably, recall, precision, and F1 score are computed based on exact match, where the entire biasing phrase must be recognized without any token errors. In addition, the real-time factor (RTF) is used to evaluate the impact of PSC-Joint on the inference speed.

B. Experimental Setup

Following the official configuration and recipe from FunASR [64], we use 80-dimensional FBANK features as input and Chinese characters as ASR vocabulary tokens to train the ASR backbone and contextual ASR models on the AISHELL-1 and KeSpeech Phase-1 corpora. The model includes a 12-layer encoder based on Conformer [7] and a 6-layer parallel decoder based on Transformer [6]. The parallel decoder consists of multiple blocks stacked with self-attention, feedforward networks [65], and cross-attention. The cross-attention module incorporates a transformer-style multi-head attention layer with an embedding size of 256 and four attention heads. More details can be found in the recipe[3]. The token-level scorer closely resembles the parallel decoder in the baseline, with the sole distinction of an additional feed-forward layer inserted before the output layer. The bias encoder consists of a single-layer LSTM with 256 hidden units. The token embedding layer shares parameters with the pre-trained embedding layer from the ASR backbone. In the

---

[3]https://github.com/alibaba-damo-academy/FunASR/tree/v0.2.0/egs/aishell/paraformer/conf

list-level scorer, the dimension of the feed-forward layer is set to 512. The phrase-level scorer does not contain trainable parameters, and its calculation process is defined in Eq. (10) and Eq. (11).

Considering the extreme imbalance between positive and negative labels in the list-level correlation, the hyperparameters α and γ of focal loss are set to 0.75 and 2.0, respectively. During inference, the hyperparameter ω of triangular window smoothing is set to 0.6. In the purification mechanism, the size of group $group\_size$, the number of rounds n_r, the threshold for list-level correlation $thres\_list$, and the number of biasing phrases being kept at each group n_top are set to 75, 2, 0.5, 10, respectively. Note that the parameters of the pre-trained ASR backbone are frozen during the training stage of SemCorPredict. SpecAugment [66] is applied for the data augmentation of the AISHELL-1 task. Following the configuration of FunASR [64], the ASR backbone of the AISHELL-1 task is trained for 50 epochs. Considering the scale of the data, the training epoch of the KeSpeech ASR backbone model is set to 150. According to preliminary experiments, the SemCorPredict module is trained for 50 epochs on both AISHELL-1 and KeSpeech tasks.

C. Results on AISHELL-1

Table II compares the baseline, three non-autoregressive contextual ASR models and the PSC-Joint model, all trained on the AISHELL-1 corpus. The baseline does not utilize the biasing information and serves as the backbone of other contextual models. For a fair comparison, ColDec, FineCoS, SeACo, and the PSC-Joint models share the same pre-trained backbone, and all these models conduct collaborative decoding with it through interpolation weighting. The ColDec, FineCoS and SeACo models are trained using a single contextual cross-entropy loss, and the biasing phrase is incorporated through a single bias probability. This can be regarded as using a single probability to represent list-level, phrase-level, and token-level correlations, serving as a baseline to highlight the effectiveness of our explicit supervision of three semantic correlations and their joint modeling. To purify the biasing list, the FineCoS model uses phrase selection (PS), while the SeACo model employs attention score filtering (ASF). The "SC-Joint" model indicates that no purification mechanism is applied, and models with the "-P" suffix incorporate the post-processing discussed in the Section III-D. To assess the robustness of models to varying-length lists, results on several biasing lists are reported.

Comparing each column in Table II, it can be observed that almost all contextual ASR models surpass the backbone model on F1 score, and the PSC-Joint-P (GCP) model achieves 25.57%, 18.23%, and 18.64% relative F1 score improvement over ColDec, FineCoS, and SeACo on the "NE-101" list, respectively. As the biasing list grows longer, it becomes more challenging to detect biasing phrases. Compared with the ColDec model, the SC-Joint model performs better on F1 score as the length of the biasing lists increases, demonstrating that the proposed semantic correlation joint modeling method can significantly enhance the robustness of contextual ASR

TABLE II
COMPARISON OF CONTEXTUAL ASR MODELS TESTED ON VARYING-LENGTH LISTS: **CER // RECALL|PRECISION|F1 SCORE** (%). (AISHELL-1)

| Models	Pur.	NE-51	NE-101	NE-201	NE-401	NE-601	NE-801	NE-1001	NE-1196
Baseline	-	5.05 76.36|**100.0**|86.60	5.05 71.49|**99.41**|83.17	5.05 77.13|**99.07**|86.73	5.05 77.99|**99.68**|87.51	5.05 83.08|**99.40**|90.51	5.05 80.21|**99.54**|88.83	5.05 77.75|**99.31**|87.21	5.05 79.90|**99.19**|88.50
ColDec [30]	✗	5.06 84.55|98.94|**91.18**	5.06 74.47|98.31|84.75	5.09 80.22|98.88|88.58	5.13 80.46|99.39|88.93	5.14 85.05|99.32|91.63	5.16 82.37|99.59|90.17	5.17 79.98|99.09|88.52	5.16 81.47|99.03|89.40
FineCoS [33]	PS	5.07 84.55|97.89|90.73	5.07 76.60|98.36|86.12	5.08 79.49|98.65|88.04	5.07 79.80|99.08|88.40	5.10 85.05|99.08|91.53	5.09 82.11|99.23|89.87	5.10 80.01|99.01|88.50	5.09 81.52|99.03|89.43
SeACo [49]	ASF	5.05 87.27|95.05|91.00	5.06 78.72|94.87|86.05	5.03 81.49|97.82|88.91	5.02 80.21|98.98|88.62	5.01 85.05|99.17|91.57	5.00 81.96|99.46|89.87	4.99 79.69|99.21|88.39	4.98 81.40|99.03|89.35
SC-Joint	✗	5.31 **89.09**|90.74|89.91	5.18 85.53|93.06|**89.14**	5.13 85.12|98.32|91.25	5.04 85.57|98.48|91.57	4.95 88.05|99.01|93.21	4.95 85.10|99.13|91.58	4.87 83.65|99.05|90.70	4.80 84.98|98.90|91.41
SC-Joint-P	✗	**5.04** **89.09**|90.74|89.91	**5.02** 85.53|93.06|**89.14**	**5.00** 85.12|98.32|91.25	4.96 85.66|98.58|91.66	4.93 88.22|99.02|93.31	4.92 85.32|99.13|91.71	4.86 83.84|99.06|90.82	4.84 85.20|98.90|91.54
PSC-Joint	OCP	5.41 **89.09**|83.76|86.34	5.18 **86.38**|91.86|89.04	5.32 86.39|96.95|91.36	5.23 86.48|98.13|91.94	5.14 89.10|98.52|93.57	5.10 86.26|98.38|91.92	5.09 85.41|98.60|91.53	4.89 86.24|98.54|91.98
PSC-Joint-P	OCP	5.05 **89.09**|83.76|86.34	5.03 **86.38**|91.86|89.04	5.07 86.57|96.95|91.47	4.95 86.64|98.22|92.07	**4.92** 89.31|98.53|93.69	4.90 86.41|98.38|92.01	4.82 85.70|98.60|91.70	4.79 86.48|98.59|92.14
PSC-Joint	GCP	5.31 **89.09**|90.74|89.91	5.26 **86.38**|91.03|88.65	5.28 86.93|96.57|91.50	5.20 87.30|97.60|92.17	5.15 89.89|98.40|93.95	5.10 87.19|98.15|92.35	4.98 85.82|98.28|91.63	4.87 86.99|98.52|92.39
PSC-Joint-P	GCP	**5.04** **89.09**|90.74|89.91	5.03 **86.38**|91.03|88.65	**5.00** **87.11**|96.58|**91.60**	4.94 **87.39**|97.70|**92.25**	**4.92** **90.02**|98.40|**94.02**	**4.88** **87.34**|98.15|**92.43**	**4.81** **86.08**|98.29|**91.78**	**4.76** **87.18**|98.52|**92.51**
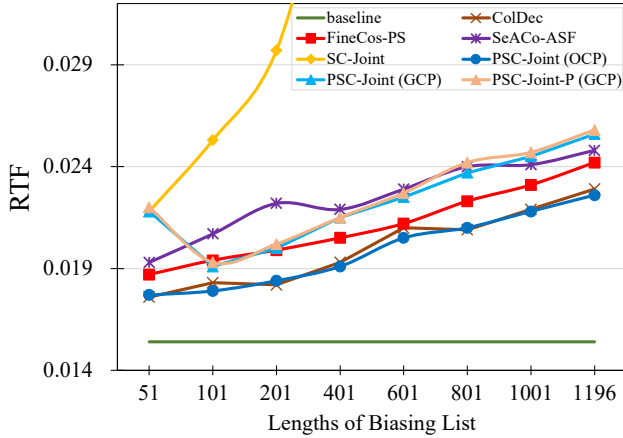
Fig. 4. Comparison of contextual ASR models tested on varying-length biasing lists: RTF.

against changes in the biasing information volume. Even with PS and ASF purification methods, the performance improvements of the FineCoS and SeACo models decrease when the length of the list increases. Nevertheless, PSC-Joint-P (GCP) achieves average F1 score improvements of 23.08% and 21.99% across varying-length biasing lists compared with FineCoS and SeACo. Assisted by the post-processing strategy, PSC-Joint-P (GCP) achieves lower CER on varying-length biasing lists while maintaining the F1 score improvement.

Fig. 4 illustrates the real-time factor (RTF) of the baseline and contextual ASR models tested on varying-length biasing lists. When comparing the inference speed of the PSC-Joint (GCP) model and the PSC-Joint-P (GCP) model, we observe that the negative effect caused by the post-processing is negligible, indicating that the post-processing alleviates the over-biasing problem at a low cost. In addition, the inference speeds of the FineCoS, SeACo, and PSC-Joint-P (GCP) are almost the same, and their RTFs increase nearly linearly with the list length.

We further conduct a modular analysis on PSC-Joint-P with the results in Table II and Fig. 4. Despite lacking purification, SC-Joint exhibits robustness across varying list lengths, which benefits from the joint modeling across multi-level semantic correlations. However, as the length of the biasing list increases, the RTF of the SC-Joint increases significantly. Compared with SC-Joint, PSC-Joint (OCP) and PSC-Joint (GCP) benefit from the purification, achieving superior performance and faster inference speed.

A potential risk of applying purification to biasing lists is the unintended removal of a target biasing phrase that appears in the text sentence. To compare the OCP and GCP mechanisms, we introduce a new metric, retention rate, which indicates the average proportion of target biasing phrases in each test sentence that are successfully retained in the purified phrase list after the purification process. The random grouping strategy, based on the principle of divide and conquer, reduces the competition among similar phrases at the beginning of purification. Although it introduces some additional computations, GCP achieves a retention rate of 92.06% on the "NE-1196" biasing list, outperforming the 86.26% retention rate achieved by OCP. Therefore, PSC-Joint (GCP) performs better across varying-length biasing lists than PSC-Joint (OCP), benefiting from the divide-and-conquer strategy.

D. Ablation Study

Table III presents an ablation study of the PSC-Joint model, evaluating the effectiveness of multi-level semantic correlation losses on biasing lists of different lengths. To perform dynamic interpolation (defined in Eq. (23)), the list-level loss must be retained. Considering that the token- and phrase-level correlation can be derived from the backbone recognition results and the attention module, respectively, their corresponding losses may optionally be excluded during training. Specifically, E1-E4 and E5-E8 present the ablation results of SC-Joint without or with purification, respectively.

TABLE III
ABLATION STUDY OF PSC-JOINT MODEL TESTED ON VARYING-LENGTH BIASING LISTS: **CER // RECALL|PRECISION|F1 SCORE** (%).

| Exp.	Token level	Phrase level	GCP	NE-201	NE-601	NE-1196
E1	✓	✓		5.13 85.12|98.32|91.25	4.95 88.05|99.01|93.21	4.80 84.98|98.90|91.41
E2	✓			5.34 87.30|90.75|88.99	5.18 89.93|97.95|93.77	4.91 86.57|98.51|92.16
E3		✓		5.09 84.75|98.11|90.94	4.97 86.22|99.23|92.27	4.90 83.04|99.02|90.33
E4				5.41 87.48|93.59|90.43	5.05 89.52|98.35|93.72	4.81 85.70|98.77|91.78
E5	✓	✓	✓	5.28 86.93|96.57|91.50	5.15 89.89|98.40|93.95	4.87 86.99|98.52|92.39
E6	✓		✓	5.42 87.11|90.74|88.89	5.50 90.23|97.25|93.61	5.30 86.77|97.61|91.87
E7		✓	✓	5.20 85.48|96.12|90.49	5.04 88.97|98.34|93.42	4.85 86.36|98.59|92.07
E8			✓	5.60 88.75|91.23|89.97	5.38 90.31|97.12|93.59	5.14 87.28|98.04|92.35

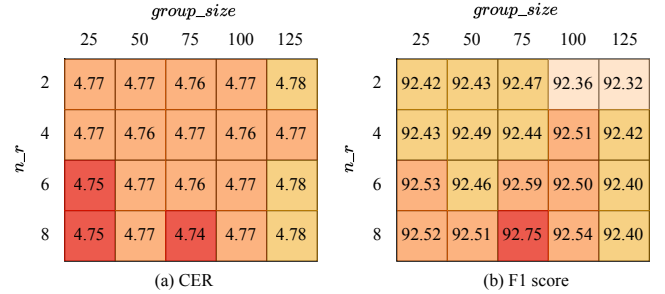Fig. 5. Sensitivity analysis of hyperparameter tuning for PSC-Joint-P (OCP) in terms of (a) CER (%) and (b) F1 score (%).

Fig. 6. Sensitivity analysis of hyperparameter tuning for PSC-Joint-P (GCP) in terms of (a) CER (%) and (b) F1 score (%).

We begin by analyzing the effect of token- and phrase-level losses without GCP. Compared to E1, removing the phrase-level loss (E2) leads to a noticeable precision drop (from 98.32% to 90.75% on NE-201) and a CER increase (from 4.95% to 5.18% on NE-601), suggesting the emergence of over-biasing errors. In contrast, removing the token-level loss while retaining the phrase-level loss (E3) improves precision (from 98.90% to 99.02% on NE-1196) but reduces recall (from 85.12% to 84.75% on NE-201), indicative of under-biasing. When both token- and phrase-level losses are removed (E4), model performance becomes unstable across biasing lists and metrics. For instance, while CER consistently increases compared to E1 (4.95% to 5.05% on NE-601), F1 scores show mixed results, decreasing on NE-201 (91.25% to 90.43%) but slightly increasing on NE-601 and NE-1196. These results reflect the difficulty of balancing over-biasing and under-biasing without the guidance of token- and phrase-level supervision.

The challenge is further amplified when introducing the GCP, which relies heavily on predicted list- and phrase-level correlations. Under this setting, the removal of both token- and phrase-level losses (E8) leads to considerable performance degradation compared to E5 (F1 drops from 91.50% to 89.97% and CER rises from 5.28% to 5.60% on NE-201), highlighting the negative effect of inaccurate biasing phrase selection on purification performance. In summary, the results clearly demonstrate that (1) token- and phrase-level losses are complementary for mitigating over- and under-biasing, and (2) their presence is essential for fully exploiting the purification capability of GCP.

In addition, GCP retains a small number of biasing phrases that are highly correlated with ASR intermediate representations, leading to over-biasing, which slightly degrades CER and precision. This issue can be mitigated by the proposed post-processing strategy, as shown in the last row of Table II.
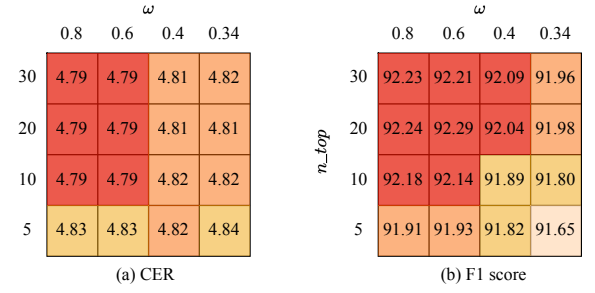
E. Evaluation on Hyperparameter Sensitivity

To explore the impact caused by hyperparameter tuning, we analyze hyperparameter sensitivity for PSC-Joint-P on the AISHELL-1 task. The threshold for list-level correlation, $thres\_list$, is set to 0.5 without tuning, which is the median of the score interval $[0, 1]$. The rest hyperparameters (the triangular window weight ω, purification group size $group\_size$, purification rounds n_r, and the number of biasing phrases being kept n_top) are split for the control variable study. In the PSC-Joint-P with the once competitive mechanism, $group\_size$ equals the length of the biasing list and n_r is 1, thus only ω and n_top need to be tuned. Fig. 5 shows the tuned performance of the PSC-Joint-P (OCP) represented in the form of a heatmap, in which the ω is adjusted from 0.34 to 0.8 and the n_top is adjusted from 5 to 30. Overall, the fluctuations of CER and F1 score are minimal, indicating that the PSC-Joint-P (OCP) is insensitive to ω and n_top. Thus, the ω is set to 0.6 and the n_top is set to 10 in our experiments.

With the determined ω and n_top, we further explore the hyperparameter sensitivity of PSC-Joint-P (GCP) on n_r and $group\_size$. From Fig. 6, it is clear that the PSC-Joint-P (GCP) model is insensitive to the rounds and group size. Considering the computational cost, the hyperparameter n_r is set to two. Another hyperparameter $group\_size$ is set to 75, closing to the average length of the biasing list during training. Notably, all experiments of PSC-Joint-P adopt the mentioned hyperparameter configuration, including the KeSpeech task.

F. Evaluation on KeSpeech

To explore the effect of the data scale, we conduct several experiments on the 895-hour data set, KeSpeech. In addi-

TABLE IV
COMPARISON OF CONTEXTUAL ASR MODELS TESTED ON THE VARYING-LENGTH BIASING LISTS OF THE DEVELOPMENT AND TEST SETS: **CER //
RECALL|PRECISION|F1** (%). (KESPEECH TASK)

Models	Pur	AISHELL-NER Dev				AISHELL-NER Test			
		NE-51	NE-801	NE-1601	NE-2204	NE-51	NE-601	NE-1196	PER-414
Baseline	-	4.76 79.83|**100.0**|88.78	4.76 80.66|**98.83**|88.83	4.76 78.17|**98.60**|87.20	4.76 79.40|**98.81**|88.05	5.07 71.82|**100.0**|83.60	5.07 81.33|**99.44**|89.48	5.07 77.89|**99.05**|87.20	5.07 45.51|**96.43**|61.83
ColDec [30]	✗	4.77 90.56|100.0|95.05	4.75 85.47|98.33|91.45	4.74 82.22|98.28|89.54	4.74 82.41|98.61|89.78	5.07 86.36|97.94|91.79	5.01 86.59|98.95|92.36	4.98 82.17|98.78|89.71	4.96 61.69|97.34|75.52
FineCoS [33]	PS	4.88 94.42|99.10|96.70	5.54 85.68|95.44|90.30	5.78 82.10|95.59|88.34	5.80 82.71|96.79|89.20	5.29 90.00|94.29|92.09	5.59 87.68|98.13|92.61	5.80 83.36|97.93|90.06	5.45 61.91|95.00|74.97
SeACo [49]	ASF	4.86 93.99|79.64|86.22	4.89 87.95|95.44|91.54	4.80 84.49|96.61|90.14	4.76 84.32|97.59|90.47	5.24 90.00|69.72|78.57	5.10 88.93|96.38|92.50	5.01 84.23|97.43|90.35	5.06 65.39|91.37|76.23
PSC-Joint-P	GCP	**4.74** **97.00|97.84|97.41**	**4.62** **92.02|95.63|93.79**	**4.49** **90.32|96.18|93.16**	**4.37** **90.12|97.26|93.55**	**5.05** **92.73|91.07|91.89**	**4.81** **93.69|96.97|95.30**	**4.61** **90.90|97.26|93.97**	**4.73** **80.67|94.35|86.98**

tion to the AISHELL-NER test set, the biasing lists of the AISHELL-NER development set are used as a supplement for evaluation, since it is not seen during the model training on KeSpeech. All model architectures, experimental settings, and hyperparameters are the same as the AISHELL-1 task. As shown in Table IV, all contextual ASR models achieve better performance in terms of F1 score, benefiting from the larger training data scale. Compared with ColDec, FineCoS, and SeACo, the PSC-Joint-P (GCP) model achieves 34.49%, 33.40%, and 43.76% average relative F1 score improvements across varying-length biasing lists. For longer biasing lists, such as "NE-2204" (Dev set), the relative F1 score improvements of PSC-Joint-P (GCP) attach to 36.89%, 40.28%, and 32.32% when compared to ColDec, FineCoS and SeACo, respectively. This demonstrates the strong robustness of PSC-Joint-P (GCP) in handling varying lengths of biasing lists.

The "PER-414" biasing list has the highest bias OOV rate and lowest inclusion rate, which poses a significant challenge for the contextual ASR. PSC-Joint-P (GCP) outperforms the backbone with a relative improvement of 65.59% on F1 score. For the issue of over-biasing, our PSC-Joint-P (GCP) attains a superior balance between the under- and over-biasing, i.e., lower CER and higher F1 score. This indicates that the most relevant biasing information is more accurately identified.

G. Instance Analysis

To obtain a deeper understanding of the semantic correlation joint modeling, we analyze an instance by plotting the multi-level semantic correlations, the joint modeling results, the backbone recognition results, and contextualized results. Fig. 7 shows that the transcription has 11 tokens and two biasing phrases, "迟名" (Chi Ming) and "常鑫" (Chang Xin). The NAR model employs greedy search decoding, i.e., only the most relevant token will be output at each decoding step. However, "迟" (Chi) and "常" (Chang) are not the most relevant tokens at the seventh and ninth decoding steps in the backbone recognition results. Instead, "驰" (Chi) and "长" (Chang), which have the same pronunciation, are output. In the list-level correlation, the correlation score of indices from seven to ten are the highest, corresponding to the top-one biasing phrase "迟名" (Chi Ming), "迟名" (Chi Ming), "常鑫" (Chang Xin), and "常鑫" (Chang Xin) in the phrase-level correlation, respectively. After joint modeling, only a few tokens satisfy
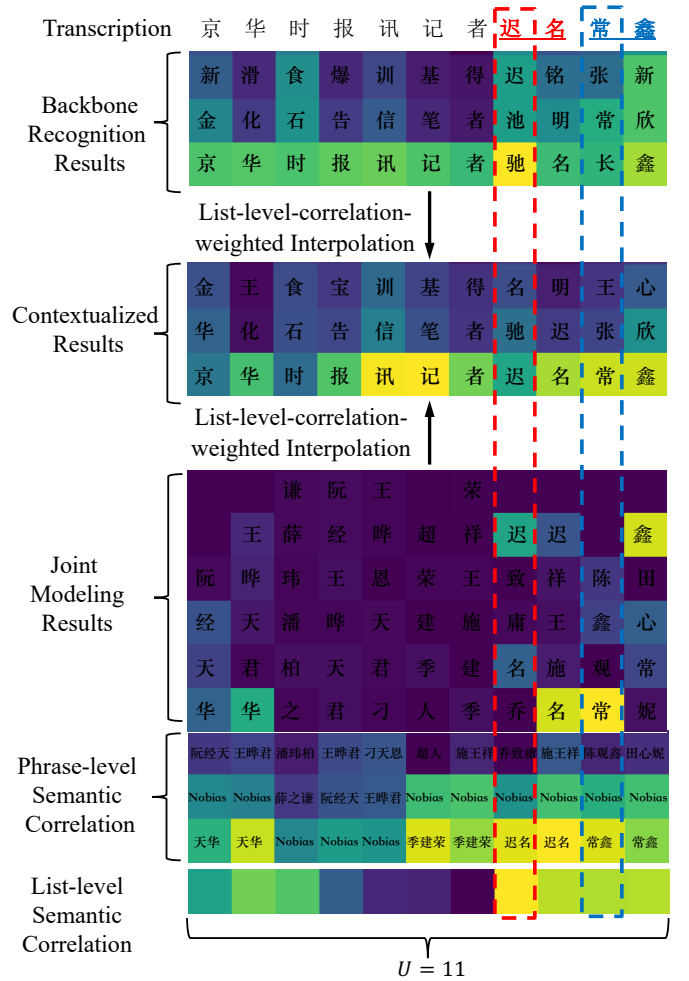

Fig. 7. An instance. Note that, for demonstration, only key elements of the backbone recognition results, joint modeling results, and phrase-level correlation are illustrated, rather than the entire matrices. And lighter color indicates higher score. The red dashed box and the blue dashed box highlight the biasing process of tokens in two person names, respectively.

the cross-granularity consistency of the most relevant biasing information, achieving higher scores in the joint modeling results. Subsequently, the contextualized results are obtained by interpolating the backbone recognition and joint modeling results with the list-level correlation. In this manner, the right

tokens in the seventh and ninth decoding steps are retrieved.

V. CONCLUSION

The cross-attention-based contextual ASR is sensitive to variations of biasing information volume. In this study, we propose the PSC-Joint approach to reduce the sensitivity by identifying and integrating the most relevant biasing information into the contextual ASR. Specifically, we first define three semantic correlations at various granularities from coarse to fine: list-level, phrase-level, and token-level. Then, based on the cross-granularity consistency of the most relevant biasing information, the intersection of three semantic correlations is calculated through joint modeling. In addition, PSC-Joint purifies the irrelevant biasing phrases in a group and competitive manner to reduce the computational cost of the semantic correlation joint modeling. Experiments evaluated on AISHELL-1, KeSpeech, and AISHELL-NER datasets demonstrate that PSC-Joint significantly improves the F1 score on varying-length biasing lists compared with baselines, enhancing the robustness of contextual ASR to variations in biasing information volume. Our PSC-Joint approach, combined with post-processing, also achieves a better trade-off between under- and over-biasing when introducing biasing information to ASR. Furthermore, the PSC-Joint approach speeds up the inference through biasing-phrase purification.

VI. ACKNOWLEDGMENTS

REFERENCES

[1] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, vol. 148, 2006, pp. 369–376.

[2] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, vol. 32, 2014, pp. 1764–1772.

[4] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NIPS*, 2015, pp. 577–585.

[5] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016, pp. 4960–4964.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and et.al., "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[7] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, and et.al., "Conformer: Convolution-augmented transformer for speech recognition," in *INTERSPEECH*, 2020, pp. 5036–5040.

[8] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 325–351, 2024.

[9] P. S. Aleksic, M. Ghodsi, A. H. Michaely, C. Allauzen, K. B. Hall, B. Roark, and et.al., "Bringing contextual information to google speech recognition," in *INTERSPEECH*, 2015, pp. 468–472.

[10] K. Hall, E. Cho, C. Allauzen, F. Beaufays, N. Coccaro, K. Nakajima, and et.al., "Composition-based on-the-fly rescoring for salient n-gram biasing," in *INTERSPEECH*, 2015, pp. 1418–1422.

[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] 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*. IEEE, 2019, pp. 6186–6190.

[13] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and et.al., "Shallow-fusion end-to-end contextual biasing," in *INTERSPEECH*, 2019, pp. 1418–1422.

[14] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, and et.al., "Streaming end-to-end speech recognition for mobile devices," in *ICASSP*, 2019, pp. 6381–6385.

[15] R. Huang, O. Abdel-hamid, X. Li, and G. Evermann, "Class lm and word mapping for contextual biasing in end-to-end asr," in *INTERSPEECH*, 2020, pp. 4348–4351.

[16] Z. Meng, Y. Wu, N. Kanda, L. Lu, X. Chen, G. Ye, and et.al., "Minimum word error rate training with language model fusion for end-to-end speech recognition," in *INTERSPEECH*, 2021, pp. 2596–2600.

[17] I. Nigmatulina, S. R. Madikeri, E. Villatoro-Tello, P. Motlícek, J. Zuluaga-Gomez, K. Pandia, and et.al., "Implementing contextual biasing in GPU decoder for online ASR," in *INTERSPEECH*, 2023, pp. 4494–4498.

[18] Z. Lei, E. Pusateri, S. Han, L. Liu, M. Xu, T. Ng, and et.al., "Personalization of ctc-based end-to-end speech recognition using pronunciation-driven subword tokenization," in *ICASSP*, 2024, pp. 10 096–10 100.

[19] G. Yang, Z. Ma, Z. Gao, S. Zhang, and X. Chen, "Ctc-assisted llm-based contextual asr," in *SLT*, 2024, pp. 126–131.

[20] G. Yang, Z. Ma, F. Yu, Z. Gao, S. Zhang, and X. Chen, "Mala-asr: Multimedia-assisted llm-based asr," *arXiv preprint arXiv:2406.05839*, 2024.

[21] D. Le, M. Jain, G. Keren, S. Kim, Y. Shi, J. Mahadeokar, and et.al., "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," in *INTERSPEECH*, 2021, pp. 1772–1776.

[22] D. Le, G. Keren, J. Chan, J. Mahadeokar, C. Fuegen, and M. L. Seltzer, "Deep shallow fusion for RNN-T personalization," in *SLT*, 2021, pp. 251–257.

[23] G. Sun, C. Zhang, and P. C. Woodland, "Tree-constrained pointer generator for end-to-end contextual speech recognition," in *ASRU*, 2021, pp. 780–787.

[24] P. Harding, S. Tong, and S. Wiesler, "Selective biasing with trie-based contextual adapters for personalised speech recognition using neural transducers," in *INTERSPEECH*, 2023, pp. 256–260.

[25] G. Sun, C. Zhang, and P. C. Woodland, "Minimising biasing word errors for contextual ASR with the tree-constrained pointer generator," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 345–354, 2023.

[26] H. Futami, E. Tsunoo, Y. Kashiwagi, H. Ogawa, S. Arora, and S. Watanabe, "Phoneme-aware encoding for prefix-tree-based contextual asr," in *ICASSP*, 2024, pp. 10 641–10 645.

[27] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: End-to-end contextual speech recognition," in *SLT*, 2018, pp. 418–425.

[28] U. Alon, G. Pundak, and T. N. Sainath, "Contextual speech recognition with difficult negative training examples," in *ICASSP*, 2019, pp. 6440–6444.

[29] M. Jain, G. Keren, J. Mahadeokar, G. Zweig, F. Metze, and Y. Saraf, "Contextual RNN-T for open domain ASR," in *INTERSPEECH*, 2020, pp. 11–15.

[30] M. Han, L. Dong, S. Zhou, and B. Xu, "Cif-based collaborative decoding for end-to-end contextual speech recognition," in *ICASSP*, 2021, pp. 6528–6532.

[31] Z. Liu, T. Li, and P. Zhang, "RNN-T based open-vocabulary keyword spotting in mandarin with multi-level detection," in *ICASSP*, 2021, pp. 5649–5653.

[32] F. Chang, J. Liu, M. Radfar, A. Mouchtaris, M. Omologo, A. Rastrow, and et.al., "Context-aware transformer transducer for speech recognition," in *ASRU*, 2021, pp. 503–510.

[33] M. Han, L. Dong, Z. Liang, M. Cai, S. Zhou, Z. Ma, and et.al., "Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection," in *ICASSP*, 2022, pp. 8532–8536.

[34] T. Munkhdalai, K. C. Sim, A. Chandorkar, F. Gao, M. Chua, T. Strohman, and et.al., "Fast contextual adaptation with neural associative memory for on-device personalized speech recognition," in *ICASSP*, 2022, pp. 6632–6636.

[35] K. Huang, A. Zhang, Z. Yang, P. Guo, B. Mu, T. Xu, and et.al., "Contextualized end-to-end speech recognition with contextual phrase prediction network," in *INTERSPEECH*, 2023, pp. 4933–4937.

[36] M. J. R. Bleeker, P. Swietojanski, S. Braun, and X. Zhuang, "Approximate nearest neighbour phrase mining for contextual speech recognition," in *INTERSPEECH*, 2023, pp. 939–943.

[37] 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 *INTERSPEECH*, 2023, pp. 3257–3261.

[38] B. Naowarat, P. Harding, P. D'Alterio, S. Tong, and B. A. S. Hasan, "Effective training of attention-based contextual biasing adapters with synthetic audio for personalised ASR," in *INTERSPEECH*, 2023, pp. 1264–1268.

[39] Y. Xu, B. Liu, Q. Huang, X. Song, Z. Wu, S. Kang, and et.al., "Cb-conformer: Contextual biasing conformer for biased word recognition," in *ICASSP*, 2023, pp. 1–5.

[40] S. Tong, P. Harding, and S. Wiesler, "Slot-triggered contextual biasing for personalized speech recognition using neural transducers," in *ICASSP*, 2023, pp. 1–5.

[41] M. A. Jalal, P. P. Parada, G. Pavlidis, V. Moschopoulos, K. Saravanan, C. Kontoulis, and et.al., "Locality enhanced dynamic biasing and sampling strategies for contextualASR," in *ASRU*, 2023, pp. 1–8.

[42] T. Xu, Z. Yang, K. Huang, P. Guo, A. Zhang, B. Li, and et.al., "Adaptive contextual biasing for transducer based streaming speech recognition," in *INTERSPEECH*, 2023, pp. 1668–1672.

[43] A. Alexandridis, K. M. Sathyendra, G. P. Strimel, F. Chang, A. Rastrow, N. Susanj, and et.al., "Gated contextual adapters for selective contextual biasing in neural transducers," in *ICASSP*, 2023, pp. 1–5.

[44] X. Fu, K. M. Sathyendra, A. Gandhe, J. Liu, G. P. Strimel, R. McGowan, and et.al., "Robust acoustic and semantic contextual biasing in neural transducers for speech recognition," in *ICASSP*, 2023, pp. 1–5.

[45] Z. Wu, T. Munkhdalai, P. Rondon, G. Pundak, K. C. Sim, and C. Li, "Dual-mode nam: Effective top-k context injection for end-to-end asr," in *INTERSPEECH*, 2023, pp. 221–225.

[46] Y. Sudo, M. Shakeel, Y. Fukumoto, Y. Peng, and S. Watanabe, "Contextualized automatic speech recognition with attention-based bias phrase boosted beam search," in *ICASSP*, 2024, pp. 10 896–10 900.

[47] J. Tang, K. Kim, S. Shon, F. Wu, and P. Sridhar, "Improving asr contextual biasing with guided attention," in *ICASSP*, 2024, pp. 12 096–12 100.

[48] X. Yang, W. Kang, Z. Yao, Y. Yang, L. Guo, F. Kuang, and et.al., "Promptasr for contextualized asr with controllable style," in *ICASSP*, 2024, pp. 10 536–10 540.

[49] X. Shi, Y. Yang, Z. Li, Y. Chen, Z. Gao, and S. Zhang, "Seaco-paraformer: A non-autoregressive asr system with flexible and effective hotword customization ability," in *ICASSP*, 2024, pp. 10 346–10 350.

[50] Y. Higuchi, S. Watanabe, N. Chen, T. Ogawa, and T. Kobayashi, "Mask CTC: non-autoregressive end-to-end ASR with CTC and mask predict," in *INTERSPEECH*, 2020, pp. 3655–3659.

[51] L. Dong and B. Xu, "CIF: continuous integrate-and-fire for end-to-end speech recognition," in *ICASSP*, 2020, pp. 6079–6083.

[52] N. Chen, S. Watanabe, J. Villalba, P. Zelasko, and N. Dehak, "Non-autoregressive transformer for speech recognition," *IEEE Signal Process. Lett.*, vol. 28, pp. 121–125, 2021.

[53] X. Song, Z. Wu, Y. Huang, C. Weng, D. Su, and H. M. Meng, "Non-autoregressive transformer ASR with ctc-enhanced decoder input," in *ICASSP*, 2021, pp. 5894–5898.

[54] J. Nozaki and T. Komatsu, "Relaxing the conditional independence assumption of ctc-based ASR by conditioning on intermediate predictions," in *INTERSPEECH*, 2021, pp. 3735–3739.

[55] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *INTERSPEECH*, 2022, pp. 2063–2067.

[56] L. Qian, H. Zhou, Y. Bao, M. Wang, L. Qiu, W. Zhang, and et.al., "Glancing transformer for non-autoregressive neural machine translation," in *ACL/IJCNLP*, 2021, pp. 1993–2003.

[57] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C. Chiu, and et.al., "Minimum word error rate training for attention-based sequence-to-sequence models," in *ICASSP*, 2018, pp. 4839–4843.

[58] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[59] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2999–3007.

[60] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, and et.al., "Supervised contrastive learning," in *NeurIPS*, 2020.

[61] B. Chen, G. Xu, X. Wang, P. Xie, M. Zhang, and F. Huang, "AISHELL-NER: named entity recognition from chinese speech," in *ICASSP*, 2022, pp. 8352–8356.

[62] 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 *O-COCOSDA*, 2017, pp. 1–5.

[63] Z. Tang, D. Wang, Y. Xu, J. Sun, and et.al., "Kespeech: An open source speech dataset of mandarin and its eight subdialects," in *NeurIPS Datasets and Benchmarks*, 2021.

[64] Z. Gao, Z. Li, J. Wang, H. Luo, X. Shi, M. Chen, and et.al., "Funasr: A fundamental end-to-end speech recognition toolkit," in *INTERSPEECH*, 2023, pp. 1593–1597.

[65] Z. Gao, S. Zhang, M. Lei, and I. McLoughlin, "SAN-M: memory equipped self-attention for end-to-end speech recognition," in *INTERSPEECH*, 2020, pp. 6–10.

[66] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and et.al., "Specaugment: A simple data augmentation method for automatic speech recognition," in *INTERSPEECH*, 2019, pp. 2613–2617.