ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448143

CROSS-MODAL ALIGNMENT FOR END-TO-END SPOKEN LANGUAGE UNDERSTANDING BASED ON MOMENTUM CONTRASTIVE LEARNING

Beida Zheng[1,2],Mijit Ablimit[1,2],Askar Hamdulla[1,2,3,]*

[1]School of Computer Science and Technology, Xinjiang University, Urumqi 830017, China
[2]Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi 830017, China
[3]School of Future Technology, Xinjiang University, Urumqi 830017, China

ABSTRACT

The end-to-end spoken language understanding system extracts the semantic intent directly from an input speech. It effectively avoids problems such as semantic drift in traditional cascade models. However, the lack of semantically labeled speech data makes the model training process difficult. Several recent multi-modal research perspectives have demonstrated that aligning speech and text embeddings based on space distance can improve the model's performance. In this study, inspired by the work related to contrastive learning, a speech and text aligning method using momentum contrast learning is proposed, and a momentum distillation method is also used in the model to learn from imperfectly matched speech and text data. The proposed method has improved intent detection accuracy by 2.14% and 5.98% on Fluent Speech Command and SmartLights datasets.

Index Terms— cross-modal, end to end, spoken language understanding, momentum contrast learning, momentum distillation

1. INTRODUCTION

The main task of spoken language understanding(SLU) is to perform semantic understanding from an input speech signal, such as intent detection, slot prediction[1], etc. SLU is the core of user interaction technologies like voice assistants (such as Siri, Google Home, Xiao Du, etc.). The traditional SLU system consists of an automatic speech recognition system(ASR) cascaded with a natural language understanding system(NLU)[2–4]. Traditional methods have some limitations, such as each module being optimized separately, and the objectives of the separate optimization may not be consistent with the final optimization objective. Secondly, there is the problem of semantic drift; the text recognized by the ASR has errors, which inevitably impact the final task. Then, there are high latency, overly complex models, and an inability to fully utilize the unique rhythmic information in speech[5–7].

The end-to-end(E2E) approach directly performs semantic understanding of the input speech signal, which can effectively avoid the limitations of traditional methods[1, 2, 5, 7]. However, this method needs a lot of high-quality speech data with semantic labels, which is one of the main reasons for the limitations in developing E2E SLU. One solution to this issue is to perform data augmentation or pre-train some modules. For example, [2] attempted to synthesize text data from NLU into speech data, and [6] experimented pre-training the ASR module using additional speech data. However, these data augmentation approaches used in the above methods also add considerable cost of time and computation power.

Another solution is to utilize transcripts of speech. For example, [7] combines pre-training and cross-modal sharing classification layers. [8–10] uses knowledge distillation with a text prediction model as the teacher model, and a speech prediction model as the student model. [11, 12] integrates ASR and NLU into an E2E model through an interface. [13] representation of text as the same frame-level structure as speech to mitigate the need for speech data. [14–17] takes a multi-modal perspective and tries to align speech embeddings and text embeddings based on spatial distance using different loss functions, and the experimental results show that this alignment operation helped to improve the performance of E2E SLU. However, the above methods cannot effectively utilize much non-parallel data, resulting in insufficient model generalization.

Second, although transcribed text can accurately describe speech information, there are many ways of text description for the same semantic speech. Therefore, relying only on one hot label training, all speech-to-text mismatches are penalized, which is detrimental to the overall optimization of the model. After our study, existing E2E SLU studies have not yet considered this issue. How to train effectively from such imperfectly matched speech-to-text data is also a vital point of this task.

We propose a cross-modal momentum contrastive (CMMC) learning E2E SLU architecture. It uses momentum contrast learning to align speech and text embeddings, and momentum distillation to enable the model to learn from imperfectly matched speech and text data. Our main contributions in this paper are as follows:

·We experimentally demonstrate that aligning speech and

*Corresponding author: askar@xju.edu.cn.

text embeddings via momentum contrast loss is effective for E2E SLU.

·We use momentum distillation to make the model learn from imperfectly matched speech and text data.

·The proposed method consistently improves two spoken language understanding datasets, Fluent Speech Commands and SmartLights.

2. RELATED WORK

Recently, helping to understand the semantics of speech using pre-trained language models is one of the essential approaches to improving the performance of E2E SLU models. Embedding the speech and text in spatial distance-based alignment can enhance the performance of E2E SLU understanding[14–17]. Contrastive learning has demonstrated promising performance in aligning multi-modal embeddings [18–20]. Second, studying E2E SLU from only the speech transcript is detrimental to optimizing the whole model. We use momentum distillation[18] to enable the model to learn from imperfectly matched speech and text data.

2.1. Contrastive Learning

Contrastive learning has received much attention recently, and contrastive loss can measure the similarity of sample pairs in an embedded space [20, 21]. For example, [19] uses pairwise contrastive loss to learn the embedding alignment of whole sentences with each image in the same batch during pre-training. [22] designed a multi-task learning framework combining self-supervision and CLIP pre-training. [23] pointed out the two main limitations of contrastive learning, batch size and coding consistency, and proposed momentum contrast learning to address the above limitations.

2.2. Knowledge Distillation

Knowledge distillation aims to make the student model learn from the teacher model, thus improving the student model's ability[24], and most of the methods match the predictions of the two models. Most current teacher models use pre-trained models[8–10]. Online distillation involves training multiple models simultaneously and using their ensemble as a teacher model, a faster and more efficient training method [18]. In this paper, to learn from unmatched speech and text data, we use the momentum distillation[18] to learn from pseudo-labels generated by the momentum model.

3. PROPOSED METHOD

In this section, we describe the proposed CMMC model. Figure 1 shows the overall framework. The whole model architecture consists of five parts: speech encoding module, text encoding module, cross-modal momentum contrast learning module, momentum distillation module, and cross-modal shared classification layer. We use the Conformer[25] to encode the speech to obtain the speech embedding and the pre-trained Bert[26] to encode the transcript to get the text embedding. The text and speech embeddings are then aligned using momentum contrast learning, and learning from imperfectly matched speech and text data uses momentum distillation. Finally, intent detection through cross-modal shared classification layers.

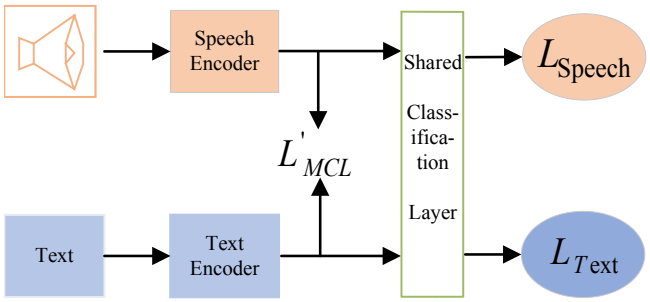

Fig.1.The overall framework of CMMC

Our proposed model optimization has three parts: speech intent classification loss, text intent classification loss, and momentum contrastive loss. Next, we will introduce the components of CMMC in detail.

3.1. Speech Encoding Module

Conformer[25] consists of convolutional downsampling, multi-head attention, and residual modules. It has achieved good results in many tasks related to speech. In CMMC, the extracted speech features FilterBank (Fbank) fed into the Conformer, and maximum pooling of the output of the last Conformer layer in the time dimension yields a fixed-dimension speech embedding.

3.2. Text Encoding Module

Bert[26] has achieved advanced performance in many NLP tasks. We use pre-trained Bert as the encoding module of the text. The output of the last encoder layer of the [CLS] is the text embedding vector.

3.3. Momentum Contrast Learning Module

Inspired by contrast learning [19], we use contrast loss to align speech and text embeddings over spatial distances. The cosine distance to compute the similarity of speech-text embeddings within a batch to obtain a semantic similarity matrix, which then aligns the speech embeddings and text embeddings using infoNCE loss. We jointly learn to optimize the speech encoder and text encoder to maximize the cosine similarity of the matched speech-text pair embedding and minimize the cosine similarity of mismatched speech-text pair embedding. Building on the method proposed in [19], we introduced momentum contrast learning [23]. The whole process of momentum contrast learning in Figure 2:

Momentum contrast learning separates the dictionary size from the batch size by introducing a queue so that the number of negative samples is no longer limited to the batch size. Also, to maintain consistency in encoding, we keep two momentum update encoders for the speech and text. Their updated rules are as follows.

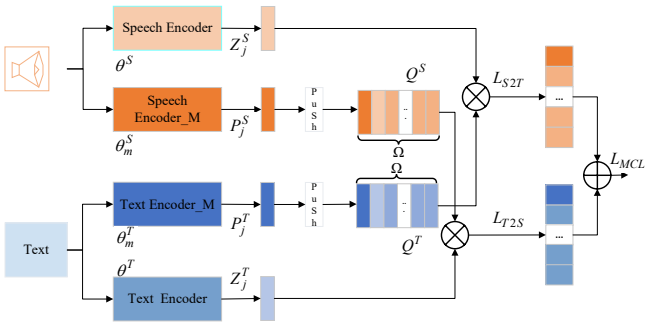

Fig.2.Cross-modal momentum contrastive learning.

$$\theta_m^S = k \cdot \theta_m^S + (1-k) \cdot \theta^S \tag{1}$$

$$\theta_m^T = k \cdot \theta_m^T + (1-k) \cdot \theta^T \tag{2}$$

Where θ^S and θ^T denote the speech encoder and text encoder, θ_m^S and θ_m^T denote the momentum encoders for speech and text. And k is a hyperparameter. It controls the update ratio of two momentum encoders.

The CMMC model has two queues (Q^S and Q^T) to store the speech and text embeddings generated by the momentum encoder. The Ω is the maximum capacity of the queue. The specific iterative update rules for the queue are as follows, set the batch size (B).The data of the nth batch will by θ_m^S and θ_m^T to obtain $N_n^S(N_n^S = \theta_m^S(x_j^S)|x_j^S \in B_n^S)$ and $N_n^T(N_n^T = \theta_m^T(x_j^T)|x_j^T \in B_n^T)$, where x_j denotes the jth piece of data in the nth batch, and S and T denote speech and text. Then pushing N_n^S and N_n^T into Q_n^S and Q_n^T. When the number in the queue reaches Ω, the N_n^S and N_n^T generated afterward, follow the first-in-first-out rule to update the queue.

After obtaining the positive samples $P_j^S(P_j^S = \theta_m^S(x_j^S))$ $P_j^T(P_j^T = \theta_m^T(x_j^T))$ from $\theta_m^S(\theta_m^T)$ that uniquely match the text(speech). For the nth iteration, the constructed speech-to-text similarity loss (L_{S2T}) is the infoNCE loss between the speech embedding $Z_j^S(Z_j^S = \theta^S(x_j^S))$ and all positive/negative samples of Q^T in the queue:

$$L_{S2T} = -\sum_j log \frac{exp(z_j^S \cdot p_j^T/\tau)}{exp(z_j^s \cdot p_j^T/\tau) + \sum_{\mu^T \in Q^T} exp(z_j^S \cdot \mu^T/\tau)} \tag{3}$$

Where μ^T denotes the negative sample embedding of the text that does not match the speech, the τ hyperparameter denotes the temperature, and $\cdot$ denotes the dot product. Similarly, the text-to-speech similarity loss (L_{T2S}) is defined as:

$$L_{T2S} = -\sum_j log \frac{exp(z_j^T \cdot p_j^S/\tau)}{exp(z_j^T \cdot p_j^S/\tau) + \sum_{\mu^S \in Q^S} exp(z_j^T \cdot \mu^S/\tau)} \tag{4}$$

Where μ^S denotes the negative sample embedding of speech that does not match the text. The cross-modal momentum contrast loss (L_{MCL}) is defined as:

$$L_{MCL} = \frac{1}{2}(L_{S2T} + L_{T2S}) \tag{5}$$

3.4. Momentum Distillation Module

The momentum model is an evolving teacher model obtained by averaging the model parameters[18]. During the training process, the main objective is to match the predictions of the base model with those of the momentum model. Specifically, we first compute the speech-to-text and text-to-speech embedding similarity losses (L'_{S2T} and L'_{T2S}) generated by the momentum encoder. It is computing as follows:

$$L'_{S2T} = -\sum_j log \frac{exp(p_j^S \cdot p_j^T/\tau)}{exp(p_j^S \cdot p_j^T/\tau) + \sum_{\mu^T \in Q^T} exp(p_j^S \cdot \mu^T/\tau)} \tag{6}$$

$$L'_{T2S} = -\sum_j log \frac{exp(p_j^T \cdot p_j^S/\tau)}{exp(p_j^T \cdot p_j^S/\tau) + \sum_{\mu^S \in Q^S} exp(p_j^T \cdot \mu^S/\tau)} \tag{7}$$

Then we use Kullback-Leibler(KL) to calculate the distance between the model prediction (L_{S2T} and L_{T2S}) and the pseudo-target (L'_{S2T} and L'_{T2S}) generated by the momentum model. The sum of the KL of the two as the total KL(L_{KL}^{all}):

$$L_{KL}^{all} = KL(L'_{S2T}(S), L_{S2T}(S)) + KL(L'_{T2S}(T), L_{T2S}(T)) \tag{8}$$

The cross-modal momentum contrast loss function(L_{MCL}) is then modified as L'_{MCL}:

$$L'_{MCL} = (1-\alpha) \cdot L_{MCL} + \frac{\alpha}{2} L_{KL}^{all} \tag{9}$$

Where α is a hyperparameter to control the weights between different losses.

3.5. Cross-Modal Sharing Intent Classification Layer

The cross-modal shared intent classification layer is fully connected. The loss function uses the cross-entropy loss. L_S and L_T denote speech and text intent classification loss. L_S and L_T are defined as follows:

$$L_S = CE(y_{pre}^S, y_{true}^S) \tag{10}$$

$$L_T = CE(y_{pre}^T, y_{true}^T) \tag{11}$$

Where CE denotes the cross-entropy loss function, y_{pre}^S and y_{pre}^T denote the intent categories predicted for speech and text. y_{true}^S and y_{true}^T is the true label. the loss function L of the proposed CMMC model consists of embedding align loss L'_{MCL}, speech intent classification loss L_S, and text intent classification loss L_T:

$$L = L'_{MCL} + L_S + L_T \tag{12}$$

4. EXPERIMENT

4.1. Dataset

We train and evaluate our proposed CMMC model using two publicly available and commonly used datasets.

Fluent Speech Commands (FSC) is one of the largest English datasets proposed by [27] for training E2E SLU models. It consists primarily of spoken commands from the virtual assistant. Each order includes three slots for action, object, and location.

SmartLights[28] is a smaller dataset of spoken English language understanding, making the prediction task more challenging. It is about spoken commands for smart light-related operations. The details of the above two data sets are in Table 1:

Table 1. Introduction of the dataset.

Datasets	FSC	SmartLights
Number of Speakers	97	52
Number of train audio files	23132	1328
Number of valid audio files	3118	166
Number of test audio files	3793	166
Number of intents	31	6
Total duration[hours]	19	5.5
Average length[seconds]	2.3	3.4

4.2. Train and Related Hyperparameter Details

We use two layers of Conformer[25] as speech encoder, with 512 hidden cells per layer. The text encoder uses 12 layers Bert pre-trained in [26], Where the shared classification is a fully connected layer with a received input of 768.

We used the Adam optimizer[29]. The speech encoder learning rate initially to 1e-3. For the FSC dataset, the text encoder learning rate is 3e-5, and for SmartLights, the text encoder is 5e-5. The batch size for training is 32, the epoch for FSC training is 20, and the epoch for SmartLights training is 50. The queue size Ω is 65536, α is 0.4, and the momentum parameter k is 0.994.

4.3. Experimental Results

The first results of the experiments are on the FSC dataset, as shown in Table 2:

Table 2. CMMC compared to previous work on FSC.

	E2E	Multi-modal	FSC(%)
Cao et al. [1]	$\surd$	$\times$	99.00
Finstreder (Quartznet) [3]	$\times$	$\times$	99.20
FANS[5]	$\surd$	$\times$	99.00
Won et al.[10]	$\surd$	$\surd$	98.98
Bhuvan et al.[16]	$\surd$	$\surd$	97.65
Loren et al.[27]	$\surd$	$\times$	98.80
Mohamed et al.[30]	$\surd$	$\times$	97.82
Cha, et al.[31]	$\surd$	$\surd$	99.18
Reptile[32]	$\surd$	$\times$	99.20
CMMC(ours)	$\surd$	$\surd$	**99.24**

We mainly compare with three types of models: conventional model (ASR+NLU), E2E (without transcript), and E2E transcript binding (Multi-modal). The results show in Table 2. The results show that our model achieves 99.24% in intent classification accuracy, which is very competitive with the main models nowadays.

Secondly, to verify the generalizability of our proposed model, we conducted experiments on the more challenging SmartLights dataset, and the results show in Table 3.

As with the FSC dataset, the comparison models chosen were the three abovementioned models. As can be seen from the results in Table 3, the intention classification accuracy of our proposed method on the SmartLights dataset is 77.84%, and our model's competitiveness still exists, indicating the effectiveness of our proposed method.

Table 3. CMMC compared to previous work on SmartLights.

	E2E	Multi-modal	SmartLights(%)
Lugosch et al.[2]	$\surd$	$\times$	71.40
Bermuth[4]	$\times$	$\times$	54.50
Bhuvan et al.[16]	$\surd$	$\surd$	74.10
Chao et al.[33]	$\times$	$\times$	67.98
CMMC(ours)	$\surd$	$\surd$	**77.84**

4.4. Ablation Experiment

Table 4. Baseline and ablation results.

Model	FSC(%)	SmartLights(%)
Baseline	97.10	71.86
Multi-modal	98.52	73.65
+CL	98.71	74.63
+MCL	99.08	75.46
+MD	99.24	77.84

To further validate the contribution of our proposed model, we trained a baseline E2E SLU model using only speech data and performed ablation experiments based on it step by step. The results show in Table 4. See 4.2 for all the details of the baseline model. Table 4 shows consistent improvement trends for both datasets. The introduction of transcript information improves the intention classification accuracy. With the addition of contrast loss and momentum contrast loss further improving the intent classification, it reflects that spatially aligning speech embedding and text embedding can enhance the performance of E2E SLU. Compared to learning from uniquely matched speech or text datasets using contrast learning or momentum contrast learning, momentum distillation further improves the intention classification accuracy on both datasets, showing that our proposed learning from imperfectly matched speech and text data is effective. Our proposed method improves by 2.14% and 5.98% relative to the baseline on both datasets.

5. CONCLUSION

This study presents the CMMC model, a multi-modal architecture for E2E SLU tasks. We use Conformer as a speech encoder and Bert as a text encoder to learn alignment of speech and text embeddings through momentum contrast. Proposed models improved by learning from imperfectly matched speech and text data using momentum distillation. Through a series of ablation experiments, the experimental results on two public datasets demonstrate the effectiveness of our proposed method. In the future, we will extend our approach to joint task models (intents detection and slot filling) for E2E SLU.

6. ACKNOWLEDGEMENT

This research received funding from the Natural Science Foundation of China (62341607).

References

[1] Y. Cao, N. Potdar, and A.R. Avila, "Sequential end-to-end intent and slot label classification and localization," *arXiv preprint arXiv:2106.04660,2021.*

[2] L. Lugosch, B.H. Meyer, et al., "Using speech synthesis to train end-to-end spoken language understanding models," in *ICASSP*, 2020, pp. 8499–8503.

[3] D. Bermuth, A. Poeppel, and W. Reif, "Finstreder: simple and fast spoken language understanding with finite state transducers using modern speech-to-text models," *arXiv preprint arXiv:2206.14589,2022*.

[4] D. Bermuth and A.and others Poeppel, "Jaco: An offline running privacy-aware voice assistant," in *HRI,2022*, pp. 618–622.

[5] M. Radfar, A. Mouchtaris, S. Kunzmann, et al., "Fans: Fusing asr and nlu for on-device slu," *arXiv preprint arXiv:2111.00400,2021*.

[6] R. Price, "End-to-end spoken language understanding without matched language speech model pretraining data," in *ICASSP,2020*, pp. 7979–7983.

[7] M. Kim, G. Kim, S-W. Lee, et al., "St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding," in *ICASSP,2021*, pp. 7478–7482.

[8] P. Denisov and N. T. Vu, "Pretrained semantic speech embeddings for end-to-end spoken language understanding via cross-modal teacher-student learning," *arXiv preprint arXiv:2007.01836,2020.*

[9] S. Kim, G. Kim, S. Shin, et al., "Two-stage textual knowledge distillation for end-to-end spoken language understanding," in *ICASSP,2021*, pp. 7463–7467.

[10] W. I. Cho, D. Kwak, J. W. Yoon, et al., "Speech to text adaptation: Towards an efficient cross-modal distillation," *arXiv preprint arXiv:2005.08213,2020*.

[11] M. Rao, P. Dheram, G. Tiwari, et al., "Do as i mean, not as i say: Sequence loss training for spoken language understanding," in *ICASSP,2021*, pp. 7473–7477.

[12] S. Seo, D. Kwak, and B. Lee, "Integration of pre-trained networks with continuous token interface for end-to-end spoken language understanding," in *ICASSP,2022*, pp. 7152–7156.

[13] S. Thomas, G. Saon, et al., "Towards reducing the need for speech training data to build spoken language understanding systems," in *ICASSP,2022*, pp. 7932–7936.

[14] W. Wang, S. Ren, Y. Qian, et al., "Optimizing alignment of speech and language latent spaces for end-to-end speech recognition and understanding," in *ICASSP,2022*, pp. 7802–7806.

[15] Y. Huang, H.-K. Kuo, S. Thomas, et al., "Leveraging unpaired text data for training end-to-end speech-to-intent systems," in *ICASSP,2020*, pp. 7984–7988.

[16] B. Agrawal, M. Müller, S. Choudhary, et al., "Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding," in *ICASSP*, 2022, pp. 7157–7161.

[17] C.-I. Lai, y-s. Chuang, et al., "Semi-supervised spoken language understanding via self-supervised speech and language model pretraining," in *ICASSP,2021*, pp. 7468–7472.

[18] J. Li, R. Selvaraju, et al., "Align before fuse: Vision and language representation learning with momentum distillation," *NIPS,2021*, vol. 34, pp. 9694–9705.

[19] S. Mo, J. Xia, and I. Markevych, "Cavl: Learning contrastive and adaptive representations of vision and language," *arXiv preprint arXiv:2304.04399,2023*.

[20] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[21] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR,2006*, 2006, vol. 2, pp. 1735–1742.

[22] N. Mu, A. Kirillov, et al., "Slip: Self-supervision meets language-image pre-training," in *European Conference on Computer Vision*, 2022, pp. 529–544.

[23] K. He, H. Fan, Y. Wu, et al., "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[25] A. Gulati, J. Qin, C.-C. Chiu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[26] T. Wolf, L. Debut, V. Sanh, et al., "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[27] L. Lugosch, M. Ravanelli, P. Ignoto, et al., "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[28] A. Coucke, A. Saade, A. Ball, et al., "Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces," *arXiv preprint arXiv:1805.10190*, 2018.

[29] D. P Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] M. Mhiri, S.l Myer, and V. S. Tomar, "A low latency asr-free end to end spoken language understanding system," *arXiv preprint arXiv:2011.04884*, 2020.

[31] S. Cha, W. Hou, et al., "Speak or chat with me: End-to-end spoken language understanding system with flexible inputs," *arXiv preprint arXiv:2104.05752*, 2021.

[32] Y. Tian and P. J. Gorinski, "Improving end-to-end speech-to-intent classification with reptile," *arXiv preprint arXiv:2008.01994*, 2020.

[33] C.-W. Huang and Y.-N. Chen, "Learning asr-robust contextualized embeddings for spoken language understanding," in *ICASSP,2020*, pp. 8009–8013.