

Calibrate and Refine! A Novel and Agile Framework for ASR-error Robust Intent Detection

Peilin Zhou^{1,*}, Dading Chong², Helin Wang², Qingcheng Zeng¹

¹Zhejiang University, Hangzhou, China

²Peking University, Shenzhen, China

zhoupalin@gmail.com, 1601213984@pku.edu.cn, wanghl15@pku.edu.cn, qingchengzeng@outlook.com

Abstract

The past ten years have witnessed the rapid development of text-based intent detection, whose benchmark performances have already been taken to a remarkable level by deep learning techniques. However, automatic speech recognition (ASR) errors are inevitable in real-world applications due to the environment noise, unique speech patterns and etc, leading to sharp performance drop in state-of-the-art text-based intent detection models. Essentially, this phenomenon is caused by the semantic drift brought by ASR errors and most existing works tend to focus on designing new model structures to reduce its impact, which is at the expense of versatility and flexibility. Different from previous one-piece model, in this paper, we propose a novel and agile framework called CR-ID for ASR error robust intent detection with two plug-and-play modules, namely semantic drift calibration module (SDCM) and phonemic refinement module (PRM), which are both model-agnostic and thus could be easily integrated to any existing intent detection models without modifying their structures. Experimental results on SNIPS dataset show that, our proposed CR-ID framework achieves competitive performance and outperform all the baseline methods on ASR outputs, which verifies that CR-ID can effectively alleviate the semantic drift caused by ASR errors.

Index Terms: intent detection, human-computer interaction, spoken language understanding

1. Introduction

Intent detection (ID), as one of the key tasks in spoken language understanding, aims to identify users' intents from their utterances. Driven by advances in deep learning technology, the ID research has entered into a stage of rapid development. Specifically, many classical methods like convolutional neural network (CNN) [1, 2, 3], recurrent neural network (RNN) [4, 5, 6], graph neural network (GNN) [7] and self-attention mechanism [8, 9] have been explored for this task and obtained superb performance on benchmark datasets. Moreover, pre-trained language models [10] have also been utilized to better understand the meaning of the user sentences and thus could help to classify the intents more accurately. Notwithstanding the favorable results of these models, they often assume that automatic speech recognition (ASR) never makes any mistakes. The training and testing of these ID models are both conducted on error-free manual transcriptions, rather than ASR outputs.

Unfortunately, this overly idealistic setting would make it hard to deploy existing ID models in real-world applications, where ASR errors are unavoidable due to the complex conditions like environment noise and diverse speaking styles or accents. As shown in the right side of Figure 1, although pre-

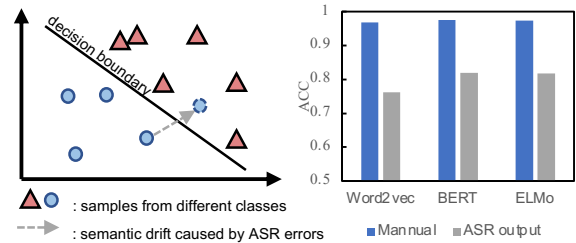


Figure 1: *Semantic drift problem (Left) and the comparison of different ID models' performance on manual transcriptions and ASR outputs (Right)*

trained language models (LMs) like BERT [11] and ELMo [12] could provide more robust representations compared with static embeddings like Word2Vec [13], they still suffer from sharp performance drop when tested on ASR outputs. This is because the original representations of user utterances are prone to be distorted by ASR errors (as shown in the left side of Figure 1), which is named as semantic drift problem in this paper.

Recently, several studies were introduced to mitigate such semantic drift problem. [14, 15, 16] proposed to remove the ASR component and extract semantics directly from the speech signals in an end-to-end manner. Following this trend, [17] applied the mask strategy to audio frames and utilized large-scale unsupervised pre-training technique to learn acoustic representations for SLU. However, compared with pipeline-based methods, these end-to-end models are less interpretable and more data-hungry. In addition, the annotation process of audio data is usually both expensive and time-consuming, which is impractical for industrial applications. Therefore, some researchers proposed to utilize text and speech together for SLU systems. For example, [18] proposed a novel ST-BERT and designed two new cross-modal language modeling tasks to better learn the semantic representations of speech and text modalities. [19] suggested to carry out both speech and language understanding tasks during pre-training and constructed a novel speech-language joint pre-training framework for SLU. Though achieving excellent performance, they still require pre-training with large-scale datasets, which are not available in some data scarce domains.

Another branch of ASR-error robust research is to reduce the impact of semantic drift by considering the acoustic similarity between words [20] or directly injecting phoneme information to the modeling process [21], which had a similar motivation with ours. But most of them only focused on designing new model structures for specific scenario, and usually show poor compatibility with other methods. So far, designing a both versatile and flexible model has still not been well explored in

* Corresponding Author.

this research field.

To overcome above-mentioned limitations, we propose a novel and agile framework called Calibration and Refinement for Intent Detection (CR-ID). Different from previous solutions, our approach decouples the semantic calibration and intent classification process, thus any existing text-based intent detection models could be incorporated into this framework and become more robust to ASR errors. Specifically, we design two plug-and-play modules to calibrate the semantic drift and refine the calibrated representation with phonemic information, which provides useful signals for the intent classification process. Our proposed framework will be further detailed in section 2 and our main contributions could be summarized as follows:

- We propose the CR-ID framework, which could effectively reduce the impact of semantic drift on existing text-based intent detection models without any structural modifications.
- We design two plug-and-play modules, namely SDCM and PRM, to calibrate both word-level and sentence-level representation for ASR outputs and utilize the phonemic information to refine and enrich the calibrated representations.
- We conduct comprehensive experiments on SNIPS dataset and the results show that, compared with the best baseline model, the intent accuracy and Macro-F1 score of our proposed CR-ID are increased by 1.99% and 1.86% respectively, which demonstrates the effectiveness of CR-ID on boosting the robustness of existing ID model.

2. The Proposed Approach

In this section, we present our CR-ID, which is able to effectively and flexibly alleviate the semantic drift problem without changing the structure of classical text-based ID models. The overall architecture of CR-ID is illustrated in Figure 2, consisting of three main modules: *Semantic Drift Calibration Module* (Sec. 2.1), *Phonemic Refinement Module* (Sec. 2.2), and *Intent Detection Module* (Sec. 2.3).

2.1. Semantic Drift Calibration Module

SDCM aims to calibrate the distorted representations of ASR outputs and minimize the negative impacts brought by semantic drift. To achieve this, inspired by the great success of pretrained language models (PLM) and finetuning techniques, we propose to adopt two PLM finetuning strategies, namely confusion-aware finetuning and task-adaptive finetuning, which are transformed from [20] and [22].

For confusion-aware finetuning, we first use both minimum edit-distance (MED) and word confusion network (WCN) to extract acoustic confusion, which are introduced by [20]. Due to space limitation, readers could check the details from their paper. Taking the two different utterances x_1 and x_2 as an example, we use $C = \{c_1, c_2, \dots, c_{|C|}\}$ to denote the set of all acoustic confusions, where $c = \{w_{t_1}^{x_1}, w_{t_2}^{x_2}\}$ consists of two acoustically similar words $w_{t_1}^{x_1}$ and $w_{t_2}^{x_2}$. Finally, we propose a new confusion loss to minimize the mean square error (MSE) between the word-level representations and sentence-level representations generated by pretrained language model as follows:

$$L_{ca} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^1 \text{MSE}(h_{t_1,i}^{x_1}, h_{t_2,i}^{x_2}) + \text{MSE}(h^{x_1}, h^{x_2}) \quad (1)$$

Task-adaptive finetuning is a widely used technique especially when domain mismatch problem happens, which could effectively adapt pretrained LM from general corpus to the target data. For example, given a pre-trained ELMo model and a sentence $x = \{w_1, w_2, \dots, w_{|x|}\}$, we could directly use the pretraining loss of ELMo as the task-adaptive loss, which could be written as:

$$L_{ta} = \frac{1}{|x|} \sum_{t=1}^{|x|} -\log p(w_t | w_{<t}) - \log p(w_t | w_{>t}) \quad (2)$$

where $p(w_t | w_{<t})$ and $p(w_t | w_{>t})$ denote the probabilities of w_t calculated from forward and backward directions.

Eventually, we jointly finetune the LM using above-mentioned two strategies in a multi-task learning manner and the final loss is as follows:

$$L = L_{ta} + \lambda L_{ca} \quad (3)$$

where λ represents a balancing hyperparameter to control the contribution of each finetuning strategy.

2.2. Phonemic Refinement Module

Phoneme is the smallest pronunciation unit in speech and the phoneme sequence of each word can represent its acoustic information to some extent. Therefore, we design PRM to refine and enrich the calibrated representation by injecting phonemic information into the modeling process.

Firstly, each word w_t in the ASR output x_{asr} will be transformed into a phoneme sequence $P_t = \{p_1, p_2, \dots, p_{N_{w_t}}\}$ via a grapheme to phoneme (G2P) conversion algorithm [23], which highly depends on the pronunciation dictionary. In this paper we adopt CMU pronunciation dictionary [24] constructed by Carnegie Mellon University, which includes 39 types of phonemes and covers more than 130,000 words as well as their corresponding pronunciation information. Figure 2 also shows the process of converting the word "find" into a phoneme sequence "F,AY1,N,D". Note that for vowels like "AY", there is a stress marker behind them indicating which stress types it belongs to. Generally, "0" represents no stress, "1" and "2" represent primary stress and secondary stress respectively, which could provide fine-grained acoustic information for intent detection process. Then, each phoneme sequence will be mapped into the embedding space and be further encoded by a BiLSTM layer as follows:

$$H_{w_t} = BiLSTM([e_{p_1}, e_{p_2}, \dots, e_{N_{w_t}}]), \quad (4)$$

where e_{p_i} denotes the embedding of the phoneme p_i and H_{w_t} represents the hidden representation matrix for word w_t .

In the end, average pooling method is conducted on these hidden representation matrices to obtain the final acoustic embedding of the whole sentence x_i :

$$H_{x_{asr}}^{acoustic} = [h_{w_1}, h_{w_2}, \dots, h_{w_N}], \quad (5)$$

$$h_{w_t} = Average(H_{w_t}), \quad (6)$$

2.3. Intent Detection Module

As shown in Figure 2, the intent detection module is decoupled from other modules, making it possible to incorporate any existing text-based intent detection model to our proposed CR-ID

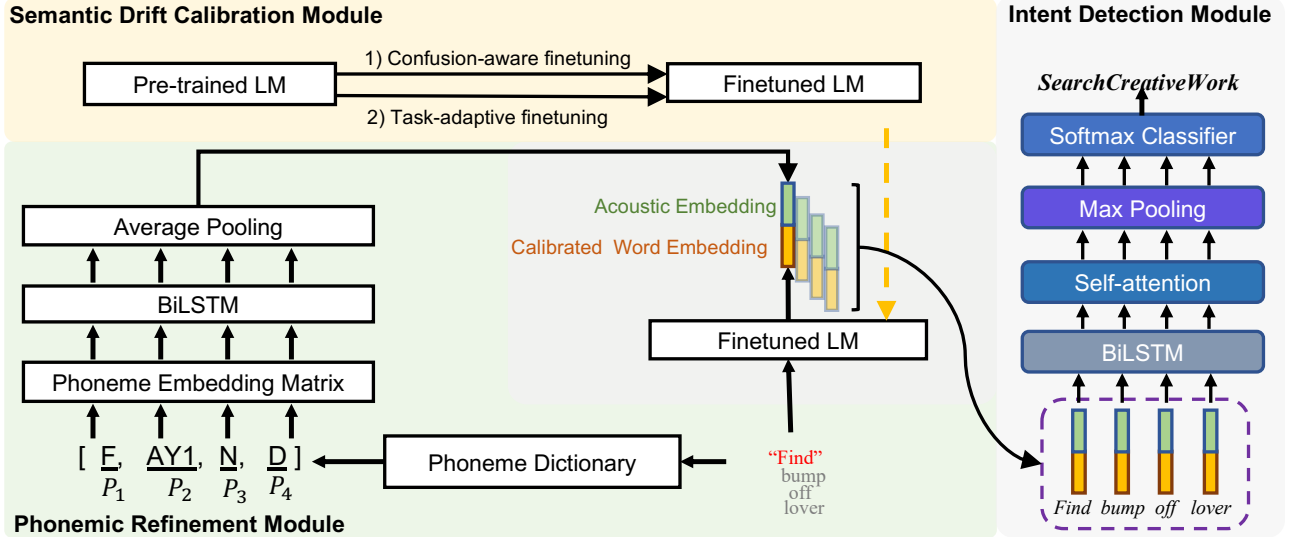


Figure 2: Overview of the proposed CR-ID framework.

Table 1: Overall performance on manual and ASR output. Bold scores represent the highest results of all methods.

Model	Manual		ASR output	
	ACC%	Macro-F1%	ACC%	Macro-F1%
Random	96.87	96.91	78.60	79.87
GloVe	97.15	97.18	77.12	77.70
Word2Vec	96.73	96.81	76.14	77.02
FastText	97.01	97.00	79.15	79.48
BERT (w/o Fine-tuning)	96.43	96.44	80.40	80.81
BERT (w Fine-tuning)	97.59	97.70	82.01	82.80
ELMo (w/o Fine-tuning)	96.70	96.77	80.60	81.61
ELMo (w Fine-tuning)	97.28	97.31	81.69	82.24
SpokenVec (MED)	97.01	97.21	88.52	89.23
SpokenVec (WCN)	97.04	97.12	89.55	89.97
CR-ID (MED)	97.42	97.50	90.85	91.32
CR-ID (WCN)	97.14	97.23	91.54	91.83

framework. Therefore, we adopt a self-attentive intent classification model inspired by [8] as the ID module. The input of ID module is the concatenation of the calibrated word embedding generated by SDCM and the acoustic embedding generated by PRM. BiLSTM is used to model the long-term dependency in the utterance and the self attention mechanism is adopted to capture the key information from the calibrated and refined representations. Max pooling is utilized to obtain the final sentence-level representation, which is further fed to a softmax classifier to predict user’s intent.

3. Experiments

3.1. Dataset

In [20], the authors used three datasets, namely SNIPS, ATIS and Smartlight, for their experiments. However, both ATIS (with confusion words) and Smartlight are not available for public because of the copyright issue. Therefore, for fair comparison with the method proposed in [20], we directly use their released version of SNIPS dataset to conduct all the experiments. Different from the original SNIPS dataset, [20] extracted con-

fusion words via the two strategies introduced in Sec2.1 and added them to the original dataset, which is convenient for researchers to reproduce their results and make improvements on it. The readers could check the details of this dataset in <https://github.com/MiuLab/SpokenVec>.

3.2. Baselines and Implementation Details

A number of ASR error robust ID models have been proposed in the past few years. We do not compare with all of them because many previous methods are not directly comparable due to the use of different model architectures. Hence, we select SpokenVec [20] and construct several baselines that are fair (use the same information, similar architectures, etc.) to compare with. Specifically, we use the intent detection module (introduced in Sec 2.3) as the base model, because the self-attentive intent detection model has already achieved comparative performance on SNIPS dataset according to [25]. Then we incorporate it with different word embedding techniques as the baselines.

Static Word Embedding. We use three pre-trained static word embeddings, Word2Vec [13], GloVe [26] and FastText [27], as the embedding matrix to help encode sentences. We also use a randomly initialized embedding matrix as a comparison.

Contextual Word Embedding. We evaluate two pretrained language models, ELMo and BERT, to obtain contextual word embeddings. And each LM is evaluated with fixed and unfixed parameters respectively.

Implementation Details. For the ID base model, the dimension of BiLSTM and self-attention layer are all set to 300, the number of heads is set to 8, the batch size is 64. All ID base models are trained on the manual transcribed training set for 50 epochs using Adam optimizer with learning rate as $3e-4$, and then tested on the manual transcriptions and ASR outputs respectively. For the SDCM, for fair comparison with SpokenVec, we follow its setting and adopt ELMo as the pretrained LM. We train the SDCM for 10 epochs with the batch size of 32, learn rate of Adam set to $1e-4$. For PRM, the dimension of the embedding and BiLSTM hidden layers are all set to 50 and the PRM is jointly trained with ID base model.

Table 2: Ablation study

Model	Manual		ASR output	
	ACC%	Macro-F1%	ACC%	Macro-F1%
Full	97.14	97.23	91.54	91.83
w/o acoustic embedding	96.71	96.81	90.55	90.87
w/o confusion-aware finetuning strategy	97.15	97.21	88.60	89.01
w/o task-adaptive finetuning strategy	97.28	97.41	82.12	83.14

Table 3: Performance comparison of using different distance functions as the confusion loss. The metric is accuracy.

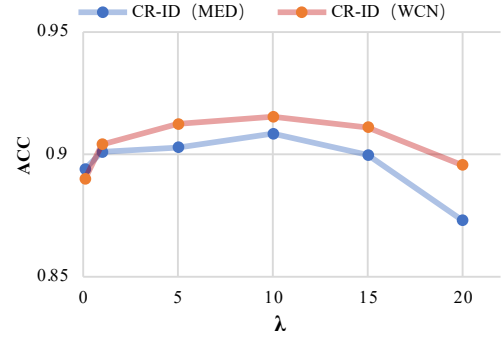
Test data type	Confusion extraction type	Type of loss function			
		Cosine	L1	MSE	Triplet
ASR	MED	90.56	84.90	90.85	88.88
ASR	WCN	90.82	88.26	91.54	90.10
Manual	MED	96.55	92.71	97.42	96.40
Manual	WCN	97.30	96.86	97.14	96.73

3.3. Results and Analysis

The overall performance of the baselines and CR-ID are summarized in Table 1. Note that as introduced in Sec 2.1 and Sec 3.1, the confusion word pairs could be generated by minimum edit distance (MED) or word confusion network (WCN). Hence, for SpokenVec and our proposed CR-ID, we also report the performance variations using different confusion extraction methods in Table 1. Here are some observations from the Table 1: when testing on the manual transcriptions, the performance scores of all methods are very close, and the method based on contextual word embedding is slightly better than the static counterparts. However, the performances of all baselines except for SpokenVec drops sharply on the ASR output, demonstrating the necessity to reduce the negative impacts caused by semantic drift problem. CR-ID (WCN) achieved the best performance in terms of both Accuracy and Macro-F1. Specifically, compared with the best static word embedding based baselines, the Accuracy and Macro-F1 of CR-ID (WCN) are increased by 12.39% and 11.96% respectively; compared with the best contextual word embedding based baseline model, the performance are improved by 9.5% and 9% respectively; even compared with the SpokenVec, which is a very strong baseline, the performance gains still achieve 1.99% and 1.86% respectively, demonstrating the effectiveness of our propose CR-ID framework.

3.4. Ablation Study

In order to figure out the contribution of different modules in our proposed CR-ID, we conduct ablation study for each plug-and-play module, as shown in Table.2: 1) CR-ID w/o acoustic embedding, which only use SDCM for calibration; 2) CR-ID w/o confusion-aware finetuning strategy, where only task-adaptive finetuning and PRM are reserved. 3) CR-ID w/o task-adaptive finetuning strategy, where only confusion-aware finetuning and PRM are reserved. We observe that all these components contribute to performance improvements when testing on the ASR outputs. Specifically, task adaptive fine-tuning strategy contributes the most to the robustness of ID module. When this strategy is removed from the CR-ID, the accuracy and Macro-F1 are decreased by 9.42% and 8.69% respectively. And the confusion-aware finetuning strategy take the second place. Without it, the accuracy and Macro-F1 are decreased by 2.94% and 2.82% respectively. Without acoustic embedding, the accuracy and Macro-F1 are decreased by 0.99% and 0.96%

Figure 3: Parameter sensitivity of λ

respectively. Therefore, the combination of SDCM and PRM could significantly improve the robustness of ID module to ASR errors.

In addition, we also explore the effect of different distance functions in the confusion loss on model’s performance. Here we select three classical distance functions (Equation 7,8,9) to substitute the MSE in Equation 1, and the results are shown in Table 3. We observe that MSE achieves the best performance under most experimental settings, which is the reason why we finally choose MSE distance for the confusion-aware finetuning strategy.

$$L_{\cos} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^1 \left(1 - \frac{h_{t_1, i}^{x_1} \cdot h_{t_2, i}^{x_2}}{\|h_{t_1, i}^{x_1}\| \|h_{t_2, i}^{x_2}\|} \right) + \left(1 - \frac{h^{x_1} \cdot h^{x_2}}{\|h^{x_1}\| \|h^{x_2}\|} \right) \quad (7)$$

$$L_{l1} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^1 \|h_{t_1, i}^{x_1} - h_{t_2, i}^{x_2}\|_1 + \|h^{x_1} - h^{x_2}\|_1 \quad (8)$$

$$L_{\text{triplet}} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^1 \text{triplet}(h_{t_1, i}^{x_1}, h_{t_2, i}^{x_2}, h_{t_3, i}^{x_3}) + \text{triplet}(h^{x_1}, h^{x_2}, h^{x_3}) \quad (9)$$

$$\text{triplet}(a, p, n) = \max\{d(a, p) - d(a, n) + \text{margin}, 0\}$$

$$d(x_i, y_i) = \|x_i - y_i\|_p$$

3.5. Hyperparameter Sensitivity

In this section, we aim to analyze the effect of the balancing hyperparameter λ (in the Equation 3) on the performance of CR-ID. The results are illustrated in Figure 3. It can be observed that for both CR-ID (MED) or CR-ID (WCN), when λ increases from 0.1 to 10, the model performance is slightly improved, but when the λ gets larger (e.g. larger than 15), the performance of the model begins to decline. Therefore, for all the CR-ID related experiments, we set λ to 10 to better balance the impact of task-adaptive finetuning and confusion-aware finetuning on model optimization.

4. Conclusion

In this paper, we propose a novel and agile framework, called CR-ID, for ASR error robust intent detection. Two plug-and-play modules, namely SDCM and PRM, are designed to calibrate both word-level and sentence-level representation for ASR outputs and utilize the phonemic information to refine and enrich the calibrated representations. Experimental results on SNIPS dataset show that our proposed CR-ID outperform all baseline models on the ASR outputs, demonstrating that our proposed framework could effectively reduce the impact of semantic drift on existing text-based intent detection models and boost their robustness to ASR errors.

5. References

- [1] G. Tür, L. Deng, D. Hakkani-Tür, and X. He, “Towards deeper understanding: Deep convex networks for semantic utterance classification,” in *ICASSP*. IEEE, 2012, pp. 5045–5048.
- [2] P. Xu and R. Sarikaya, “Convolutional neural network based triangular CRF for joint intent detection and slot filling,” in *ASRU*. IEEE, 2013, pp. 78–83.
- [3] C. Zhang, W. Fan, N. Du, and P. S. Yu, “Mining user intentions from medical queries: A neural network based heterogeneous jointly modeling approach,” in *WWW*. ACM, 2016, pp. 1373–1384.
- [4] S. V. Ravuri and A. Stolcke, “Recurrent neural network and LSTM models for lexical utterance classification,” in *INTERSPEECH*. ISCA, 2015, pp. 135–139.
- [5] B. Liu and I. Lane, “Attention-based recurrent neural network models for joint intent detection and slot filling,” in *INTERSPEECH*. ISCA, 2016, pp. 685–689.
- [6] Y. Wang, Y. Shen, and H. Jin, “A bi-model based RNN semantic frame parsing model for intent detection and slot filling,” in *NAACL-HLT (2)*. Association for Computational Linguistics, 2018, pp. 309–314.
- [7] J. Hu, G. Wang, F. H. Lochovsky, J. Sun, and Z. Chen, “Understanding user’s query intent with wikipedia,” in *WWW*. ACM, 2009, pp. 471–480.
- [8] S. Yolchuyeva, G. Németh, and B. Gyires-Tóth, “Self-attention networks for intent detection,” in *RANLP*. INCOMA Ltd., 2019, pp. 1373–1379.
- [9] M. Chen, J. Zeng, and J. Lou, “A self-attention joint model for spoken language understanding in situational dialog applications,” *CoRR*, vol. abs/1905.11393, 2019.
- [10] Q. Chen, Z. Zhuo, and W. Wang, “BERT for joint intent classification and slot filling,” *CoRR*, vol. abs/1902.10909, 2019.
- [11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT (1)*. Association for Computational Linguistics, 2019, pp. 4171–4186.
- [12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *NAACL-HLT*. Association for Computational Linguistics, 2018, pp. 2227–2237.
- [13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *ICLR (Workshop Poster)*, 2013.
- [14] Y. Chen, R. Price, and S. Bangalore, “Spoken language understanding without speech recognition,” in *ICASSP*. IEEE, 2018, pp. 6189–6193.
- [15] P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. J. Moreno, R. Prabhavalkar, Z. Qu, and A. Waters, “From audio to semantics: Approaches to end-to-end spoken language understanding,” in *SLT*. IEEE, 2018, pp. 720–726.
- [16] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, “Speech model pre-training for end-to-end spoken language understanding,” in *INTERSPEECH*. ISCA, 2019, pp. 814–818.
- [17] P. Wang, L. Wei, Y. Cao, J. Xie, and Z. Nie, “Large-scale unsupervised pre-training for end-to-end spoken language understanding,” in *ICASSP*. IEEE, 2020, pp. 7999–8003.
- [18] M. Kim, G. Kim, S. Lee, and J. Ha, “St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding,” in *ICASSP*. IEEE, 2021, pp. 7478–7482.
- [19] Y. Chung, C. Zhu, and M. Zeng, “SPLAT: speech-language joint pre-training for spoken language understanding,” in *NAACL-HLT*. Association for Computational Linguistics, 2021, pp. 1897–1907.
- [20] C. Huang and Y. Chen, “Learning asr-robust contextualized embeddings for spoken language understanding,” in *ICASSP*. IEEE, 2020, pp. 8009–8013.
- [21] Q. Chen, W. Wang, and Q. Zhang, “Pre-training for spoken language understanding with joint textual and phonetic representation learning,” in *Interspeech*. ISCA, 2021, pp. 1244–1248.
- [22] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” in *ACL (1)*. Association for Computational Linguistics, 2018, pp. 328–339.
- [23] J. Park, Kyubyong Kim, “g2pe,” <https://github.com/Kyubyong/g2p>, 2019.
- [24] J. Kominek and A. W. Black, “The CMU arctic speech databases,” in *SSW*. ISCA, 2004, pp. 223–224.
- [25] L. Qin, T. Xie, W. Che, and T. Liu, “A survey on spoken language understanding: Recent advances and new frontiers,” in *IJCAI*. ijcai.org, 2021, pp. 4577–4584.
- [26] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*. ACL, 2014, pp. 1532–1543.
- [27] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, “Fasttext.zip: Compressing text classification models,” *CoRR*, vol. abs/1612.03651, 2016.