SEMI-SUPERVISED SPOKEN LANGUAGE UNDERSTANDING VIA SELF-SUPERVISED SPEECH AND LANGUAGE MODEL PRETRAINING

Cheng-I Lai*, Yung-Sung Chuang†, Hung-Yi Lee‡, Shang-Wen Li‡, James Glass*

* MIT Computer Science and Artificial Intelligence Laboratory
† College of Electrical Engineering and Computer Science, National Taiwan University ‡ Amazon AI

ABSTRACT

Much recent work on Spoken Language Understanding (SLU) is limited in at least one of three ways: models were trained on oracle text input and neglected ASR errors, models were trained to predict only intents without the slot values, or models were trained on a large amount of in-house data. In this paper, we propose a clean and general framework to learn semantics directly from speech with semi-supervision from transcribed or untranscribed speech to address these issues. Our framework is built upon pretrained end-to-end (E2E) ASR and self-supervised language models, such as BERT, and fine-tuned on a limited amount of target SLU data. We study two semi-supervised settings for the ASR component: supervised pretraining on transcribed speech, and unsupervised pretraining by replacing the ASR encoder with self-supervised speech representations, such as wav2vec. In parallel, we identify two essential criteria for evaluating SLU models: environmental noise-robustness and E2E semantics evaluation. Experiments on ATIS show that our SLU framework with speech as input can perform on par with those using oracle text as input in semantics understanding, even though environmental noise is present and a limited amount of labeled semantics data is available for training.

Index Terms— spoken language understanding, speech representation learning, semi-supervised learning, speech recognition.

1. INTRODUCTION

Spoken Language Understanding (SLU)1 is at the front-end of many modern intelligent home devices, virtual assistants, and socialbots [1, 2]: given a spoken command, an SLU engine should extract relevant semantics from spoken commands for the appropriate downstream tasks. Since SLU tasks such as the Airline Travel Information System (ATIS) [3], the field has progressed from knowledge-based [5] to data-driven approaches, notably those based on neural networks. In the seminal paper on ATIS by Tur et al. [3], incorporating linguistically motivated features for NLU and improving ASR robustness were underscored as the research emphasis for the coming years. Now, a decade later, we should ask ourselves again, how much has the field progressed, and what is left to be done?

Self-supervised language models (LMs), such as BERT [6], and end-to-end SLU [7, 8, 9] appear to have addressed the problems posed in [3]. As shown in Figure 1 we can examine past SLU work from the angle of how they constructed the input/output pairs. In

1SLU typically consists of Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU). ASR maps audio to text, and NLU maps text to semantics. Here, we are interested in learning a mapping directly from raw audio to semantics.

2Semantic acquisition is commonly framed as Intent Classification (IC) and Slot Labeling/Filling (SL), see [1, 2, 3].

The case for semi-supervised SLU. Neural networks benefit from large quantities of labeled training data, and one can train end-to-end SLU models with them [2, 7, 8, 17]. However, curating labeled IC/SL data is expensive, and only a limited amount of labels are available. Semi-supervised learning could be a useful scenario for training SLU models for various domains whereby model components are pretrained on large amounts of unlabeled data and then fine-tuned with target semantic labels. While [9, 14, 17] have explored this pretraining then fine-tuning scheme, they did not take advantage of the generalization capacity of contextualized LMs, such as BERT, for learning semantics from speech. Notably, self-supervised speech representation learning [21, 22, 23, 24, 25] provides a clean and general learning mechanism for downstream speech tasks, yet the semantic transferrability of these representations are unclear. Our focus is on designing a better learning framework distinctly for semantic understanding under limited semantic labels, on top of ASR and BERT. We investigated two learning settings for the ASR component: (1) pretraining on transcribed speech with ASR subword prediction, and (2) pretraining on untranscribed speech data with contrastive losses [23, 24].

The key contributions of this paper are summarized as follows:

Fig. 1. Comparison of input/output pairs of our proposed framework with past work, which are categorized as one of: (A) NLU, which assumes oracle text as input instead of speech, (B) predicting intent only from speech, ignoring their slot values, and (C) predicting text, intent, and slots from speech. (D) Our work predicts text, intent, and slots from speech while taking advantage of unlabeled data.
We introduce a semi-supervised SLU framework for learning semantics from speech to alleviate: (1) the need for a large amount of in-house, homogenous data [5][7][13][17], (2) the limitation of only intent classification [5][9][13] by predicting text, slots and intents, and (3) any additional manipulation on labels or loss, such as label projection [26], output serialization [7][18][19]. ASR n-best hypothesis, or ASR-robust training losses [1][2][27]. Figure 2 illustrates our approach.

We investigate two learning settings for our framework: supervised pretraining and unsupervised pretraining (Figure 3), and evaluated our framework with a new metric, the slot edit $F_1$ score, for end-to-end semantic evaluation. Our framework improves upon previous work in Word Error Rate (WER) and IC/SL on ATIS, and even rivaled its NLU counterpart with $\theta_{BERT}$. The main benefit this formulation brings is that now $S$ and $I$ do not solely depend on an ASR top-1 hypothesis $W^*$ during training, and the end-to-end objective is thus,

$$L_{SLU}(\theta_{SLU}) = L_{ASR}(\theta_{SLU}) + L_{SLU}(\theta_{SLU}).$$ (2)

The ASR objective $L_{ASR}$ is formulated to maximize sequence-level log-likelihood, and $L_{ASR}(\theta_{SLU}) = L_{ASR}(\theta_{BERT})$. Before writing down $L_{SLU}$, we describe a masking operation because ASR and BERT typically employ different subword tokenization methods.

Differentiate Through Subword Tokenizations To concatenate $\theta_{ASR}$ and $\theta_{BERT}$ outputs along the hidden dimension, we need to make sure they have the same length along the token dimension. We stored the first indices where $W$ are broken down into subword tokens into a matrix: $M^a \in \mathbb{R}^{N^a \times N}$ for $\theta_{ASR}$ and $M^b \in \mathbb{R}^{N^b \times N}$ for $\theta_{BERT}$, where $N$ is the number of tokens for $W$ and $S$, $N^a$ is the number of ASR subword tokens, and $N^b$ for BERT. Let $H^a$ be the $\theta_{ASR}$ output matrix before softmax, and similarly $H^b$ for $\theta_{BERT}$. The concatenated matrix $H^\text{cat} \in \mathbb{R}^{N\times(N^a+F_a+H_a)}$ is given as $H^\text{cat} = \text{concat}([M^a \times H^a, M^b \times H^b])$, where $F_a$ and $F_b$ are hidden dimensions for $\theta_{ASR}$ and $\theta_{BERT}$.

$$L_{SLU} = E \left[ \ln P(S \mid H^\text{cat}; \theta_{SLU}) + \ln P(I \mid H^\text{cat}; \theta_{IC}) \right],$$ (3)

where the sum of cross entropy losses for IC and SL are maximized, and $\theta_{ASR}$ and $\theta_{BERT}$ are updated through $H^\text{cat}$. Ground truth $W$ is used as input to $\theta_{BERT}$ instead of $W^*$ due to teacher forcing.

### 2.1. End-to-End: Joint E2E ASR and BERT Fine-Tuning.

As illustrated in Figure 2, $\theta_{SLU}$ consists of a pretrained E2E ASR $\theta_{ASR}$ and a pretrained deep contextualized LM $\theta_{NLU}$, such as BERT, and is fine-tuned jointly for $W$, $S$ and $I$ on $D$. The choice of E2E ASR over hybrid ASR here is because the errors from $S$ and $I$ can be back-propagated through $A$: following [10], we have $S$ predicted via an additional CRF linear layer on top of $W$ and $I$ is predicted on top of the BERT output of the [CLS] token. The additional model parameters for predicting SL and IC are $\theta_{SL}$ and $\theta_{IC}$, respectively, and we have $\theta_{SLU} = \{\theta_{BERT}, \theta_{IC}; \theta_{SL}\}$. During end-to-end fine-tuning, outputs from $\theta_{ASR}$ and $\theta_{BERT}$ are concatenated to predict $S$ and $I$ with loss $L_{SLU}$, while $W$ is predicted with loss $L_{ASR}$. The main benefit this formulation brings is that now $S$ and $I$ do not solely depend on an ASR top-1 hypothesis $W^*$ during training, and the end-to-end objective is thus,

$$L_{SLU}(\theta_{SLU}) = L_{ASR}(\theta_{SLU}) + L_{SLU}(\theta_{SLU}).$$ (2)

For inference, an input audio sequence $a = a_1T$ and the sets of all possible word tokens $W$, slots $S$, and intents $I$ are given. We are then interested in decoding for its target word sequence $w^* = w_{1:N}$, slots sequence $s^* = s_{1:N}$, and intent label $i^*$. Having obtained $\theta_{SLU}$, the decoding procedure for the end-to-end approach is,

$$w^* = \arg \max_{w_n \in W} \prod_{n=1}^{N} p(w_n \mid w_{n-1}, a; \theta_{ASR}),$$ (4)

$$i^*, s^* = \arg \max_{i \in I} p(i \mid w^*, a; \theta_{SLU}), \arg \max_{s_n \in S} \prod_{n=1}^{N} p(s_n \mid w^*, a; \theta_{SLU})$$ (5)

This two step decoding procedure, first $w^*$ then $(i^*, s^*)$ is necessary given that no explicit serialization on $W$ and $S$ are imposed, as in [7][15]. While decoding for $(i^*, s^*)$, additional input $a$ is given and we have $w^*$ instead of $w_n$ given the context from self-attention in BERT. Note that here and throughout the work, we only take the top-1 hypothesis $w^*$ (instead of top-N) to decode for $(i^*, s^*)$. 

---

3We abuse some notations by representing models by their model parameters, e.g. $\theta_{ASR}$ for the ASR model and $\theta_{BERT}$ for BERT.
3. LEARNING WITH LESS SUPERVISION

Our semi-supervised framework relies on pretrained ASR and NLU components. Depending on the accessibility of the data, we explored two levels of supervision\[^1\]. The first setting is where an externally transcribed corpus is available, and we utilized transfer learning for initializing the ASR. The second setting is where external audio is available but not transcriptions, and in this case, the ASR is initialized with self-supervised learning. In both settings, BERT is pretrained with MLM and NSP as described in \[^6\]. Figure 3 distinguishes the two learning settings.

(A.1) Unsupervised Pretrain: BERT

(A.2) Supervised Pretrain on Librispeech: ASR

(A.3) SLU Fine-tune: ASR and BERT

(B.1) Unsupervised Pretrain: wav2vec and BERT

(B.2) SLU Fine-tune: ASR and BERT

Fig. 3. Two semi-supervised settings: (A) additional transcribed speech is available. \(\theta_{ASR}\) is pretrained and fine-tuned for ASR. (B) additional audio is available but without transcription. \(\theta_{ASR}\) encoder is replaced with a pretrained wav2vec \[^{24,23}\] before fine-tuning.

3.1. Transfer Learning from a Pretrained ASR

Following \[^{9,17,18}\], \(\theta_{ASR}\) is pretrained on an external transcribed speech corpus before fine-tuning on the target SLU dataset.

3.2. Unsupervised ASR Pretraining with wav2vec

According to UNESCO, 43% of the languages in the world are endangered. Supervised pretraining is not possible for many languages, as transcribing a language requires expert knowledge in phonetics, morphology, syntax, and so on. This partially motivates the line of self-supervised learning work in speech, that powerful learning representations require little fine-tuning data. Returning to our topic, we asked, how does self-supervised learning help with learning semantics?

Among many others, wav2vec 1.0 \[^{24}\] and 2.0 \[^{23}\] demonstrated the effectiveness of self-supervised representations for ASR. They are pretrained with contrastive losses \[^{22}\], and differ mainly by their architectures. We replaced \(\theta_{ASR}\) encoder with these wav2vec features, and appended the \(\theta_{ASR}\) decoder for fine-tuning on SLU.

4. EXPERIMENTS

Datasets ATIS \[^3\] contains 8hr of audio recordings of people making flight reservations with corresponding human transcripts. A total of 5.2k utterances with more than 600 speakers are present. Note that ATIS is considerably smaller than those in-house SLU data used in

\[^{2}https://github.com/EFord36/normalise\]

\[^{3}Code: Semi-Supervised-Spoken-Language-Understanding-PyTorch\]
idea of training BERT with audio-text pairs fitting as another baseline for our end-to-end approach. We modified the pretraining and fine-tuning setup described in [32] for SLU. Audio MLM (c.f. MLM in BERT [6]) pretrains \( \theta_{BERT} \) by mapping masked audio segments to text. This pretraining step gradually adapts the original BERT to a phonetic-semantic joint embedding space. Then, \( \theta_{NLU} \) is fine-tuned by mapping unmasked audio segments to IC/SL. Figure 3 illustrates the audio-text and audio-IC/SL pairs for SpeechBERT. Unlike the end-to-end approach, \( \theta_{ASR} \) is kept frozen throughout SpeechBERT pretraining and fine-tuning.

### 4.4. Main Results on Clean ATIS

We benchmarked our proposed framework with several prior works, and Table 4 presents their WER, slots edit F1 and intent F1 results. JointBERT [10] is our NLU baseline, where BERT is jointly fine-tuned for IC/SL, and it gets around 95% slots edit F1 and over 98% for IC F1. Since JointBERT has access to the oracle text, this is the upper bound for our SLU models with speech as input. CLM-BERT [26] explored using in-house conversational LM for NLU. We replicated \([18]\), where an LAS [33] directly predicts interleaving words and slots tokens (serialized output), and optimized with CTC over words and slots. We also experimented with a Kaldi hybrid ASR.

Both our proposed end-to-end and baselines approaches surpassed prior SLU work. We hypothesized the performance gain originates from our choices of (1) adopting pretrained E2E ASR and BERT, (2) applying text-norm on target transcriptions for training the ASR, and (3) end-to-end fine-tuning text and IC/SL.

**Table 1.** WER, slots edit and intent F1 on ATIS. ASR is pretrained on Librispeech 960h (LS-960). Results indicate our semi-supervised framework is effective in data scarcity setting, exceeding prior work in WER and IC/SL while approaching the NLU upperbound.

### 4.5. Environmental Noise Augmentation

A common scenario where users utter their spoken commands to SLU engines is when environmental noises are present in the background. Nonetheless, common SLU benchmarking datasets like ATIS, SNIPS [2], or FSC [9] are very clean. To quantify model robustness under noisy settings, we augmented ATIS with environmental noise from MS-SNSD. Table 5 reveals that those work well on clean ATIS may break under realistic noises, and although our models are trained with SpecAugment, there is still a 4-27% performance drop from clean test.

We followed the noise augmentation protocol in [29], where for each sample, five noise files are sampled and added to the clean file with SNR levels of [0, 10, 20, 30, 40]dB, resulting in a five-fold augmentation. We observe that augmenting the training data with a diverse set of environmental noises work well, and there is now minimal model degradation. Our end-to-end approach reaches 95.46% for SL and 97.4% for IC, which is merely a 1-2% drop from clean test, and almost a 40% improvement over hybrid ASR+BERT.

#### 4.6. Effectiveness of Unsupervised Pretraining with wav2vec

Table 6 shows the results on different ASR pretraining strategies: unsupervised pretraining with wav2vec, transfer learning from ASR, and no pretraining at all. We extracted both the latent vector \( z \) and context vector \( c \) from wav2vec 1.0. To simplify the pipeline and in contrast to \([24]\), we pre-extracted the wav2vec features and did not fine-tune wav2vec with \( \theta_{SLU} \) on ATIS. We also chose not to decode with a LM to be consistent with prior SLU work. We first observed the high WER for latent vector \( z \) from wav2vec 1.0, indicating they are sub-optimal and merely better than training from scratch by a slight margin. Nonetheless, encouragingly, context vector \( c \) from wav2vec 1.0 gets 67% slots and 90% intent F1. To improve the results, we added subsampling layers \([33]\) on top of the wav2vec features to downsample the sequence length with convolution. The motivation here is \( c \) and \( z \) are comparably longer than the normal ASR encoder outputs. With sub-sampling, \( c \) from wav2vec 1.0 now achieves 85.64% for SL and 95.67% for IC, a huge relative improvement over training ASR from scratch, and closes the gap between unsupervised and supervised pretraining for SLU.

### 5. Conclusions and Future Work

This work attempts to respond to a classic paper “What is left to be understood in ATIS?” [35], and to the advancement put forward by contextualized LM and end-to-end SLU up against semantics understanding. We showed for the first time that an SLU model with speech as input could perform on par with NLU models on ATIS, entering the 5% “corpus errors” range [34, 35]. However, we collectively believe that there are unsolved questions remaining, such as the prospect of building a single framework for multi-lingual SLU [35], or the need for a more spontaneous SLU corpus that is not limited to short segments of spoken commands.

**Acknowledgments** We thank Nanxin Chen, Erica Cooper, Alexander H. Liu, Wei Fang, and Fan-Keng Sun for their comments on this work.
6. REFERENCES

[1] Jianyu Cui, Michelle Cohn, Yi Yang, Chun-Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhownick, et al., “Grunrock: A social bot for complex and engaging long conversations,” arXiv preprint arXiv:1910.03042, 2019.

[2] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Dommouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibault Lavril, et al., “Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces,” arXiv preprint arXiv:1805.10190, 2018.

[3] Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck, “What is left to be understood in atis?,” in 2010 IEEE Spoken Language Technology Workshop. IEEE, 2010, pp. 19–24.

[4] Charles T Hemphill, John J Godfrey, and George R Doddington, “The atis spoken language systems pilot corpus,” in Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24–27, 1990, 1990.

[5] Stephanie Seneff, “Tina: A natural language system for spoken language applications,” Computational linguistics, vol. 18, no. 1, pp. 61–86, 1992.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[7] Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters, “From audio to semantics: Approaches to end-to-end spoken language understanding,” in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 720–726.

[8] Dmitry Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio, “Towards end-to-end spoken language understanding,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 5754–5758.

[9] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, “Speech model pre-training for end-to-end spoken language understanding,” arXiv preprint arXiv:1904.03670, 2019.

[10] Qian Chen, Zhu Zhuo, and Wen Wang, “Bert for joint intent classification and slot filling,” arXiv preprint arXiv:1902.10909, 2019.

[11] Jie Zhou and Wei Xu, “End-to-end learning of semantic role labeling using recurrent neural networks,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1127–1137.

[12] Su Zhu and Kai Yu, “Encoder-decoder with focus-mechanism for sequence labeling based spoken language understanding,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017, pp. 5675–5679.

[13] Chao-Wei Huang and Yun-Nung Chen, “Learning asr-robust contextualized embeddings for spoken language understanding,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 8009–8013.

[14] Pengwei Wang, Liancheng Wei, Yong Cao, Jinghui Xie, and Zaiqing Nie, “Large-scale unsupervised pre-training for end-to-end spoken language understanding,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7999–8003.

[15] Won Ik Cho, Donghyun Kwak, Jiwon Yoon, and Nam Soo Kim, “Speech to text adaptation: Towards an efficient cross-modal distillation,” arXiv preprint arXiv:2005.08213, 2020.

[16] Martin Radfar, Athanasios Mouchtaris, and Siegfried Kunzmann, “End-to-end neural transformer based spoken language understanding,” arXiv preprint arXiv:2008.10984, 2020.

[17] Milind Rao, Anirudh Raju, Pranav Dheram, Bach Bui, and Ariya Rastrow, “Speech to semantics: Improve asr and nlu jointly via all-neural interfaces,” arXiv preprint arXiv:2008.06173, 2020.

[18] Natalia Tomashenko, Antoine Caubrière, Yannick Estève, Antoine Laurent, and Emmanuel Morin, “Recent advances in end-to-end spoken language understanding,” in International Conference on Statistical Language and Speech Processing. Springer, 2019, pp. 44–55.

[19] Sahar Ghannay, Antoine Caubrière, Yannick Estève, Antoine Laurent, and Emmanuel Morin, “End-to-end named entity extraction from speech,” arXiv preprint arXiv:1805.12045, 2018.

[20] Taku Kudo and John Richardson, “Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing,” arXiv preprint arXiv:1808.06226, 2018.

[21] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, “An unsupervised autoregressive model for speech representation learning,” arXiv preprint arXiv:1904.03240, 2019.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, “Representation learning with contrastive predictive coding,” arXiv preprint arXiv:1807.03748, 2018.

[23] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” arXiv preprint arXiv:2006.11477, 2020.

[24] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, “wav2vec2: Unsupervised pre-training for speech recognition,” arXiv preprint arXiv:1904.05862, 2019.

[25] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, “Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6419–6423.

[26] Jin Cao, Jun Wang, Wael Hamza, Kelly Vanee, and Shang-Wen Li, “Style attuned pre-training and parameter efficient fine-tuning for spoken language understanding,” arXiv preprint arXiv:2010.04355, 2020.

[27] Chia-Hsuan Lee, Yun-Nung Chen, and Hung-Yi Lee, “Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 7300–7304.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015, pp. 5206–5210.

[29] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Siriram Srinivasan, and Johannes Gehmke, “A scalable noisy speech dataset and online subjective test framework,” arXiv preprint arXiv:1909.08050, 2019.

[30] Shizhi Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, “Hybrid ctc/attention architecture for end-to-end speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, 2017.

[31] Daniel S Park, William Chan, Yu Zhang, Cheng-Cheng Chiu, Barrett Zoph, Ekin D Cubuk, and Quoc V Le, “Spectaegument: A simple data augmentation method for automatic speech recognition,” arXiv preprint arXiv:1904.08779, 2019.

[32] Yung-Sung Chung, Chi-Liang Liu, and Hung-Yi Lee, “Speechbert: Cross-modal pre-trained language model for end-to-end spoken question answering,” arXiv preprint arXiv:1910.11559, 2019.

[33] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4960–4964.

[34] Frédéric Béchet and Christian Raymond, “Is atis too shallow to go deeper for benchmarking spoken language understanding models?,” 2018.

[35] James Glass, Giovanni Flammini, David Goodine, Michael Phillips, Joseph Polifroni, Shinusue Sakai, Stephanie Senef, and Victor Zue, “Multilingual spoken-language understanding in the mit voyager system,” Speech communication, vol. 17, no. 1-2, pp. 1–18, 1995.