ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414922

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[2] from spoken commands for the appropriate downstream tasks. Since SLU tasks such as the Airline Travel Information System (ATIS) [4], 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 [10], Intent Classification (IC) and Slot Labeling (SL) are jointly predicted on top of BERT, discarding the need of a Conditional Random Fields (CRF) [11]. However, these NLU works [10, 12, 13] usually ignore ASR or require an off-the-shelf ASR during testing. A line of E2E SLU work does take speech as input, yet it frames slots as intents and therefore their SLU models are really designed for IC only [8, 9, 14, 15, 16]. Another line of E2E SLU work jointly predicts text and IC/SL from speech, yet it either requires large amounts of in-house data, or restricts the pretraining scheme to ASR subword prediction [7, 17, 18, 19]. In contrast, we would desire a framework that predicts text, intents, and slots from speech, while learning with limited semantics labels by pretraining on unlabeled data.

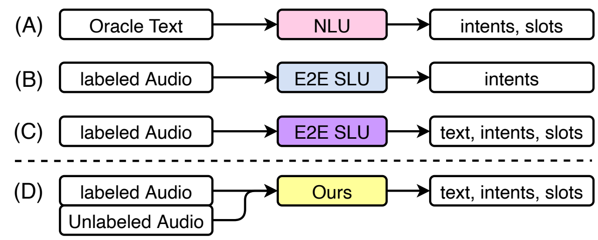


Fig. 1. Comparison of input/output pairs of our proposed framework with past work, which are categorized as one of: **(A)** NLU, which assumes oracle text as input instead of speech, **(B)** predicting *intent only* from speech, ignoring their slot values, and **(C)** predicting text, intent, and slots from speech. **(D)** Our work predicts text, intent, and slots from speech while taking advantage of *unlabeled* data.

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 often 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, 18] 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:

The fourth author contributed to the work before joining Amazon.

[1]SLU 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.

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

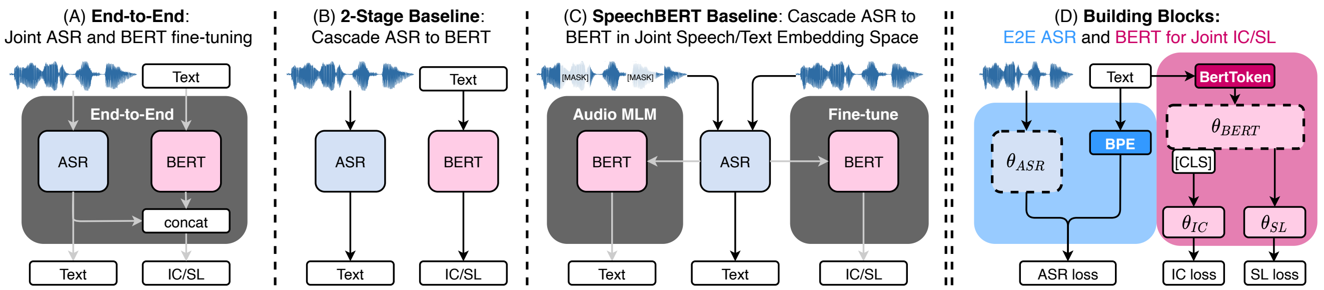


Fig. 2. Our proposed semi-supervised learning framework with ASR and BERT for joint intent classification (IC) and slot labeling (SL) directly from speech. **(A)** shows the end-to-end approach, in which E2E ASR and BERT are trained jointly by predicting text and IC/SL. **(B)** shows the 2-Stage baseline, where text and IC/SL are obtained successively. **(C)** shows the SpeechBERT baseline, where BERT is adapted to take audio as input by first pretraining with Audio MLM loss and then fine-tuning for IC/SL. A separate pretrained ASR is still needed for (B) and (C). **(D)** shows the ASR (θ_{ASR}) and NLU ($\{\theta_{BERT}, \theta_{IC}, \theta_{SL}\}$) building blocks used in (A)-(C). Note that θ_{ASR} and θ_{BERT} have different subword tokenizations: SentencePiece (BPE) [20] and BertToken. Dotted shapes are pretrained. Figure best viewed in colors.

- 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 [2, 7, 8, 17], (2) the limitation of only intent classification [8, 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 [13, 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 oracle text input [10]. In addition, it is trained with noise augmentation such that it is robust to real environmental noises.

2. PROPOSED LEARNING FRAMEWORK

We now formulate the mapping from speech to text, intents, and slots. Consider a target SLU dataset $\mathcal{D} = \{\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{S}, \boldsymbol{I}\}_{i=1}^{M}$, consisting of M i.i.d. sequences, where $\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{S}$ are the audio, word and slots sequences, and $\boldsymbol{I}$ is their corresponding intent label. Note that $\boldsymbol{W}$ and $\boldsymbol{S}$ are of the same length, and $\boldsymbol{I}$ is a one hot vector. We are interested in finding the model θ^*_{SLU} with loss,

$$\mathcal{L}_{SLU}(\theta_{SLU}) = \mathbb{E}_{(\boldsymbol{A},\boldsymbol{W},\boldsymbol{S},\boldsymbol{I})\sim\mathcal{D}}\Big[\ln P(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{I} \mid \boldsymbol{A}; \theta)\Big] \quad (1)$$

We proceed to describe an end-to-end implementation of θ_{SLU}[3].

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

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

$$\mathcal{L}_{SLU}(\theta_{SLU}) = \mathcal{L}_{ASR}(\theta_{SLU}) + \mathcal{L}_{NLU}(\theta_{SLU}). \quad (2)$$

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

Differentiate Through Subword Tokenizations To concatenate θ_{ASR} and θ_{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 $\boldsymbol{W}$ are broken down into subword tokens into a matrix: $M^a \in \mathbb{R}^{N^a \times N}$ for θ_{ASR} and $M^b \in \mathbb{R}^{N^b \times N}$ for θ_{BERT}, where N is the number of tokens for $\boldsymbol{W}$ and $\boldsymbol{S}$, N^a is the number of ASR subword tokens, and N^b for BERT. Let H^a be the θ_{ASR} output matrix before softmax, and similarly H^b for θ_{BERT}. The concatenated matrix $H^{cat} \in \mathbb{R}^{N \times (F_a + F_b)}$ is given as $H^{cat} = \text{concat}(\left[(M^a)^T H^a, (M^b)^T H^b\right], \text{dim=1})$, where F_a and F_b are hidden dimensions for θ_{ASR} and θ_{BERT}. $\mathcal{L}_{NLU}$ is then,

$$\mathcal{L}_{NLU} = \mathbb{E}\Big[\ln P(\boldsymbol{S} \mid H^{cat}; \theta_{SL}) + \ln P(\boldsymbol{I} \mid H^{cat}; \theta_{IC})\Big], \quad (3)$$

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

2.2. Inference

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

$$\boldsymbol{w}^* = \underset{w_n \in \mathcal{W}}{\text{argmax}} \prod_{n=1}^{N} p(w_n \mid w_{n-1}, \boldsymbol{a}; \theta^*_{ASR}), \quad (4)$$

$$\boldsymbol{i}^*, \boldsymbol{s}^* = \underset{i \in \mathcal{I}}{\text{argmax}}\, p(i \mid \boldsymbol{w}^*, \boldsymbol{a}; \theta^*_{SLU}), \underset{s_n \in \mathcal{S}}{\text{argmax}} \prod_{n=1}^{N} p(s_n \mid \boldsymbol{w}^*, \boldsymbol{a}; \theta^*_{SLU}) \quad (5)$$

This two step decoding procedure, first $\boldsymbol{w}^*$ then $(\boldsymbol{i}^*, \boldsymbol{s}^*)$ is necessary given that no explicit serialization on $\boldsymbol{W}$ and $\boldsymbol{S}$ are imposed, as in [7, 18]. While decoding for $(\boldsymbol{i}^*, \boldsymbol{s}^*)$, additional input $\boldsymbol{a}$ is given and we have $\boldsymbol{w}^*$ 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 $\boldsymbol{w}^*$ (instead of top-N) to decode for $(\boldsymbol{i}^*, \boldsymbol{s}^*)$.

[3] We abuse some notations by representing models by their model parameters, e.g. θ_{ASR} for the ASR model and θ_{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 level of supervision[4]. The first setting is where an external 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.

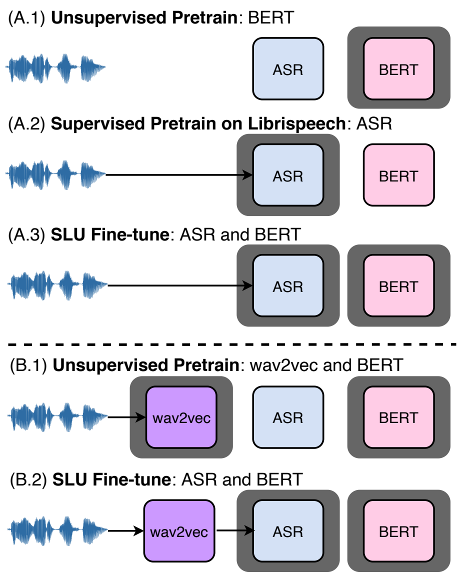


Fig. 3. Two semi-supervised settings: **(A)** additional transcribed speech is available. θ_{ASR} is pretrained and fine-tuned for ASR. **(B)** additional audio is available but without transcription. θ_{ASR} encoder is replaced with a pretrained wav2vec [24, 23] before fine-tuning.

3.1. Transfer Learning from a Pretrained ASR

Following [9, 17, 18], θ_{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?

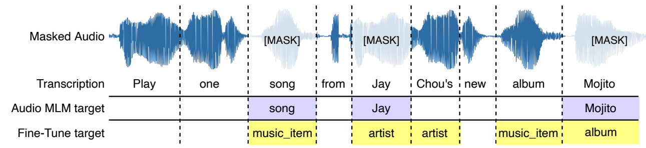
Among many others, wav2vec 1.0 [24] and 2.0 [23] demonstrated the effectiveness of self-supervised representations for ASR. They are pretrained with contrastive losses [22], and differ mainly by their architectures. We replaced θ_{ASR} encoder with these wav2vec features, and appended the θ_{ASR} decoder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

4. EXPERIMENTS

Datasets ATIS [4] 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

[4]In either settings, the amount of IC/SL annotations remains the same.

Fig. 4. Our SpeechBERT [32] pretraining and fine-tuning setup.



[2, 7, 8, 17], justifying our limited semantics labels setup. Waveforms are sampled at 16kHz. For the unlabeled semantics data, we selected Librispeech 960hr (LS-960) [28] for pretraining. Besides the original ATIS, models are evaluated on its noisy copy (augmented with MS-SNSD [29]). We made sure the noisy train and test splits in MS-SNSD do not overlap. Text normalization is applied on the ATIS transcription with an open-source software[5]. Utterances are ignored if they contain words with multiple possible slot tags.

Hyperparameters All speech is represented as sequences of 83-dimensional Mel-scale filter bank energies with F_0, computed every 10ms. Global mean normalization is applied. E2E ASR is implemented in ESPnet, where it has 12 Transformer encoder layers and 6 decoder layers. The choice of the Transformer is similar to [16]. E2E ASR is optimized with hybrid CTC/attention losses [30] with label smoothing. The decoding beam size is set to 5 throughout this work. We **do not** use an external LM during decoding. SpecAugment [31] is used as the default for data augmentation. A SentencePiece (BPE) vocabulary size is set to 1k. BERT is a bert-base-uncased from HuggingFace. Code will be made available[6].

4.1. E2E Evaluation with Slots Edit F_1 score.

Our framework is evaluated with an end-to-end evaluation metric, termed the slots edit F_1. Unlike slots F_1 score, slots edit F_1 accounts for instances where predicted sequences have different lengths as the ground truth. It bears similarity with the E2E metric proposed in [7, 17]. To calculate the score, the predicted text and oracle text are first aligned. For each slot label $v \in \mathcal{V}$, where $\mathcal{V}$ is the set of all possible slot labels except for the "O" tag, we calculate the <u>insertion</u> (false positive, FP), <u>deletion</u> (false negative, FN) and <u>substition</u> (FN and FP) of its slots value. Slots edit F_1 is the harmonic mean of precision and recall over all slots:

$$\text{slots edit } F_1 = \frac{\sum_{v \in \mathcal{V}} 2 \times \text{TP}_v}{\sum_{v \in \mathcal{V}} \Big[(2 \times \text{TP}_v) + \text{FP}_v + \text{FN}_v \Big]} \tag{6}$$

4.2. End-to-End 2-Stage Fine-tuning

An observation from the experiment was that ASR is much harder than IC/SL. Therefore, we adjusted our end-to-end training to a two-stage fine-tuning: pretrain ASR on LS-960, then fine-tine ASR on ATIS, and lastly jointly fine-tune for ASR and IC/SL on ATIS.

4.3. Baselines: Alternative θ_{SLU} Formulations

Two variations for constructing θ_{SLU} are presented (refer to Figure 2). They will be the baselines to the end-to-end approach.

2-Stage: Cascade ASR to BERT A natural complement to the E2E approach is to *separately* pretrain and fine-tune ASR and BERT. In this case, errors from $\boldsymbol{S}$ and $\boldsymbol{I}$ cannot be back-propagated to θ_{ASR}.

SpeechBERT : BERT in Speech-Text Embed Space Another sensible way to construct θ_{SLU} is to somehow "adapt" BERT such that it can take audio as input and outputs IC/SL, while not compromising its original semantic learning capacity. SpeechBERT [32] was initially proposed for spoken question answering, but we found the core

[5]https://github.com/EFord36/normalise
[6]Code: Semi-Supervsied-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 θ_{BERT} by mapping *masked audio segments* to text. This pretraining step gradually adapts the original BERT to a phonetic-semantic joint embedding space. Then, θ_{NLU} is fine-tuned by mapping *unmasked audio segments* to IC/SL. Figure 4 illustrates the audio-text and audio-IC/SL pairs for SpeechBERT. Unlike the end-to-end approach, θ_{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 1 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 F_1 and over 98% 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 word 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 hypothesize 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 F_1 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.

Frameworks	Unlabeled Semantics Data	ATIS clean test WER	slots edit F_1	intent F_1
NLU with Oracle Text				
JointBERT [10]		-	95.64	98.99
Proposed				
End-to-End w/ 2-Stage	LS-960	2.18	**95.88**	97.26
2-Stage Baseline	LS-960	1.38	93.69	97.01
SpeechBERT Baseline	LS-960	1.4	92.36	**97.4**
Prior Work				
ASR-Robust Embed [13]	WSJ	15.55	-	95.65
Kaldi Hybrid ASR+BERT	LS-960	13.31	85.13	94.56
ASR+CLM-BERT [26]	in-house	18.4.	93.8[7]	97.1
LAS+CTC [18]	LS-460	8.32	86.85	-

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 2 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.

[7]For [26], model predictions are evaluated only if its ASR hypothesis and human transcription have the same number of tokens.

Table 2. Noise augmentation effectively reduces model degradation.

Frameworks	ATIS noisy test WER	slots edit F_1	intent F_1
Kaldi Hybrid ASR+BERT	44.72	69.55	88.94
Proposed w/ Noise Aug.			
End-to-End w/ 2-Stage	3.6	**95.46**	**97.40**
2-Stage Baseline	3.5	92.52	96.49
SpeechBERT Baseline	3.6	88.7	96.15
Proposed w/o Noise Aug.			
End-to-End w/ 2-Stage	9.62	91.54	96.14
2-Stage Baseline	8.98	90.09	95.74
SpeechBERT Baseline	9.0	81.72	94.05

4.6. Effectiveness of Unsupervised Pretraining with wav2vec

Table 3 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 $\boldsymbol{z}$ and context vector $\boldsymbol{c}$ 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 θ_{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 $\boldsymbol{z}$ from wav2vec 1.0, indicating they are sub-optimal and merely better than training from scratch by a slight margin. Nonetheless, encouragingly, context vector $\boldsymbol{c}$ from wav2vec 1.0 gets 67% slots and 90% intent F_1.

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 $\boldsymbol{c}$ and $\boldsymbol{z}$ are comparably longer than the normal ASR encoder outputs. With sub-sampling, $\boldsymbol{c}$ 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.

Table 3. Effectiveness of different ASR pretraining strategies for our 2-Stage baseline. Results with wav2vec 2.0 [23] is omitted since they are not much better. Setup is visaulized in Figure 3.

Frameworks	ATIS clean test WER	slots edit F_1	intent F_1
Proposed 2-Stage w/o ASR Pretraining			
2-Stage Baseline	58.7	29.22	82.08
Proposed 2-Stage w/ Transfer Learning from ASR			
2-Stage Baseline	1.38	93.69	97.01
Proposed 2-Stage w/ Unsupervised Pretraining			
wav2vec1.0 [24] $\boldsymbol{z}$ + 2-Stage	54.2	35.04	83.68
wav2vec1.0 [24] $\boldsymbol{c}$ + 2-Stage	30.4	67.33	89.86
wav2vec1.0 [24] $\boldsymbol{c}$ + subsample + 2-Stage	13.2	**85.64**	**95.67**

5. CONCLUSIONS AND FUTURE WORK

This work attempts to respond to a classic paper ”What is left to be understood in ATIS? [3]”, 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 [3, 34]. 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] Dian Yu, Michelle Cohn, Yi Mang Yang, Chun-Yen Chen, Weiming Wen, Jiaping Zhang, Mingyang Zhou, Kevin Jesse, Austin Chau, Antara Bhowmick, et al., "Gunrock: 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 Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut 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-Tür, 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] Dmitriy 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 labelling 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, Liangchen 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 Caubriere, Yannick Esteve, 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, "wav2vec: 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, Sriram Srinivasan, and Johannes Gehrke, "A scalable noisy speech dataset and online subjective test framework," *arXiv preprint arXiv:1909.08050*, 2019.

[30] Shinji 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, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[32] Yung-Sung Chuang, 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 Flammia, David Goodine, Michael Phillips, Joseph Polifroni, Shinsuke Sakai, Stephanie Seneff, and Victor Zue, "Multilingual spoken-language understanding in the mit voyager system," *Speech communication*, vol. 17, no. 1-2, pp. 1–18, 1995.