Language Technologies Institute
Carnegie Mellon University

ESPnet-SLU:
Advancing Spoken Language Understanding through ESPnet

University of Stuttgart
Institute for Natural Language Processing

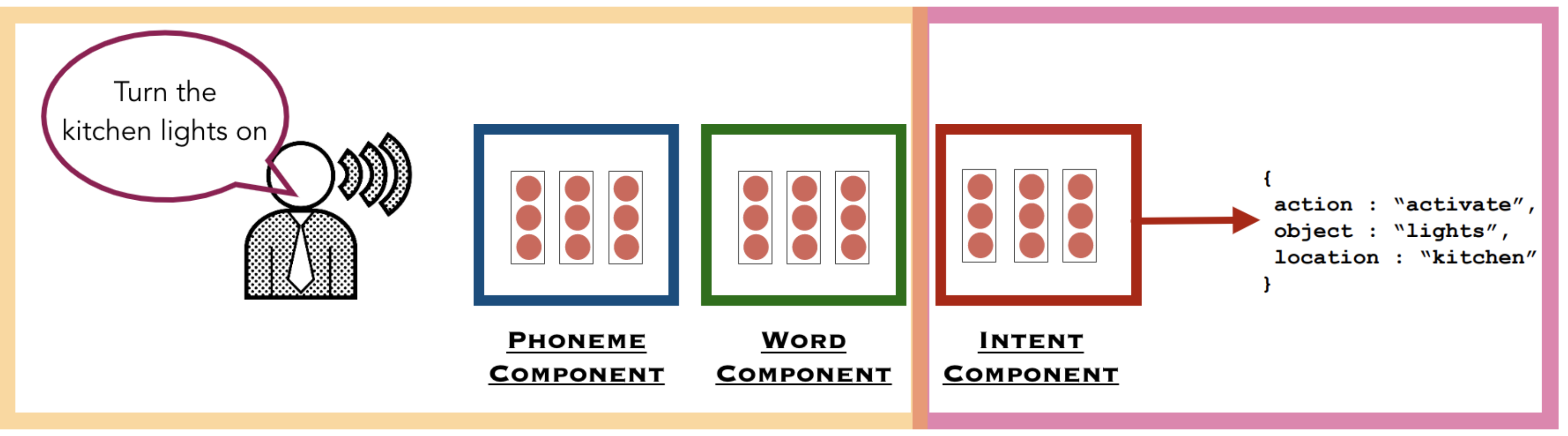
zoom

Siddhant Arora, Siddharth Dalmia, Pavel Denisov, Xuankai Chang, Yushi Ueda
Yifan Peng, Yuekai Zhang, Sujay Kumar, Karthik Ganesan, Brian Yan
Ngoc Thang Vu, Alan W Black, Shinji Watanabe
siddhana@cs.cmu.edu

Spoken Language Understanding

Definition: As ASR systems get better, there is increasing interest of using ASR output for downstream NLP tasks.

Example: Spoken Language Understanding (Intent Prediction)



```
{
 action : “activate”,
 object : “lights”,
 location : “kitchen”
}
```

Applications:

1. Intent Classification : Spoken Utterance → Executable Intent
2. Slot Filling : User Command → Associated Entities
3. Emotion Recognition : Understanding emotion behind a utterance
4. Dialogue Act Classification : Modeling the topic of a conversation

Motivation & Design

With the increase in SLU datasets and methodologies growing need for an open-source SLU toolkit!

Design Features of our SLU Toolkit

	Alexa[9]	Lugosch[3]	CoraJung [25]	SpeechBrain[26]	ESPnet-SLU
BiLSTM based encoder	✓	✓	✓	✓	✓
Transformer based encoder				✓	✓
Conformer based encoder				✓	✓
Classifier	✓			✓	
RNN based decoder		✓	✓	✓	✓
Transformer based decoder				✓	✓
Multi tasking with ASR?					✓
Supports multi tasking with NLU?	✓		✓		
Pretrained ASR model?		✓	✓	✓	✓
Pretrained NLU model?	✓		✓		✓
Other task?				✓	✓
SLU on languages besides English?					✓
Context from previous utterances?					✓
Tasks in pipeline manner?					✓
Provide pretrained model		✓		✓	✓

At a Glance

ESPnet-SLU is a new End to End Spoken Language Understanding toolkit built on an already existing open-source speech processing toolkit ESPnet which cover all the experiment processes for various Spoken Language Understanding Tasks.

Contribution: A Unified Pipeline for SLU Model

1. Standardize the pipeline of building an SLU model
2. Incorporate pretrained ASR like Hubert, Wav2vec2 and NLU models like BERT, MPNet as feature extractors
3. Implementations of various speech processing tasks that can be used in a pipeline manner
4. Provide easy access to trained models

,

(1) Supported Tasks and Datasets

Task	Dataset	Metric	Paper Results	ESPnet-SLU
IC	SLURP [4]	Acc.	78.3	86.3
	FSC [3]	F1	98.8	99.6
	FSC Unseen (S) [3, 40]	Acc.	94.2	98.6
	FSC Unseen (U) [3, 40]	Acc.	88.3	86.4
	FSC Challenge (S) [3, 40]	Acc.	92.3	97.5
	FSC Challenge (U) [3, 40]	Acc.	78.3	78.5
	SNIPS [13]	F1	91.7	91.7
	HarperValleyBank [41]	Acc	45.5	47.1
	Grabo [12, 42]	Acc.	94.5	97.2
	CAT-SLU MAP [27, 43]	Acc.	79.8	78.9
	Speech Commands [44]	Acc.	88.2	98.4
SF	SLURP [4]	SLU-F1	70.8	71.9
DA	Switchboard [45, 46]	Acc.	68.7	67.5
	HarperValleyBank [41]	Acc.	45.5	47.1
ER	IEMOCAP [6, 47]	5-fold Acc.	67.6	69.4

Recipes for over 10 SLU corpora, for multiple languages and task types, with performance nearing or exceeding the prior SOTA.

(2) Using ASR and NLU pretrained models for SLU

	Model	IC (F1)
Baseline	Pipeline ASR+NLU w/ synthetic data [4]	74.6
	+ Additional ASR data [4]	78.3
	E2E-SLU w/ Pretraining + synthetic data [26]	75.1
ESPnet-SLU	E2E-SLU w/ Conformer Encoder	76.4
	+ Pretrained ASR HuBERT [19]	77.0
	+ synthetic data	**86.3**
Ablations for Pretrained ASR	+ VQ-APC [22]	82.1
	+ HuBERT [19]	83.3
	+ Wav2vec2 [20]	83.3
	+ TERA [21]	83.5
Ablations for Pretrained NLU	+ MPNET [24]	82.5
	+ BERT [23]	85.7

Our Toolkit can compare the utility of different pretrained ASR and NLU systems as feature extractors!

(3) ASR Multi-Tasking can improve SLU performance

	Model	IC (% Acc)
Baseline	E2E-SLU [3]	96.6
	+ Pretraining ASR [3]	98.8
	Pretrained E2E-SLU + data augmentation [26]	99.6
ESPnet-SLU	Tsf. Encoder w/ Full Intent Decoding	93.5
	+ SpecAug Data Augmentation	98.9
	+ ASR Multi-tasking	99.4
	+ Pretrained ASR HuBERT	**99.6**

(4) Speech Enhancement Frontend Improves Noisy IC

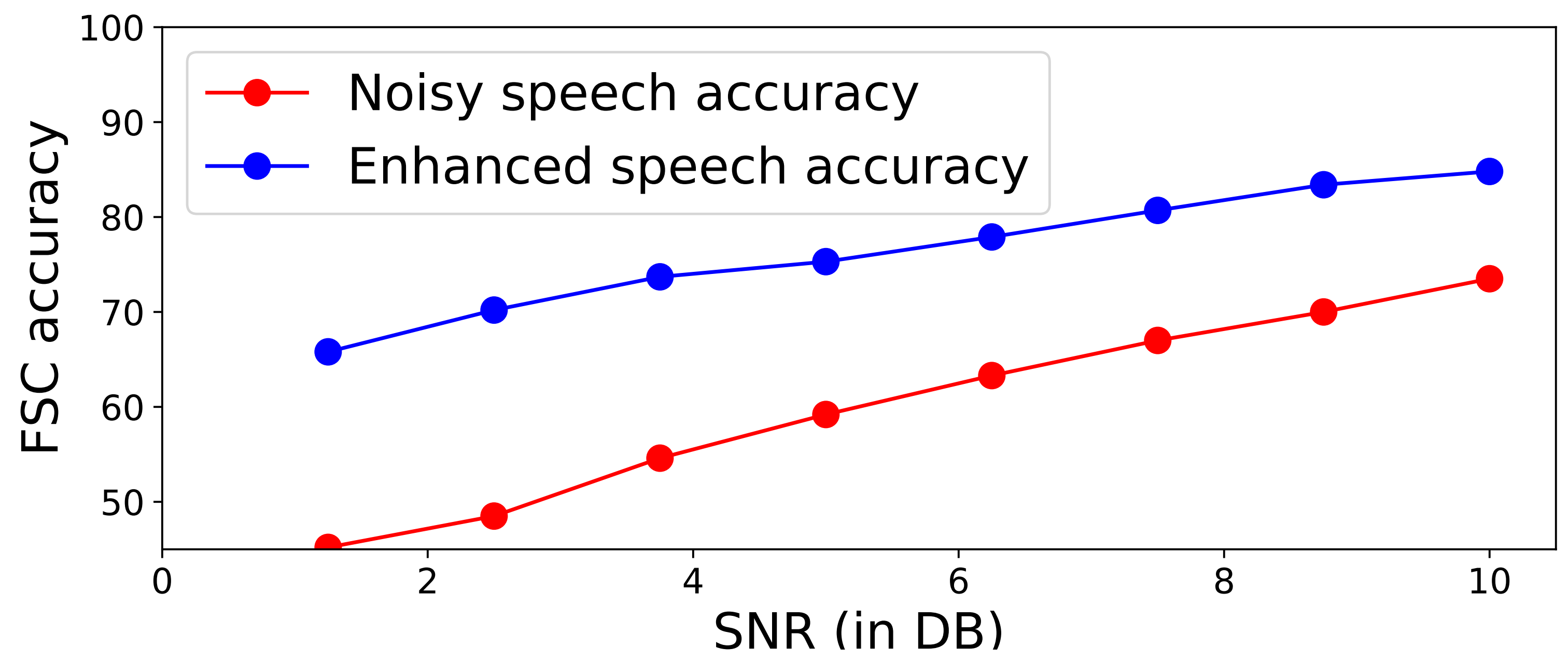


Figure: IC accuracy on the FSC dataset against the SNR of noisy speech.