# Learning ASR-Robust Contextualized Embeddings for Spoken Language Understanding

**Chao-Wei Huang** **Yun-Nung (Vivian) Chen**

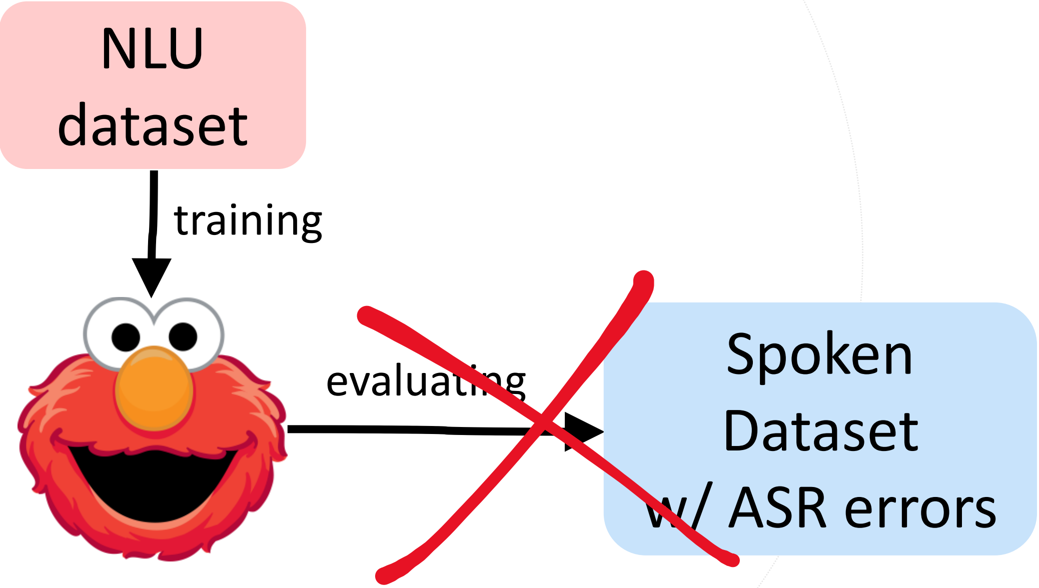
National Taiwan University

r07922069@ntu.edu.tw y.v.chen@ieee.org

Code available at https://github.com/MiuLab/SpokenVec

# Highlights

- Contextualized embeddings like ELMo do not transfer well to spoken domain w/ ASR errors.



NLU dataset

training

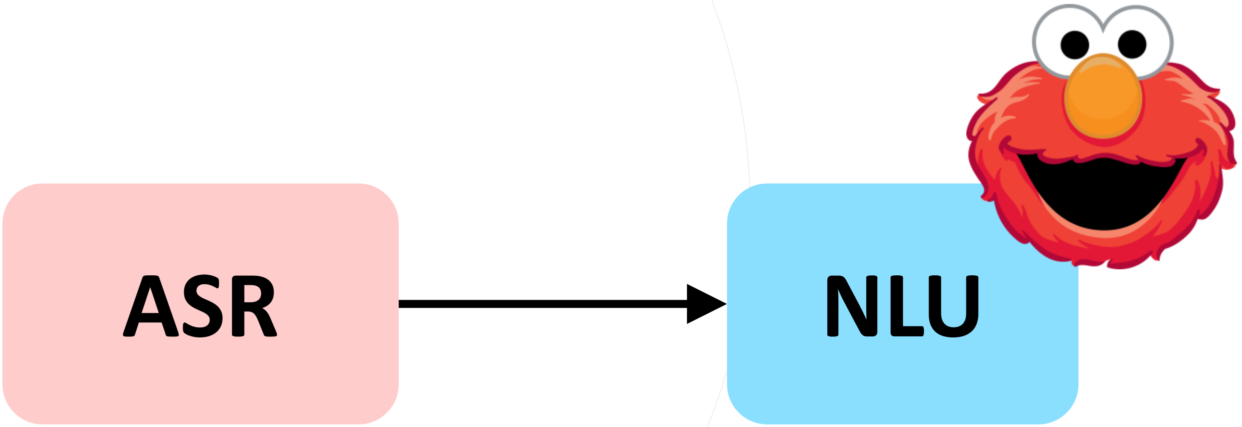
evaluating

Spoken Dataset w/ ASR errors

- We propose a fine-tuning method to tackle this problem.

# Motivation: Bridge between ASR and NLU

- Intuitive way for SLU: **pipelined approach**



ASR

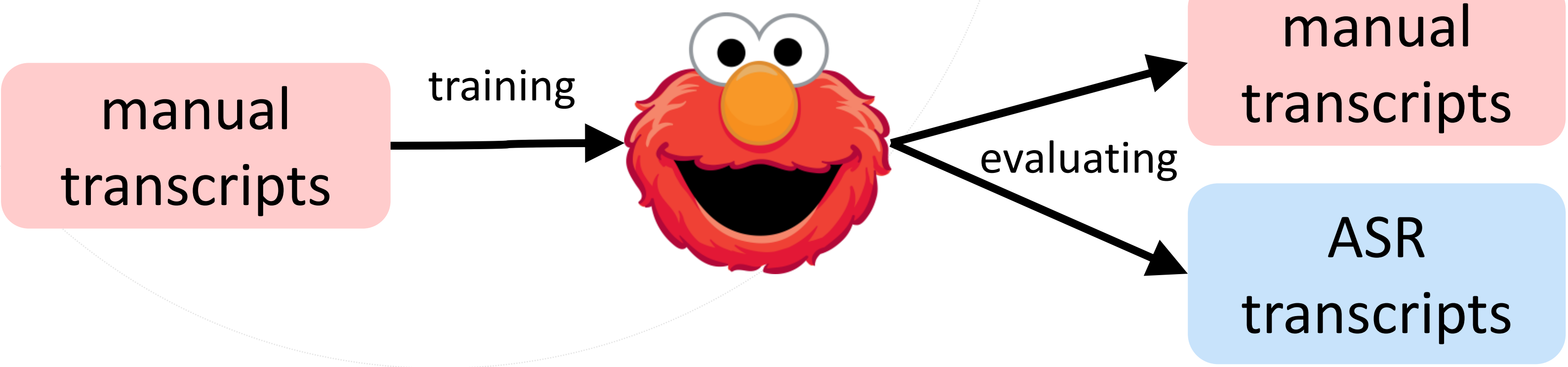
NLU

- ASR errors affects downstream tasks

# LU models

- The SOTA LU models are usually pre-trained LMs
- But they are pre-trained on written text

Do they transfer well to spoken domain w/ ASR errors?



manual transcripts

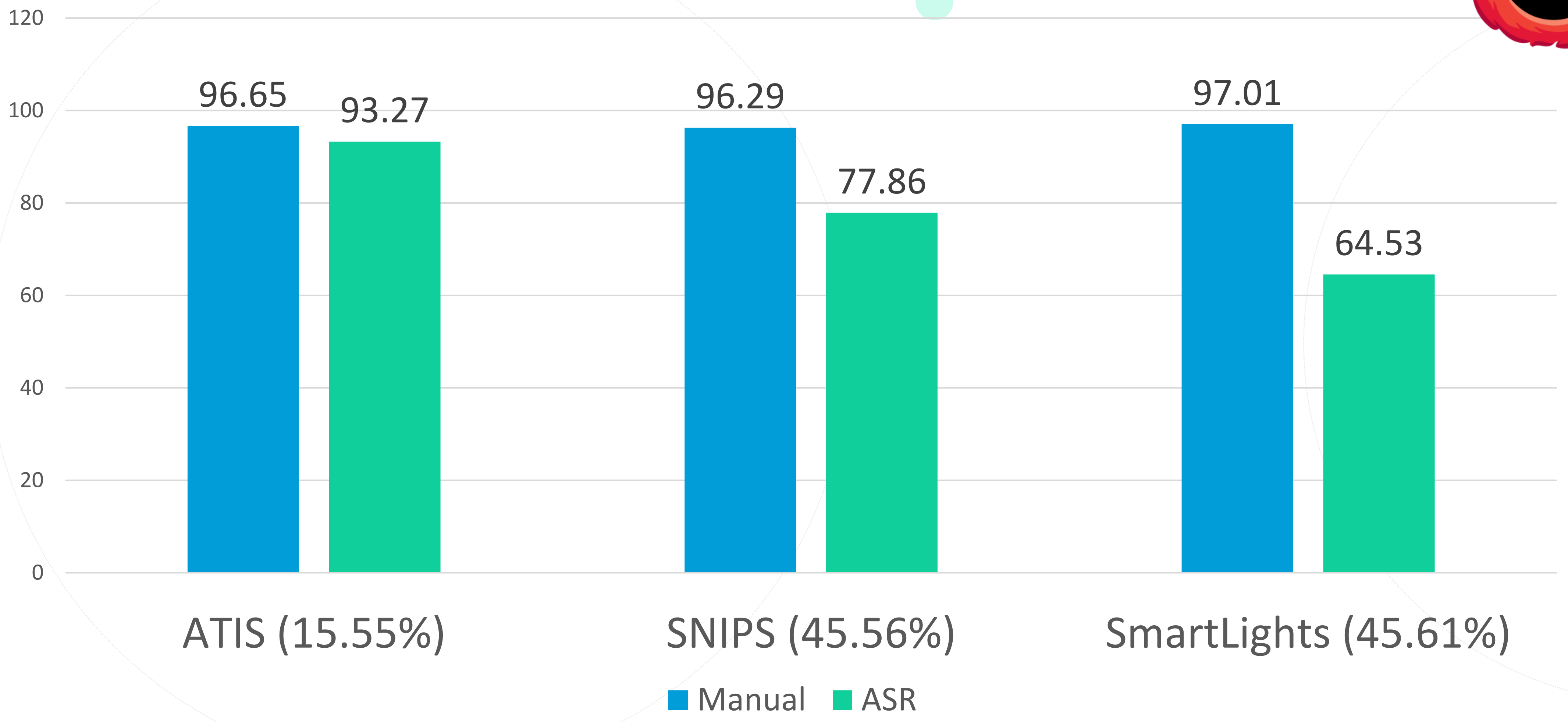
training

evaluating

manual transcripts

ASR transcripts

# Do they transfer well to spoken domain?



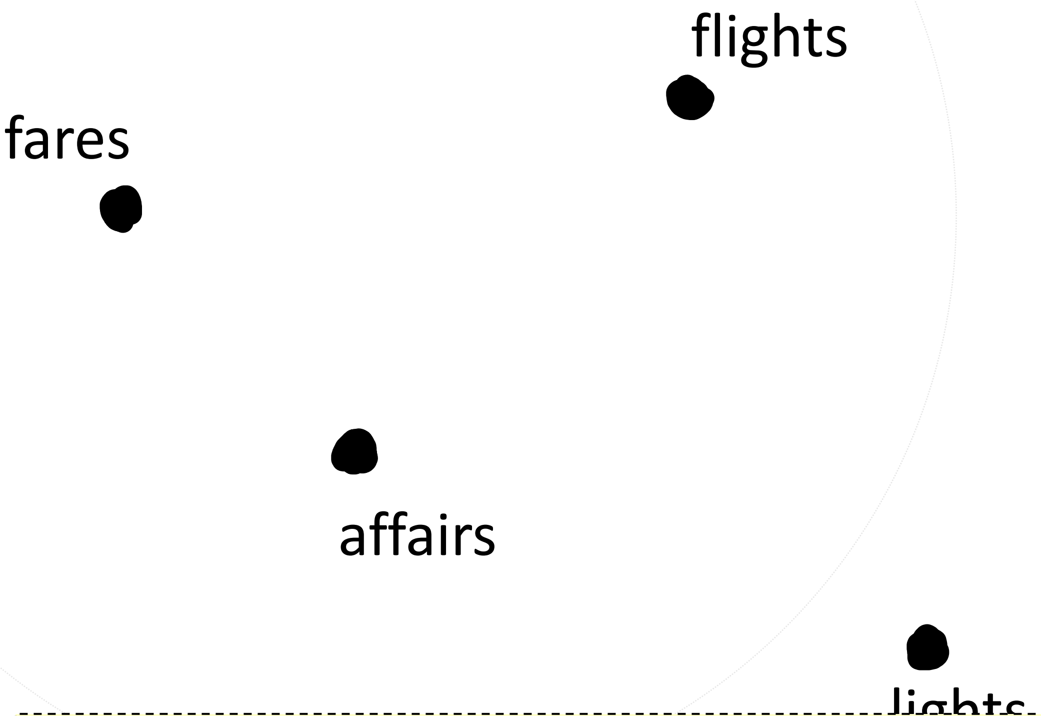
How can we transfer them to spoken domain?

# Our method: Additional fine-tuning stage

- LM pre-training
  - Same as ELMo
- LM fine-tuning
  - make the embeddings acoustic-aware
- Training target task classifier
  - Pre-trained LM is used as a feature extractor

# Make them acoustic-aware!

- Force embeddings of acoustically similar words to be closer



Confusion loss

$$\mathcal{L}_{\text{conf}} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^{1} 1 - \frac{h_{t_1,i}^{x_1} \cdot h_{t_2,i}^{x_2}}{\|h_{t_1,i}^{x_1}\| \, \|h_{t_2,i}^{x_2}\|}$$

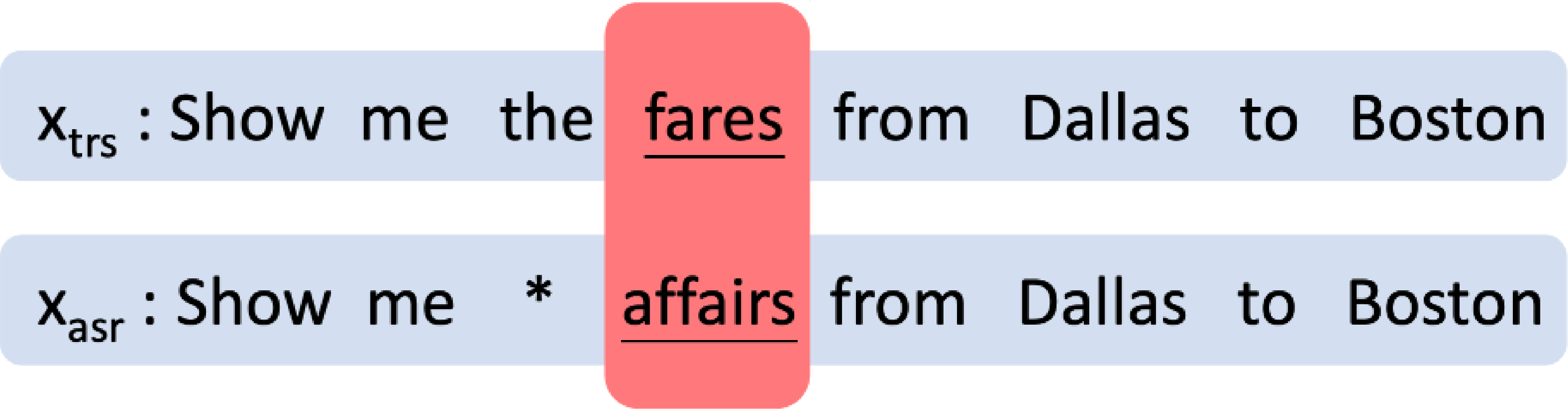
How to determine which words to bring closer?

# How to determine which words to bring closer?

- Case 1: we have paired ASR and manual transcripts (supervised)
- Case 2: we only have some ASR transcripts (unsupervised)

# How to determine which words to bring closer?

- Case 1: we have paired ASR and manual transcripts (supervised)

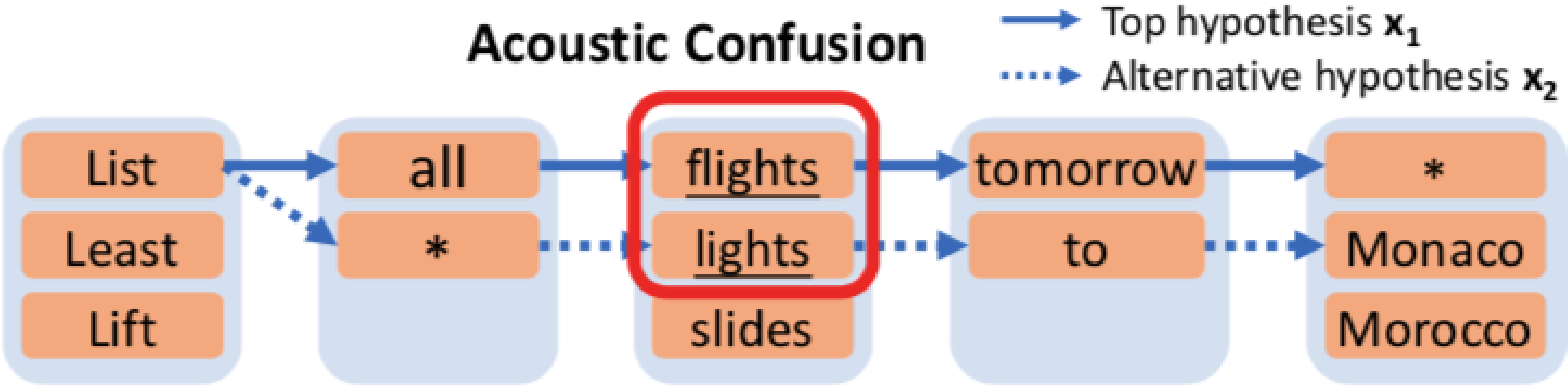


$x_{trs}$ : Show me the fares from Dallas to Boston

$x_{asr}$ : Show me * affairs from Dallas to Boston

# How to determine which words to bring closer?

- Case 2: we only have some ASR transcripts (unsupervised)



# Fine-tuning LMs (ELMo)

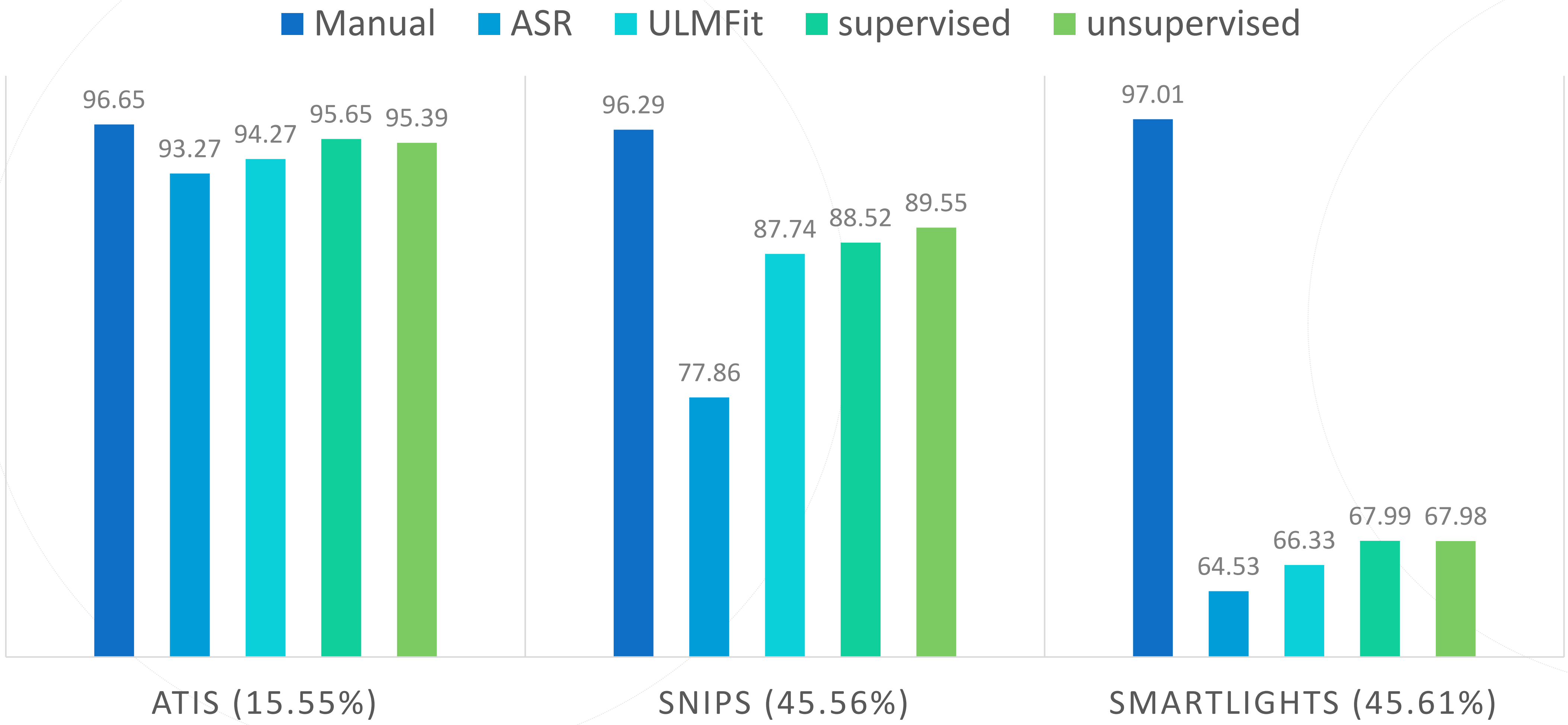
- **ULMFit**:
  Fine-tune w/ LM objective helps domain transfer (Howard and Ruder, 2018)

$$\mathcal{L}_{\text{LM}} = \frac{1}{|x|} \sum_{t=1}^{|x|} - \log p(w_t \mid w_{<t}) - \log p(w_t \mid w_{>t}),$$

$$\mathcal{L}_{\text{conf}} = \frac{1}{|C|} \sum_{c \in C} \sum_{i=0}^{1} 1 - \frac{h_{t_1,i}^{x_1} \cdot h_{t_2,i}^{x_2}}{||h_{t_1,i}^{x_1}|| \, ||h_{t_2,i}^{x_2}||}$$

$$\mathcal{L}_{\text{FT}} = \mathcal{L}_{\text{LM}} + \beta \mathcal{L}_{\text{conf}},$$

# Results


Manual
ASR
ULMFit
supervised
unsupervised
96.65
93.27
94.27
95.65
95.39
96.29
77.86
87.74
88.52
89.55
97.01
64.53
66.33
67.99
67.98
ATIS (15.55%)
SNIPS (45.56%)
SMARTLIGHTS (45.61%)

# Conclusions

- Contextualized models (ELMo) do not transfer well to spoken domain with ASR errors.
- We introduce an additional fine-tuning stage to make embeddings more acoustic-aware.
- We achieve this by forcing embeddings of acoustically similar words to be closer. We propose two methods to extract these pairs
- The experiment results show that our method can make contextualized embeddings more robust to ASR errors.

# Thanks for listening!

Code available at https://github.com/MiuLab/SpokenVec

Chao-Wei Huang

r07922069@ntu.edu.tw

Yun-Nung (Vivian) Chen

y.v.chen@ieee.org