Weak Supervision Techniques towards Enhanced ASR Models in Industry-level CRM Systems

Zhongsheng Wang[1]*, Sijie Wang[1]*, Jia Wang[2], Yung-I Liang[2], Yuxi Zhang[2], and

Jiamou Liu[1]

jiamou.liu@auckland.ac.nz

[1] School of Computer Science, The University of Auckland, New Zealand

[2] Atom Intelligence, Hong Kong SAR, China

31st International Conference on Neural Information Processing (ICONIP 2024) New Zealand

December 3, 2024

About Ourselves

- Associate Professor at UoA
- Leader of LIU AI LAB
- Lead 15+ doctoral students on Web3, AI Agent and multimodality
- Lab Website: https://www.liuailab.org/

Jiamou Liu

- Founder and CEO of Atom-Intelligence Group
- Big data and artificial intelligence service provider
- Company Website: https://atom-intelligence.com/

Jia Wang

- VP of data consulting at Atom-Intelligence Group
- Lecturer of “Digital Transformation in Marketing” at Singapore Management University
- International Mentor of Startupbootcamp Fashion Tech

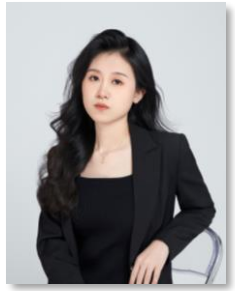
Yung-I Liang

- To start PhD at UoA
- Team member at LIU AI LAB
- Interested in LLM Agent and NLP
- Personal Website: https://www.wzs010429.github.io

Zhongsheng Wang

- Team member at LIU AI LAB
- Data scientist
- Personal Website: https://github.com/swan387

Sijie Wang



- Senior PM of data science at Atom-Intelligence Group
- 6+ years experience of project management and data analysis, rich experience in projects related to data science and AI

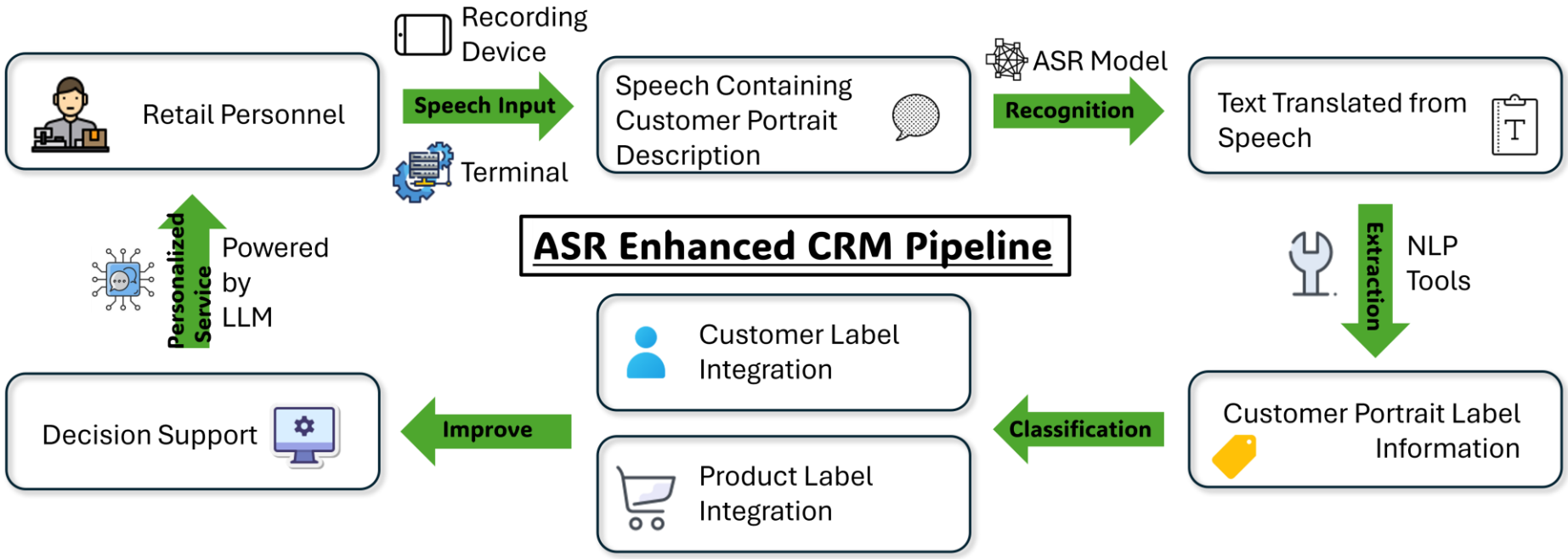
Yuxi Zhang

LIU AI Lab: https://www.liuailab.org
Atom-Intelligence Group: https://atom-intelligence.com/

Content

- Automatic Speech Recognition Enhanced Customer Relationship Management Pipeline
- Difficulties of ASR Model in Industry
- ASR Problem Definition
- Weak Supervision ASR
- Synthetic Data Generation
- Experimental Results
- Detailed analysis

ASR Enhanced CRM Pipeline



1. **Voice Input Capture**: Retail describe customer information
2. **Speech-to-Text Conversion**: Convert speech into text using ASR technology
3. **Data Extraction and Classification**: Extract key customer portrait labels
4. **Data Integration and Analytics**: Integration of fragmented user data
5. **Decision Support**: Personalized customization and recommendations for users
6. **Iterative Learning and Improvement**: Iterative data improvement

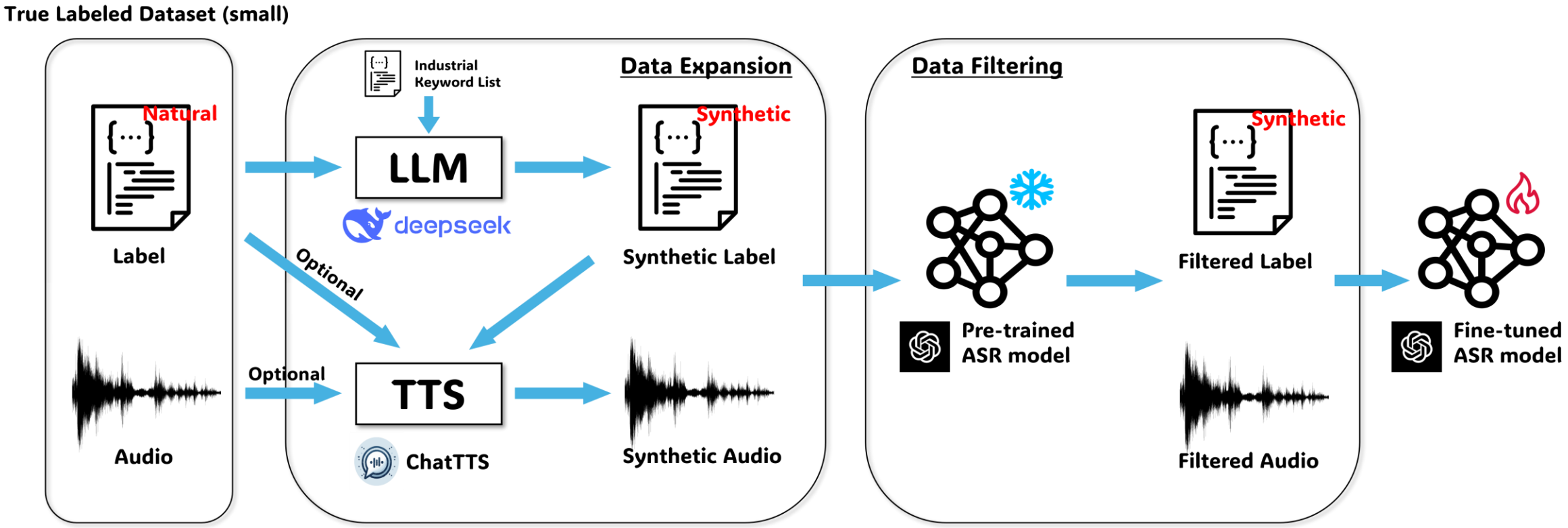
Difficulties of ASR Model in Industry

- Lack of industry user data to fine-tune an ASR model
- High cost of manual labeling
- Low recognition accuracy of proper nouns
- Too many voice variations: accent, noise

ASR Problem Definition

- Provide an ASR model suitable for the CRM pipeline, addressing practical challenges to provide accurate transcription into text data.
- The key challenge is the **lack of real domain-specific labeled data**, in the form of voice-text pairs, for training an accurate ASR model.
- Given a small amount of real audio data $\mathcal{D}_r = \{(\mathbf{x}_i^r, \mathbf{y}_i^r)\}_{i=1}^{N_r}$, $\mathbf{x}_i^r$ is the real audio sample, and $\mathbf{y}_i^r$ is the corresponding text label, with K = {k_1, k_2, . . . , k_m} representing the keyword lists. Use a pre-trained large language model M_{LLM} to generate synthetic text labels $\hat{\mathbf{T}} = \{\hat{\mathbf{y}}_j^s\}_{j=1}^{N_s}$.

Weak Supervision ASR



Original Data | Synthetic Data Generation | Synthetic Data Filtering | Model Training

Synthetic Data Generation (Prompt)

System prompt:

```
Please refer to the following examples and consider the existing
    ↪ product information. Take a deep breath and think carefully—
    ↪ can you provide additional examples? These examples should
    ↪ closely align with the original samples. For product-related
    ↪ information, please use the product list we have provided.
```

User prompt:

```
Below is the product information for your reference:
Attribute: {SOCIAL}; Brand: {BRAND}; Pattern: {LINES};
Material: {MATERIAL}; Product: {NICKNAME}; Series: {SERIES};
Type: {TYPE};
Here are some examples: {SENTENCE}
Based on the template in the examples, please generate {s} new
    ↪ sentences. The content and style should be aligned with the
    ↪ examples.
```

Synthetic Data Generation

Dataset	Samples	Duration (seconds, avg ± std)	Audio length (minutes)
GUCCI100	100	8.58±4.07	14.87
LV100	100	8.92±3.39	14.31
Test Set	1000	8.74±3.34	145.46
LVChatTTS	10000	8.81±3.21	1468.98
GUCCIChatTTS	10000	8.98±3.12	1496.14

GUCCI100 & LV100 & Test Set: Raw data in real industry for different use in this project

LVChatTTS & GUCCIChatTTS: Synthetic data generated using method mentioned in this project

Category	Samples	Example
SERIES	408	objets nomades
TYPE	273	装饰品 (decorations)
BRAND	92	balenciaga
MATERIAL	42	empreinte
NICKNAME	42	水桶包 (bucket bag)
LINES	19	monogram
SOCIAL	3	保值 (value preservation)

Statistics of different types of keywords

Experimental Results

Model	CER	CER_cn	CER_oth	WER	WER_cn	WER_oth
whisper-medium	0.54125	0.50068	0.70524	1.48446	0.98699	0.94388
medium-GUCCI100	0.38658	0.30772	0.45721	0.85850	0.69	0.36364
medium-GUCCIChatTTS	0.21169	0.11968	0.38572	1.30250	0.72773	0.73488
medium-LV100	0.42269	0.34472	0.67785	0.69223	0.47222	0.58127
medium-LVChatTTS	0.21234	0.14897	0.34662	1.13762	0.70	0.88423
medium-GUCCCI&LV	0.20007	0.19243	0.22237	0.77381	0.68225	0.73338
whisper-large-v2	0.11958	0.06540	0.24480	1.56604	0.58	0.36471
v2-GUCCI100	0.12333	0.06677	0.18500	0.70142	0.35	0.35498
v2-GUCCIChatTTS	0.08795	0.07739	0.14031	0.71664	**0.33921**	0.44834
v2-LV100	0.07996	0.05031	**0.12779**	0.99662	0.44989	0.57324
v2-LVChatTTS	0.07743	0.05552	0.17980	0.83076	0.42877	**0.29445**
v2-GUCCCI&LV	**0.07390**	**0.04593**	0.13383	**0.62264**	0.44	0.31461
whisper-large-v3	0.18793	0.09832	0.22471	1.17422	0.82	0.99314
v3-GUCCI100	0.14223	0.07774	0.16648	1.22439	0.897	0.88442
v3-GUCCIChatTTS	0.08732	0.07002	0.09956	1.04436	0.66793	0.98092
v3-LV100	0.11339	0.11042	0.13398	1.11147	0.87	0.93732
v3-LVChatTTS	0.11271	0.10887	0.16643	0.90742	0.72	0.80452
v3-GUCCCI&LV	0.09332	0.07741	0.13346	0.80201	0.69	0.78147

CER: Character Error Rate

WER: Word Error Rate

Lower is better

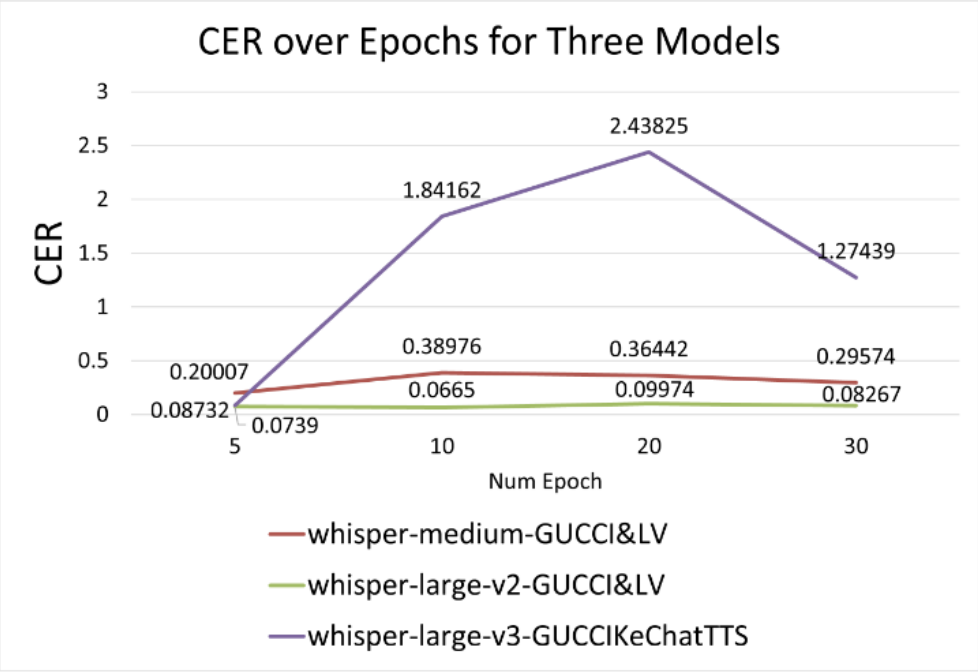
GUCCI&LV: Use two types of synthetic data to fine-tune the ASR model at the same time

Detailed Analysis

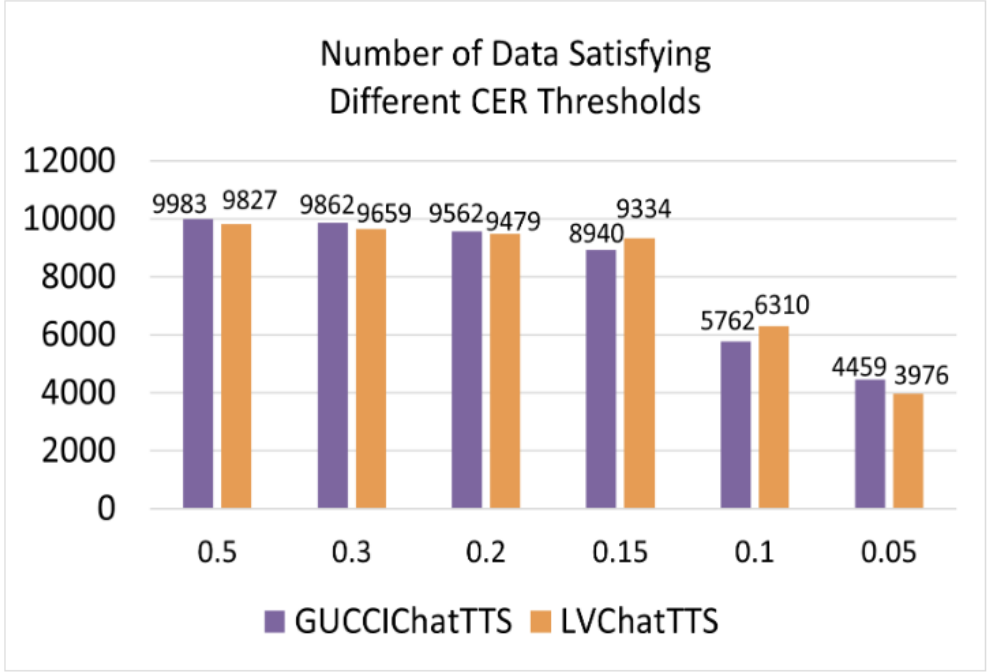
Original Label	Finetuned Model	Inference Result
苏州人，休闲年轻，时尚，喜欢老花，喜欢小包，可推荐男装，喜欢宽松款式。	whisper-medium	苏州，休闲年轻时尚，喜欢老化西湖小包，可推荐男装
	whisper-large-v2	苏州人，休闲，年轻时尚，喜欢老化喜欢小包，可推荐男装。
	whisper-large-v3	苏州人，休闲年轻时尚，喜欢老花喜欢小包，可推荐男装装装装装装装装装装装...
无锡人，喜欢男款西装裤，偏爱复古一点的风格。	whisper-medium	巫溪人，喜欢男款西装，偏复古一点的风格
	whisper-large-v2	无锡人，喜欢男款，西装，裤，偏爱复古一点的风格
	whisper-large-v3	无锡人喜欢阿阿阿阿阿阿阿阿阿阿阿阿阿 i 阿 i 阿 i 阿 i 阿 i 阿 i 阿 i 阿 i 阿...

After fine-tuning, whisper-large-v3 unexpectedly shows worse performance and endless repetition.

Detailed Analysis (Cont.)



CER indicator results after further fine-tuning the three models with 10, 15,20, 30 epoch



Number of remaining data in the two synthetic datasets after filtering with different CER indicators

Finally, we decided to fine-tune for 5 epochs and the CER limit for data filtering was 0.15

Thanks For Listening!

Q & A

Jiamou Liu

jiamou.liu@auckland.ac.nz

School of Computer Science
The University of Auckland