Audio LLM research weekly report – Week 3

Hongyu Jin

ERC task and benchmark

Table 2: The results of Automatic Speech Recognition (ASR), Speech-to-Text Translation (S2TT), Speech Emotion Recognition (SER), Vocal Sound Classification (VSC), and AIR-Bench chat benchmark. Note that for Qwen2-Audio, the results for Fleurs are zero-shot, whereas the results for Common Voice are not zero-shot.

Task	Dataset	Model	Performance	
			Metrics	Results
ASR	**Librispeech** *dev-clean | dev-other | test-clean | test-other*	SpeechT5 (Ao et al., 2021)	WER ↓	2.1 | 5.5 | 2.4 | 5.8
		SpeechNet (Chen et al., 2021)		- | - | 30.7 | -
		SLM-FT (Wang et al., 2023b)		- | - | 2.6 | 5.0
		SALMONN (Tang et al., 2024)		- | - | 2.1 | 4.9
		SpeechVerse (Das et al., 2024)		- | - | 2.1 | 4.4
		Qwen-Audio (Chu et al., 2023)		1.8 | 4.0 | 2.0 | 4.2
		Qwen2-Audio		**1.3** | **3.4** | **1.6** | **3.6**
	Common Voice 15 *en | zh | yue | fr*	Whisper-large-v3 (Radford et al., 2023)	WER ↓	9.3 | 12.8 | 10.9 | 10.8
		Qwen2-Audio		**8.6** | **6.9** | **5.9** | **9.6**
	Fleurs *zh*	Whisper-large-v3 (Radford et al., 2023)	WER ↓	7.7
		Qwen2-Audio		**7.5**
	Aishell2 *Mic | iOS | Android*	MMSpeech-base (Zhou et al., 2022)	WER ↓	4.5 | 3.9 | 4.0
		Paraformer-large (Gao et al., 2023)		- | **2.9** | -
		Qwen-Audio (Chu et al., 2023)		3.3 | 3.1 | 3.3
		Qwen2-Audio		**3.0** | 3.0 | **2.9**
S2TT	**CoVoST2** *en-de | de-en | en-zh | zh-en*	SALMONN (Tang et al., 2024)	BLEU ↑	18.6 | - | 33.1 | -
		SpeechLLaMA (Wu et al., 2023a)		- | 27.1 | - | 12.3
		BLSP (Wang et al., 2023a)		14.1 | - | - | -
		Qwen-Audio (Chu et al., 2023)		25.1 | 33.9 | 41.5 | 15.7
		Qwen2-Audio		**29.9** | **35.2** | **45.2** | **24.4**
	CoVoST2 *es-en | fr-en | it-en |*	SpeechLLaMA (Wu et al., 2023a)	BLEU ↑	27.9 | 25.2 | 25.9
		Qwen-Audio (Chu et al., 2023)		39.7 | **38.5** | 36.0
		Qwen2-Audio		**40.0** | **38.5** | **36.3**
SER	**Meld**	WavLM-large (Chen et al., 2022)	ACC ↑	0.542
		Qwen-Audio (Chu et al., 2023)		**0.557**
		Qwen2-Audio		0.553

Table 1: Multi-task pre-training dataset.

Types	Task	Description	Hours
Speech	ASR	Automatic speech recognition (multiple languages)	30k
	S2TT	Speech-to-text translation	3.7k
	OSR	Overlapped speech recognition	<1k
	Dialect ASR	Automatic dialect speech recognition	2k
	SRWT	English speech recognition with word-level timestamps	10k
		Mandarin speech recognition with word-level timestamps	11k
	DID	Dialect identification	2k
	LID	Spoken language identification	11.7k
	SGC	Speaker gender recognition (biologically)	4.8k
	ER	Emotion recognition	<1k
	SV	Speaker verification	1.2k
	SD	Speaker diarization	<1k
	SER	Speech entity recognition	<1k
	KS	Keyword spotting	<1k
	IC	Intent classification	<1k
	SF	Slot filling	<1k
	SAP	Speaker age prediction	4.8k
	VSC	Vocal sound classification	<1k
Sound	AAC	Automatic audio caption	8.4k
	SEC	Sound event classification	5.4k
	ASC	Acoustic scene classification	<1k
	SED	Sound event detection with timestamps	<1k
	AQA	Audio question answering	<1k
Music&Song	SID	Singer identification	<1k
	SMER	Singer and music emotion recognition	<1k
	MC	Music caption	25k
	MIC	Music instruments classification	<1k
	MNA	Music note analysis such as pitch, velocity	<1k
	MGR	Music genre recognition	9.5k
	MR	Music recognition	<1k
	MQA	Music question answering	<1k

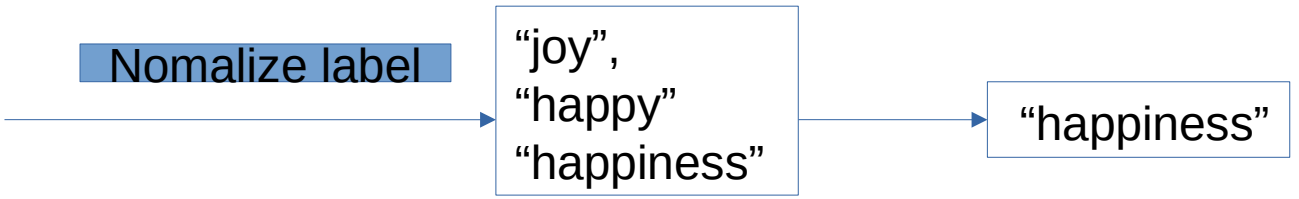
ERC task and benchmark

Table 3: Comparison results (%) on different methods. The best scores are in bold.

Methods	# Param.	MELD		IEMOCAP		EmoryNLP		Avgerage	
		Acc	w-F1	Acc	w-F1	Acc	w-F1	Acc	w-F1
bc-LSTM	1.2M	65.87	64.87	63.08	62.84	40.85	36.84	56.60	54.85
ICON	0.5M	-	-	64.00	63.50	-	-	-	-
MTL	1.2M	62.45	61.90	-	-	36.36	35.92	49.40	48.91
DialogXL	510M	-	62.41	-	65.94	-	34.73	-	54.36
TODKAT	330M	67.24	65.47	61.11	61.33	42.38	38.69	56.91	55.16
CoG-BART	415.1M	64.95	63.82	65.02	64.87	40.94	37.33	56.97	55.34
DialogueRNN	9.9M	65.96	65.30	64.85	64.65	43.66	37.54	58.16	55.83
DialogueGCN	2.1M	63.62	62.68	62.49	62.11	36.87	36.43	54.33	53.14
DialogueCRN	3.3M	66.93	65.77	67.39	67.53	41.04	38.79	58.45	57.36
RGAT	13M	-	60.91	-	65.22	-	34.42	-	53.52
DAG-ERC	9.5M	63.75	63.36	66.54	66.53	39.64	38.29	56.64	56.06
CauAIN	6.1M	65.85	64.89	65.08	65.01	43.13	37.87	58.02	55.92
COIN	0.6M	-	-	66.05	65.37	-	-	-	-
GraphCFC	0.6M	-	58.86	-	68.91	-	-	-	-
SACL-LSTM	2.6M	**67.51**	**66.45**	**69.08**	**69.22**	**42.21**	**39.65**	**59.60**	**58.44**
LLaMA-7B	2.1M	15.09	16.02	19.32	18.24	17.78	17.40	17.40	17.22
Alpaca	2.1M	19.22	18.37	20.35	19.16	17.95	17.33	19.17	18.29
LLaMA 2-7B	4.2M	23.71	24.12	26.73	24.35	25.50	17.27	25.31	21.91
DialogueLLM	4.2M	**71.96**	**71.90**	**70.62**	**69.93**	**41.88**	**40.05**	**61.49**	**60.52**
Improve △		↑ 6.59%	↑ 8.20%	↑ 2.22%	↑ 1.03%	↓ 0.78%	↑ 1.00%	↑ 3.17%	↑ 3.56%

Dataset

IEMOCAP
MELD
MOSEI
EmoWOZ
CRRMA-D



Dataset format


```
{
    "id": "sample_001",
    "audio": {
      "path": "path/to/audio1.wav",
    },
    "text_context": [
      {"speaker": "User", "text": "I just finished my project presentation."},
      {"speaker": "Assistant", "text": "How did it go?"}
    ],
    "target_utterance": {
      "text": "It was terrible. Nobody seemed interested and my boss barely looked up.",
      "emotions": {
        "primary": "disappointed",
        "secondary": "frustrated",
      }
    },
    "desired_response": {
      "text": "I'm sorry to hear your presentation didn't get the reception you hoped for. That must be really disappointing, especially after all your preparation. Did you get any feedback that might be helpful, or would you like to talk about what aspects felt particularly frustrating?",
      "response_strategy": "validation_then_exploration",
      "emotion_awareness": "acknowledges_disappointment_and_frustration"
    }
  },
```

THE UNIVERSITY OF MELBOURNE

What about records of semi-autonomous bodies or subsidiary companies?

Label: Afraid
Situation: Speaker felt this when...
"I've been hearing noises around the house at night"
Conversation:
Speaker: I've been hearing some strange noises around the house at night.
Listener: oh no! That's scary! What do you think it is?
Speaker: I don't know, that's what's making me anxious.
Listener: I'm sorry to hear that. I wish I could help you figure it out

Label: Proud
Situation: Speaker felt this when...
"I finally got that promotion at work! I have tried so hard for so long to get it!"
Conversation:
Speaker: I finally got promoted today at work!
Listener: Congrats! That's great!
Speaker: Thank you! I've been trying to get it for a while now!
Listener: That is quite an accomplishment and you should be proud!

Figure 2: Two examples from EMPATHETICDIALOGUES training set. The first worker (the speaker) is given an emotion label and writes their own description of a situation when they've felt that way. Then, the speaker tells their story in a conversation with a second worker (the listener).

Comfort user?
Fight back?
Make a joke?

THE UNIVERSITY OF MELBOURNE

How to make a prompt template

Prompt Template

Instruction $\mathcal{I}$

Now you will play the role of Yuan Bao in the dialogue.Below is the background information for your dialogue with Xia Bing

Background $\mathcal{B}$

Xia Bing went to work and Saw Yuan Bao,who hadn't returned all night.She blocked his way and questioned about the reason for not coming back the previous night.....

Historical Dialogue $\mathcal{H}$

Here is the historical dialogue between you and Xia Bing.

Xia Bing:"Has your life changed since getting married and having children util now ?"
Yuan Bao:"Yes, I haven't changed."
Xia Bing:"You should eat, drink, and sleep."
Yuan Bao:"That's right."

Current Utterance u

Xia Bing:"For me, from top to bottom, inside and out, everything about me has changed except for my name, Xia Bing. I've lost everything."

Video Clip E_V

Here is a video clip of Xia Bing speaking.

<video token> <video token> <video token>.....

Demonstration Retrieval $\mathcal{R}$

Please choose an appropriate emotion from <happy, surprised, sad, disgusted, fearful, angry, neutral> to continue the conversation.

Demonstration:
digusted:"Alright, alright, no need to get worked up!"

For negative emotions (sadness, anger, fear, etc.):
Validation/Acknowledgment - Recognizing the emotion without judgment
Empathetic response - Showing understanding of the feeling
Problem-solving - Offering solutions or asking how you can help
Redirection - Gently shifting focus to more constructive thoughts
De-escalation - Calming techniques for intense emotions
Normalization - Reassuring that such feelings are normal
For positive emotions (happiness, excitement, pride, etc.):
Celebration/Amplification - Reinforcing and joining in the positive emotion
Inquiry - Asking to learn more about the positive experience
Gratitude expression - Expressing appreciation for sharing good news
Reciprocal enthusiasm - Matching their energy level
Future-oriented discussion - Building on the positive momentum
For neutral or mixed emotions:
Clarification - Seeking to better understand the emotional state
Open-ended questioning - Inviting deeper exploration
Mindful presence - Simply being present without directing the conversation
Summarization - Reflecting back the content and emotional undertones
I'll provide examples of these strategies in the context of dialogue responses to different emotions.

Evaluation

The SECEU test

Item 1
Story: The airplane model that Wang made fell from the sky one minute after take-off. When she inspected the model, she found a part that could possibly be improved. At this moment, she would feel:

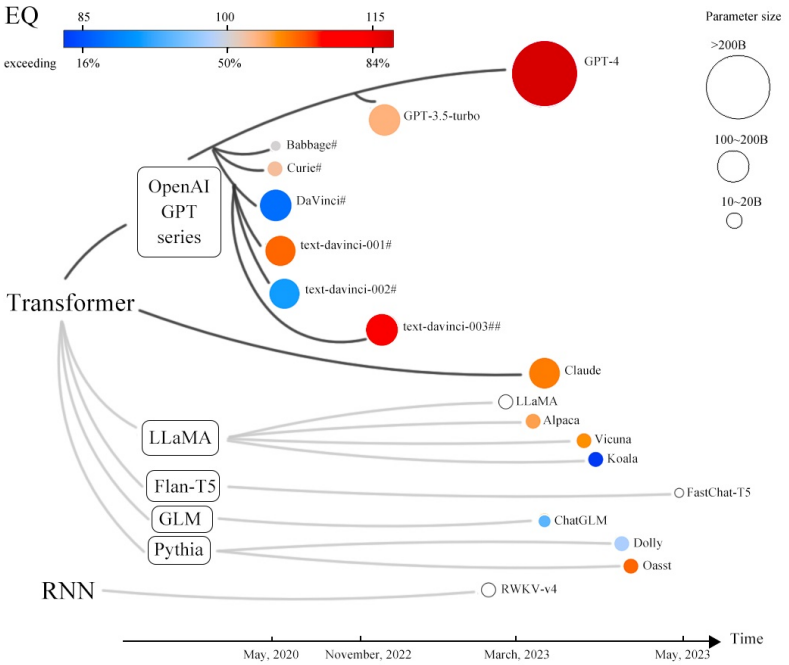
Options:	(1) Expectant	(2) Excited	(3) Joyful	(4) Frustrated
	3.56	3.09	1.78	1.57

Item 2
Story: Although Aunt Li pays close attention to her lifestyle and diet, she still got a malignant tumor. The chances of curing the tumor now are quite slim, and Aunt Li can do nothing about it. She would feel:

Options:	(1) Desperate	(2) Fear	(3) Helpless	(4) Sad
	3.45	2.04	2.17	2.33

The Standard Scores:

40 × 4 matrix



THE UNIVERSITY OF
MELBOURNE