Turn-taking and Backchannel Prediction with Acoustic and Large Language Model Fusion

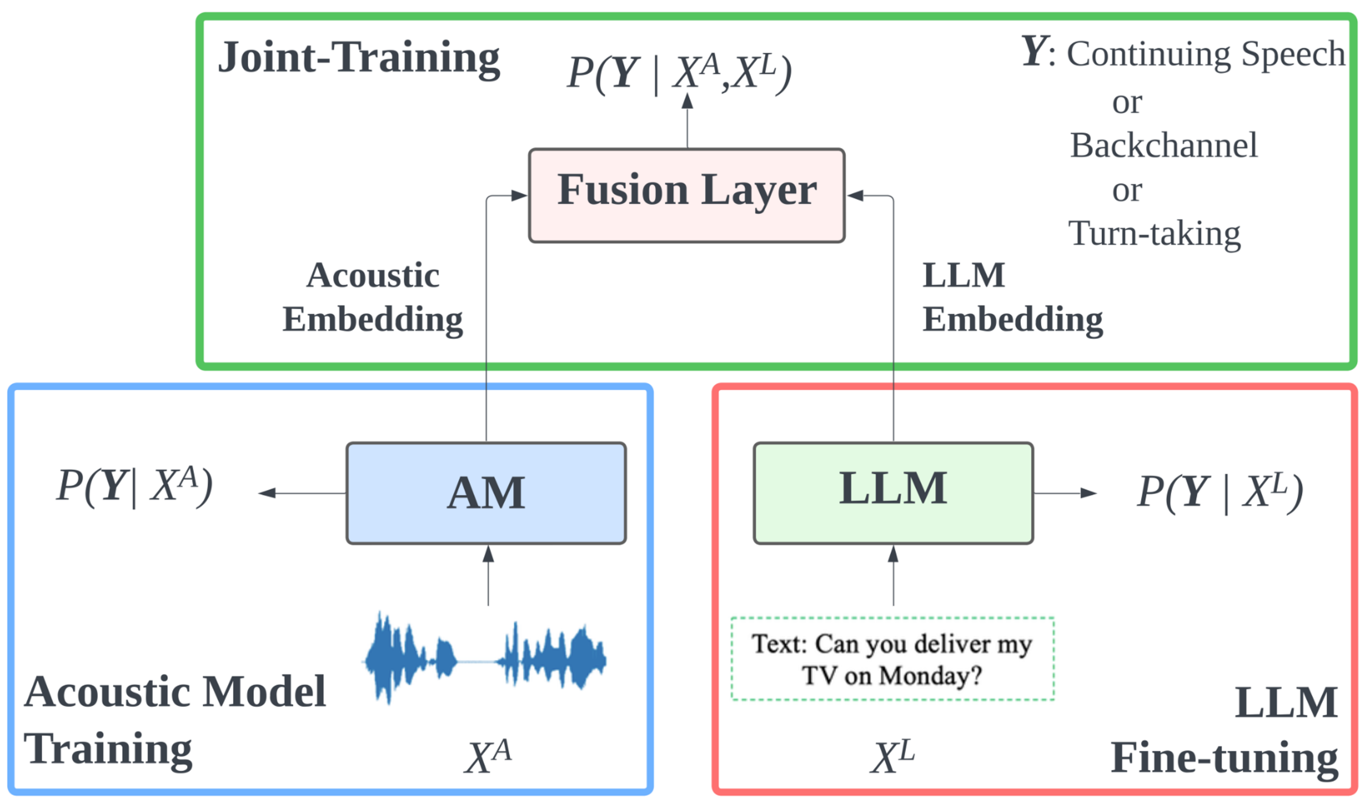
Jinhan Wang[1], Long Chen[2], Aparna Khare[2], Anirudh Raju[2], Pranav Dheram[2], Di He[2], Minhua Wu[2], Andreas Stolcke[2], Venkatesh Ravichandran[2]
[1] University of California, Los Angeles, USA
[2] Amazon Alexa AI, USA

ICASSP 2024 KOREA

I. Introduction

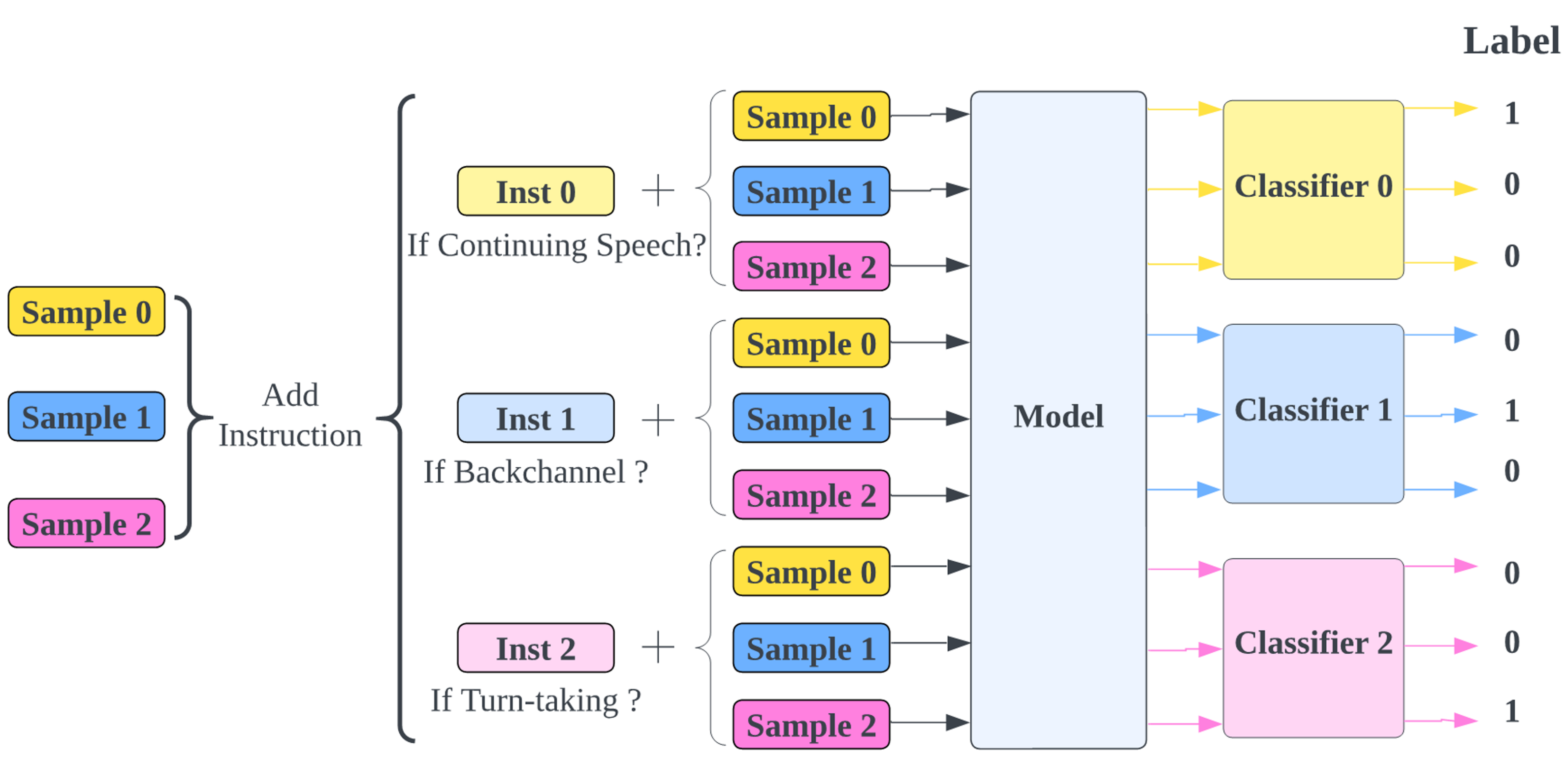
- Objective: More natural Voice Assistant System
  - Human-human like conversational experience.
  - Chat instead of query-answer / no push-to-talk.
  - Proper backchanneling and turn-taking.
- Motivations:
  - Large language models (LLM) promise to better capture formal dependencies and meaning relations in language.
  - Fusion of LLM and acoustic models (AM) for dialogue modeling has not been extensively studied.
- Contribution:
  - Propose a fusion framework for turn-taking and backchannel prediction with LLM and acoustic models.
  - A novel instruction fine-tuning method in multi-task manner to unlock LLM's power rather than text encoding.

II. Multi-Modal Fusion



- Joint Modeling: embedding late fusion with classifier head
- Fusion Options (Opt):
  - Opt1: load pre-trained LLM; trainable: LM, classifier
  - Opt2: load fine-tuned LLM and freeze; trainable: classifier

III. Multi-task Instruction Fine-tuning



- Augment each sample 3 times with 3 instructions (Inst).
- *Classifier i* 's update is only triggered by samples with *Inst i*, where $i \in \{0, 1, 2\}$
- Shared LLM backbone model.

IV. Experiments

- Dataset: Switchboard
  - 2438 dialogues, two speakers each, ~260hrs
  - Continuing speech/ Backchannel / Turn-taking is labeled by self-defined heuristics at each token end.
  - Statistics (Cont vs Back vs Turn tokens)
    - Train: 71k (downsample) vs 56k vs 86k
    - Dev: 6k (downsample) vs 5k vs 7k
    - Test: 123k vs 2.4k vs 3.2k
- Model:

Model		#param	Fine-tune	Trainable
AM	HuBERT	95M	Frozen	0%
LM	GPT2	117M	Full	100%
	Redpajama	3B	LoRA	0.4%

- Evaluation: Area-under-the-curve(AUC), Equal Error Rate(EER) for each class and in average.

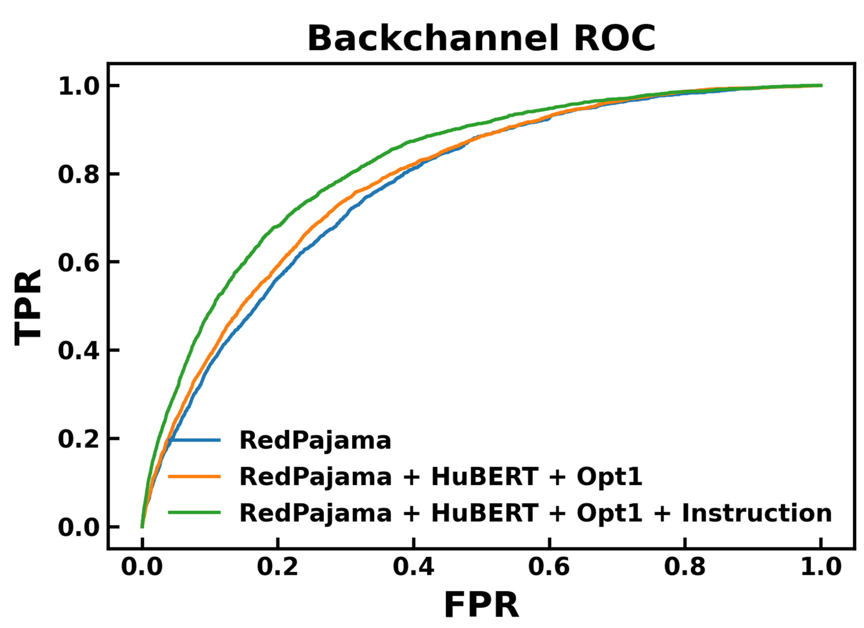
V. Results and Discussion

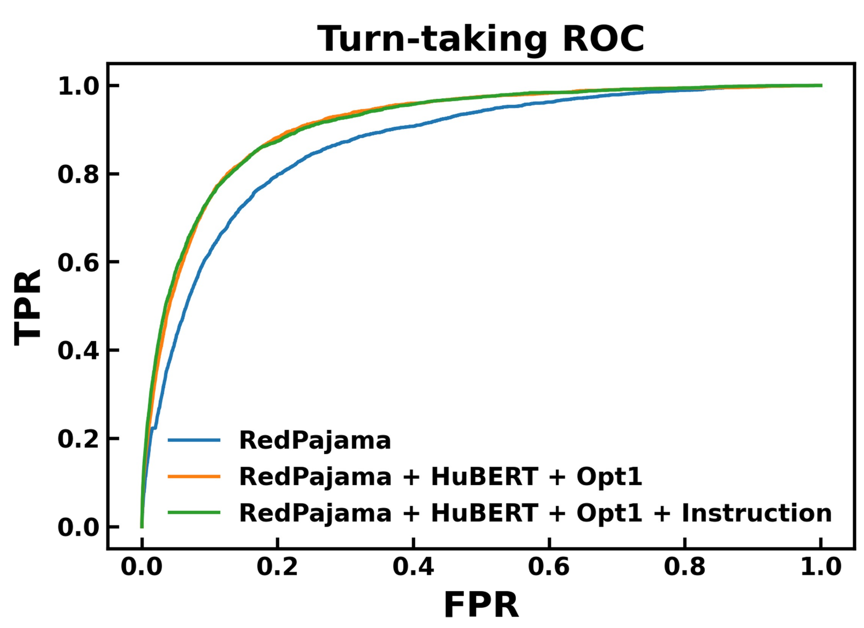
Table 1: Results for single modality and fusion models.

Method	AUC(Cont)	AUC(Back)	AUC(Turn)	AUC(avg)	EER(avg)
HuBERT	0.7323	0.6455	0.7401	0.7060	34.87
GPT2	0.8510	0.7744	0.8623	0.8292	24.47
+ HuBERT Opt1	0.8783	0.7798	0.884	0.8474	22.63
+ HuBERT Opt2	0.8778	**0.7862**	0.8859	0.8500	22.77
RedPajama	0.8629	0.7739	0.8685	0.8351	23.60
+ HuBERT Opt1	**0.8992**	**0.7862**	**0.9116**	**0.8657**	**20.33**
+ HuBERT Opt2	0.8982	0.7743	0.9006	0.8577	21.57

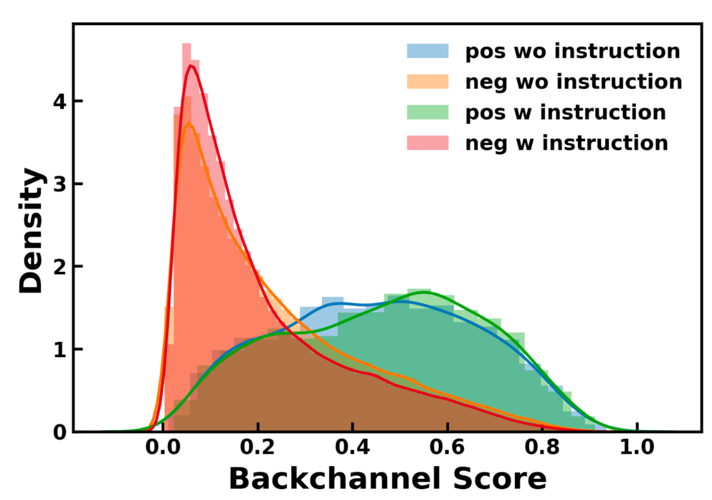
Table 2: Results with multi-task instruction fine-tuning.

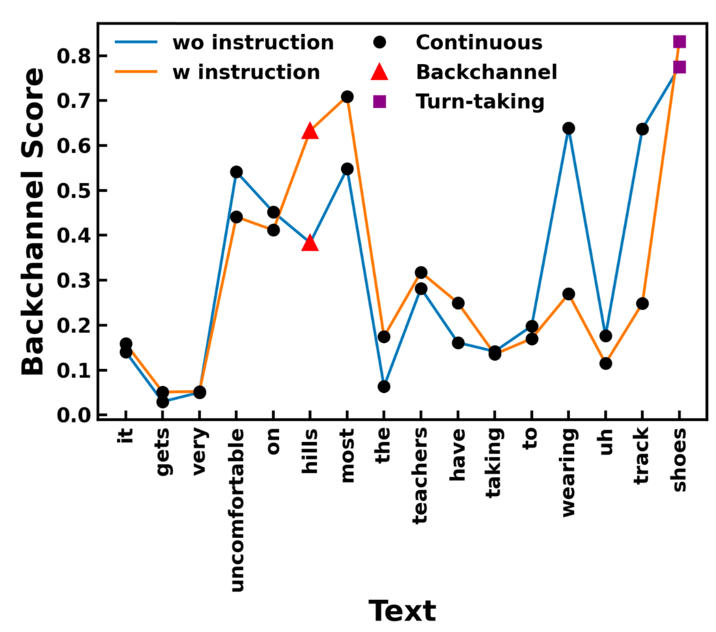
Method	AUC(Cont)	AUC(Back)	AUC(Turn)	AUC(avg)	EER(avg)
GPT2	0.8416	0.7863	0.8582	0.8287	24.13
+ HuBERT Opt1	0.8726	0.7901	0.8766	0.8464	22.50
+ HuBERT Opt2	0.8806	0.7838	0.8890	0.8511	22.23
RedPajama	0.8668	0.8097	0.8796	0.8520	21.80
+ HuBERT Opt1	0.9000	**0.8229**	0.9127	0.8785	19.50
+ HuBERT Opt2	0.8980	0.8182	0.9129	0.8764	19.60
RedPajama + History	0.8747	0.8074	0.8912	0.8578	21.63
+ HuBERT Opt1	**0.9029**	0.8184	**0.9197**	**0.8803**	**19.30**





- Redpajama > GPT2. It benefits more from Instruction fine-tuning.
- Turn-taking prediction benefits remarkably from fusion.
- Multi-task Instruction fine-tuning improves backchannel the most.
- Including dialogue history in instruction only improves continuing speech and turn-taking prediction. Backchannel is more local.





- Pos & Neg backchannel scores are pushed to the range ends.
- Backchannel relay most on syntactic context. Instruction helps.