# Livia Qian

KTH Royal Institute of Technology
Lindstedtsvägen 24, floor 5
Stockholm, Sweden
114 28

liviaq@kth.se

## 1 Research interests

My research interests lie generally in the area of **natural language processing (NLP)**, **text-to-speech (TTS)** and **automatic speech recognition (ASR)**, with a special focus on **dialogue modeling** with both **supervised** and **unsupervised learning**.

With respect to spoken dialogues, I focus on creating representations that can **model prosody**, turn shifts, pauses, backchannels and improve the accuracy of two-channel speech generation, as well as solve downstream tasks like emotion and laughter detection. I am also interested in **online ASR** and **speech synthesis** for dialogues.

With respect to written dialogues, I aim to improve spoken dialogue representations with text and other modalities through **multimodal learning**, as is common in ASR. I also try to improve language models using other modalities like speech, among others (e.g., vision and motion), and solve less-researched dialogue tasks using language models (like **reference resolution (RR)** and **coreference resolution (CRR)**).

### 1.1 Resolving References in Visually-Grounded Dialogue via Text Generation

RR is about finding linguistic elements that are semantically related or refer to the same entity. CRR is similar but instead of connecting textual elements, it resolves references to entities in other modalities. In our case, we tried to identify references to images in dialogues. This is challenging because information can be scattered across the dialogue history and pronouns are especially hard to resolve, e.g., when used by multiple speakers.

We used the dataset *A Game of Sorts* by Willemsen et al. (2022). It consists of dialogues in which pairs of speakers were tasked with ranking a set of images based on predefined criteria. The points at which the speakers referred to specific images were marked during data collection, thus making it possible to connect the mentions and utterances to the images.

The modeling consists of two steps: first, we fine-tuned a language model (GPT-2 and GPT-3) to summarize what the speakers said about the different images up to a certain point in a dialogue; in each case, the reference(s) in the most recent utterances are pointed out to indicate which image(s) we are interested in generating the summaries for, for which the model is supposed to use the previous mentions belonging to the very same image(s). After this, the caption-like summaries were passed to a vision-language model to identify the most likely images. The correctness of the generated summaries were compared against simple pronoun substitutions and the state-of-the-art models in CRR.

The paper (Willemsen et al., 2023) has been accepted at SIGDIAL 2023. We showed that discourse processing is possible to frame as a causal language learning problem and that large language models can be fine-tuned to generate referent descriptions.

### 1.2 ASR on multi-channel dialogues

The field of ASR is widely addressed by the research community, but does not work well on dialogues where there is overlapping speech or different adversarial effects (e.g., backchannels). I am working on two-speaker ASR by conditioning the speech representations of the two channels on each other before connecting them with their respective transcripts. There is a possibility for extending this to on-the-fly (online) speech recognition.

The resulting models and representations could be used to solve different downstream tasks, e.g., turn taking and emotion detection, as well as speech synthesis and NLP-related tasks. Such models could improve the conversational skills of social robots or the time-alignment of video transcripts. These tasks have not received not much attention in text-based large language models (LLMs) as the main focus has been laid on content and memory improvement, but with the emergence of multimodal NLP and LLM-based chatbots, aspects like turn taking could be crucial to improve the flow of the conversation.

### 1.3 Dialogue speech synthesis

I might also consider working on speech synthesis. Some issues within this field are that there is usually not enough data to train on and that the generated speech is not expressive enough. Although current systems are enough for many use-cases, there is room for improvement in many aspects, e.g., delivery, prosody expression, controllable pauses and long-term effects. I think that speech synthesis may benefit from the speech representations from my other projects, especially in relation to dialogues where these aspects are highly dependent on context.

*Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems*, pages 33–34
September 11–12, 2023. ©2023 Association for Computational Linguistics

## 2 Spoken dialogue system (SDS) research

**Where do you think the field of dialogue research will be in 5 to 10 years?** Dialogue research is becoming more and more important as TTS, speech-to-text (STT) and language models work fairly well on monologues, continuous text and short snippets (e.g., sentences) but do not usually take into account data more complex that these. Dialogues can also provide context that is often missing from monologues (e.g., disambiguation of terms and affirmation) which can aid in better response generation when it comes to speech synthesis, for example. What I think will improve in the upcoming years is the connection between content and prosody as well as other non-verbal cues.

**What do you think this generation of young researchers could accomplish in that time?** Young researchers can improve the fine-grained details to make conversations more human-like, with respect to both content and the naturalness of non-verbal signs. Another thing that is worth looking into is how to make use of long-term dependencies (similarly to how language models do) and representing dialogue history (previous dialogue sessions) by compressing previous information.

**What kind of questions need to be investigated to get the field to that point?** I think it is really important to create high-quality representations similar to wav2vec and HuBERT but for dialogues and dialogue context, either with the inclusion of prosody or separately from it.

**What are the most important things for users of SDSs?** Natural-sounding speech, informativeness and correctness of generated responses, identifying turn shifts in online systems, knowledge and visual grounding, correct reaction to human emotions, multimodal cues.

**Is there a difference between SDS research in academia and industry?** Industry and commissioned research are more profit-oriented than general academic research, and as such, they focus on hands-on and immediate applications. This narrows down the possible research questions, creates the need for patents and NDAs and requires system integration (e.g., in video games).

**Will SDSs be more widely used in the future? How? In what scenarios?** As TTS and ASR are considered to be solved for many use-cases, the areas where SDS can be used becomes more and more niche. Nevertheless, as mentioned before, the dialogue system of robots could greatly benefit from this field, as well as interactive GPS systems, general-use chatbots, personal assistants, video game scripts and video transcriptions (especially for interviews and movie subtitles).

## 3 Suggested topics for discussion

- How to keep up with the rapid development of large language models? How can SDS benefit from this? How can we improve models to keep up the pace?
- Can SDS research focus on paralinguistic elements while making use of the recent advancements in NLP to focus on the purely linguistic aspects? What kind of linguistic concepts could be useful to address and apply?
- Multi-model learning: how can SDS be combined with other modalities (e.g., vision, text, gestures)? What is the state-of-the-art in these topics?
- Modular vs. holistic models.
- Augmenting spoken dialogue data: how could we address the lack of data as a common problem in this field? What about speaker variation and prosody?
- Standardized frameworks and libraries. Common ways to do ablation studies.
- Frameworks and repositories for setting up user studies and data collection. Useful datasets.

## References

Bram Willemsen, Dmytro Kalpakchi, and Gabriel Skantze. 2022. Collecting Visually-Grounded Dialogue with A Game Of Sorts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 2257–2268. https://aclanthology.org/2022.lrec-1.242.

Bram Willemsen, Livia Qian, and Gabriel Skantze. 2023. Resolving References in Visually-Grounded Dialogue via Text Generation (under review). In *SIGdial, ACL Anthology*.

## Biographical sketch

Livia Qian is a PhD student at KTH Royal Institute of Technology in Stockholm, Sweden. Her research interests include NLP, ASR and representation learning. Her PhD project is about representing spoken and written dialogues using machine learning models. Her current work focuses on multi-channel ASR for improving turn-shift detection and backchannel transcription, among others.

Livia has a Bachelor's degree in Computer Science a Master's degree in Machine Learning. Before joining academia, she worked at an IT company as a software developer. Her extra-curricular interests include board games, language learning and swimming.