MoTT: A Speech Dataset for Modular Composition of Turn-Taking Conversations

Giulio Salada
Dept. of Computer Science
University of Milan
Milan, Italy
giulio.salada@studenti.unimi.it

Davide Fantini
Dept. of Computer Science
University of Milan
Milan, Italy
davide.fantini@unimi.it

Federico Avanzini
Dept. of Computer Science
University of Milan
Milan, Italy
federico.avanzini@unimi.it

Giorgio Presti
Dept. of Computer Science
University of Milan
Milan, Italy
giorgio.presti@unimi.it

2025 Immersive and 3D Audio: from Architecture to Automotive (I3DA) | 979-8-3315-5828-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/I3DA65421.2025.11202114

***Abstract*—Among the numerous speech datasets in the literature, only a minority concerns conversational data, and even fewer datasets isolate the elements occurring in turn-taking conversations. To address this gap, this paper presents MoTT, an English speech dataset composed of questions, answers, reciprocal questions, and backchannel responses recorded by eight participants. The questions and answers pertain to ten topics and were recorded in two takes. The voice directivity pattern was simultaneously captured at frontal and lateral positions by two microphones. The MoTT dataset was designed to provide interchangeable conversational elements and enable their modular composition to obtain fictional but plausible and convincing conversations. As a result, multiple virtual speakers engage in a turn-taking conversation that emulates real-world interactions, with spatial audio techniques employed to enhance realism by arranging the speakers in the auditory scene. This dataset offers a valuable resource for studies in immersive spatial audio, human-computer interaction, and auditory scene analysis. The dataset is therefore well-suited for experiments that necessitate the simulation of ecologically valid conversations, as the one described in the use case reported in this paper.**

***Index Terms*—Dataset, speech, audio recording, turn-taking.**

I. Introduction

Human conversations are typically characterized by a structured pattern of interaction known as *turn-taking* [1], [2]. In this type of organization, participants take turns to talk, thereby alternating between the roles of speaker and listener. In virtual acoustic environments (VAEs), the user is acoustically immersed in a simulated room characterized by virtual sound sources rendered in different positions [3], [4]. The use of speech data as virtual sources constitutes a common application scenario [5], [6]. A more realistic VAE is attained when virtual speakers interact in a turn-taking conversation [7]–[11]. From an acoustic perspective, these scenarios can be constructed through a variety of approaches to spatial audio rendering. For example, in audio augmented virtuality (AAV) [6], the VAE is derived from auditory content that has been captured from the real world. To this end, spatial room impulse responses (SRIRs) provide comprehensive acoustic modeling of the captured environment by encoding the directions of arrival of direct and reflected sound waves reaching the listening point. Speech recordings can then be convolved with the SRIR measurements, thereby simulating virtual speakers at various positions in the VAE. This approach engenders an immersive and ecologically valid acoustic experience, wherein users can explore the auditory scene as if they were in the place where the SRIR was recorded. From a communication perspective, the mechanisms governing turn-taking must be understood and modeled to simulate interactions between virtual speakers that comply with the dynamics of human conversations. The simulation of virtual speakers conversing with one another is typically based on audio recordings from speech datasets. However, despite the plethora of speech datasets proposed in the literature, only a minority is conversation-oriented, i.e., collecting recordings related to human conversations. The majority of these datasets capture entire conversations, disregarding the collection of isolated conversational elements that can be flexibly rearranged to generate artificial yet natural-sounding dialogues.

To address this gap, this paper presents MoTT (Modular Turn-Taking), an English speech dataset specifically designed to isolate key components of human conversations. A total of eight participants were assigned the task of articulating a question and its corresponding answer for each of ten predefined topics. The MoTT dataset includes two takes of these recordings. Furthermore, participants were instructed to utter other speech elements that typically occur in a conversation. These included backchannel responses and reciprocal questions, defined as short generic queries used to redirect the question back. The objective of the recorded speeches is to serve as interchangeable elements that can be assembled in a modular fashion. The proper arrangement of these elements enables the construction of fictional conversations that closely mimic the mechanisms of natural human interaction. In VAEs, this results in virtual speakers engaging in realistic conversations, thereby enhancing the listener's sense of presence in the acoustic simulation. For this reason, MoTT is particularly well-suited to be employed in combination with SRIR datasets [24]–[27] for the simulation of ecologically valid VAEs from both acoustic and human interaction perspectives. Therefore, MoTT offers a significant resource for research and applications in immersive spatial audio as well as human-computer interaction, and auditory scene analysis. The dataset

This work is part of SONICOM, a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101017743.

TABLE I: Comparison of some of the published conversation-oriented speech datasets.

Dataset	Measurement context	Speech type	Language	Number of speakers	Measurement environment	Size	Recording device	Sample rate [kHz]	Additional data
ISL [12]	In-the-wild	Real meeting sessions in given scenarios	English	∼6.4	Office	104 sessions (103 hours)	Lavalier and table mic.	16	Video, transcript
ICSI [13]	In-the-wild	Real meeting sessions	English	12	Meeting room	75 sessions (72 hours)	Head-worn and table mic.	16	Video, transcript and annotations
AMI [14]	In-the-wild and laboratory	Real meeting sessions and elicited meetings with roles and tasks	English	4	3 meeting rooms	100 hours	Omni and headset mic. (close-talking), circular mic. arrays (far-field), and manikin (binaural)	16	Video, transcript, annotations, and whiteboard activity
CHIL [15]	In-the-wild	Real lectures and meetings	English	10–20 3–5	5 meeting rooms	86 sessions	64-channel and 4-channel mic. arrays, table and lavalier mic.	44.1	Transcript, video, and annotations
CCDb [16]	Laboratory	Natural conversations with suggested topics	English	2	Laboratory	30 dialogues (5 hours)	Unspecified microphone	44.1	Transcript, video, and annotations
ACE Challenge [17]	Laboratory	Single words and numbers, answers to open-ended questions	English	14	Anechoic chamber	50 utterances	Monophonic mic.	48	No
NoXi [18]	Laboratory	Screen-mediated natural conversations	7 languages	2	Laboratory	84 dialogues (25 hours)	Headset mic.	48	Transcript, video, depth images and annotations
CHiME-6 [19]	In-the-wild	Real dinner parties	English	4	Kitchen, dining and living rooms	20 sessions (50 hours)	4-channel mic. arrays and binaural mic.	16	Transcript, video, and annotations
ODSQA [20]	Laboratory	Reading existing documents and questions	Chinese	20	?	3,654 questions	?	16	Transcript
AISHELL-4 [21]	In-the-wild	Real meeting sessions	Mandarin	4–8	10 meeting rooms	211 sessions (120 hours)	8-channel mic. array	48	Transcript and voice activity
Adigwe and Klabbers [22]	Laboratory	Single sentences, scripted and elicited dialogs	English	2	Recording studio	23 dialogues	?	48	None
DailyTalk [23]	Laboratory	Scripted dialogs	English	2	Studio	2,541 dialogues (20 hours)	Monophonic mic.	44.1	Transcript, topic, and emotion
MoTT (Ours)	Laboratory	Open-ended answers and questions, reciprocal questions, and backchannel responses	English	8	Acoustically treated chamber	Two takes of 10 questions and answers pairs ($>$ 2 hours)	Frontal and lateral monophonic mic.	48	Transcript

is therefore suitable for experiments requiring the simulation of ecologically valid conversations, as illustrated by the use case described in this paper. The MoTT dataset is publicly available on Zenodo [28].

This paper is organized as follows. After the present introduction (Section I), Section II provides an overview of the related datasets proposed in the literature. Section III describes the MoTT dataset including the technical setup, the experimental procedure, and the postprocessing operations. Section IV presents a use case of MoTT, in which the speech recordings were assembled to generate realistic conversations for an experiment in an AAV framework. Section V concludes the paper.

II. Related Work

A plethora of speech datasets have been proposed in the extant literature. The variations between these datasets pertain to a multitude of factors, including but not limited to language, speech type (e.g., reading, conversation), number of involved speakers, measurement context (laboratory or in the wild), and technical setup. In the following, an array of these datasets will be presented, with a particular focus on those oriented towards conversations, such as the dataset described in this paper. Table I presents a comparative analysis of the characteristics of these datasets.

The majority of conversation-oriented speech datasets consist of recordings of conversations occurring between people in uncontrolled settings. AISHELL-4 [21] includes 211 real meeting sessions captured in different rooms with Mandarin speakers talking to each other. The recordings were conducted with a circular microphone array. Transcriptions and speaker activity were provided as additional data. The dataset was intended for speech front-end processing, speech recognition, and speaker diarization. The CHIL corpus [15] collects audio and video streams of 86 real lectures and meetings. The dataset is accompanied by several manual annotations, including speaker turns and identities, acoustic conditions, named entities, and video annotations. The AMI meeting corpus [14] collects audio and video recordings of both real meeting sessions and meeting elicitation scenarios in which speakers were assigned roles and tasks. The provided manual annotations included named entities, dialogue acts, topics, summaries, and emotions. A similar dataset that was collected in a working context is the NIST meeting room corpus [29]. With regard to alternative contexts, the CHiME-6 dataset [19] comprises audio and video recordings of 20 dinner parties

that occurred in domestic environments. The ISL meeting corpus [12] includes audio and video recordings of meetings in which participants conversed within predefined scenarios. The CCDb [16] is an audiovisual dataset comprising 30 dyadic conversations with annotations for facial expressions, verbal and non-verbal utterances, and speech transcriptions. While recorded under controlled laboratory conditions, the dataset captures non-scripted conversations. Similarly, the NoXi dataset [18] provides audiovisual recordings and annotations of multi-lingual spontaneous dyadic conversations across a wide range of topics. However, a key distinction from the previous datasets is its exclusive focus on screen-mediated interactions.

These datasets captured uncontrolled conversations that occurred in the wild. For the purpose of modular composition of turn-taking, it is ideal to record speech data under controlled settings, preferably in anechoic chambers, to enable the later integration of acoustic cues. This is the case of the corpus of the ACE challenge [17]. Besides providing SRIRs, this corpus includes English speeches recorded in an anechoic chamber at a sample rate of 48 kHz and 16-bit depth. The speech content includes single words and numbers, as well as some longer answers to predefined questions. However, the recording of the questions was not part of the dataset. The speech dataset DailyTalk [23] includes more than two thousand dialogues recorded by two English-fluent speakers. These were sampled from another dataset of written dialogues. The dataset was intended for conversational text-to-speech. Similarly, the ODSQA dataset [20] includes audio recordings of Chinese speakers reading text documents and related questions retrieved from an existing dataset. However, some details about the dataset are missing, such as the measurement room and recording devices. Further, Adigwe and Klabbers [22] recruited two professional voice talents to read scripted dialogues out loud from existing sources. They considered different recording setups, including single sentences, the whole scripted dialogues, and semi-spontaneous dialogues within given scenarios. However, the dataset is not publicly available.

Other worth citing speech datasets oriented to conversation scenarios used low-quality recording devices. These datasets encompass those that collect audio recordings of telephone conversations, such as DSTC2 [30] and SpokenWOZ [31].

At the time of writing, generative artificial intelligence (AI) is capable of producing convincing and highly realistic voices that may render most speech datasets obsolete for many of the use cases for which they were recorded. Also, directivity, microphone distance, and flat microphone response can be corrected after generation, but at the cost of increasing experimental variables and complexity (to our knowledge, no publicly available AI is currently able to natively account for these factors). Nevertheless, we believe that AI will give real voice datasets a new value as training and test sets. This encompasses the training of both generative models and models capable of discriminating between real and synthetic speech.

III. Dataset

A. Technical Setup

1) Room Setup: The speech recordings for the MoTT dataset were conducted in an acoustically-treated chamber at the University of Milan (Italy), which is shown in Fig. 1. This chamber is an environment where sound reflections are significantly damped by the walls, although it is not a fully anechoic room. In addition, sound absorption panels were placed on the window located at one of the room's walls as well as at the room's corners. This configuration favored the attenuation of sound reflections and resonances, which is needed to obtain speech audio signals that are as dry as possible.

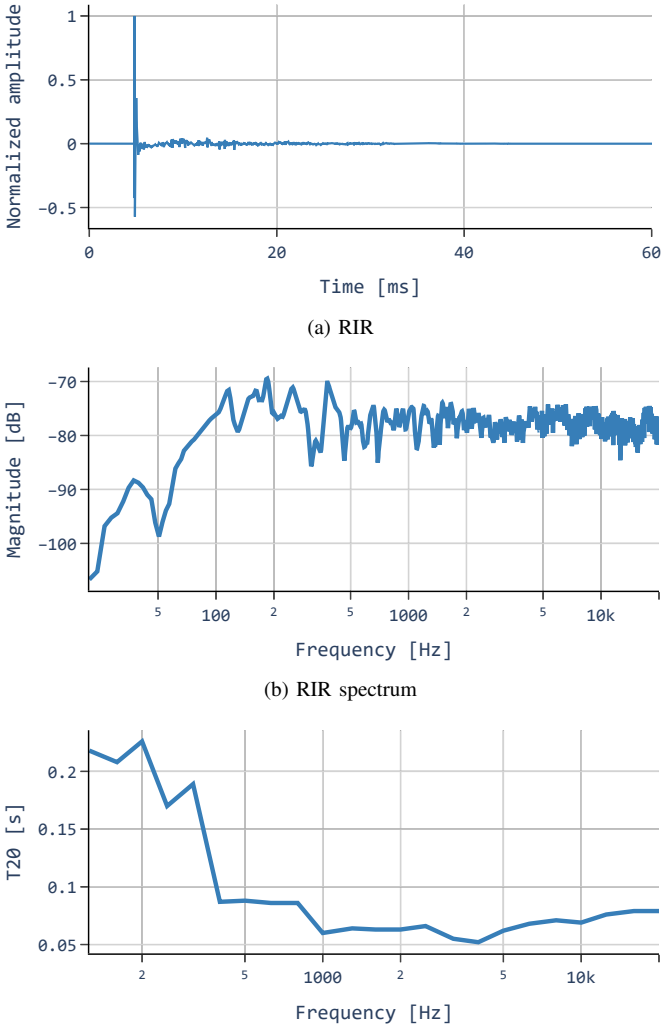
2) Recording Setup: Two Beyerdynamic MM1 monophonic omnidirectional microphones[1] were utilized to capture the participant's utterances at a distance of 50 cm and at the same height as the participant's mouth. The first microphone was positioned in front of the participant. The second microphone was positioned on the participant's right at an angle of $90°$ relative to the first microphone. This configuration was conceived to sample the directivity pattern of the participant's voice at two positions, thereby enabling the simulation of a listener positioned on the right of the speaker. The microphones were connected to a laptop via the Motu UltraLite-MK3 Hybrid audio interface, providing a gain of $+28$ dB. The digital audio workstation (DAW) Reaper was used to record the audio signals coming from the Motu interface at a sample rate of 192 kHz and a bit depth of 24 bits. For each participant, a single track was recorded for the entire session. The resulting audio files are in stereo format, with the first channel (left) storing the sound captured by the frontal microphone, while the right channel (right) storing the lateral microphone's output.

3) RIR Measurement: The same measurement setup was employed to record room impulse responses (RIRs) of the room. The RIRs were measured according to the logarithmic sine sweep technique [32]. The sweeps were reproduced in the room by a Focal Alpha 65 loudspeaker and captured by the same microphones that were employed to record the speeches. The microphones were in the same positions as during the speech recording, while the loudspeaker was placed at the position occupied by the head of the participant. Fig. 2 show the RIR and the one-third octave band T_{20} for the RIR recorded by the frontal microphone. The broadband T_{20} is 77 ms, while the broadband EDT is 2 ms. The values of these reverberation features were computed according to the ISO 3382 standard [33] using the Aurora plugin [34][2]. Furthermore, for each microphone, three further RIRs were measured, each with the loudspeaker facing the microphone but with slight deviations from the ideal position. These recordings are intended to capture variations in the speaker's head position while talking.

[1]See product manual for sensitivity and other specifications.

[2]https://aurora-plugins.com/

Fig. 1: The acoustically-treated chamber where the speeches were recorded along with the measurement setup. The participant was seated on the red chair, while the experimenter was seated on the blue one.



(a) RIR

(b) RIR spectrum

(c) Reverberation time T_{20}

Fig. 2: RIR recorded by the frontal microphone in the acoustically treated chamber. (a) Time-domain response of the RIR. (b) Magnitude spectrum in dB of the RIR between 20 Hz and 20 kHz. (c) One-third octave band reverberation time T_{20} of the RIR. The frequency axis has a logarithmic scale.

B. Recording Session

1) Experimental Procedure: During the recording session, four element types of turn-taking conversations were collected from the participants. These types included question, answer, reciprocal question, and backchannel response. Fig. 3 shows the experimental procedure adopted to collect these elements. The first step in the procedure is the question-and-answer session that covers ten distinct topics. The experimenter posed a question related to the first topic, and the participant replied with the first *answer*. The answers were considered correct if they exceeded 15 seconds; otherwise, the participant was requested to repeat the answer. The targeted range for the duration of the answer was considered to be between 30 and 50 seconds. The responses falling between 15 and 30 seconds, as well as those that exceeded 50 seconds, were deemed acceptable. However, the experimenter encouraged the participant to provide a longer or shorter response, respectively, to the subsequent question. The participant was requested to pause for 5 seconds following the end of their answer. Thereafter, the participant was instructed to formulate a *reciprocal question*, defined as a generic and succinct query to redirect the question back, thereby maintaining the flow of the conversation. Examples of reciprocal questions include "And you?", "How about you?", and "What about you?". The participant was then asked to wait for an additional 5 seconds before repeating the original *question* posed by the experimenter. The participant was encouraged to reformulate the original question by employing alternative words that would convey the same meaning. This procedure was repeated for each of the ten topics.

In the subsequent step of the recording session, the participant was instructed to record a series of approximately

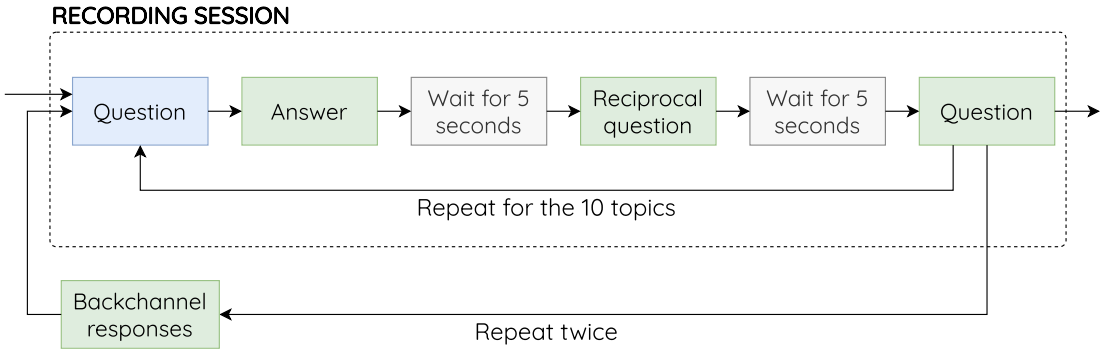


Fig. 3: Experimental procedure adopted to collect the conversational elements for the MoTT speech dataset. The blue box is related to the experimenter, while the green boxes are related to the participant.

ten *backchannel responses*. These were defined as non-verbal sounds employed as interjections during a conversation to express a spontaneous feeling or reaction. Examples of backchannel responses include "oh", "m-hm", "uh", "um" as well as coughing. The participant was instructed to utter nearly neutral backchannel responses and refrain from expressing a clear intention in their voice. The objective of this directive was to elicit backchannel responses that could be adapted to a broader range of conversational scenarios. Subsequently, the question-and-answer session was repeated to record a second take. The order of the questions posed to the participants differed from the first session. This was intended to dissuade participants from providing the same answers as in the first session.

Throughout the recording session, the experimenter ensured that the procedure was followed correctly, through the provision of proper instructions to the participant. These instructions included: indicating the 5 seconds that the participant was required to wait, ensuring that the participant repeated the questions with the original meaning, and checking the given answers for their duration and content. In the event of any issue, the participant was requested to repeat the recording.

2) Questions: The speech data collection encompassed ten open-ended questions, each addressing a distinct topic. Table II presents the ten selected topics, along with the corresponding ten questions posed to the participants. The questions were conceived to be broad in nature and to resemble topics that can plausibly arise in a real conversation, such as hobbies, music, travels, and so on. Additionally, the questions were designed to be easy to understand and to reply with fictional responses. Further, the questions are not interrelated, and thus, they can be assembled in any order to simulate a conversation.

3) Instructions for Participants: The experimental protocol was designed to isolate individual conversational elements, which can be coherently reproduced next to each other. During the recording session, the experimenter avoided any overlap with the participant's utterances. Prior to the beginning of the recording session, the course of the experiment was explained to each participant. Participants were instructed to respond to the questions with fictional, yet plausible answers. Further, participants were informed of the question topics, but not the precise questions, at the time of their recruitment. This approach was adopted to allow participants to familiarize themselves with the subject matter and be adequately prepared when the actual questions were posed.

TABLE II: Topics and related questions which were asked to the participants during the speech data collection of MoTT.

Topic	Question
Music preferences	What kind of music do you listen to?
Cooking	Do you have a favorite dish? What do you like about it?
Hobbies	What do you do in your free time? What are your hobbies?
Travel	What is the most beautiful place you have visited?
Job	What is your job or field of study? What do you like about it?
Preferred school subject	What was your favourite school subject at primary school? Can you tell me something about it?
Languages	How many languages do you speak and which foreign language would you like to learn?
Entertainment	Can you tell me something about the last book, movie, or TV series you watched?
Animals	Can you tell me something about an animal that you like or don't like?
Superpowers	What superpowers would you like to have, and why?

4) Participants: A total of eight participants were recruited to collect the speech recordings for the dataset. The participants were divided equally between males and females. While none of the participants were native English speakers, all of them were fluent in English. The majority of the participants were native Italian speakers, but native French and Brazilian speakers were also recruited. This provides a variety of accents in the English speeches included in the dataset. The English language was selected as it is the most widely spoken international language.

C. Dataset Creation

1) Postprocessing: As previously stated, a single stereo audio track was recorded for the entire recording session of

a participant. Left channel refers to the frontal microphone, while the right channel is that of the lateral microphone. This track underwent manual editing before its inclusion in the MoTT dataset. First, extraneous components were removed, including the experimenter's instructions, silences, redundant speech data resulting from repetitions, and private information given by the participant to prevent doxing. Subsequently, each conversational element recorded during the session was isolated and assigned to the corresponding type, including question, answer, reciprocal question, and backchannel responses. These elements were edited to include a minimal amount of silence at the beginning and end. A high-pass filter with a cutoff frequency of 80 Hz was applied to the audio signals to remove low-frequency noise.

The MoTT dataset also includes the transcript of questions, answers, and reciprocal questions. The transcript was automatically obtained using Google Speech Recognition through the Python library SpeechRecognition[3]. Subsequently, the transcriptions underwent manual review to rectify errors and supplement punctuation.

2) Dataset Organization: The collected speech data are provided in the MoTT dataset as audio files in WAVE format. The naming convention employed for these file is: `IdSex_TypeNum_Take.wav`. `Id` is a two-character identifier that denotes the participant with a number from *00* to *07*. `Sex` denotes the sex of the participant with a single character that can be either *M* for males or *F* for females. `Type` denotes the element type with a single character that can be *Q* for questions, *A* for answers, *R* for reciprocal questions, and *B* for backchannel responses. `Num` is a two-character counter identifying different audio files belonging to the same participant and to the same type. If `Type` is *Q* or *A*, then `Num` corresponds to the identifier of the topic among those reported in Table II. `Take` is an optional suffix that is used only when `Type` is *Q* or *A* to distinguish between the two takes of collected questions and answers. As shown in Fig. 3, the reciprocal questions were measured in two takes as well. However, the two takes of the reciprocal questions were not distinguished in the nomenclature since they are not related to the topics.

IV. Use Case

The MoTT dataset described in this paper can be utilized for a variety of purposes. As an example, this section details a use case in which we are currently employing MoTT for research purposes. Specifically, we are conducting an alternative version of an experiment previously presented in the literature [6]. The experiment is conducted within the framework of AAV, in which the VAE is derived from auditory content captured from the real world, which is augmented by synthetic sound generation. In the considered use case, SRIRs represent real-world capture, while speech data are used for sound augmentation. Within the AAV framework, the experiment evaluates the *co-immersion* criterion, which is operationalized by asking listeners to identify a virtual sound source rendered with a different acoustic approach among reference sources. In the original experiment, Fantini et al. [6] investigated the scenario of virtual speakers talking simultaneously, which were rendered at various locations within the VAE using dynamic spatial audio techniques. The authors evaluated the co-immersion of two simplified late reverberation approaches to render the speakers. One involved static late reverberation derived from binaural downmixing of Ambisonics SRIRs, while the other utilized an artificial reverberator with parameters automatically tuned by a previously proposed reverb matching method [35]. The authors reported the use of concurrent speech as a limitation of the experiment, which made it challenging for listeners to concentrate on individual speeches and their acoustic properties.

In the experiment's alternative version that we are conducting, our aim is to assess the co-immersion of conversations rather than concurrent speech. Conversations represent a more ecologically-valid scenario, which is commonly encountered in everyday life and VAEs. For this purpose, we utilized the MoTT dataset to create turn-taking conversations in a modular way. The experiment included trials with two to four speakers, a parameter regarded as scene complexity. Therefore, we chose four speakers from MoTT, consisting of two males and two females, to represent the experiment's virtual characters. The dataset's diverse voice timbres and accents enable the construction of dialogues between heterogeneous and well-characterized virtual speakers, which facilitates their identification by participants during the experimental task. For each level of complexity, we generated three distinct conversations involving two, three, or four characters based on the complexity level. Each conversation begins with one character posing a question about one of the ten topics, followed by a response from another character on that topic. The rest of the conversation is further developed similarly, with questions and answers from various speakers seamlessly joined together. The answers from MoTT were adequately shortened to avoid long utterances and to increase the perceived conversation's interactivity. We also included reciprocal questions that either followed an answer, came before a question, or occurred in both positions. Additionally, backchannel responses from other characters were incorporated throughout a character's utterances to enhance the interactivity and realism of the conversation. Each conversation was designed to last between one and a half to two minutes. In creating these conversations, we aimed to ensure a balanced distribution of speaking time among characters and diverse topics. As a result, the experiment's participants attend a turn-taking conversation where multiple characters engage in turns by asking and answering to questions spanning different topics. The realism of this conversation is enhanced by the use of dynamic spatial audio techniques, which provide an acoustical rendering of characters at various locations within the VAE. Compared to the speech data from the original experiment, the MoTT dataset is associated with a greater degree of perceived auditory externalization, according to preliminary findings.

[3]https://pypi.org/project/SpeechRecognition/

The conversations generated for the experiment are included in the MoTT dataset [28].

V. Conclusion

This paper presented MoTT, an English speech dataset for modular turn-taking composition. The dataset is composed of questions, answers, reciprocal questions, and backchannel responses recorded by eight participants. The questions and answers are related to ten topics selected to cover different topics that can occur in a conversation. The speech data included in the dataset are therefore intended to be interchangeable elements allowing their modular composition. A natural-sounding conversation can be simulated by adequately assembling these elements as if multiple speakers are talking to each other. The MoTT dataset can find several applications, such as the design of ecological experiments in which the participant listens to fictional, but realistic, conversations. This is exemplified by the use case of the experiment described in Section IV, where spatial audio techniques enhance the perceived realism of the conversation.

Future works include the recruitment of further participants to obtain a larger dataset with more diverse voices and accents. A limitation of the described dataset is the employed recording room which is not fully anechoic. This could affect the simulation of an artificial reverberation effect on the recorded speeches given that they already encode the acoustic cues of the recording room.

References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] J. M. Wiemann and M. L. Knapp, "Turn-taking in conversations," *Journal of Communication*, vol. 25, no. 2, pp. 75–92, 1975.

[3] D. R. Begault, *3-D sound for virtual reality and multimedia*. Academic Press, 1994.

[4] M. Geronazzo and S. Serafin, Eds., *Sonic Interactions in Virtual Environments*, 1st ed., ser. Human–Computer Interaction Series. Cham: Springer, 2023.

[5] S. A. Wirler, N. Meyer-Kahlen, and S. J. Schlecht, "Towards transfer-plausibility for evaluating mixed reality audio in complex scenes," in *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020, pp. 1–10.

[6] D. Fantini, G. Presti, M. Geronazzo, R. Bona, A. G. Privitera, and F. Avanzini, "Co-immersion in audio augmented virtuality: the case study of a static and approximated late reverberation algorithm," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 11, pp. 4472–4482, 2023.

[7] R. Kilgore, M. Chignell, and P. Smith, "Spatialized audioconferencing: what are the benefits?" in *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research*, ser. CASCON '03. IBM Press, 2003, p. 135–144.

[8] J. Ahrens, M. Geier, A. Raake, and C. Schlegel, "Listening and conversational quality of spatial audio conferencing," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*. Audio Engineering Society, 2010, pp. 1–13.

[9] J. Skowronek and A. Raake, "Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing," in *Interspeech 2011*, 2011, pp. 829–832.

[10] ——, "Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls," *Speech Communication*, vol. 66, no. C, p. 154–175, 2015.

[11] J. Fels, C. A. Ermert, J. Ehret, C. Mohanathasan, A. Bönsch, T. W. Kuhlen, and S. J. Schlittmeier, "Listening to, and remembering conversations between two talkers: Cognitive research using embodied conversational agents in audiovisual virtual environments," in *Fortschritte der Akustik - DAGA 2021, Wien, Austria]*. Deutsche Gesellschaft für Akustik e.V. (DEGA), 2021, pp. 1328–1331.

[12] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: the impact of meeting type on speech style," in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 301–304.

[13] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The ICSI meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[14] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," in *Proc. International Conference on Methods and Techniques in Behavioral Research*. Noldus Information Technology, 2005, pp. 1–4.

[15] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia *et al.*, "The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language resources and evaluation*, vol. 41, pp. 389–407, 2007.

[16] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vendeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (CCDb): A database of natural dyadic conversations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 277–282.

[17] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.

[18] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The NoXi database: multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 350–359.

[19] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. Interspeech 2018*, Hyderabad, India, 2018, pp. 1561–1565.

[20] C.-H. Lee, S.-M. Wang, H.-C. Chang, and H.-Y. Lee, "ODSQA: Open-domain spoken question answering dataset," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 949–956.

[21] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu *et al.*, "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Proceedings of Interspeech 2021*, 2021, pp. 3665–3669.

[22] A. Adigwe and E. Klabbers, "Strategies for developing a conversational speech dataset for text-to-speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022. International Speech Communications Association, 2022, pp. 2318–2322.

[23] K. Lee, K. Park, and D. Kim, "Dailytalk: Spoken dialogue dataset for conversational text-to-speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] H. Lee and D. Johnson, "An open-access database of 3D microphone array recordings," in *Audio Engineering Society Convention 147*. Audio Engineering Society, october 2019, pp. 1–6.

[25] G. Götz, S. J. Schlecht, and V. Pulkki, "A dataset of higher-order ambisonic room impulse responses and 3D models measured in a room with varying furniture," in *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. IEEE, 2021, pp. 1–8.

[26] F. Miotello, P. Ostan, M. Pezzoli, L. Comanducci, A. Bernardini, F. Antonacci, and A. Sarti, "HOMULA-RIR: A room impulse response dataset for teleconferencing and spatial audio applications acquired through higher-order microphones and uniform linear microphone arrays," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 795–799.

[27] F. Denti, D. Fantini, F. Avanzini, and G. Presti, "PAN-AR: A multimodal dataset of higher-order ambisonics room impulse responses, ambient noise and spherical pictures," in *Proceedings of the 19th International*

Audio Mostly Conference (AM '24). New York, NY, USA: Association for Computing Machinery, 2024.

[28] G. Salada, D. Fantini, F. Avanzini, and G. Presti, "MoTT: A speech dataset for modular composition of turn-taking conversations," Aug. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.14929530

[29] V. Stanford, J. Garofolo, O. Galibert, M. Michel, and C. Laprun, "The nist smart space and meeting room projects: signals, acquisition annotation, and metrics," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. IEEE, 2003, pp. IV–736.

[30] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*. Philadelphia, PA, U.S.A.: Association for Computational Linguistics, 2014, pp. 263–272.

[31] S. Si, W. Ma, H. Gao, Y. Wu, T.-E. Lin, Y. Dai, H. Li, R. Yan, F. Huang, and Y. Li, "SpokenWOZ: A large-scale speech-text benchmark for spoken task-oriented dialogue agents," *Advances in Neural Information Processing Systems*, vol. 36, pp. 39 088–39 118, 2023.

[32] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES 108th convention*. Audio Engineering Society, 2000, pp. 1–24. [Online]. Available: https://aes2.org/publications/elibrary-page/?id=10211

[33] ISO, "International Standard ISO/DIS 3382-1: Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces," International Organization for Standardization, Standard ISO 3382-1:2009, 2009. [Online]. Available: https://www.iso.org/standard/40979.html

[34] A. Farina, "Auralization software for the evaluation of a pyramid tracing code: results of subjective listening tests," in *ICA95 (International Conference on Acoustics), Trondheim (Norway)*. Acoustical Society of Norway, 1995, pp. 26–30.

[35] R. Bona, D. Fantini, G. Presti, M. Tiraboschi, J. I. Engel Alonso-Martinez, and F. Avanzini, "Automatic parameters tuning of late reverberation algorithms for audio augmented reality," in *Proceedings of the 17th International Audio Mostly Conference*. New York, NY, USA: Association for Computing Machinery, 2022, pp. 36–43.