提示引导的轮流预测

Koji Inoue, Mikey Elmers, Yahui Fu, Zi Haur Pang, Divesh Lala, Keiko Ochi, Tatsuya Kawahara

Graduate School of Informatics, Kyoto University, Japan,

Correspondence: inoue@sap.ist.i.kyoto-u.ac.jp

摘要

轮流对话预测模型是口语对话系统和会话机器人中的重要组成部分。近期的方法利用基于变压器的架构来连续实时地预测语音活动。在本研究中，我们提出了一种新型模型，该模型能够通过文本提示动态控制轮流对话的预测。这种方法允许通过指令如“更快”或“更平静”，进行直观且明确的控制，并根据对话伙伴和语境动态调整。所提出的模型基于基于变压器的语音活动投影（VAP）模型，将文本提示嵌入同时整合到通道内变压器和跨通道变压器中。我们使用超过 950 小时的人际口语对话数据评估了该方法的可行性。由于现有数据集中没有可供本研究使用的文本提示数据，我们利用了一个大型语言模型（LLM）来生成合成的提示句子。实验结果表明，所提出的模型提高了预测准确性，并能根据文本提示有效地调整轮流对话的时间行为。

1 介绍

轮流发言，管理说话者之间的转换，是顺畅自然的人类沟通的基础 (Skantze, 2021)。准确预测轮流发言在开发口语对话系统和会话机器人中尤为重要，因为它通过最小化不适当的打断和减少响应延迟直接影响交互质量 (Ter Maat et al., 2011; Khouzaimi et al., 2015; Tisserand et al., 2024; Addlesee and Papaioannou, 2025; Skantze and Irfan, 2025; Inoue et al., 2025b)。最近的进步，尤其是基于变压器的模型如 TurnGPT (Ekstedt and Skantze, 2020) 和语音活动投影（VAP）(Ekstedt and Skantze, 2022)，显著提高了连续实时语音活动预测的有效利用对话历史的能力。

在本文中，我们介绍了一种基于变换器的新模型，该模型能够根据文本提示动态调整其轮流预测。我们的方法将提示嵌入集成到 VAP 中的通道内和跨通道变换器架构中，通过简单的指令如“更快”或“更平静的”，实现明确且直观的控制。先前的研究表明，在对话中轮流行为的变化取决于个体属性，例如年龄、性别以及外向性和内向性等性格特征 (Anderson and Leaper, 1998; Levinson and Torreira, 2015; Su et al., 2016; Liesenfeld et al., 2020; Lourenço et al., 2023)。因此，我们提出的模型的适应性有助于根据对话环境、用户类型和系统配置进行此类响应式交互。据我们所知，这是首次明确使用文本提示指导轮流预测的工作。

2 提出的方法

我们提出的模型基于 VAP 架构，该架构预测二元对话的未来语音活动 (Ekstedt and Skantze, 2022)。我们的方法的关键创新在于集成了文本提示，允许对模型的轮流行为进行动态控制，如图 1 所示。

本文已被接受在 2025 年 SIGdial 话语与对话会议（SIGDIAL 2025）上发表，并代表了作者的工作版本。

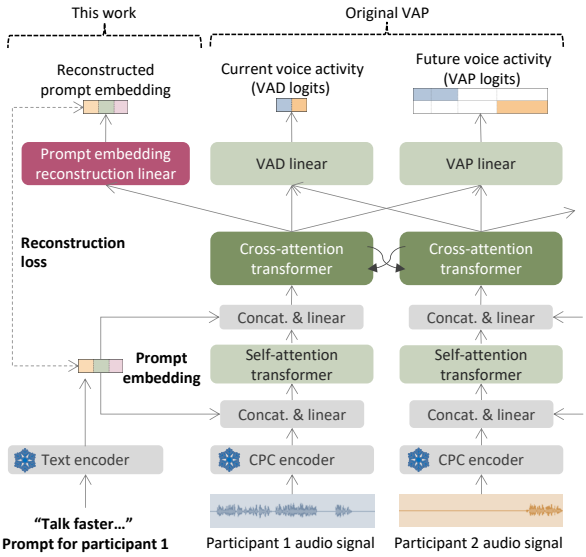

图 1: 所提出模型的架构。提示处理（左侧所示）也适用于参与者 2。

2.1 基础 VAP 模型架构

我们的模型基础是 VAP 架构，该架构将立体音频波形作为输入，其中每个通道对应一个参与者，类似于全双工系统。每个音频通道由使用对比预测编码（CPC）(Riviere et al., 2020) 的预训练音频编码器进行处理。在训练 VAP 时，编码器参数保持不变。每个通道的编码特征分别由自注意力变换器独立处理，以建模说话者的时序依赖关系。来自各通道变换器的输出然后被送入交叉注意力变换器。该组件通过允许每个通道的表现形式关注另一个来模拟两个参与者之间的交互。从交叉注意力变换器得到的最终表示通过两个独立的线性层进行预测：

- **蒸汽对数几率单位** 表示一个在 256 个离散状态上的概率分布。每个状态对应于四个未来时间间隔（0-200 毫秒，200-600 毫秒，600-1200 毫秒，1200-2000 毫秒）中每个说话人的预测语音活动。这构成了主要的 VAP 目标，被表述为一个多类分类问题。

- **语音活动检测（VAD）对数几率分数** 是当前时间帧内每个说话人当前语音活动的概率。这作为辅助任务。

对于轮流预测，预测的 VAP 状态概率（由 VAP 对数几率得出）被汇总以表示近期和长期未来的语音活动：

- p_{now}：表示每个参与者近期未来（0-600 毫秒，结合前两个区间）预测到的语音活动的综合概率。

- p_{future}：表示每个参与者长期未来（600-2000 毫秒，结合最后两个区间）预测的语音活动的聚合概率。

2.2 提示整合

为了便于基于提示的控制，我们通过独立地为每个参与者集成文本提示嵌入来扩展 VAP 架构。这些嵌入编码了提供的轮流行为，例如“更快的响应”或“更平静”。在本研究中，我们使用了 Sarashina-Embedding-v1-1B[1] 生成 1792 维的归一化句子向量。由于这些嵌入的维度根据嵌入模型的不同而变化（我们的案例是 1792 维），我们应用线性变换来标准化它们的维度，使其符合变压器输入的要求（例如 256 维）。

受提示引导的文本到语音框架 (Guo et al., 2023) 启发，我们在架构中的两个战略点将提示嵌入与 VAP 特征表示在每个时间帧中进行连接：

- **后音频编码器：** 提示嵌入最初与 CPC 音频编码器的输出连接。随后，一个线性投影调整这个连接向量以匹配通道式变压器的输入维度。

- **自注意力后处理：** 通道注意力变换器之后，提示嵌入再次被拼接。另一个线性层将该拼接表示的维度降低到与交叉注意力变换器输入维度对齐。

然后，生成的条件提示向量按照原始 VAP 架构通过交叉注意力转换器进行处理。

VAP 和 VAD 预测的最终输出来自于交叉注意力变换器生成的表示。此外，该模型还包

[1] https://huggingface.co/sbintuitions/sarashina-embedding-v1-1b

括一个辅助线性头，负责从交叉注意力层的特定参与者输出中重构原始提示嵌入。这种重构构成了额外的训练目标，确保提示信息在整个网络中的稳健传播。

所提出的模型的训练损失包括三个部分：

$$L = L_{\text{vap}} + L_{\text{vad}} + L_{\text{prompt}}\,, \tag{1}$$

其中 L_{vap} 和 L_{vad} 对应于原始 VAP 框架中的损失，而 L_{prompt} 表示输入与重构提示嵌入之间的均方误差（MSE）。通过将提示嵌入与初始音频特征和中间自注意力表示进行拼接，我们全面地对模型的预测处理进行了条件设置——从单通道分析到跨通道交互——基于提供的行为提示。

3 数据集

本节描述了本研究中使用的数据集以及生成所提出方法文本提示的方法。

3.1 口语对话语料库

本研究利用了以下三种类型的日语口语对话数据集。这些数据集的总时长合计约为 953.5 小时。整个数据集在会话级别上被随机分为训练集、验证集和测试集，比例为 8:1:1。通过纳入包含各种对话任务的数据而不仅仅是简单的闲聊，我们的目标是训练一个可以适应各种情况的提示可控模型。

在线对话数据集 该数据集是为此次研究新收集的。它包含由 109 名参与者使用在线会议工具进行的自由形式闲聊对话。每位参与者参与了多次时长约为 20 到 30 分钟的会话，且限定每对参与者仅交互一次。总体而言，该数据集包括 2,166 个对话，总计 803 小时。Silero VAD[2] 被用于标注话语段。

旅行社任务对话 (Inaba et al., 2022) 这包括模拟旅行社内客户与员工之间互动的任务导向对话。这些对话也使用在线会议工具进行了录制。完整数据集包含 329 个对话，总计 115.5 小时。

[2] https://github.com/snakers4/silero-vad

人机对话 这包括人类与机器人 ERICA 之间的对话，这些对话是通过巫师控制的偶人（WOZ）方法收集的 (Inoue et al., 2025a)。它涵盖了各种任务，包括 ERICA 作为倾听者的专注聆听对话、工作面试以及初次见面的对话。这些任务的总时长约为 35 小时。

3.2 提示生成

尽管文本提示对于控制我们提出的方法中的轮流说话风格至关重要，但此类信息通常在现有数据集中缺失。因此，我们使用大型语言模型（LLM）从对话数据中合成了提示句子。具体而言，我们为 GPT-4.1[3] 提供了语音活动详情，包括说话人身份和每个 20 秒音频段的发言时间（开始和结束时间），这些信息与 VAP 模型训练输入格式相匹配。

基于这些信息，我们指示了 LLM 首先使用链式思考方法（附录 A）生成每个参与者轮流发言风格的描述性印象。随后，LLM 产生了旨在复制该特定风格的相应提示（指令）。下面提供了一个实际生成输出的例子：

> **参与者 *A* 的印象**: 较少发言但一旦开始往往会持续较长一段时间，给人一种平静的印象，在说话前有明显的停顿。
>
> **参与者 *A* 的提示**: 请平静地说话，并有意识地停顿，一次传递大量内容。在您的陈述前后包括短暂的沉默。

使用这种方法生成的提示句用于训练和评估所提出的模型。

4 实验

我们首先通过将所提出的方法与不利用提示输入的原始 VAP 模型进行比较，对轮流预测性能进行了定量评估。评估指标包括 VAP 测试损失（L_{vap}）以及在相互沉默期间预测轮换或保持的平衡准确性 (Inoue et al., 2024)。模型

[3] https://platform.openai.com/docs/models/gpt-4.1

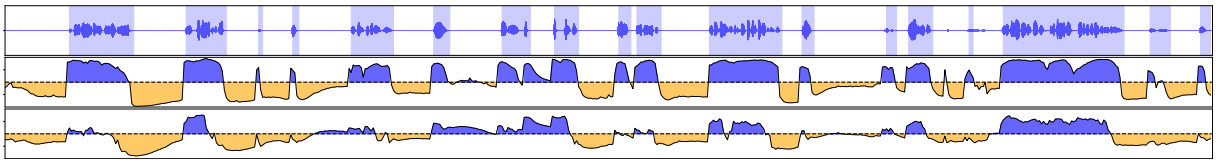

图 2: 所提方法的示例输出。顶部行是用户音频输入的波形表示。中间和底部行分别指示 p_{now} 和 p_{future} 的概率，蓝色表示用户，黄色表示系统。

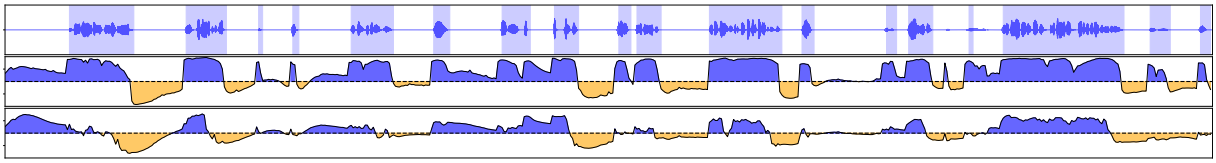

图 3: 示例输出，其中输入提示与图 2 中呈现的相反。

表 1: 换手预测性能。预测值/小时是换手或保持预测的平衡准确性（%）。

方法	VAP loss (↓)	S/H pred. (↑)
Original VAP	2.431	77.17
Proposed	2.346	79.80

配置和训练参数与原作中描述的一致 (Ekstedt and Skantze, 2022)。

结果汇总在表 1 中。所提出的方法通过结合提示信息，对 VAP 损失和平衡准确率都取得了改进。这些发现表明，生成的提示句子有效提高了换位预测的准确性，说明架构成功集成了额外的提示数据。

接下来，我们定性评估了输入提示文本如何影响所提模型的输出。为了模拟用户系统对话并评估系统何时应该发言，我们将测试音频的一个声道用作用户的音频输入，并将另一个声道设置为零以表示系统的静默。我们准备了两个文本提示：一个代表对话中用户的指令或假设意图，另一个指定系统的行文设置。

图 2 说明了一个示例输出，突出了系统应该进行其回合的时间（黄色区域）。在这个例子中，用户提示是“在说话之前，请短暂停顿并仔细思考，然后礼貌地开始说话。不要急于回答；保持冷静和稳定的语速。”相反，系统提示指令为，“以良好的节奏说话，并在对方说完后立即回应。尽量更频繁地发言并主导对话。”系统的提示鼓励立即回应和频繁的轮流发言，正如预测的系统回合频率增加和时间更快（黄色部分）所反映的那样。

图 3 展示了另一种场景，其中用户与系统之间的提示被互换，使用相同的音频输入。现在系统被指示缓慢且有目的地作出回应，观察到的轮换次数少于之前的例子，这与给定的行为提示一致。

5 结论

我们引入了一种新型的提示引导式轮流预测模型，该模型根据集成到 VAP 模型中的文本提示动态调整其行为。我们的实验使用了约 950 小时的日语对话数据，结果表明该模型能够根据不同的提示有效调节轮流行为，提高了预测准确性和适应性。

未来的工作包括将这种方法整合到实用的对话系统和机器人中，以验证其效用，特别是通过自动推断用户提示或配置系统提示。这包括根据用户的特征（如年龄：儿童、成人或老年人）或个性来调整交互。此外，验证生成的提示的适当性并扩展提示的变化，以及将工作扩展到包含诸如人类对齐和多语言能力 (Inoue et al., 2024) 等方面，仍然是至关重要的下一步。

致谢

此项工作得到了 JST PRESTO JPMJPR24I4、JST Moonshot R&D JPMJPS2011 以

及 JSPS KAKENHI JP23K16901 的支持。

限制

我们研究的一个显著限制是依赖于合成生成的文本提示，因为现有的对话数据集中缺乏自然提示。这种人工场景可能会限制模型在现实世界情境中有效性的泛化能力。此外，实际实现需要将人类（用户或开发人员）的反馈纳入提示，以有效地反映现实世界的使用场景并增加模型的可控性。另外，我们目前的评估主要基于日语对话数据集，这可能限制了我们的研究发现应用于其他语言和文化背景的能力。

伦理考虑

我们的研究涉及会话行为的自动化调节，这引发了关于用户自主权和同意的伦理考量。此类系统的实现必须确保透明度，让用户能够清楚地了解并控制他们的会话风格和个人数据是如何被解释和使用的。在自动推断用户的特征（如年龄或语言背景）时应仔细考虑以避免偏见或刻板印象，确保与不同用户群体之间的互动是公平和尊重的。

References

Angus Addlesee and Ioannis Papaioannou. 2025. Building for speech: designing the next-generation of social robots for audio interaction. *Frontiers in Robotics and AI*, 11:1–5.

Kristin J. Anderson and Campbell Leaper. 1998. Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how. *Sex Roles*, 39(3-4):225–252.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *International Conference on Empirical Methods in Natural Language Processing (EMNLP) Findings*, pages 2981–2990.

Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised learning of turn-taking events. In *INTERSPEECH*, pages 5190–5194.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. PromptTTS: Controllable text-to-speech with text descriptions. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Michimasa Inaba, Yuya Chiba, Ryuichiro Higashinaka, Kazunori Komatani, Yusuke Miyao, and Takayuki Nagai. 2022. Collection and analysis of travel agency task dialogues with age-diverse speakers. In *International Conference on Language Resources and Evaluation (LREC)*, pages 5759–5767.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Multilingual turn-taking prediction using voice activity projection. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 11873–11883.

Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2025a. Yeah, Un, Oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In *North American Chapter of Association for Computational Linguistics (NAACL)*, pages 7171–7181.

Koji Inoue, Yuki Okafuji, Jun Baba, Yoshiki Ohira, Katsuya Hyodo, and Tatsuya Kawahara. 2025b. A noise-robust turn-taking system for real-world dialogue robots: A field experiment. *arXiv preprint, arXiv:2503.06241*.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefevre. 2015. Optimising turn-taking strategies with reinforcement learning. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 315–324.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

Andreas Liesenfeld, Gabriella Parti, Yu-Yin Hsu, and Chu-Ren Huang. 2020. Predicting gender and age categories in english conversations using lexical, non-lexical, and turn-taking features. In *Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 22–30.

Vânia Lourenço, Joana Serra, Joana Coutinho, and Alfredo F. Pereira. 2023. Turn-taking in free-play interactions: A cross-sectional study from 3 to 5 years. *Cognition*, 239:105568.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Gabriel Skantze and Bahar Irfan. 2025. Applying general turn-taking models to conversational human-robot interaction. In *International Conference on Human-Robot Interaction (HRI)*, pages 859–868.

Ming-Hsiang Su, Chung-Hsien Wu, and Yu-Ting Zheng. 2016. Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):733–744.

Mark Ter Maat, Khiet P Truong, and Dirk Heylen. 2011. How agents' turn-taking strategies influence impressions and response behaviors. *Presence: Teleoperators and Virtual Environments*, 20(5):412–430.

Lucien Tisserand, Brooke Stephenson, Heike Baldauf-Quilliatre, Mathieu Lefort, and Frédéric Armetta. 2024. Unraveling the thread: understanding and addressing sequential failures in human-robot interaction. *Frontiers in Robotics and AI*, 11:1–19.

A 使用的提示

生成第 3.2 节中描述的轮流行为提示所使用的提示如下：

Prompt Generation

以下是记录了对话过程中 A 和 B 两人每次发言的开始时间和结束时间，单位为秒。根据说话量、讲话时长、开始说话所需的时间、沉默时长以及语速等因素，将评估 A 和 B 两人的印象。然后，基于这些印象，设计提示以使 AI 模仿每个人的语速风格。

输出应严格按照以下格式，包含四行：
对人物 A 的印象：<描述人物 A 印象的句子>
对人物 B 的印象：<描述人物 B 印象的句子>
提示给人物 A：<针对人物 A 的提示语句>
给人物 B 的提示：<用于人物 B 的提示句>

请注意，这是原文为日语的提示的翻译版本。提示后面跟着对话上下文数据，其中包括双方发言的时间信息。