

中华人民共和国国家标准

GB/T 44089—2024

信息技术 全双工语音交互系统 通用技术要求

Information technology—General technical requirements of
full duplex speech interaction system

2024-05-28 发布

2024-05-28 实施

国家市场监督管理总局
国家标准化管理委员会 发布

目 次

前言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 缩略语 2

5 系统参考功能框架及交互过程 2

 5.1 系统参考功能框架 2

 5.2 系统交互过程 4

6 功能要求 5

 6.1 核心要求 5

 6.2 声学处理层 6

 6.3 语音识别层 6

 6.4 对话处理层 6

 6.5 语音合成层 6

7 性能要求 6

 7.1 语音识别层 6

 7.2 对话处理层 7

 7.3 语音合成层 7

 7.4 交互响应时间 7

附录 A（资料性） FDX 语音交互过程案例 8

 A.1 车载终端场景 8

 A.2 智能客服场景 8

 A.3 智慧办公场景 9

 A.4 智能家居场景 10

参考文献 11

前 言

本文件按照 GB/T 1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：中国电子技术标准化研究院、科大讯飞股份有限公司、美的集团(上海)有限公司、深圳市腾讯计算机系统有限公司、北京百度网讯科技有限公司、中国电信集团有限公司、小米通讯技术有限公司、中移(杭州)信息技术有限公司、青岛海尔科技有限公司、福州数据技术研究院有限公司、深圳云天励飞技术股份有限公司、北京电信规划设计院有限公司、思必驰科技股份有限公司、杭州方得智能科技有限公司、羚羊工业互联网股份有限公司、合肥智能语音创新发展有限公司、深圳市矽赫科技有限公司、上海智能制造功能平台有限公司、北京捷通华声科技股份有限公司、马上消费金融股份有限公司。

本文件主要起草人：董建、徐洋、贾一君、刘颖、宋文林、何永春、于磊、苏丹、袁杰、鄂磊、蔡亚森、梅林海、赵培、刘聪、杨震、雷宗、龚晟、樊帅、洪鹏达、黄超、李林璐、方斌、陈明、胡国平、杨一帆、刘志强、毕盛楠、丁强、高羽、李旭。

信息技术 全双工语音交互系统 通用技术要求

1 范围

本文件规定了全双工语音交互系统的参考功能框架、交互过程,以及功能要求、性能要求。
本文件适用于全双工语音交互系统的设计、开发、应用、测试和维护。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

全双工 full duplex

能够同时双向传递数据的通信方法。

3.2

功能单元 functional unit

能够完成特定目标的硬件或软件实体。

3.3

语音识别 speech recognition

将人类的声音信号转化为文字或者指令的过程。

[来源:GB/T 21023—2007,3.1]

3.4

语义理解 semantic understanding

使功能单元理解人说话的意图。

[来源:GB/T 36464.1—2020,3.11]

3.5

语音合成 speech synthesis

通过机械的、电子的方法合成人类语言的过程。

[来源:GB/T 21024—2007,3.1]

3.6

话术 telephony

交互过程中使用的具有一定逻辑的对话文本内容。

3.7

对话管理 dialogue management

跟进当前的对话状态和上下文输入,对对话的状态进行更新,同时依据对话处理逻辑生成需要实施的对话动作。

4 缩略语

下列缩略语适用于本文件。
AI:人工智能(Artificial Intelligence)
FDX:全双工(Full Duplex)
MOS:平均意见得分(Mean Opinion Score)
VAD:声音活动检测(Voice Activity Detection)

5 系统参考功能框架及交互过程

5.1 系统参考功能框架

5.1.1 概述

图 1 所示的 FDX 语音交互系统的参考功能框架包括交互层、知识和数据资源层、AI 和机器学习层和基础层。

- a) 交互层包括声学处理层、语音识别层、对话处理层、语音合成层。交互层的主要功能是将输入信号通过声学处理层以及语音识别层识别为纯文本,通过对话处理层理解输入信号的真实意图,并生成交互回复语,最后通过语音合成层将交互回复语合成为语音音频作为输出信号。
- b) 知识和数据资源层主要为交互层提供必备的数据资源和知识库。
- c) AI 和机器学习层主要为交互层提供模型推理、在线数据挖掘、数据分析等能力;基础层包括云服务、终端和边缘计算,提供硬件计算资源,是 AI 和机器学习算法的运行载体,同时负责保障 FDX 语音交互过程中每个模块的能力调用、系统稳定。

层是指完成一大类功能能力的单元集合体。这些层可以根据其输入、输出及其意图或功能来描述。每层及其组件都可以单独使用和测试。所有层可以集成在一起,使用户能够与功能单元进行对话,帮助用户满足自己的需求。

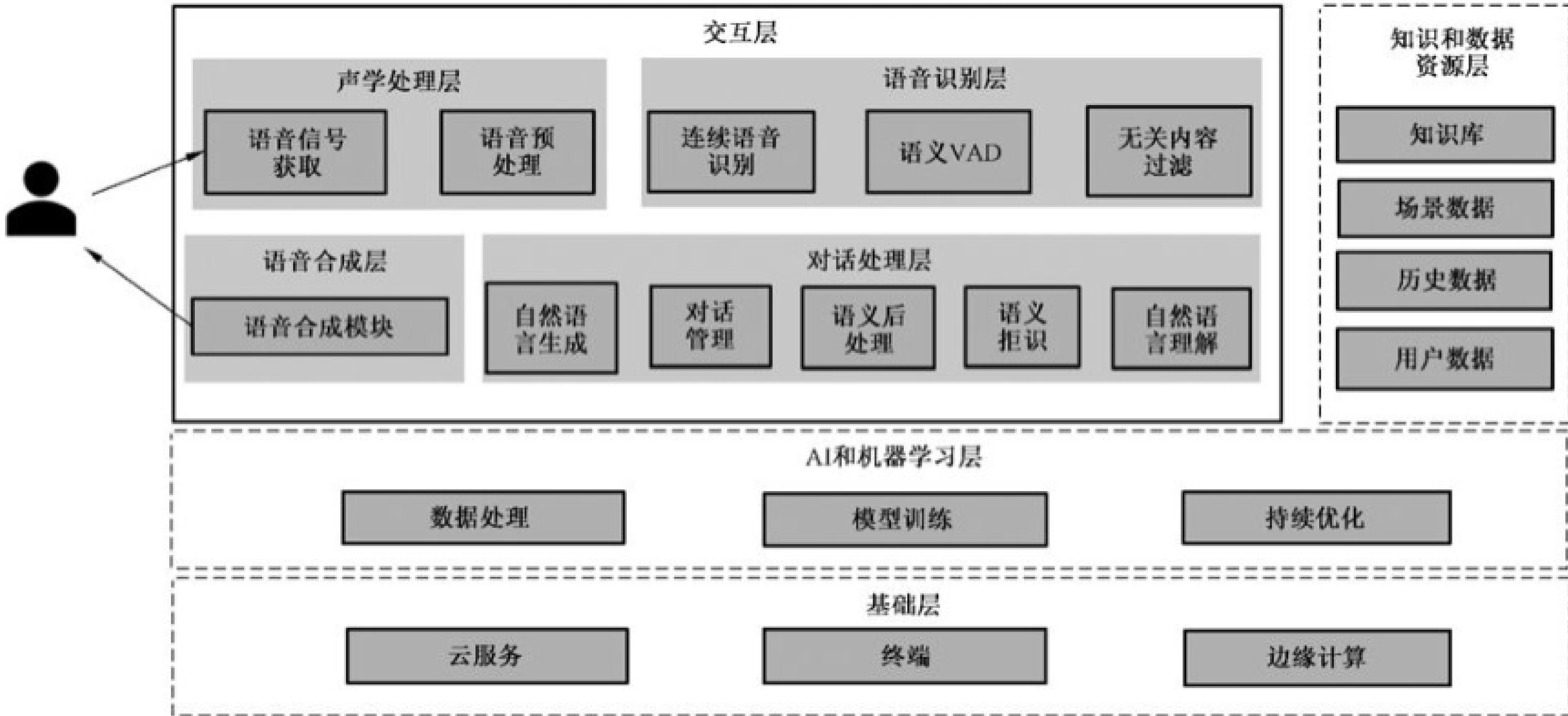


图 1 FDX 语音交互系统的参考功能框架

5.1.2 交互层

5.1.2.1 声学处理层

声学处理层包括语音信号获取和语音预处理。

- a) 语音信号获取是指使用麦克风或麦克风阵列提供连续音频采集。
- b) 语音预处理是指对采集的语音信号进行以下预处理中的一项或多项：语音增强、声源定位、去混响、去噪、回声消除和语音源信号提取。

5.1.2.2 语音识别层

语音识别层包括连续语音识别、语义 VAD、无关内容过滤。

- a) 连续语音识别是指将用户连续的语音信号转化为文字或者指令的过程。
- b) 语义 VAD 是指通过对语音蕴含的语义进行理解，得到语音活动帧的判别结果，比如用户说“我想听(停顿 1 s)XX 的歌曲”，通过对语音片段“我想听”的语义进行理解(此句话还未表达完整，缺少宾语信息)，以此判别后续仍有语音活动帧。
- c) 无关内容过滤是指通过对语音信号进行分析与决策，过滤无效的语音输入，比如，场景噪声、回声等。

5.1.2.3 对话处理层

对话处理层包括自然语言理解、语义拒识、语义后处理、对话管理和自然语言生成。

- a) 自然语言理解将文本或语音转换为内部描述，该内部描述为输入的结构化语义的表达。
- b) 语义拒识是指系统通过自然语言理解技术，能够区分系统当前状态下不应处理的输入信息，不应处理的输入信息包括与交互任务以及对话主题或上下文无关的内容。
- c) 语义后处理是指系统对输入信号进行自然语言理解之后，对理解的结果进行后续再处理，比如：在对输入语音“明天的天气”进行自然语言理解之后，还需要计算出“明天”对应的具体日期值。
- d) 对话管理是指系统跟进当前的对话状态和上下文输入，对对话的状态进行更新，同时依据对话处理逻辑生成需要实施的对话动作。
- e) 自然语言生成是指系统根据对话管理得到的对话动作，生成合适的自然语言文本。

5.1.2.4 语音合成层

通过语音合成将文本合成语音。

5.1.3 知识和数据资源层

知识和数据资源层包括场景和语境理解所需的相关知识和数据，场景和语境是指不同的场景或语言上下文信息。知识和数据资源层包括知识库、场景数据、历史数据、用户数据。

5.1.4 AI 和机器学习层

AI 和机器学习层使用基于机器学习的 AI 方法进行数据处理、模型训练和持续优化。

5.1.5 基础层

基础层使用云服务和/或终端和/或边缘计算的方式来提供 FDX 语音交互能力，其中语音识别、对话管理、文本合成等组件可使用云服务进行处理。

5.2 系统交互过程

FDX 语音交互系统的交互过程示例如图 2 所示,交互过程用于表示用户与 FDX 语音交互系统之间的语音流传输。FDX 语音交互系统与一般的半双工语音交互系统的交互过程至少存在以下区别。

- a) 一次唤醒多次交互:FDX 语音交互系统只需在对话开始时唤醒一次,用户能连续对话(语音采集设备在预设的时长内没有有效人声输入,则停止采集,进入休眠状态)。如图 2 所示,用户通过输入语音信号“XXX”唤醒机器,然后进行了三次连续对话。一次完整对话过程会依次执行语音识别、语义理解、对话管理、自然语言生成和语音合成等功能单元,整个处理链路还依赖场景和语境,知识和数据,以及各模块的实现计算方法。FDX 语音交互系统通过对输入的语音信号或其他输入信息进行处理,最终输出合成的语音或者其他信息与指令动作。FDX 语音交互系统可持续接收输入的各类信号,包括但不限于语音信号、信息和请求等,将有用的信号转录为文本,从转录文本中提取语义信息,根据语义信息对交互任务进行预测和决策,根据预测和决策向用户提供输出信号,输出信号包括但不限于合成语音、回答、信息和行为等。
- b) 用户语音动态结束判别:在用户与 FDX 语音交互系统交互过程中,针对用户输入停顿,FDX 语音交互系统应能够实现智能等待,从而实现连续对话,其中,语义 VAD 是指通过对声音蕴含的语义进行理解,得到语音活动帧的判别结果。如图 2 所示,比如用户说“我想听(停顿 1 s) XX 的歌曲”,通过对语音片段“我想听”的语义进行理解(此句话还未表达完整,缺少宾语信息),以此判别后续仍有语音活动帧。即 FDX 语音交互系统忽略中间的 1 s 停顿,持续收音,并根据对话上下文进行语义理解。
- c) 上行/下行信道并行处理:FDX 语音交互系统中用户和机器应能够同时相互通信,即上行信道(输入自然语音)和下行信道(输出人工语音)应能够在相同的时间间隔内接收和发送语音信号。用户应能够随时自由打断功能单元的讲话,机器可以在用户说话或保持沉默时管理节奏或给出提示。使得在任一时刻,FDX 语音交互系统可以同时处理输入输出信号,实现双工通信交互。如图 2 所示,比如用户说“合肥今天的天气”,FDX 语音交互系统会依次对“合肥今天的天气”执行语音识别、语义理解、对话管理和自然语言生成,在生成交互回复语“合肥今天……”被打断暂停播放的同时,可持续监听接下来用户输入的语音信号“上海呢”。系统此时可在不影响上一轮交互回复语经过语音合成模块,生成合成后音频的同时,对本轮输入信号“上海呢”进行语音识别、语义理解、对话管理、自然语言生成和语音合成等处理,实现了同时处理上下行通道。

FDX 语音交互系统应能够根据用户的状态和场景,对用户的意图进行一定程度的预测,控制对话的节奏,并主动给出反馈和信息,引导用户下一步的行动。不同应用场景下 FDX 语音交互过程案例见附录 A。

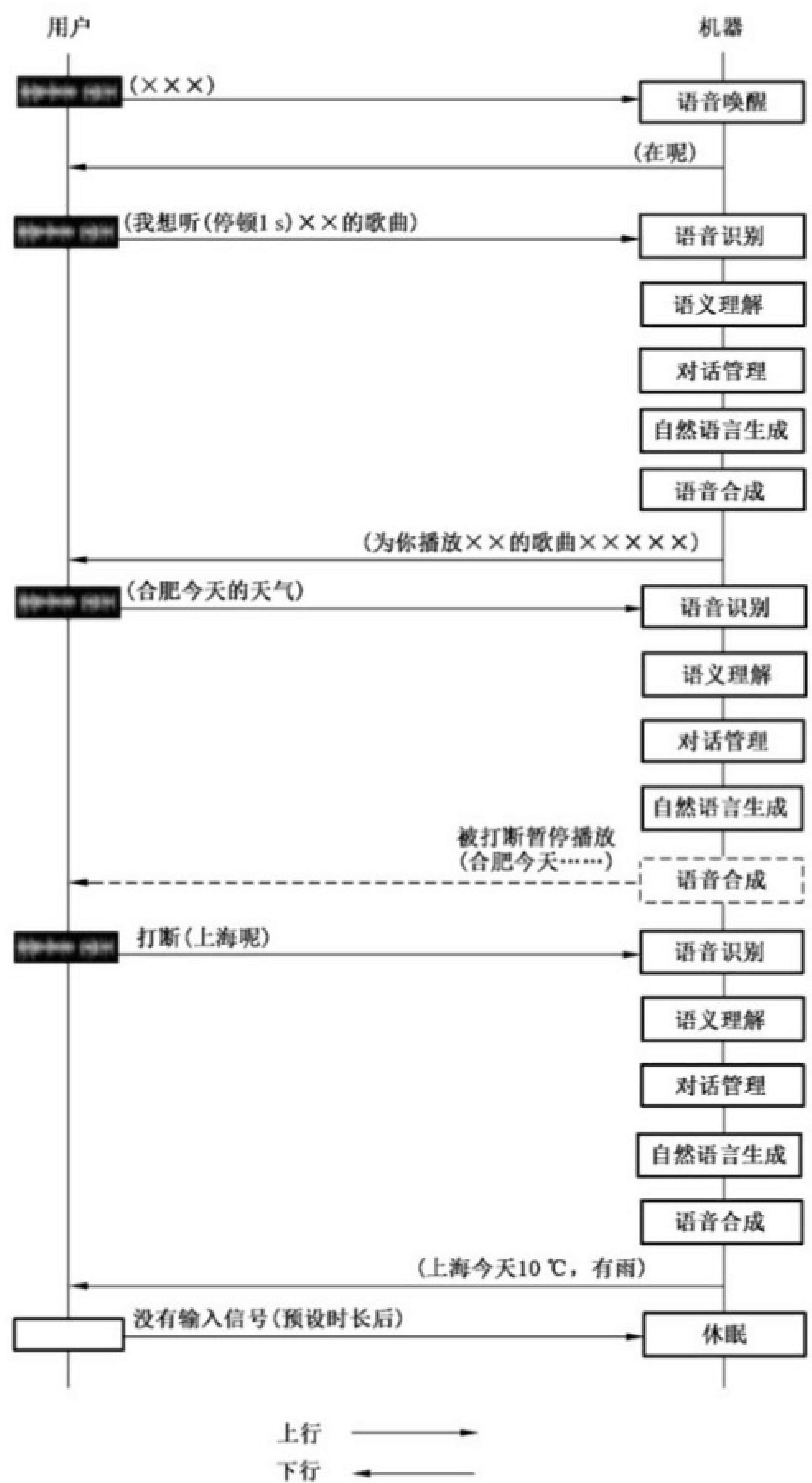


图 2 FDX 语音交互系统的交互过程示例

6 功能要求

6.1 核心要求

- 系统支持以下功能：
- a) 用户进行一次唤醒(即触发用户界面的语音控制操作)可完成整个对话流程,即系统应只需要在通话开始时触发,对话过程中不必触发；
 - b) 在整个交互过程中可根据需要随时打断,即系统应能在广播或讲话过程中的任何时刻被用户中断干预,并能在被用户中断后继续进行对话；
 - c) 应对连续音频流进行 VAD,能实现连续语音识别,并根据对话上下文的语义理解进行用户意

图预测和语音交互；

- d) 应能结合知识库、场景数据、历史数据、用户数据进行对话管理,如:在用户不说话的时候可以结合之前的响应上下文与常识实施合理的主动对话,在用户说话的时候选择静默模式。

6.2 声学处理层

声学处理层支持以下功能:

- a) 应能够使用麦克风或麦克风阵列提供连续音频采集、语音增强、声源定位、去混响、去噪、回声消除和语音源提取功能;
- b) 应能够实现近场音频采集和远场音频采集,其中,近场通常是指话筒与语音源之间的距离在 3 m 以及 3 m 以内,远场是指距离大于 3 m;
- c) 应能够通过计算麦克风阵列与用户之间的平面角、方位角、俯仰角和距离来定位用户,并提高语音信号的信噪比。

6.3 语音识别层

语音识别层支持以下功能:

- a) 应支持多种不同使用场景下语音识别、纠错的能力;
- b) 应支持断句、音频按交互分段;
- c) 应能够从连续语音流中检测多个语音片段的起点和终点;
- d) 应能够设置两个语音片段之间的静音等待时间并调整 VAD 的灵敏度;
- e) 应根据语句和场景的语义拒绝对不当内容的识别;
- f) 应支持中文;
- g) 应支持方言和/或多语种。

6.4 对话处理层

对话处理层支持以下功能:

- a) 应能够理解用户的意图,并根据知识和数据对未来的会话内容做出一定程度的预测;
- b) 应能够提供推理功能,包括空间推理、时间推理、常识推理、计算策略应用或任何形式的推理;
- c) 应能够生成用于形成人工语音的文本,文本的内容可以包括:简单的回复文本、基于预定义模板的回复文本、通过理解并响应用户意图的回复文本,合理指导或建议的回复文本;
- d) 应能够跟踪会话状态、管理会话策略、根据用户意图改变或进行会话主题;

6.5 语音合成层

语音合成层支持以下功能:

- a) 应支持多语种;
- b) 应能够处理连续的语音流;
- c) 应能够模拟目标说话人的语音特征,输出具有目标说话人听觉感知特征的语音;
- d) 应能够调整输出语音的声韵、速度、音调、语调,其中,声韵是指声母韵母发音。

7 性能要求

7.1 语音识别层

语音识别层的性能指标包括句识别正确率和字识别正确率:

- a) 在低噪声环境(信噪比 10 dB 以上)下,语音识别句识别正确率应大于或等于 84%,语音识别

字识别正确率应大于或等于 95%；

- b) 在高噪声环境(信噪比 10 dB 及以下)下,语音识别句识别正确率应大于或等于 75%,语音识别字识别正确率应大于或等于 88%。

注：字识别正确率和句识别正确率的测试集构建、测试方法和指标计算方法参考 GB/T 36464.1—2020。

7.2 对话处理层

对话处理层应满足以下要求。

- a) 在有限域的交互场景或者其他特定场景中,应支持基于语义的播报打断：
 - 1) 在检测到用户输入部分有效信息但仍需要其他信息时,回复反馈语；
 - 2) 在进行目标场景的交互时,系统对目标场景用户语句意图理解的精确率大于或等于 90%,召回率大于或等于 90%；系统对目标场景用户语句中的关键信息提取的精确率大于或等于 90%,召回率大于或等于 90%,关键信息是指输入语句中满足系统正确响应用户请求的所有必要槽值信息。
- b) 为了保障用户端到端交互体验,考虑噪声环境对理解结果的影响,应满足：
 - 1) 在低噪声环境(信噪比 10 dB 以上)下,非人机交互响应率小于或等于 6%,非人机交互响应率是指非人机交互场景下系统给出话术响应的数量占机器收音成功的所有非人机交互话术数量的比例；
 - 2) 在高噪声环境(信噪比 10 dB 及以下)下,非人机交互响应率小于或等于 10%。

注：精确率和召回率的测试集构建、测试方法和指标计算方法参考 GB/T 41813.2—2022。

7.3 语音合成层

MOS 应大于或等于 4.2,其中,MOS 的量化规则见表 1。

表 1 MOS 的量化规则

MOS	规则
5 分(非常自然)	和播音员真人发声非常接近,达到可以以假乱真的程度。总体听感清晰、流畅,评测者乐于接受
4.5 分(较自然)	发音清晰、可懂。总体听感流畅,评测者愿意接受,没有明显韵律错误
4 分(自然)	勉强接受,没有明显的分词错误,在语气节奏处理上没有大问题
3 分(一般)	基本能接受(打分的一个分界线分),但语气节奏处理上问题较多,音节之间不流畅感较重。测听人不太愿意接受,有明显疲劳感
2 分(不自然)	一些关键词听不清楚,评测人员不愿意接受
1 分(非常不自然)	发音不清晰,听不懂,机器音质。只能表达断续、个别的语音信息,无法猜测句意,不能接受

7.4 交互响应时间

全双工语音交互响应时间应不超过 1.5 s,响应时间为从用户语音输入结束到系统合成语音响应的
时间。

附录 A
(资料性)
FDX 语音交互过程案例

A.1 车载终端场景

车载终端交互场景下用户和机器交互过程中使用 FDX 语音交互的案例如图 A.1 所示。

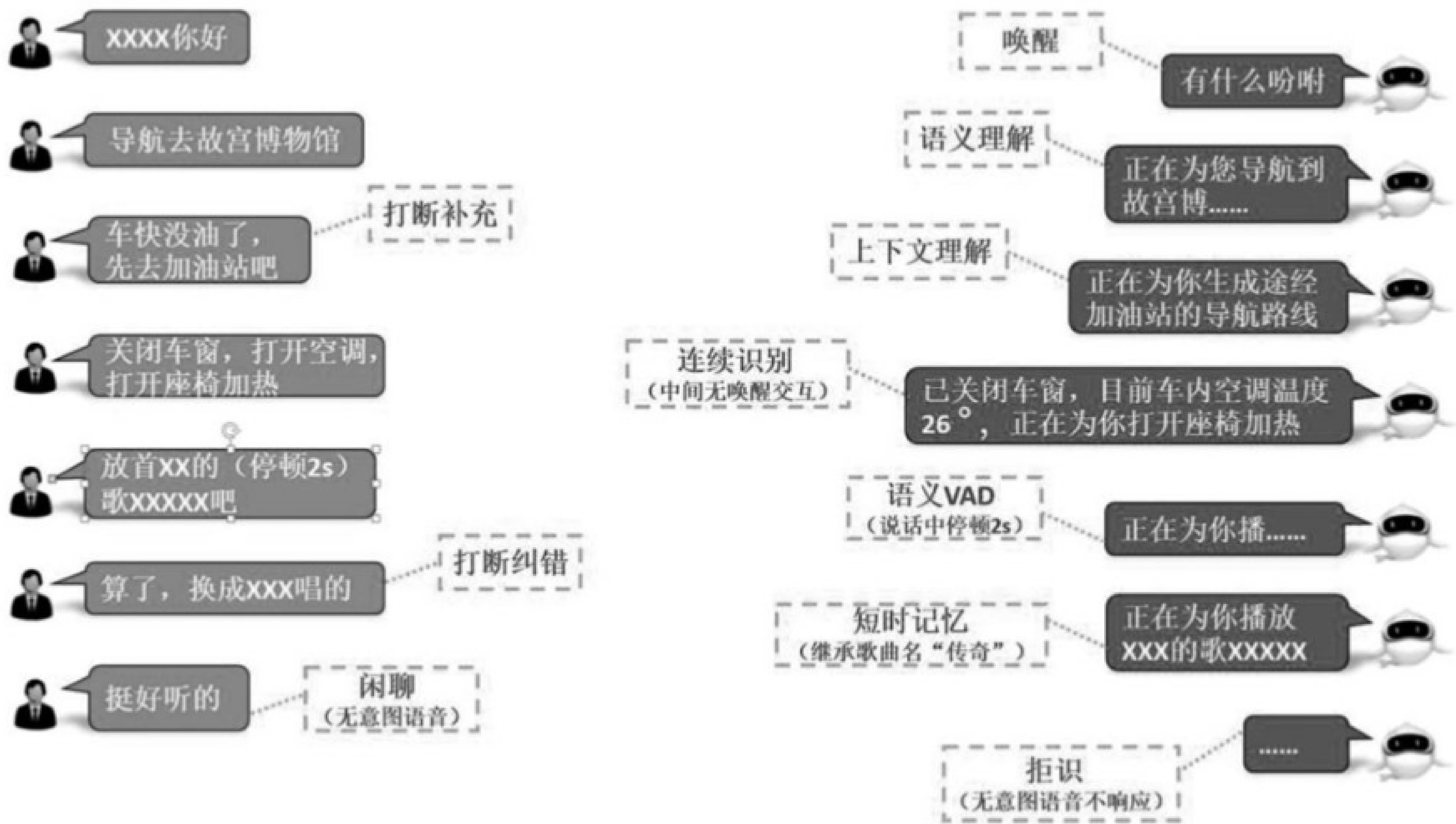


图 A.1 车载终端 FDX 语音交互过程

A.2 智能客服场景

智能客服交互场景下用户和机器交互过程中使用 FDX 语音交互的案例如图 A.2 所示。

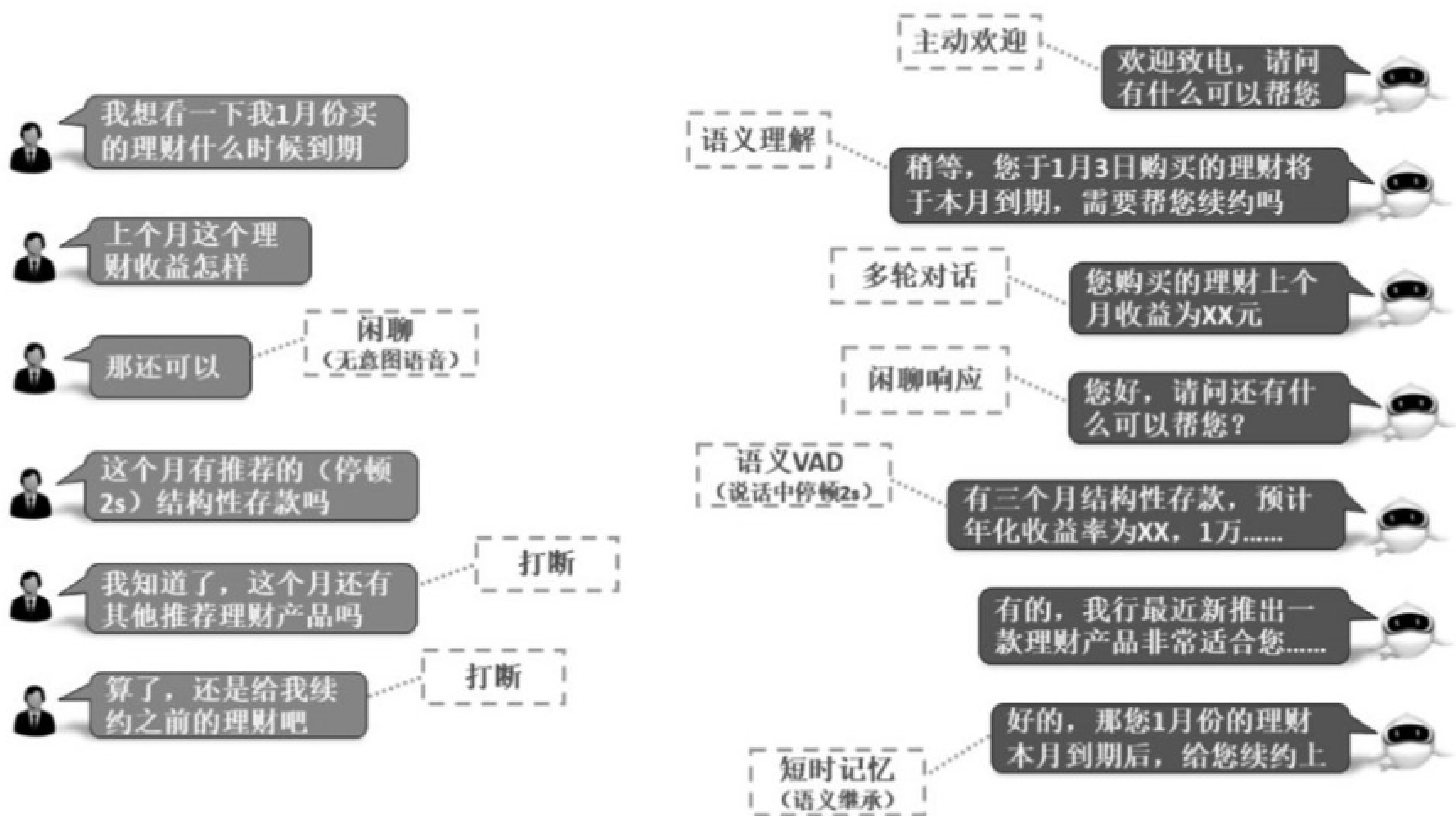


图 A.2 智能客服 FDX 语音交互过程

A.3 智慧办公场景

智慧办公交互场景下用户和机器交互过程中使用 FDX 语音交互的案例如图 A.3 所示。



图 A.3 智慧办公 FDX 语音交互过程

A.4 智能家居场景

智能家居交互场景下用户和机器交互过程中使用 FDX 语音交互的案例如图 A.4 所示。

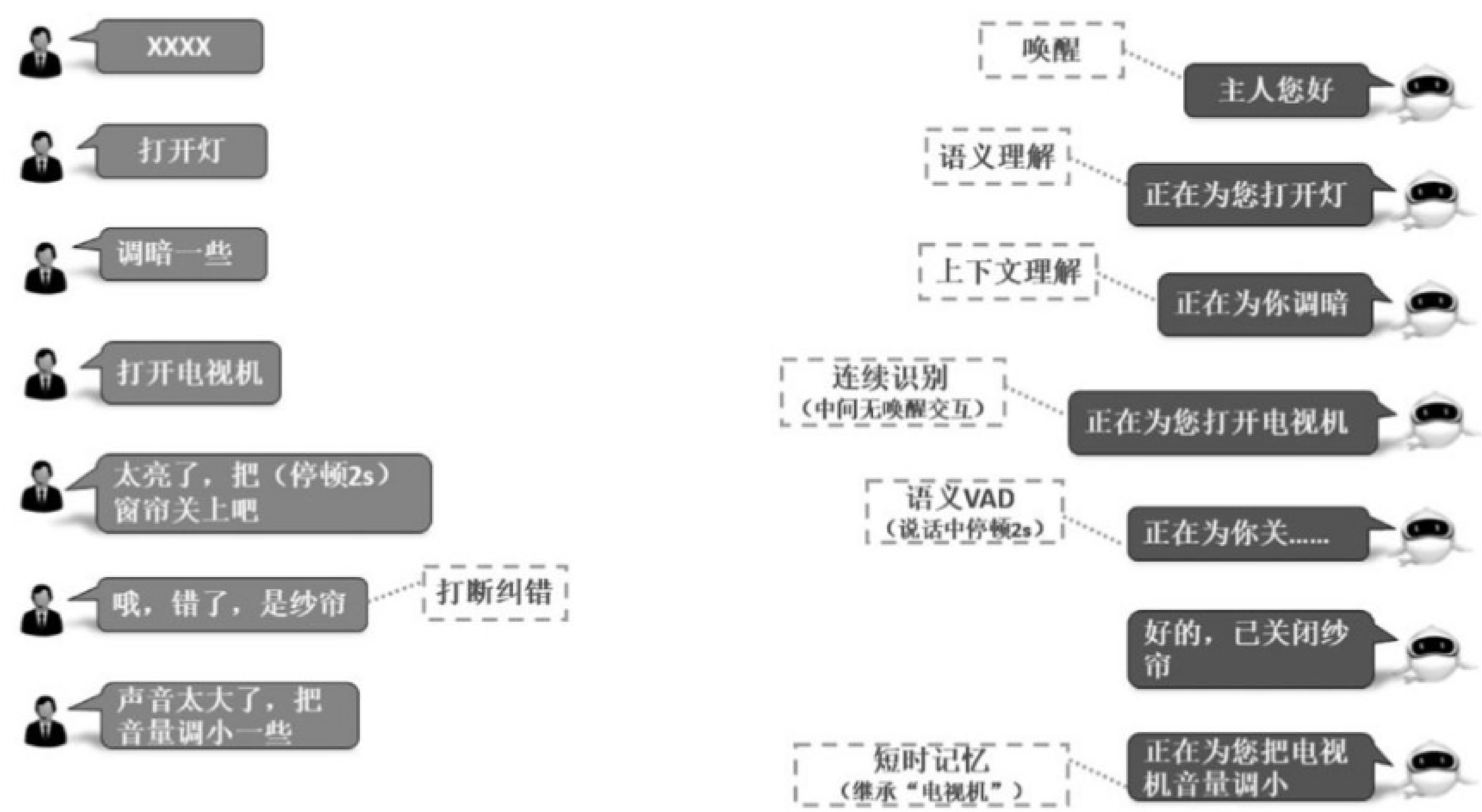


图 A.4 智能家居 FDX 语音交互过程

参 考 文 献

- [1] GB/T 21023—2007 中文语音识别系统通用技术规范
 - [2] GB/T 21024—2007 中文语音合成系统通用技术规范
 - [3] GB/T 36464.1—2020 信息技术 智能语音交互系统 第1部分:通用规范
 - [4] GB/T 41813.2—2022 信息技术 智能语音交互测试方法 第2部分:语义理解
-

中 华 人 民 共 和 国
国 家 标 准
信息技术 全双工语音交互系统
通用技术要求

GB/T 44089—2024

*

中国标准出版社出版发行
北京市朝阳区和平里西街甲2号(100029)
北京市西城区三里河北街16号(100045)

网址:www.spc.net.cn

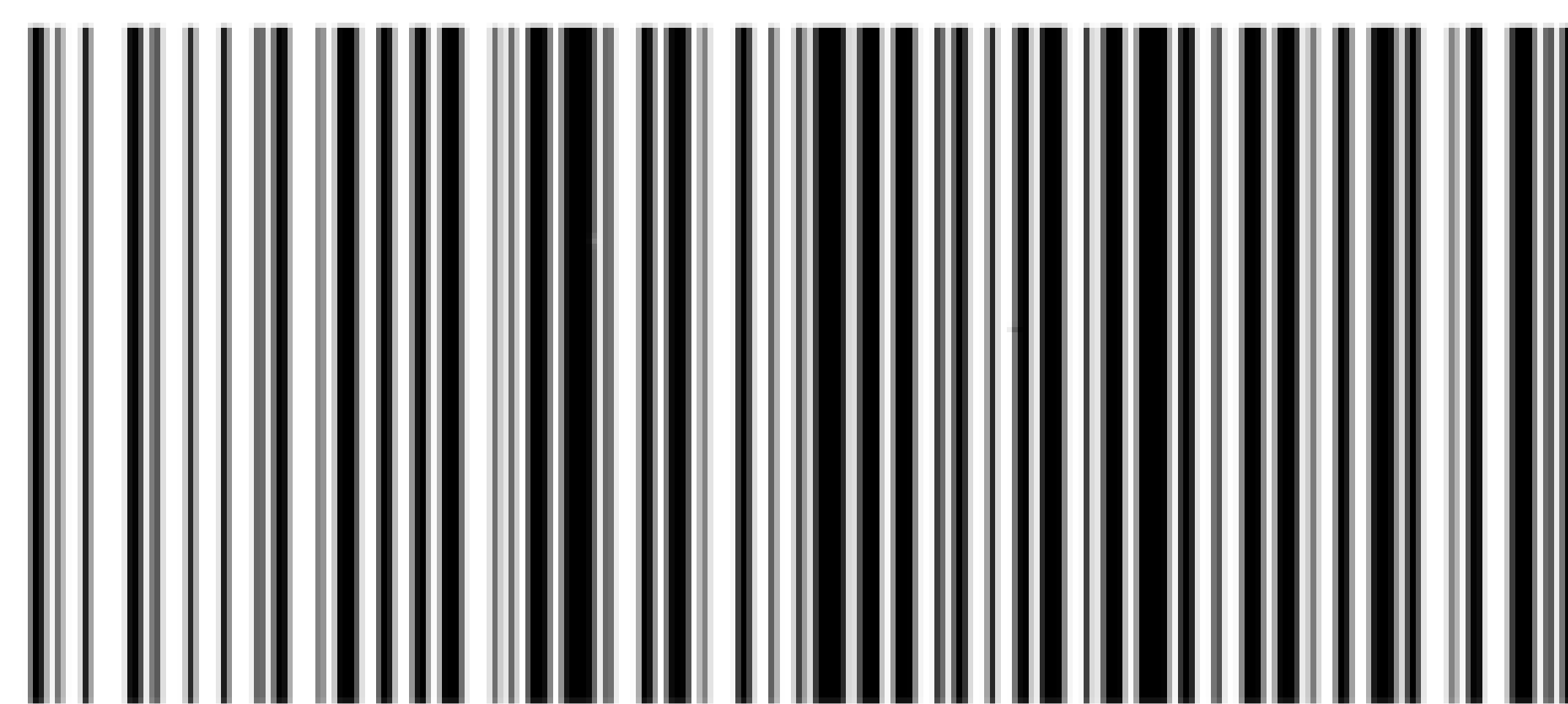
服务热线:400-168-0010

2024年5月第一版

*

书号:155066·1-76374

版权专有 侵权必究



GB/T 44089-2024