基于多模态信息融合的语音意图理解方法

郑彬彬 贾 珈 蔡莲红

(清华大学计算机科学与技术系，北京 100084)

摘 要：为从语音中获取包括字面含义和说话人情绪状态在内的全面意图信息，提出了一种基于多模态信息融合的语音意图理解方法，并对其中的关键词抽取、命令解析、基于文本/韵律特征的情绪状态检测以及多模态信息融合等关键算法进行了设计。该方法从识别文本和语音信号中抽取不同模态的信息并进行融合，能够有效地从语音中获取丰富的意图信息，有助于建立自然的人机交互环境。

关键词：语音意图理解；多模态信息抽取；多模态信息融合

中图分类号：TP 309　　文献标识码：A

A Speech Intention Understanding Method Based on Multi-modal Information Integration

Zheng Binbin,　Jia Jia,　Cai Lianhong

(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: In order to obtain comprehensive speech intention information containing both the literal meaning and speaker's affective state, a speech understanding method based on multimodal information integration is proposed. Key algorithms including keywords extraction, command analyzing, text/prosody-based affective state determination and multimodal information integration are designed. The method is able to effectively obtain rich intention information by extracting information of different modality from recognition text and speech signal and merging them together, which is helpful to establish a natural human-computer interaction environment.

Key words: speech intention understanding; multimodal information extraction; multimodal information integration

0 引言

随着人机交互技术的迅速发展，如何使计算机具有理解话语的能力逐渐成为研究热点。意图理解系统旨在对说话人的意图进行准确地分析和理解。目前针对语音意图理解的研究主要集中于话语理解（spoken language understanding）[1]，通过分析特定领域的语音文本来获取其中的语义信息，大致可分为基于规则/文法的理解方法[2]、基于统计的理解方法[3]、基于例句的理解方法[4]。虽然上述方法能够有效地理解话语的字面意思，但普遍存在以下2个问题。

1）话语理解的研究对象是对语音进行人工转写得到的文本，而在实际应用中只能使用语音识别引擎来获取文本信息。由于自然语言中存在多种复杂的口语现象，在现有语音识别技术条件的限制下得到的识别文本不可避免地存在大量错误，这将导致理解性能大幅下降。

2）只考虑了话语的字面意思，即语言学信息。然而，语音可以传达丰富的信息，除语言学信息外，说话人的态度、情绪或者说话风格等副语言学信息也对意图的理解起着重要的作用。忽视这部分信息可能导致对说话人意图的理解产生很大偏差。

针对以上问题，本文提出了一种基于多模态信息融合的语音理解方法，该方法利用关键词检测等技术对识别文本进行分析以减少识别错误带来的影响；除关键词信息外，也从语音信号中抽取声学特征，获得多模态信息并进行融合，以获取说话人状态，最终对说话人的意图进行准确而全面地理解。

基金项目：国家自然科学基金（编号：(61003094，90920302，60931160443)）；
作者简介：郑彬彬（1985－）　女，硕士。主要研究方向：语音意图理解。
通信联系人：贾珈，助研，主要研究方向为人机语音交互，　E-mail：jjia@tsinghua.edu.cn

1 意图结构与理解框架

1.1 意图结构设计

本文考虑语音意图理解在智能家居场景下的应用。在智能家居控制场景中，说话人的意图主要是对家居设备进行命令控制。为有效表示用户意图，设计了表 1 所示的意图结构，包括命令内容、用户状态以及命令状态 3 部分。

表 1　智能家居控制场景意图结构

Table 1　Intention Structure of Home Automation

意图结构项	命令内容	用户状态	命令状态
意图结构子项	设备名称；设备属性； 设备位置；操作类别	高兴；愤怒； 悲伤；惊奇	有效性 优先级

命令内容指用户语音中包含的具体指令。智能家居控制场景中常用的设备控制命令具有统一的模式，命令内容包括设备名称、设备属性、设备位置及操作类别 4 个子项。例如命令“把卧室空调的温度升高”中，设备名称、属性、位置和操作类别分别为“空调”、“风力”、“卧室”和“升高”。

用户状态反应说话人讲话时的情绪状态，包括高兴、悲伤、愤怒和惊奇 4 种基本情感类型，意图理解结果给出用户情绪状态属于每种情感类型的置信度得分，其取值范围为 0～1。

命令状态包括命令的有效性和优先级，有效性指明该命令是否符合当前场景的设定。有效性取值为 1 表明命令有效，取值为 0 表明命令无效。优先级分为 3 个等级(0-low, 1-normal, 2-high)，反映了用户对该条命令响应时间的要求，优先级越高的命令用户要求的响应时间越短。

1.2 基于多模态信息融合的语音意图理解系统框架

为充分理解用户语音中包含的意图信息，设计了如图 1 所示的意图理解框架。

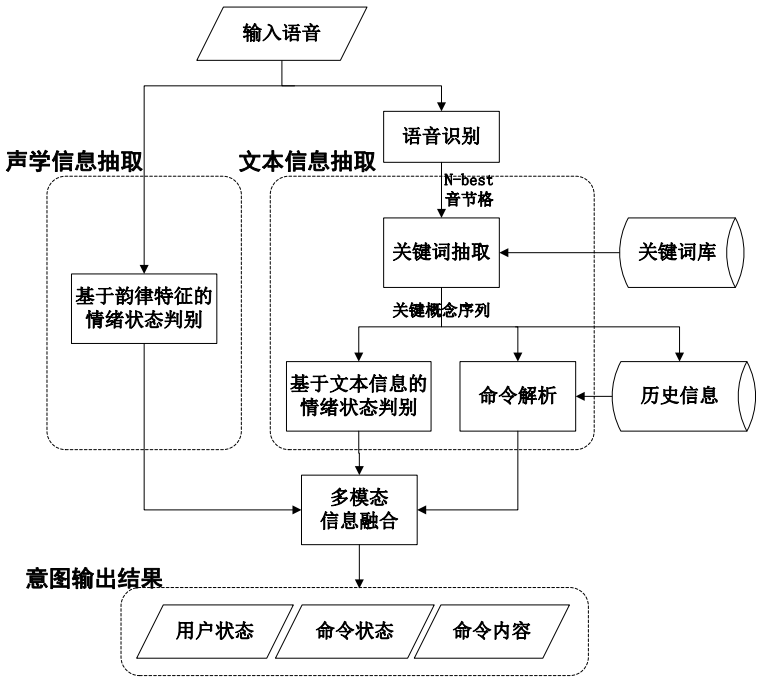

图 1　语音意图理解系统框架

Fig. 1 Framework of Speech Intention Understanding System

语音意图理解的多模态信息融合方法系统框架中的输入为用户语句，意图输出结果的形式是如前所述的意图结构。这种方法一方面从语音识别文本中抽取命令内容相关的关键信息和说话人状态判断的辅助信息；另一方面对用户语音进行声学特征分析，得到说话人状态的判别结果。最终将这两部分信息进行融合，获取到最终的用户意图。其中文本信息的获取包括关键词抽取、命令解析和基于文本

信息的情绪状态判别等主要模块；声学信息由基于韵律特征的情绪状态判别得到。将这两部分信息经过多模态信息融合模块的融合处理，得到最终的用户意图。

2 基于 N-Best 音节格的命令关键词抽取

近年来，语音识别技术的研究取得了很大进展，但自然口语的语音识别仍然存在很多问题。这是由于自然口语中存在大量的重复、遗漏和倒序等现象[5]，使得识别准确率难以提升。另一方面，要对句子进行准确的理解，并不需要考虑句子中的每个词，只需对几个蕴含关键概念的关键词汇进行理解就能把握句子的意义。关键词识别技术从语句中抽取用户关心的关键信息，能够降低对识别系统和环境噪声的要求。

2.1 关键概念及关键词集设计

由于说话人可以用不同词汇来表达同一概念，所以可选择关键概念而非关键词本身作为后续理解算法的输入。根据 1.1 节中对命令模式的分析和意图结构的设计，针对智能家居控制场景定义五类关键概念(KC，key concept)，包括设备类型（cc_device）、设备属性（cc_attri）、设备位置（cc_pos）、操作类别（cc_oper）和用户状态（user_state）。其中“user_state”描述说话人的情绪状态。

每个概念项可能对应多个关键词，根据关键概念种类，定义了 6 类关键词：“Devices”，“Attributes”，“Positions”，“Operations”，“UserStates”和“Combinations”。其中前五类关键词依次对应前述的关键概念类别，“Combinations”类的关键词可以转化为一个设备属性和一个操作类别概念。为便于后续理解，算法并不输出关键词本身，而输出对应的关键概念。表 2 和表 3 所示为关键概念及关键词集的总结。

表 2 关键概念
Table 2 Key Concepts

关键概念类别	概念项举例	概念项数量
cc_device	Door, Light, Sound	7
cc_attribute	Power, Wind, Volume	7
cc_pos	Bedroom, Kitchen	5
cc_oper	Open, Close, Up, Down	4
user_state	Happy, Sad, Angry	3

表 3 关键词集
Table 3 Keywords Set

关键词类别	举例 [关键词(概念项)]	关键词数量
Devices	电灯(Light);音响(Sound)	7
Attributes	开关(Power);音量(Volume)	15
Positions	洗手间，厕所(Washingroom)	11
Operations	开,打开(Open)	23
Userstates	高兴，开心(Happy)	60
Combinations	大点声(VolumeUp)	15

2.2 基于N-Best音节格的关键抽取算法

在命令检测和对话系统的应用中，最通用的关键词抽取方法是基于垃圾模型的方法[6]。这类方法在对限定关键词建立声学模型的同时，构建垃圾模型吸收所有其他发音，其优点是实时性好，但关键词库难以扩展。在音频内容检索的研究中，研究者主要利用语音识别引擎产生基于音节或音素的搜索网格[7]进行关键词抽取，这类方法更换词库方便，但搜索开销较大。

针对上述问题，本文设计了基于 N-Best 音节格的命令关键词抽取算法。为保证关键词库的可扩展性，并使语音识别结果有较高的稳定性，采用基于微软 SAPI5.1 的大词汇量连续语音识别引擎作为前端。为进行细粒度词汇匹配，将语音识别结果转换为拼音和声调的组合串，构建基于音节的 N-Best 搜索网格作为关键词抽取的输入。N-Best 音节格即 N 个识别得分最高的经过时间对齐的识别结果语句（实现中取 N=5），其形式见图 2。采用 N-Best 音节格代替 1-Best 的识别结果能够为关键词匹配提供更多的信息，并且这种音节格的形式简单，搜索时花销小，适合命令检测。

ba3	wo3	shi4	dian4	deng1	da3	kai1	0.2527
dang5	wo3	shi4	dian3	deng3	da3	kai1	0.2514
dang5	wo3	shi4	jian4	deng3	da3	kai1	0.2443
ba3	wo4	shi4	dian4	deng1	da3	kai1	0.2283
ba3	wo3	shi4	jian4	deng3	da3	kai1	0.2208

图 2　5-Best 音节格举例（“把卧室电灯打开”）

Fig. 2　Example of 5-Best Syllable-level Lattice(“Turn on the light in bedroom”)

关键词抽取算法的输出是一个关键概念序列，各关键概念按出现位置排列，并且每个概念项标注其匹配位置和匹配得分。

搜索算法为从 N-Best 音节格的第 1 列开始以音节为单位进行扫描，计算每个预设关键词与从当前列开始的 5 个对应音节串之间的相似度得分，对这 5 个相似度得分以识别得分加权求和作为匹配得分；记录匹配得分超过预设阈值（根据实验取 0.9）的关键词所对应的关键概念、匹配得分和匹配位置，构成关键概念序列；对得到的关键概念序列进行后处理，保证关键概念按匹配位置排列，并且前后关键词之间没有位置重叠。

为计算匹配得分，建立基于最小编辑距离（MED，minimum edit distance）[8]的词汇相似度度量。最小编辑距离是指把一个字符串转换为另一个字符串在编辑操作上所付出的最小代价，MED 越小代表 2 个字符串越接近。其中允许的编辑操作包括替换、插入和删除。采用最小编辑距离作为词汇相似度度量能细致地描述不同音节串之间的差距，并可以用成熟的动态规划算法进行求解。

在音节级的匹配中，令插入、删除和替换代价均为 1。音节级的 MED 将作为词汇级匹配中的替换代价，因此可用两音节长度的最大值进行归一化。定义拼音 PY_i 和 PY_j 间的归一化最小编辑距离为

$$\mathrm{NMED}(PY_i, PY_j) = \frac{\mathrm{MED}(PY_i, PY_j)}{\max\{\mathrm{Length}(PY_i), \mathrm{Length}(PY_j)\}}$$

在词汇级的匹配中，计算 MED 时令删除代价为无穷大，插入代价为 1，替换代价为对应音节之间的归一化编辑距离。当计算得到词汇间的最小编辑距离后，将其转化为取值在 0～1 的相似度度量，定义词汇 W_i 和 W_j 间的相似度为

$$\mathrm{WordSimilarity}(W_i, W_j) = 1 - \frac{\mathrm{MED}(W_i, W_j)}{\mathrm{Length}(W_i)}$$

3 基于概念关系图的命令解析

目前通用的命令控制系统往往只能理解符合预设格式的语音命令，限制了说话人的表达方式。为使说话人能够更自然地进行表达，本文在分析命令结构的基础上，设计了一个基于关键概念限制关系的解析算法，不仅可以适应不同的命令表达方式，还允许用户在一句话中包含多条命令。

由 1.1 节的分析可知，智能家居控制场景中的命令包含设备位置、类别、属性及操作 4 个要素，设备的位置及类别、类别与属性、属性与操作之间存在直接的限制关系，据此构造并维护一个无向图（概念关系图）来表征不同命令要素即关键类别概念之间的关系（不考虑 user_state 类概念），图 3 给

出了关键概念图的一部分。图中的节点表示关键概念项。概念关系图中的节点分为四层，只有相邻层之间存在边，代表直接的限制关系。当 2 个节点之间存在边连接时，表明这 2 个节点表示的关键概念可以在同一条命令中出现。若 1 个节点和相邻层只存在 1 条边，说明如果该节点代表的关键概念存在，可以直接推理出相邻层对应的关键概念也存在。

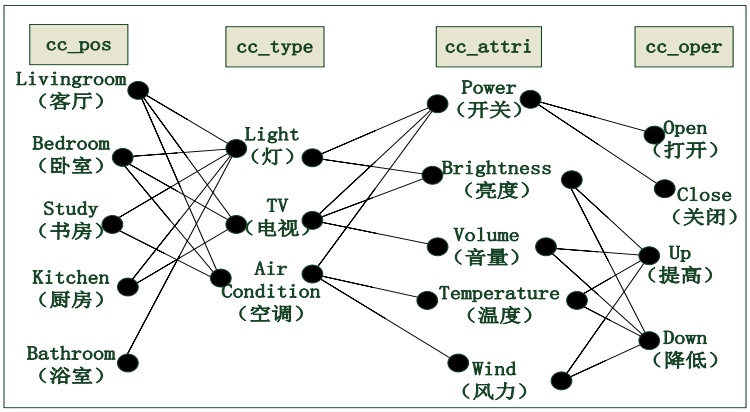

图 3　概念关系图

Fig. 3　Concept Relations Graph

命令解析算法以关键概念序列为输入，根据命令模式输出 0、1 个或多个关键词组合。利用概念关系图进行命令解析的过程如下：首先将语音中未出现但可由输入 KC 唯一确定的 KC 加入待解析序列，并令新加入的 KC 与对应的原始 KC 匹配位置相同；统计每类 KC 的个数并取其最大值 *M*，构造 *M* 条候选命令；从 KC 个数最多的一类开始，对其中的每个 KC 按照概念关系图规定的邻接关系（左/右）在输入序列中查找与其相关的 KC，直至查找到概念关系图中的第 1（4）层；若在相邻层查找关键概念时只找到 1 个，将其加入包含当前 KC 的候选命令；若找到多于一个 KC，计算这些 KC 与当前 KC 的匹配位置之间距离，则选取距离最小或未被加入候选命令的一个；最后对抽取得到的候选路径进行筛选，去除重复路径和不完整路径，输出合法路径对应的关键词组合结果。

为提高命令检出率，在进行二次识别时考虑历史信息：历史信息为用户上一次语音输入中的关键概念，在查找时若找不到相关的概念项，则用合适的历史信息概念代替。

4 说话人情绪状态判定及多模态信息的线性加权融合

说话人的情绪状态是比命令内容更高层次的语义信息，对意图理解有重要的影响。为对说话人状态进行判定，利用情绪相关的关键词和语音声学特征的多模态信息进行融合，提出了一种线性加权的多模态信息融合方法。

4.1 基于文本的情绪状态判定

命令解析模块只处理命令内容相关的关键概念，基于文本信息的情绪状态判定以关键词抽取模块得到的属于“user_state”类的关键概念为输入，输出语句属于 4 种基本情感类别的置信度得分，其值域为 0～1。当 user_state 类的关键概念被检测到，令文本属于情感类别 *k* 的置信度得分为

$$C_T(k)=\begin{cases}\dfrac{\sum_{i=1}^{N_k}\mathrm{CW}(k,i)}{\sum_{k}N_k}, & N_k>0\\ 0, & N_k=0\end{cases}$$

$$k\in\{\text{angry,happy,sad,surprise}\}$$

其中，N_k 表示检测到的第 *k* 类情感的关键词数，CW*(k,i)*表示属于 *k* 类情感的第 *i* 个词的置信度得分（关键词匹配得分）。

4.2 基于语音韵律特征的情绪状态判定

近年来，国内外已有大量语音情感识别的研究成果，其基本的研究思路是通过对语音声学特征进行分析和抽取，利用不同的模式分类方法将语音判别为某一类情感类型。本文借鉴语音情感识别的一般方法，构建了如图 4 所示的模式分类框架。

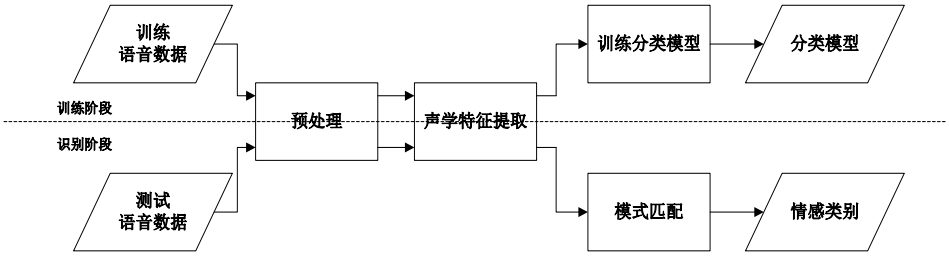

图 4　情感分类一般框架

Fig. 4　General Framework of Emotion Classification

情绪状态相关的声学特征对识别结果有重要影响。目前在语音情感识别中使用最为广泛的声学特征包括韵律特征、音质特征和频谱特征。情感语音相关的大量分析结果表明，在这几类声学特征中，语音情感信息主要体现在韵律特征的变化上[9]。语音的韵律特征包括基频、能量和语速的相关统计或时序特征。因此选用音节平均时长、平均短时能量、平均基频、最大基频和基频范围作为用于识别的声学特征。声学特征的抽取利用 praat 软件完成[10]。

有关语音情感识别的大量研究证明了支持向量机（SVM）分类方法的有效性，因此，获得特征向量后，利用 SVM 的开源工具包 libsvm[11]进行模型训练和识别。其中分类器类型选定为 C-SVM，核函数采用 3 阶的径向基函数（RBF），分类器的输出为 $C_p(k)$，为语句与第 k 类基本情感相对应的匹配得分。

4.3 多模态信息的线性加权融合

多模态信息融合进行说话人情绪状态和命令状态的最终判定，可以在特征层或决策层进行融合。由于声学特征和文本特征在形式上存在很大差异，选择在决策层进行融合。

文献[12]提出了一种求均值的方法来进行决策级的多模态信息融合，该方法简单易行，相当于对各模态信息赋予相同权重，但是忽略了不同模态本身的置信度。为描述不同模态信息的置信度，提出了一种加权求和的多模态信息融合方法。

融合算法的输入为基于文本和韵律特征的情绪状态判定模块的输出结果，二者均为输入语句属于 4 类基本情感的置信度得分，可以看作为四维的置信度得分向量，其各分量均为 0～1 之间的实数。计算语句最终属于第 k 类情感置信度得分的线性加权公式为

$$C(k)=\mu C_T(k)+(1-\mu)C_P(k),\quad k\in\{\text{angry,happy,sad,surprise}\}$$

其中μ 为加权系数，取值范围是 0～1。采用实验方法确定加权系数，对语料库中的每个训练语句进行 2 种模态置信度得分向量的抽取，选取令分类正确的语句数最多的系数值μ ，最终选定μ 为 0.45。μ 的取值也说明了情绪信息主要蕴含在语音的声学特征中，所以语音模态的置信度应当高于文本模态。

多模态信息融合模块还对命令状态进行基于规则的判定：如果命令解析模块没有输出完整命令，则将命令视为无效；否则命令有效。对于命令优先级，若最终用户状态融合结果属于愤怒情感的置信度最高，置优先级为最高级 level-2；若属于悲伤情感的置信度最高，置优先级为最低级 level-0；对其他有效命令置优先级为普通级 level-1。

5 测试与实验

5.1 命令内容检测性能测试

在命令集中选择 15 条命令进行测试，其中包含 2 个关键概念的命令有 2 条，包含 3 个关键概念的有 10 条，包含 4 个关键概念的有 3 条。对每条命令进行 60 次语音输入，不限制测试者对命令的表达方式，如“开门”也可以说成“把门打开”。

测试指标为命令的 1 次检出率和 2 次检出率。

一次检出率 P_1 定义为进行一次命令输入时的检出率，有

$$P_1 = \frac{\text{第1次输入检出正确命令数}}{\text{测试命令总数}} \times 100\%$$

当第 1 次输入未得到合法命令时，系统提示用户进行第 2 次输入，在进行命令解析时利用历史信息。二次检出率 P_2 定义为

$$P_2 = \frac{\text{2次输入至少1次检出正确命令数}}{\text{测试命令总数}} \times 100\%$$

测试结果如图 5 所示。

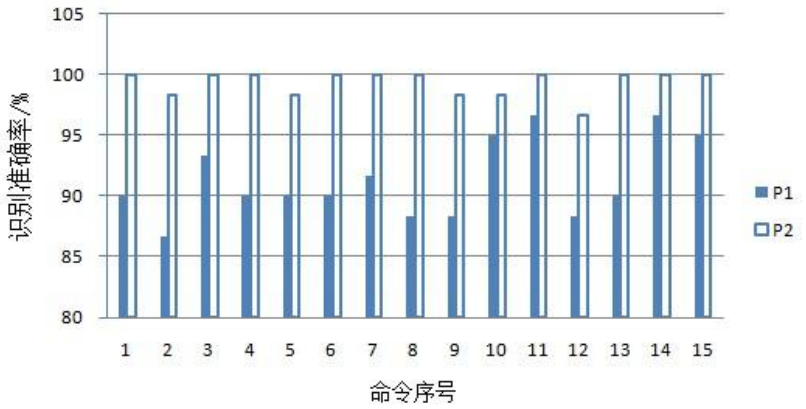

图 5　15 条测试命令的一次与二次识别准确率

Fig. 5　P_1 and P_2 of 15 test commands

15 条测试命令的平均 1 次识别准确率为 91.3%，平均 2 次识别率达到 99.3%，说明系统能有效地从用户语音中抽取命令内容。

5.2 情绪状态判定实验

为进行说话人情绪状态的判定，用一个包含 4 种情感语句的语料库进行训练和测试，这 4 种基本情感分别为愤怒、高兴、难过和惊奇。每种情感含有 220 个语句。句子文本并不限于命令内容。对文本中情感相关的关键词进行人工筛选构造关键词库。每种情感选用 200 句作为训练数据，其他句子用于测试。

定义情绪状态判别准确率为

$$P_e = \frac{\text{情绪状态判别正确的语句数}}{\text{测试语句总数}} \times 100\%$$

只采用语音声学特征时，对情感状态包含在训练数据中的 80 个测试语句，情绪状态判别准确率达到 88.8%。加入文本情绪状态信息时，情绪状态判别准确率达到 93.8%。证明了融合多模态信息能提高情绪状态判别的准确率。

6 结束语

本文提出一种基于多模态信息融合的语音意图理解方法，通过从识别文本和语音信号中抽取多模态信息进行融合进行说话人意图理解。该系统能获得包括说话人命令内容、情绪状态和命令状态在内的更加全面的用户意图信息，有利于人机交互的顺利进行。

［参考文献］(References)

[1] Wang Y Y, Deng L, Acero A. Spoken language understanding[J]. IEEE Signal Processing Magazine, 2005, 22(5): 16-31.

[2] Kimura H, Tokuhisa M, Mera K, et al. Comprehension of intentions and planning for response in dialogue[J]. Technical Report of IEICE, 1998, 98(441): 25-32.

[3] Wang Y Y. Strategies for statistical spoken language understanding with small amount of data – an empirical study[C]// Proceeding of INTERSPEECH'10. 2010: 2498-2501.

[4] Shimada K, Iwashiata K, Endo T. A case study of comprehension of several methods for corpus-based speech intention understanding[C]// Proceeding of PACLING'07. 2007: 255-262.

[5] 宗成庆，吴华，黄泰翼，等. 限定领域汉语口语对话语料分析[C]// 全国第五届计算语言联合学术会议论文集. 1999: 115-122.
Zong Chengqing, Wu Hua, Huang Taiyi, et al. Analysis of Spoken Dialog Corpus in Restrict Domain[C]// Proceeding of JSCL-99. 1999: 115-122. (in Chinese)

[6] Rose R, Paul D. A hidden markov model based keyword recognition system[C]// Proceeding of ICASSP'90. 1990: 129-132.

[7] Lin H, Stupakov A, Bilmes J. Improving multi-Lattice alignment based spoken keyword spotting[C]// Proceeding of ICASSP'09. 2009: 4877-4880.

[8] Navapro G. A guide tour to approximate string matching[J]. ACM Computing Surveys, 2001, 33: 31-88.

[9] Tao J H, Tan T, Picard R W. Feature importance analysis for emotional speech classification[C]// LNCS 3748. 2005: 449-457.

[10] Boersma P, Weenink D. Praat: doing phonetics by computer [EB/OL]. [2011-3-3]. http://www.praat.org/.

[11] Chang C C, Lin C J. LIBSVM: a library for support vector machines[EB/OL]. [2010-3-3]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[12] Lee C M, Narayanan S. Towards detection emotions in spoken dialogs[J]. IEEE Transaction on Speech and Audio Processing, 2005, 13(2): 293-302.