

# 航管模拟器的语音识别与合成技术实现

潘倩<sup>1,2</sup>, 张玲<sup>3</sup>, 胡术<sup>1,2</sup>, 李璞<sup>1,2</sup>, 李艳<sup>1,2</sup>

(1. 四川大学 计算机学院, 四川 成都 610064;

2. 四川大学 视觉合成图形图像技术国防重点学科实验室, 四川 成都 610064;

3. 四川大学 计算机基础教学实验中心, 四川 成都 610064)

**摘要:**随着航空运输业的快速发展,亟需雷达模拟机、塔台模拟机等设备培训空中管制人员,以提高空管人员的处置能力。为了提高航管雷达模拟训练的效率,基于微软语音引擎 Windows Speech SDK,成功实现了航管模拟器中的语音识别与合成技术。语音识别中,基于基本语音识别,通过在 xml 中添加语音集和语法集来提高识别率;设计并构建了动态产生配置文件的训练系统,该系统通过特定词语的语音训练生成个人 xml 文件,正式使用时只需将 xml 文件导入以减少人工操作。语音合成中,通过微软引擎实现了基本的朗读功能,并提出了使声音更接近人声的方法。鉴于空管训练的严肃性,采用录音和播放方式实现了由特定人朗读的语音合成功能。运行情况表明,所设计构建的空管训练系统较好地满足了当前空中管制训练的要求,可明显提高航管模拟器的效率。

**关键词:**航管模拟器;语音识别与合成;xml 技术;微软语音引擎

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2017)07-0131-04

doi:10.3969/j.issn.1673-629X.2017.07.030

## Implementation of Speech Recognition and Synthesis in ATC Simulator

PAN Qian<sup>1,2</sup>, ZHANG Ling<sup>3</sup>, HU Shu<sup>1,2</sup>, LI Pu<sup>1,2</sup>, LI Yan<sup>1,2</sup>

(1. College of Computer, Sichuan University, Chengdu 610064, China;

2. National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610064, China;

3. Computer Teaching Experiment Center, Sichuan University, Chengdu 610064, China)

**Abstract:** With the rapid development of the air transportation industry, it is urgent to train the air control personnel by the equipment of radar simulator and tower simulator so as to improve the handling capacity of air traffic control personnel. In order to improve the efficiency of simulation training, the speech recognition and synthesis technology in the simulator has been successfully realized based on the Microsoft speech engine Windows Speech SDK. In speech recognition, voice sets and grammar sets are added into the xml to improve the recognition rate based on basic speech recognition; a training system that dynamically generates configuration file has been designed and built through a specific word voice training to generate personal xml file. When officially using the system, the xml file must be imported into the system for reducing manual operation. In speech synthesis, the basic reading function has been realized through Microsoft engine, and the new function has been implemented that makes the voice more reality. In view of the seriousness of air traffic control training, the function of reading by specific person is realized by recording and playback. The simulation results show that the designed ATC training system has met the requirements of the current air traffic control training and improved the efficiency of the simulator.

**Key words:** ATC simulator; speech recognition and synthesis; xml technology; Microsoft speech engine

## 0 引言

航管模拟器是航管雷达模拟训练的仿真系统,该系统目前已经被普遍使用在军民航之中。

航管模拟器由模拟管制席位、模拟机长席位、中心机、服务器、数据库和通讯系统组成<sup>[1]</sup>。学员操作管制席位,教员操作机长席位<sup>[2]</sup>。系统工作流程如下:首先

收稿日期:2016-06-15

修回日期:2016-09-28

网络出版时间:2017-06-05

基金项目:中国民航创新基金项目(MHRD20150228)

作者简介:潘倩(1992-),女,硕士研究生,研究方向为计算机网络、分布式系统理论;张玲,实验师,研究方向为计算机应用、仿真与网络;胡术,博士,副教授,通讯作者,研究方向为计算机网络、操作系统及中间件。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20170605.1506.026.html>

由数据库拟定计划并保存,然后由中心机开启对应的训练计划。学员关注空域情况并发送管制命令,教员收到该命令后改变对应参数,然后将结果发送至服务器,服务器将收到的数据进行计算后发送到所有席位,各席位再更新各自的数据。

在目前的航管模拟器中,学员与教员之前交互使用声音信号进行传输,教员通过键盘输入的方式来记录管制训练人员发布的命令。这种操作方式有诸多缺点:教员的工作量很大,对教员的个人素质要求较高,不仅要听懂还要能快速录入管制命令;一旦错误录入管制命令,会使训练效果不理想<sup>[3]</sup>。

因此在航管模拟器中使用语音识别是有必要的,通过语音识别将命令转换为文字,可以实现自动模拟机长的功能,大大提升训练效率。为此,基于 Windows 的语音识别与合成引擎,设计并实现了航管模拟器中的语音识别与合成系统。为提高识别率,通过 XML 文件添加语法集和语音集,实现了语音训练系统。为适应空管训练的环境,实现了特定人朗读的语音合成系统。

## 1 SAPI

### 1.1 SAPI 简介

随着计算机技术的发展,语音合成与识别技术越来越成熟,应用越来越广泛,如苹果公司的 Siri 软件和高德地图的语音导航等。当前很多中文语音识别引擎,如科大讯飞和百度语音,都是将声音文件通过互联网传到服务商的服务器进行识别,识别后将结果返回终端设备。由于航管模拟器系统运行于离线环境,因此采用离线的微软语音识别引擎 SAPI。

SAPI,全称是 The Microsoft Speech API,即微软语音 API。使用该 API 需要调用 Speech SDK 开发包<sup>[4]</sup>。

Windows Speech SDK 有语音识别和语音合成两种引擎<sup>[5]</sup>。语音识别引擎用于识别语音(该语音既能由麦克风实时输入来获取,也能通过将录制好的. mp3 等文件导入来获取),并将语音转换为文字输出。语音合成引擎则是将文字转换为语音,并将该语音输出(该输出既能通过听筒实时输出,也能直接形成. mp3 等文件)。

SAPI 的结构如图 1 所示。应用层直接调用 API 与 SAPI 进行通信,语音引擎则调用 DDI 层和 SAPI 进行交互<sup>[6]</sup>。

目前最常用的 Windows Speech SDK 版本有三种: 5.1、5.3 和 5.4。

5.1 版本在 XP 系统和 Server 2003 系统中使用,需手动下载。XP 系统默认只有英文语音库,若使用中文语音库,需下载中文补丁包 SpeechSDK51LangPach.

exe。

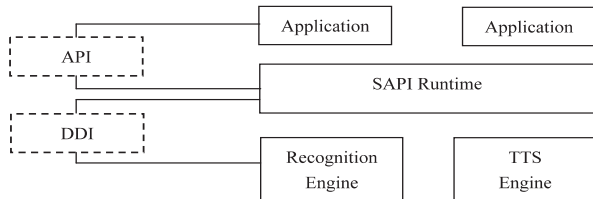


图 1 SAPI 结构图

5.3 版本在 Vista 系统和 Server 2003 系统中使用,无需手动下载,并且系统默认带有中文语音库 lili。

5.4 版本支持 Windows7 系统,也无需手动下载。Windows7 系统带有中文(lili)和英文(Anna)语音库。其中,lili 支持中英文混读。lili 与低版本的中文库不同之处在于,当中文中出现英文单词,低版本合成的结果是逐个字母朗读,而 Microsoft lili 合成的结果为朗读整个英文单词。

### 1.2 SAPI 安装和 VC 环境配置

系统使用了 Windows7 自带的 Windows Speech SDK 5.4 版本的开发包,默认路径为 C:\Windows\SysWOW64\Speech。创建项目后,需在程序中加入头文件和静态库:

```
#include "sapi. h"
#include "sphelper. h"
#pragma comment(lib, "sapi. lib")
```

## 2 语音识别的实现

### 2.1 基本语音识别

#### 2.1.1 初始化操作

为了实现基本的语音识别,需首先进行初始化:

(1) 调用 CoInitializeEx 函数进行 COM 组件的初始化。

(2) 调用 CoCreateInstance 创建语音识别引擎<sup>[7]</sup>。

(3) 调用 ISpRecognizer::CreateRecoContext 创建识别上下文。

(4) 调用 SetNotifyWindowMessage 设置识别消息。

(5) 调用 SetInterest 设置相应事件。

#### 2.1.2 设置匹配词典

识别过程中匹配的词典分为两种:听写模式和语音控制模式。其最主要的差异是,两者识别中使用的语法字典不同。前者使用的是通用型字典,所涉范围广泛,词语数量较多,适合没有特定范围的识别场合。并且,正是因为词汇量大而直接导致识别的精度较低,识别速度也较慢。而后者的字典是由程序员添加,字典格式为 xml。xml 是一种使用特定格式存储数据的文件,能够描述结构化的数据,也能够有效描述半结构化,甚至是非结构化的数据<sup>[8]</sup>。由于该模式是由程序员添加需识别词语,所以减少了识别过程中查询的词

汇数量,因此能提升识别效率。同时,因为词典中候选项极少,所以一般不会出现识别错误的情况。

听写模式加载词典调用的是 LoadDictation,而语音控制模式加载词典调用的是 LoadCmdFromFile。创建好语法规则后需调用 SetDictationState 或 SetRuleState 激活语法识别。

2.1.3 释放所有对象

调用各对象的 Release 方法。

2.2 提高识别精度

上述代码运行后,若在听写模式下,由于该模式调用的字典是通用型字典,范围太广,所以识别的精度并不高。有两种方法来解决此问题,以 Windows7 系统为例,一种是使用微软的语音识别组件,一种是在代码中添加 xml 来添加自己的语法字典和语音字典。

第一种方法共有三个部分:

(1)在语音设置中调整麦克风位置和自己声音的大小,调整好的标准是:音量计的计数稳定在一定区域<sup>[8]</sup>,如图 2 所示。

(2)训练操作系统的语音配置文件,其文本是操作系统规定的语句。通过训练,能略微提升识别效率。

(3)该步效果较为明显,也就是使用组件的语音词典功能,通过添加自己需要识别的文字并录制自己的声音来修改 Windows 识别引擎的语音库,从而大大提高识别率。



图 2 设置麦克风

通过设置组件的方式比较简洁,但是只适合一个用户使用一台计算机的场景。其原因是:通过训练和添加语音词典的方式,用户的语音配置文件将自动集成到用户账户的配置文件中,但用户账户的配置文件无法导出。为适应多人使用同一台机器进行识别,可在一台计算机上为每个学员创建一个用户账户,当不同学员使用语音识别时需要切换到各自的账户。这类方法操作起来较为繁琐,也很浪费时间。

第二种方法为通过 xml 文件定义个人的语法字典和语音字典。系统通过 xml 来添加语法集和语义集来

提高识别率<sup>[9]</sup>,原因如下:其一是管制训练人员来自不同地区,各自的说话方式不同,所以系统需要提供语音集。其二是航管命令不同于日常生活用语,词汇量较少,且有一定的语法结构,系统需要添加语法集。

经过实践证明,第一种方法由于使用了微软公司开发的语音组件,改变了识别时匹配的语音库,所以识别率很高。第二种方法由于未能真正改变匹配的语音库,识别率并没有前者高。因此,为了减少操作的复杂性,最终选择第二种方式。

2.3 添加语法集

2.3.1 语法树结构

若有一课语法树,如图 3 所示。

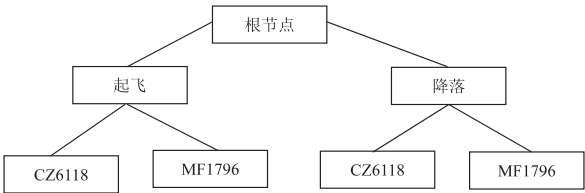


图 3 语法树结构

2.3.2 生成 BNF 语法

S: = MainRule. SubNameRule  
MainRule: = 起飞|降落  
SubNameRule: = CZ6118|MF1796

以上的 BNF 语法中,英文(如“MainRule”)表示的是变量,中文表示的是需要识别的内容,“|”表示或,“.”表示的是连接符号。

2.3.3 生成相应的 xml

在项目文件中,添加资源文件,类型为 xml,然后在界面中添加与以上 BNF 语法对应的语法数据<sup>[10]</sup>:

```
<? xml version = "1.0" encoding = "utf-8" ? >
<GRAMMAR LANGID = "804" >
<DEFINE>
<ID NAME = "VID_MainRule" VAL = "1" />
<ID NAME = "VID_SubNameRule" VAL = "2" />
</DEFINE>
<RULE ID = "VID_MainRule" TOPLEVEL = "ACTIVE">
<O>
<L>
<P>起飞</P>
<P>降落</P>
</L>
</O>
<RULEREFF REFID = "VID_SubNameRule" />
</RULE>
<RULE ID = "VID_SubNameRule">
<L>
<P>CZ6118</P>
<P>MF1796</P>
</L>
```

```
</RULE>
</GRAMMAR>
```

LANG1 \* = "804" 代表简体中文, 在 < \* >... < / \* > 中增加命令<sup>[11]</sup>。通过添加以上 xml 文件, 只有当对着麦克风朗读“起飞 CZ6118”、“起飞 MF1796”、“降落 CZ6118”或“降落 MF1796”时, 系统才能识别, 如果学员说出其他文字, 或者系统不能识别学员所说的“起飞 CZ6118”等, 也无法打印出来。

注意: 若在 xml 文件中需要实现通配的语义, 在 <P>关键字</P>中, 关键字左右添加“\*+”和“\*+”表示多个任意词<sup>[12]</sup>, 当一个语音句子中包含关键词和非关键词时, 这个句子实际上可以用规则“\*+关键词\*+”来描述, 即在关键词之前和之后可以有任意的非关键词语音<sup>[13]</sup>。

## 2.4 添加语音集

为实现向 xml 添加语音集, 需了解<P DISP = “降落”>江罗</P>的含义。该 xml 语句表示, 当计算机识别结果为“江罗”时, 会向屏幕上打印出“降落”<sup>[3]</sup>。

通过这个标签, 可实现添加个人语音集的功能。假设原本要表达的意思是“降落”, 但因为口音, 朗读出来的“降落”计算机认为更像是“江罗”, 所以识别结果是“江罗”, 但是因为添加了以上 xml 语句, 计算机会把识别出的“江罗”打印成“降落”。

所以, 假设一个人对着麦克风说三次“降落”, 计算机识别结果依次是“江罗”、“姜萝”和“角落”, 将这些都记录下来, 添加到 xml 语句中。也就是把以上语句中的“<P>降落</P>”替换为:

```
<P DISP = “降落”>江罗</P>
<P DISP = “降落”>姜萝</P>
<P DISP = “降落”>角落</P>
```

通过这种方法, 可以大大提高系统的识别率, 并且经过测试, 对于同一个词, 朗读的次数越多, 记录的个数越多, 可选的个数也越多, 识别的精度就越高。

## 2.5 动态生成 XML

除上述的识别系统外, 设计并实现了航管模拟器中的训练系统, 每个学员使用系统进行一次训练后, 能生成个人的 xml 文件, 将该 xml 文件添加到正式识别系统的配置文件中, 可大大提高识别效率。经过测试, 该训练系统可以实现预期的功能, 减少了人工操作, 满足了语音识别的需求。

# 3 语音合成的实现

## 3.1 一般语音合成

在日常生活中的很多方面都用到了语音合成技术, 如天气预报、机场广播等等。在这些日常的语音合成中, 由于对语音库的声音要求不高, 所以能直接使用

微软的语音合成引擎。基本方法如下:

### 3.1.1 基本合成

(1) 初始化语音接口<sup>[14]</sup>。

```
ISpVoice * pSpVoice;
::CoInitialize( NULL ); //初始化 COM 组件
HRESULT hr = CoCreateInstance( CLSID_SpVoice,
NULL, CLSCTX _ ALL, IID _ ISpVoice, ( void * * )
&pVoice );
```

(2) 朗读字符串内容。

```
pSpVoice -> Speak ( L “ 起飞”, SPF _ DEFAULT,
NULL );
```

//通过调用 Speak, 系统可发出“起飞”的声音

(3) 释放资源。

```
pSpVoice->Release( );
::CoUninitialize( );
```

### 3.1.2 ISpVoice 的成员函数

也可以通过调用 ISpVoice 的成员函数来使合成的声音更接近人声。例如, 使用 SetRate() 和 GetRate() 方法能设置和获取朗读的速度, 使用 SetVolume 和 GetVolume 能设置和获取语音库的音量大小等。

### 3.1.3 使声音更自然

为了使朗读的声音更自然, 可以下载 Windows8 的语音库“Microsoft HuiHui”。经过测试, 该语音库朗读出的效果比“Microsoft lili”要更符合人的说话声音。其原因是逐词朗读比逐字朗读更自然, 而 HuiHui 正是比 lili 录入了更多的词汇。

也可以尝试使用其他中文引擎, 例如科大讯飞、捷通华声等能免费下载科大讯飞语音合成引擎, 而且该引擎支持 Windows 和 Linux 等多种平台, 共有两种离线语音合成引擎, 但其中一个需要连接互联网才能使用, 而另一个可以完全在离线的条件下使用。对于两者语音合成的效果, 从测试来看, 前者朗读的声音不仅自然, 而且带有一定的情感, 非常逼真, 而后者朗读的声音也很自然, 与 HuiHui 的效果接近。

## 3.2 特定人发声的语音合成

在一些特定领域, 例如在航管模拟器中, 由于航管命令数量有限, 航管训练的场合比较严肃, 所以需要使用时特定人的声音实现语音合成。实现此功能主要是使用录音和播放的方法。

例如: 需要系统语音合成一个中文词语为“起飞”, 首先使用录音工具录制特定人朗读“起飞”的声音, 形成 .wav 文件。如“起飞 .wav”, 再使用::playSound 播放<sup>[1]</sup>。经测试, 该语音合成系统能够实现预期功能, 与上述的语音识别系统一起实现了自动机长的功能, 大大提升了训练效率。不足的是, 合成的声音不够流

(下转第 139 页)



重建过程的适用性和可行性,通过仿真实验对其进行验证。理论分析和实验结果表明,提出的算法可以很好地应用在超声图像的重建中。

参考文献:

[1] 万明习. 生物医学超声学[M]. 北京:科学出版社,2010.

[2] 万明习,宗瑜瑾,王素品,等. 生物医学超声实验[M]. 西安:西安交通大学出版社,2010.

[3] 王文博. 关于医学超声成像机理的研究[D]. 青岛:青岛大学,2006.

[4] 张仕刚,谢耀钦,包尚联. 医学影像物理学学科的现状和未来[J]. 物理,2004,33(10):753-758.

[5] Candes E J,Tao T. Near optimal signal recovery from random projection; universal encoding strategies[J]. IEEE Transactions on Information Theory,2006,52(12):5406-5425.

[6] Donoho D L. Compressed sensing[J]. IEEE Transactions on Information Theory,2006,52(4):1289-1306.

[7] 白凌云,梁志毅,徐志军. 基于压缩感知信号重建的自适应正交多匹配追踪算法[J]. 计算机应用研究,2011,28(11):4060-4063.

[8] Donoho D L. For most large underdetermined systems of linear equations,the minimal  $\ell_1$  norm solution is also the sparsest solution[J]. Communications on Pure and Applied Mathematics,

ics,2006,59(6):797-829.

[9] 李树涛,魏 丹. 压缩传感综述[J]. 自动化学报,2009,35(11):1369-1377.

[10] 邵文泽,韦志辉. 压缩感知基本理论:回顾与展望[J]. 中国图象图形学报,2012,17(1):1-12.

[11] Lobo M S,Vandenbergh L,Boyd S,et al. Applications of second-order cone programming[J]. Linear Algebra and Its Applications,1998,284(1):193-228.

[12] Elad M. Optimized projections for compressed sensing[J]. IEEE Transactions on Signal Processing,2007,55(12):5695-5702.

[13] Kingsbury N G. Complex wavelets for shift invariant analysis and filtering of complex wavelets for shift invariant analysis and filtering of signals[J]. Journal of Applied and Computational Harmonic Analysis,2001,10(3):234-253.

[14] Herrity K K,Gilbert A C,Tropp J A. Sparse approximation via iterative thresholding[C]//Proceedings of the IEEE international conference on acoustics,speech and signal processing. Washington D. C. ,USA:IEEE,2006:624-627.

[15] Candés E,Romberg J,Tao T. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information[J]. IEEE Transactions on Information Theory,2006,52(2):489-509.

(上接第 134 页)

畅,这也是今后需要改进的地方。

4 结束语

为提高航管模拟器的培训效果和效率,基于录音并播放的方法,设计并实现了语音合成系统;同时设计并构建了训练系统和识别系统,实现了较高识别率的语音识别功能。为了提高语音识别的准确率,进行了很多有益的探索,从设置组件中的语音字典到动态生成 xml 文件,从添加语义集到添加语音集,都做了大量的实际工作。经过实践证明,该系统的设计可以很好地适应当前的管制训练,完全满足了系统实际运行的需求,大大提高了航管模拟器的效率。但还有不足的地方。例如,在语音识别系统中,单个词汇识别错误应该如何修改,这个功能并没有实现。希望通过此后的学习解决该问题。

参考文献:

[1] 由 扬,徐肖豪. 空管模拟机的 IBM ViaVoice 技术实现研究[J]. 中国民航学院学报,2002,20(3):6-9.

[2] 彭志勇. 语音识别功能在 DRS2000 雷达模拟机系统中的设计与实现[D]. 成都:四川大学,2006.

[3] 李 锐. 语音技术在塔台模拟机上的应用[D]. 成都:四川

大学,2004.

[4] Microsoft Corporation. Microsoft speech SDK version 5. 1[EB/OL]. 2016. <http://www.microsoft.com/speech/download/sdk51/>.

[5] Microsoft Speech SDK (SAPI) 5. 1 Help[M]. [s. l.]:Microsoft Corporation,2001.

[6] 肖 玮. 使用 SAPI 实现语音识别与合成[J]. 现代计算机,2005(2):91-94.

[7] 尹 成. Visual C++ 2010 开发基于 Windows7 的语音识别与语音合成[J]. 程序员,2010(6):116-118.

[8] 黄妙燕,王咸锋. 基于 Microsoft 语音识别引擎的语音识别系统的设计[J]. 电脑开发与应用,2010,23(9):74-75.

[9] 黄 旭. 基于 HTK 和 Microsoft Speech SDK 的连续语音识别系统的研究及实现[D]. 厦门:厦门大学,2007.

[10] XML schema part 0:primer[EB/OL]. 2016. [http://www.w3.org/TR/xmlschema-0/Simon St. Laurent](http://www.w3.org/TR/xmlschema-0/Simon%20St.%20Laurent).

[11] Extensible Markup Language (XML) 1. 0[M]. 2nd ed. [s. l.]:[s. n.],2013.

[12] Birbeck M. XML 高级编程[M]. 北京:机械工业出版社,2002.

[13] 林 茜,欧建林,蔡 骏. 基于 Microsoft Speech SDK 的语音关键词检出系统的设计和实现[J]. 心智与计算,2007(4):433-441.

[14] 潘爱民. COM 原理与应用[M]. 北京:清华大学出版社,1999.