doi: 10.7690/bgzdh.2023.04.005

人机交互语音识别发展及军事应用分析

鹿哲源[1]，牛小明[1,2]，康　林[3]，李文才[1]，刘歆浏[1]
（1. 中国兵器装备集团自动化研究所有限公司信控中心，四川 绵阳 621000；
2. 重庆大学数学与统计学院，重庆 400044；3. 陆装驻广元地区军代室，四川 广元 628000）

摘要：针对人机交互语音识别技术军事应用的现状，介绍语音识别技术的发展历史，并对其军事应用进行分析。将关键词识别技术应用于军事场景中，介绍目前主流的几种关键词识别模型，并对其在军事领域的应用进行展望。结果表明，该分析能为语音人机交互技术应用于军事装备提供参考。

关键词：人机交互；语音识别；关键词识别
中图分类号：TJ02　　**文献标志码：**A

Human-computer Interaction Speech Recognition Development and Military Application Analysis

Lu Zheyuan[1], Niu Xiaoming[1,2], Kang Lin[3], Li Wencai[1], Liu Xinliu[1]
(1. *Weapon Equipment Information and Control Technology Innovation Center*, *Automation Research Institute Co*., *Ltd*., *of China South Industries Group Corporation*, *Mianyang* 621000, *China*;
2. *College of Mathematics and Statistics*, *Chongqing University*, *Chongqing* 400044, *China*;
3. *Military Representative Office in Guangyuan District*, *Army Equipment Department*, *Guangyuan* 628000, *China*)

Abstract: In view of the current situation of the military application of human-computer interaction speech recognition technology, the development history of speech recognition technology is introduced, and its military application is analyzed. Keyword recognition technology is applied to the military scene, several current mainstream keyword recognition models are introduced, and its application in the military field is prospected. The results show that the analysis can provide a reference for the application of voice human-computer interaction technology in military equipment.

Keywords: human-computer interaction; speech recognition; keyword recognition

0　引言

工业化时代，按钮、开关、拉杆等被应用于机器控制，是人机交互的主要手段。电子信息化时代，新增了感应式触摸屏、实体或虚拟数字键盘、软件菜单等人机交互手段，人机交互手段进一步丰富。智能化时代，基于语音、肢体动作识别的非接触式新兴人机交互手段因使用方式灵活、便捷等优势，在商用领域的发展和应用十分活跃。军事装备的人机交互手段，因在高强度对抗的战场环境中应用，不但要求方式灵活、便捷，而且更为关注交互的快速性和准确性，这是人机交互语音识别技术军事化应用发展的重点。

1　语音识别技术的发展历程

1.1　语音识别的兴起

语音识别的研究产生于 20 世纪 50 年代左右。1952 年，贝尔实验室的 3 位研究人员共同构建了一种叫做“Audrey”的系统，该系统把共振峰定位到每个单词的功率谱中进行辨识，是一种能够辨识十多个英文数码的孤立系统[1]。1956 年，美国 RCA 实验室用提取元音频谱的方法也实现了类似的成绩[2]。

1.2　语音识别的重大突破

语音识别技术在 20 世纪六七十年代前主要有 3 种方法：1) 基于语音的端点检测时间规整方法[3]，该方法利用 ALT 的过滤单元获得声音中的特征；2) 苏联学者 Vintsyuk 提出的动态时间规整(dynamic time warping，DTW)算法，该算法将语音分成短帧单独处理，能够对 200 个单词的词汇表进行操作[4]；3) 卡内基梅隆大学提出的连续语音识别方法，其目的是动态的跟踪音素。1966 年，日本名古屋大学教授板仓文忠发明了一个新型的语言编码方法，即线性预测方法，该方法通过线性预测模型的文字信息，以压缩形式描述为语言信息的频谱包络[5]。

收稿日期：2022-12-23；修回日期：2023-01-28
作者简介：鹿哲源(1998—)，男，陕西人，从事语音识别研究。E-mail: actedeer@qq.com。

进入到 20 世纪 80 年代后，语音识别的方向转向了连接词识别的问题，一系列算法被相继提出。其中包括了两级动态规划法、层建法以及帧同步法[6-8]等。同时，语音匹配的方法从基于模板匹配的方法转变为基于统计概率的方法，这归功于 HMM 模型的推广使用[9]。HMM 的引入，使研究人员可以从一个共同的概率模型中整合不同的信息资源，如声学、语音和句法。现今 HMM 模型仍被作为语音识别系统融合使用。

20 世纪 80 年代末，神经网络也被用来解决语音识别问题，并应用于分类方法。通过对 20 世纪七八十年代语言技术的突破，对 90 年代语言技术的完善，研发基于 GMM-HMM 提供了语言区分训练标准的模型自适应技术，大大提高了语言识别系统的稳定性。

1.3 深度神经网络与语音识别

在 21 世纪的前十几年里，语音识别科学研究的进展相对放缓，但深度学习的研究却逐渐引起了科学家的关注。Hiton 等[10]发现的深层置信网(deep belief network，DBN)使深入神经网络的培养成为可能性。2009 年，Mohamed 等[11]将深入神经网络运用于声学模型，并在小词汇中利用 TIMIT 提高库上的性能。2011 年，俞栋、邓力等给出了基于上下文的建议(context dependent，CD)深入神经网络马尔可夫模型[12](CD-DNN-HMM)，比较以往的 GMM-HMM 模式，显著改善了功能。然后研发人员采用声音建模功能较强的新模式，比如 CNN，LSTM，CLDNN 等替代 DNN 进一步提高了声音识别系统的表现。近年来，通过采用如 CTC 和 Encoder-Decoder 等的声学建模框架对比，新声学建模框架中也开始进一步展现 HMM 建模的新特征。

如图 1 所示，整个语言识别系统分为语言信号处理和特性抽取、声学建模(acoustic model，AM)、语言建模(language model，LM)以及解码与检索等 4 大组成部分。

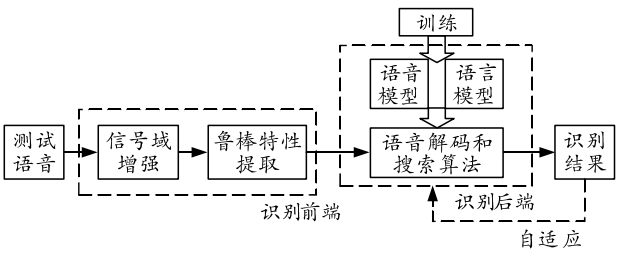

图 1 语音识别系统

语音处理与特征提取技术可以看作声音识别系统的最前端，而语音处理又称为预处理技术，一般包括了取样、数字化、声音增强、预加重、加窗分帧等流程，而目前最主要的特征提取技术则是利用梅尔的倒谱系数[13]来描述声音特征即 MFCC 法。

声学建模、语音模拟和解码都被视为语音识别系统的最后端。声音信息的形成、传递与接收都是一种非常复杂的过程，尤其是在复杂的声学环境中，在信息传递过程中很容易被各种背景噪声和干扰的声音所破坏；因此，对于复杂场景的声音信号降噪主要需要识别前端有效信息提取，有效的声音信息识别需要识别后端声学模型，解码以产生相应的识别结果。

2 语音识别军事应用现状分析

在军事应用中使用语音输入替代传统的手动操作来控制设备，作战人员可以将注意力集中于对目标的判断、攻击火力应用等关键重要操作，以充分发挥战术优势。基于语音识别的人机交互手段在军事应用领域越来越受到关注和发展。

目前，语音识别在军事应用中主要是在电子侦听、语音情报分析、网络对抗中的信息甄别、声纹身份识别等领域。因作战环境噪声大带来的识别准确率不高、快速性不适应高强度作战对抗节奏等因素，而暂未被广泛应用于军事装备的操控输入。军事装备的进一步发展离不开人机交互语音识别技术支撑。军事装备发展趋势之一就是多功能高度集成化，单一武器平台内需要单人操控的设备复杂多样，并行高效操控需求突出，语音输入是有效解决手段之一。随着无人技术发展推动，军事装备发展的又一重要趋势是大量装备无人化，同构、异构无人军事装备将集群化应用于作战，不论是人在回路中，还是人在回路上的操控方式，作战人员需同时指挥操控的装备数量及种类多，如仅采用传统操控手段将增加操控难度和作业强度，不适应高强度快节奏作战需求，而基于语音交互的操控手段可有效解决该问题。

目前，在智能手机、智能音箱、智能电视、智能驾驶座舱等商用领域，人机交互语音识别的重点是语义理解，大多需在网络后台云端支持下，才能有识别准确率保证，且识别反应时间较长(秒级以上)。语义理解虽有使用灵活性，但却不能直接应用于军事装备的语音操控输入。主要原因：1) 军事装备无线网络覆盖有限；2) 基于安全考虑而禁止应用。此外，军事装备的操控指令集相对有限，对基

于自由语义理解的操控输入需求迫切性不强，更重要的是要在轻量化(信息容量及处理速度不高)技术手段下的高效令行禁止。基于关键词的语音控制指令识别是解决这类问题的有效途径之一，它能够从连续的语音流中快速高效地关联识别出预设的控制指令，转变为设备的控制输入。

3　基于关键词识别的语音识别

关键字辨识是语音识别的组成部分，但相比于传统语音识别，关键字辨识更关注标识的内容或所指定词汇，而对于识别内容是什么并不关心。基于这种思路，关键词识别大致可分成如下几类。

3.1　补白模型

补白模型有时也被称为垃圾模型，将关键字识别问题考虑为一个逐帧的序列标记问题。关键词定为不同的标注，而一个额外的“补白”标记用来匹配所有非关键词。

在如图 2 所示的隐马尔可夫模型(hidden markovmodel，HMM)中，由于人们并不了解模子具体的状态顺序，只了解状态转化的概率，即模式的状态转化过程难以观测。

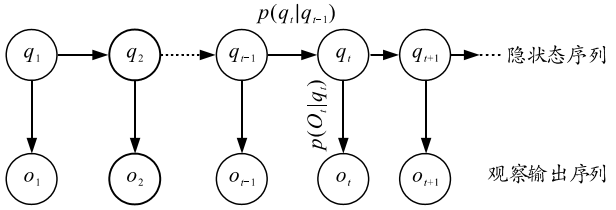

图 2　HMM 图解

补白模型通过为所有关键字创建一个隐马尔可夫模型，为非关键字额外创建一个隐马尔科夫模型，观测概率可通过混合高斯型或神经网络模型得到。陈太波等[14]利用全连接式神经网络，结合 Softmax 类别器对汉语的 408 个声调构建声调类别器，将 Softmax 类别器输出概率当作后验概率图，与隐马尔科夫补白模式(HMM/Filler)实现了首次完全融合，并针对样本量较少的问题使用基于最大后验概率的改进 HMM，即 HMM-MAP，使模型的综合识别率达到了 87.88%

3.2　基于样例关键词识别

基于样例关键词分析可将关键词的问题简单地考虑为一个匹配问题，即研究输入音频和样例词之间的相似程度，如果输入音频的词相似程度达到了指定阈值，就可以认为该音频包含指定词。

基于样例的关键词识别可分为基于 DTW 算法与基于神经网络学习的识别 2 类。

DTW 技术利用对话语音时序的缩短与扩展在时间域中对齐二者的时刻，由此使话音信息的相似性问题转变为 2 个语音特征向量的间距问题。在计算间距的方法中，间距计算方法包括欧式距离、余弦距离、对数内积距离、切比雪夫距离。这种方法属于语音识别早期使用的方法，从召回率以及精确度的数据上来看都在 40%左右，远远达不到应用的程度，李志涵[15]针对 DTW 算法提出了改进方案，如图 3 所示，通过多模板匹配以及加速算法，使模型在召回率上从 46%提升到 70%。相比于之后的算法，DTW 模型占用资源较小，虽然不作为主流的方法使用，但是在算法的融合应用以及近似查询过程中仍有一定的作用。

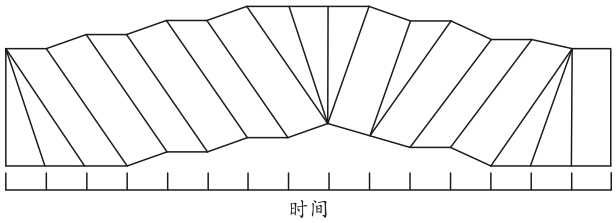

图 3　DTW 算法时序匹配

因为深度学习的流行普及，神经网络有了对特征信息强大的提取能力。基于神经网络的样例识别系统，可以利用神经网络把任何不同波长的语音序列调整在一个确定的长度向量之间，使用预设的关键字来训练神经网络，使网络中能够把相同的关键字映射到互相接近的向量，把不同的关键字音频映射到较远处的向量；最终，比较输入音频与样例词中映射向量之间的差异是确定二者是否一致的重点。李昭奇等[16]使用了如图 4 所示的 3 层的 BLSM 神经网络作为嵌入函数，并将 2 个方向网络的最后一帧输出向量拼接为嵌入式向量，并对特征提取的方式进行了改进，使模型的平均准确率(average accuracy，AP)达到了 86%。

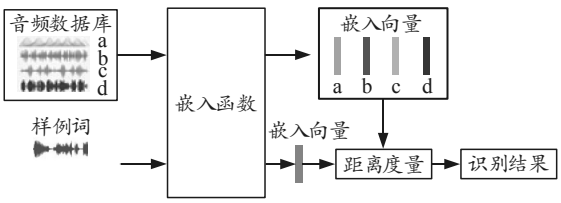

图 4　神经网络声学词嵌入样例关键词系统

3.3　基于大词汇量连续语音识别系统的关键词检测

从语言识别方法的角度出发，前面 2 个识别模型分别代表声学模型辨识的不同方式，这 2 个方法是从关键词识别的实际出发，即关键词不要求全部

辨别出内容，只要求辨别出语言或者仅仅是知道这句语言的所需语言即可，且这 2 个模式都几乎不要求语言模型的使用。而基于大词汇量连续语音识别(large vocabulary continuous speech recognition，LVCSR)系统的关键词检测则从另一个角度出发，即先通过语音识别将语音转换为文本，建立索引以供用户搜索。这种索引使得用户可以获取每个词的时间信息，并且由于本质上是个语音识别，也不需要对关键词进行单独训练，方便使用者进行识别任务。同时，LVCSR 系统在关键词检测上会遇到一些语音识别错误的情况，所以基于 LVCSR 的关键词检测还需要将识别出来的次优选结果包含，以提高其召回率。具体的方法有时间因子转换器(timed factor transducer，TFT)

3.4 基于端到端的关键词识别

随着神经网络在关键词识别领域中的广泛应用，关键词识别又有了一种新的解决思路，即抛弃传统的声学模型+语言模型的识别思路，整个系统包含特征提取、神经网络训练以及解码 3 部分。

神经网络的训练过程中，通过网络直接建立从声音到文字间的连接，端到端的训练主要有 2 种方法，其中一种是基于 CTC(connectionist temporal classification)模型[17]，可看作主动学习输入 X 与 Y 的对齐，由于 Y 的长度远小于 X，所以 CTC 引入空单元和 y_i 的重复来使 X 和 Y 彼此对应。另一种是基于注意力机制的编码-解码模型[18]。其主要思路：利用解码器(encoder)将原顺序转换成一条恒定宽度的隐层表达，而后解码器(decoder)再根据这些隐层表达产生解码顺序，而产生解码的步骤便是思考当前的输出和隐层表达中的哪一段最相关，这部分便是注意力机制，其流程如图 5 所示。

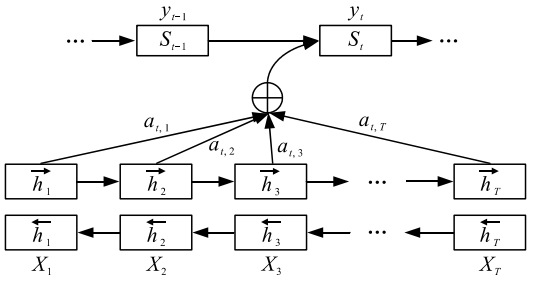

图 5 注意力机制

4 结束语

人机交互语音识别技术改变了现有单人单点串行设备操控形式，以远距离非接触、灵活、便捷的技术优势，高效支撑担任对复杂武器装备内的多设备或多装备的同时并行操控。在军事化应用发展过程中，还需关注在嘈杂的战场环境下语音提取、多人环境下的声纹识别等问题。随着语音识别技术的不断成熟演化，更高的召回率、更低的响应时间将使语音人机交互技术更广泛地应用于军事装备的操控中。

参考文献：

[1] DAVIS K H. Automatic Recognition of Spoken Digits[J]. The Journal of the Acoustical Society of America, 1952, 24(6): 669.

[2] OLSON H F, BELAR H. Phonetic Typewriter[J]. Journal of the Acoustical Society of America, 1956, 28(4): 767-767.

[3] MARTIN T B, NELSON A L, ZADELL H J. Speech recognition by feature abstraction techniques[R]. NTIs, 1964.

[4] BENESTY J, SONDHI M M, HUANG Y. Springer Handbook of Speech Processing[J]. The Journal of the Acoustical Society of America, 2008, 126(4): 653-680.

[5] Itakura F . Minimum prediction residual principle applied to speech recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1975, 23(1): 67-72.

[6] SAKOE H. Two-level DP-matching-A dynamic programming-based pattern matching algorithm for connected word recognition[J]. Acoustics, Speech and Signal Processing, IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(6): 588-595.

[7] MYERS C, RABINER L. A level building dynamic time warping algorithm for connected word recognition[J]. EEE Transactions on Acoustics, Speech, and Signal Processing, 1981, 29(2): 284-297.

[8] LEE C H, RABINERL R. A frame-synchronous network search algorithm for connected word recognition[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37(11): 1649-1658.

[9] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proc IEEE, 1989, 77(2): 257-286.

[10] HINTON G E, OSINDERO S, TEH Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[11] MOHAMED A R, DAHL G, HINTON G. Deep Belief Networks for phone recognition[J]. Proc Nips, 2009, 4(5): 1-9.

[12] DAHL G E, YU D, DENG L, et al. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 20(1): 30-42.

[13] Muda L, Begam M, Elamvazuthi I. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques[J]. Ttps, 2010, 2: 138-143.

[14] 陈太波, 张翠芳. 后验概率图与补白模型二次融合的

关键词识别[J]. 浙江大学学报(工学版), 2020, 54(6): 1170−1176.
[15] 李志涵. 基于样例模板的连续语音关键词检测技术研究[D]. 哈尔滨: 哈尔滨工程大学, 2021.
[16] 李昭奇, 黎塔. 基于wav2vec预训练的样例关键词识别[J]. 计算机科学, 2022, 49(1): 59−64.
[17] HANNUN A, CASE C, CASPER J, et al. Deep speech: Scaling up end-to-end speech recognition[J]. arXiv, 2014(12): 1412.
[18] CHOROWSKI J, BAHDANAU D, CHO, et al. End-to-end Continuous Speech Recognition using Attention-based Recurrent NN: First Results[J]. arXiv, 2014(12): 1412.

****

（上接第 16 页）

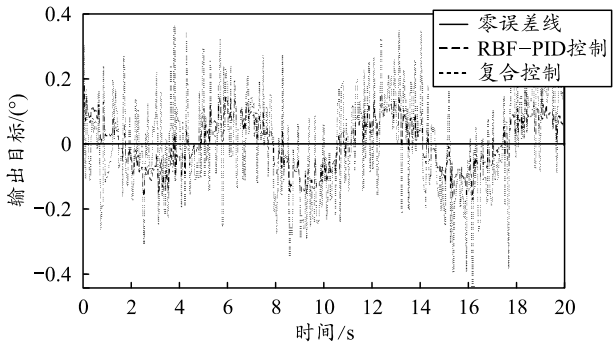

图 9　正弦跟踪误差曲线

4　结论

针对某舰载火箭炮交流伺服系统在工作过程中会受到海洋风浪环境干扰和自身扰动影响的特点，结合了滑模变结构控制、RBF 神经网络以及模糊控制的优点，提出了一种智能复合控制方法。仿真实验结果表明：该方法能够很好地解决舰载火箭炮交流伺服系统因受到内外部扰动而导致射击精度和稳定性降低的问题，提高了舰载火箭炮交流伺服系统的性能。

参考文献：

[1] 罗映红, 徐志奇, 刘丽媛, 等. 模糊 PID 控制在交流伺服电机系统中的仿真研究[J]. 煤矿机械, 2012, 33(9): 96−97.
[2] 侯润民, 刘荣忠, 侯远龙, 等. 自适应模糊小波滑模控制在交流伺服系统中的应用[J]. 兵工学报, 2014, 35(6): 769−775.
[3] 陶征勇, 童仲志, 侯远龙, 等. 基于RBF神经网络的破障武器内模 PID 控制[J]. 电气自动化, 2020, 42(5): 87−89, 95.
[4] 张世维, 林苹, 张大鹏, 等. 基于模糊 RBF 神经网络 PID 算法的无刷直流电机控制[J]. 电工技术, 2021(2): 16−18, 52.
[5] 汪圆萍, 张俊, 张兴. 永磁同步电机位置伺服系统的 RBF 神经网络滑模控制策略[J]. 湖北大学学报(自然科学版), 2021, 43(4): 429−436.
[6] 霍龙, 乐贵高, 胡健. 交流位置伺服系统反演滑模并行复合控制[J]. 机床与液压, 2012, 40(11): 85−87.
[7] 石文求. 基于 Matlab 的永磁同步电机滑模变结构控制系统仿真研究[J]. 机电信息, 2018(12): 134−135.
[8] WU H, WANG L, NIU P, et al. Global projective synchronization in finite time of nonidentical fractional-order neural networks based on sliding mode control strategy[J]. Neurocomputing, 2017, 235: 264−273.
[9] LIN L X, LIU Z, KAO Y G, et al. Observer-based adaptive sliding mode control of uncertain switched systems[J]. IET Control Theory and Applications, 2020, 14(3): 519−525.
[10] 徐晨峰, 陈强. 基于模糊控制的交流伺服系统仿真[J]. 自动化与仪器仪表, 2017(1): 107−109.
[11] 范佳慧, 李庆奎. 基于改进的模糊神经网络自适应控制[J]. 信息技术与信息化, 2021(8): 241−243.
[12] WANG Y Y, SHEN H, KARIMI H R, et al. Dissipativity-based fuzzy integral sliding mode control of continuous-time T-S fuzzy system[J]. IEEE Transactions on Fuzzy Systems, 2018, 26(3): 1164−1176.
[13] JI X W, HE X K, LV C, et al. A vehicle stability control strategy with adaptive neural network sliding mode theory based on system uncertainty approximation[J]. Vehicle System Dynamics, 2018, 56(6): 923−946.