Broadview®
www.broadview.com.cn

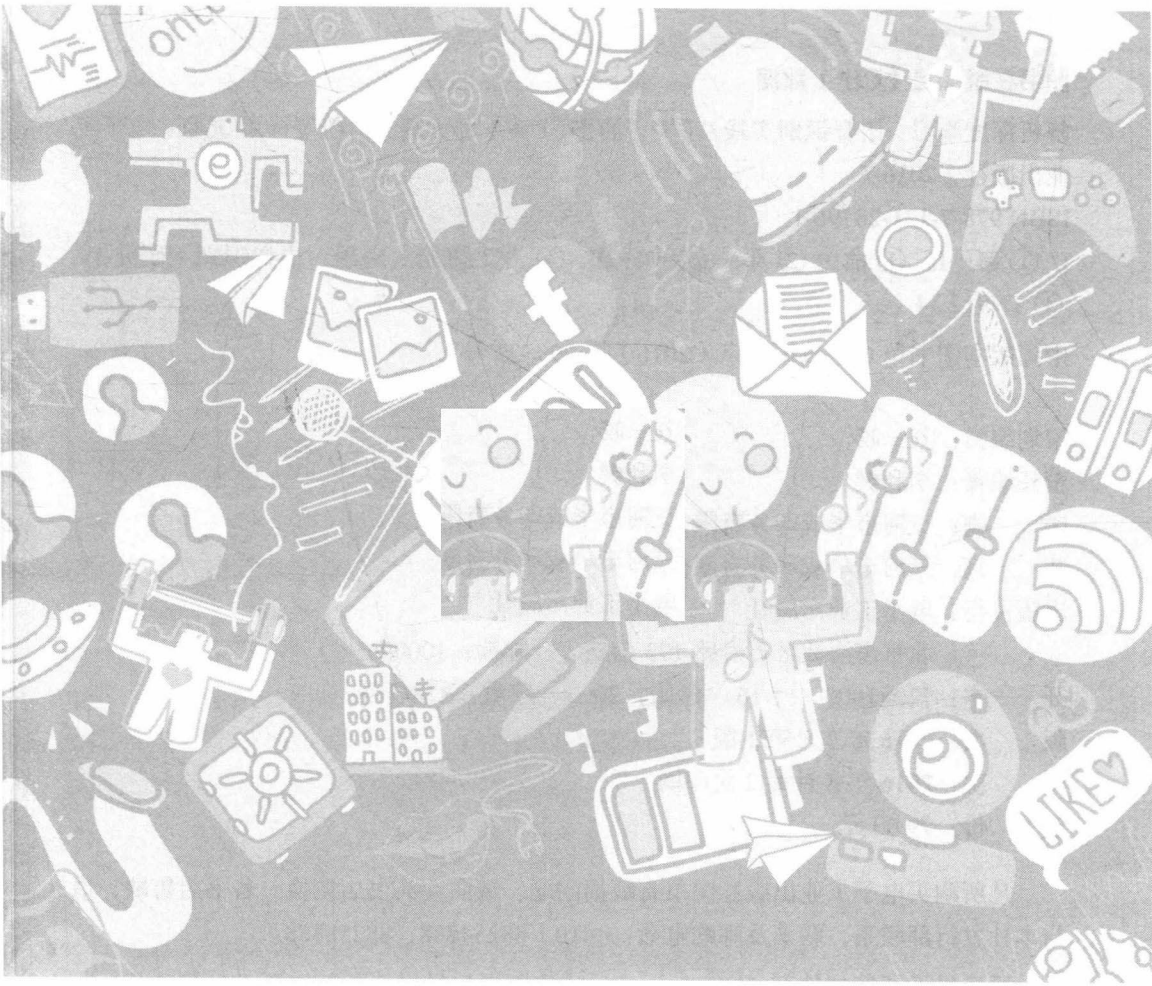
解析深度学习
语音识别实践

【美】俞栋 邓力 著
俞凯 钱彦旻 等译

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

解析深度学习
语音识别实践

【美】俞栋 邓力 著
俞凯 钱彦旻 等译



電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书是首部介绍语音识别中深度学习技术细节的专著。全书首先概要介绍了传统语音识别理论和经典的深度神经网络核心算法。接着全面而深入地介绍了深度学习在语音识别中的应用，包括“深度神经网络-隐马尔可夫混合模型”的训练和优化，特征表示学习、模型融合、自适应，以及以循环神经网络为代表的若干先进深度学习技术。

本书适合有一定机器学习或语音识别基础的学生、研究者或从业者阅读，所有的算法及技术细节都提供了详尽的参考文献，给出了深度学习在语音识别中应用的全景。

图书在版编目（CIP）数据

解析深度学习：语音识别实践 /（美）俞栋，（美）邓力著；俞凯等译.—北京：电子工业出版社，2016.7

ISBN 978-7-121-28796-1

I. ①解…II. ①俞… ②邓… ③俞…III. ①人工智能－应用－语音识别－研究 IV. ①TN912.34

中国版本图书馆 CIP 数据核字（2016）第 099823 号

策划编辑：刘　皎
责任编辑：李利健
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　印张：20　　字数：378 千字
版　　次：2016 年 7 月第 1 版
印　　次：2016 年 8 月第 2 次印刷
定　　价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）51260888-819　faq@phei.com.cn。

献给我的妻子和父母

——俞栋（Dong Yu）

献给 Lih-Yuan、Lloyd、Craig、Lyle、Arie 和 Axel

——邓力（Li Deng）

作者及译者简介

俞栋

1998 年加入微软公司，现任微软研究院首席研究员、浙江大学兼职教授和中科大客座教授。他是语音识别和深度学习方向的资深专家，出版了两本专著，发表了 150 多篇论文，是近 60 项专利的发明人及有广泛影响力的深度学习开源软件 CNTK 的发起人和主要作者之一。他在基于深度学习的语音识别技术上的工作带来了语音识别研究方向的转变，极大地推动了语音识别领域的发展，并获得 2013 年 IEEE 信号处理协会最佳论文奖。俞栋博士现担任 IEEE 语音语言处理专业委员会委员，曾担任 IEEE/ACM 音频、语音及语言处理汇刊、IEEE 信号处理杂志等期刊的编委。

邓力

世界著名人工智能、机器学习和语音语言信号处理专家，现任微软首席人工智能科学家和深度学习技术中心研究经理。他在美国威斯康星大学先后获硕士和博士学位，然后在加拿大滑铁卢大学任教获得终身正教授。其间，他还任麻省理工学院研究职位。1999 年加入微软研究院历任数职，并在 2014 年初创办深度学习技术中心，主持微软公司和研究院的人工智能和深度学习领域的技术创新。邓力博士的研究方向包括自动语音与说话者识别、口语识别与理解、语音–语音翻译、机器翻译、语言模式、统计方法与机器学习、听觉和其他生物信息处理、深层结构学习、类脑机器智能、图像语言多模态深度学习，商业大数据深度分析等。他在上述领域做出了重大贡献，是 ASA（美国声学学会）会士、IEEE（美国电气和电子工程师协会）会士和理事、ISCA（国际语音通信协会）会士，并凭借在深度学习与自动语音识别方向做出的杰出贡献荣获 2015

年度 IEEE 信号处理技术成就奖。同时，他也曾在顶级杂志和会议上发表过与上述领域相关的 300 余篇学术论文，出版过 5 部著作，发明及合作发明了超过 70 多项专利。邓力博士还担任过 IEEE 信号处理杂志和《音频、语音与语言处理学报》（*IEEE/ACM Transactions on Audio, Speech & Language Processing*）的主编。

❖ ❖ ❖

俞凯

IEEE 高级会员，上海交通大学计算机科学与工程系特别研究员。清华大学本科、硕士，英国剑桥大学工程系博士。长期从事智能语音及语言处理、人机交互、模式识别及机器学习的研究和产业化工作。他是中组部“千人计划”（青年项目）获得者，国家自然科学基金委优秀青年科学基金获得者，上海市“东方学者”特聘教授；作为共同创始人和首席科学家创立“苏州思必驰信息科技有限公司”。现任中国声学学会语音语言、听觉及音乐分会执委会委员，中国计算机学会人机交互专委会委员，中国语音产业联盟技术工作组副组长。他的研究兴趣涉及语音识别、语音合成、口语理解、对话系统、认知型人机交互等智能语音语言处理技术的多个核心技术领域，在本领域的一流国际期刊和会议上发表论文 80 余篇，申请专利 10 余项，取得了一系列研究、工程和产业化成果。在 InterSpeech 及 IEEE Spoken Language Processing 等国际会议上获得 3 篇国际会议优秀论文奖，获得国际语音通信联盟（ISCA）2013 年颁发的 2008—2012 Computer Speech and Language 最优论文奖。受邀担任 InterSpeech 2009 语音识别领域主席、EUSIPCO 2011/EUSIPCO 2014 语音处理领域主席、InterSpeech 2014 口语对话系统领域主席等。他负责搭建或参与搭建的大规模连续语音识别系统，曾获得美国国家标准局（NIST）和美国国防部内部评测冠军；作为核心技术人员，负责设计并实现的认知型统计对话系统原型，在 CMU 组织的 2010 年对话系统国际挑战赛上获得了可控测试的冠军。作为项目负责人或 Co-PI，他主持了欧盟第 7 框架 PARLANCE、国家自然科学基金委、上海市教委、经信委，以及美国通用公司、苏州思必驰信息科技有限公司的一系列科研及产业化项目。2014 年，因在智能语音技术产业化方面的贡献，获得中国人工智能学会颁发的“吴文俊人工智能科学技术奖”。

钱彦旻

上海交通大学计算机科学与工程系助理研究员，博士。分别在 2007 年 6 月和 2013 年 1 月于华中科技大学和清华大学获得工学学士和工学博士学位。2013 年 4 月起，任上海交通大学计算机科学与工程系理研究员。同时从 2015 年 1 月至 2015 年 12 月，在

英国剑桥大学工程系机器智能实验室语音组进行访问，作为项目研究员与语音识别领域的著名科学家 Phil Woodland 教授和 Mark Gales 教授开展合作研究。现为 IEEE、ISCA 会员，同时也是国际开源项目 Kaldi 语音识别工具包开发的项目组创始成员之一。此外，担任 IEEE Transactions on Audio, Speech, and Language Processing、Speech Communication、ICASSP、Interspeech、ASRU 等国际期刊和会议的审稿人。目前在国内外学术刊物和会议上发表学术论文 50 余篇，Google Scholar 总引用数近 1000 次。其中包括在语音识别领域权威国际会议 ICASSP、InterSpeech 和 ASRU 上发表论文 30 余篇，申请国家专利共 3 项，已授权 1 项。2008 年获科技奥运先进集体奖，2014 年获中国人工智能学会颁发的“吴文俊人工智能科学技术奖进步奖”。曾作为负责人和主要参与者参加了包括英国 EPSRC、国家自然科学基金、国家 863 等多个项目。目前的研究领域包括：语音识别、说话人和语种识别、自然语言理解、深度学习建模、多媒体信号处理等。

译者序

技术科学的进步历程往往是理论通过实践开辟道路的过程。尽管众多研究者将Geoffrey Hinton在2006年发表关于深度置信网络（Deep Belief Networks）的论文视为深度学习出现的重要标志，但那时，该技术还只是多层神经网络权值初始化的一种有效理论尝试，仅仅对一小部分机器学习专家产生着影响。真正让深度学习成为2013年《麻省理工学院技术评论》的十大突破性技术之首的，则是深度学习在应用领域的巨大实践成功。而语音识别正是深度学习取得显著成功的应用领域之一。

语音识别的发展自20世纪70年代采用隐马尔可夫模型（HMM）进行声学建模以来，每个时代都有经典的创新成果。如20世纪80年代的N元组语言模型，20世纪90年代的HMM状态绑定和自适应技术，21世纪第一个十年的GMM-HMM模型的序列鉴别性训练等。尽管这些技术都显著降低了语音识别的错误率，但它们都无法把语音识别推动到商业可用的级别。深度学习技术在21世纪第二个十年产生的最重大的影响，就是使得语音识别错误率在以往最好系统的基础上相对下降30%或更多，而这一下降恰恰突破了语音识别真正可用的临界点。该技术的突破伴随着并行计算基础设施的发展和移动互联网大数据的产生，其影响进一步交叠扩大，目前已经成为业界毫无争议的标准前沿技术。

本书作者俞栋博士和邓力博士正是这一突破的最早也是最主要的推动者和实践者。他们与Geoffrey Hinton合作，最早将深度学习引入语音识别并取得初步成功，后续又连续突破一系列技术瓶颈，在大尺度连续语音识别系统上取得了研究界和工业界广泛认可的突破。在几乎所有的语音识别应用深度学习的核心领域上都有这两位学者的影响。我与这两位学者相交多年，深刻地感觉到，他们在深度学习应用上的突破并非在恰当的时间接触到恰当的算法那么简单，而是来源于对语音识别技术发展历程的

不懈摸索。事实上，如作者们在本书中提到的，神经网络、层次化模型等思路在语音识别发展的历史上早已被提出并无数次验证，但都没有成功。回到深度学习成功前的十年，那时能够持续不断地在“非主流”的方向上尝试、改进、探索，是一件非常不易的事情。因此，我对二位学者一直怀有敬意。此次受他们之托，将展现深度学习在语音识别中实践历程的英文著作翻译成中文，也感到十分荣幸。

目前已有的语音识别书籍均以介绍经典技术为主，本书是首次以深度学习为主线，介绍语音识别应用的书籍，对读者了解前沿的语音识别技术以及语音识别的发展历程具有重要的参考价值。全书概要地介绍了语音识别的基本理论，主体部分则全面而详细地讲解了深度学习的各类应用技术细节，既包括理论细节，也包括工程实现细节，给出了深度学习在语音识别领域进行应用研究的全景。本书适合有一定机器学习或语音识别基础的学生、研究者或从业者阅读。由于篇幅限制，一些算法的介绍没有进行大幅展开，但所有的算法及技术细节都提供了详尽的参考文献，读者可以按图索骥。

本书的翻译是由我与钱彦旻博士共同完成的，同时，也得到了上海交通大学智能语音实验室的贺天行、毕梦霄、陈博、陈哲怀、邓威、金汶功、刘媛、谭天、童思博、项煦、游永彬、郑达、朱苏、庄毅萌的帮助，以及电子工业出版社的大力支持，在此一并表示感谢。翻译过程难免存在疏漏和错误，欢迎读者批评、指正。

俞　凯

序

本书首次专门讲述了如何将深度学习方法，特别是深度神经网络（DNN）技术应用于语音识别（ASR）领域。在过去的几年中，深度神经网络技术在语音识别领域的应用取得了前所未有的成功。这使得本书成为在深度神经网络技术的发展历程中一个重要的里程碑。作者继其前一本书 *Deep Learning: Methods and Applications* 之后，在语音识别技术和应用上进行了更深入钻研，得成此作。与上一本书不同，该作并没有对深度学习的各个应用领域都进行探讨，而是将重点放在了语音识别技术及其应用上，并就此进行了更深入、更专一的讨论。难能可贵的是，这本书提供了许多语音识别技术背景知识，以及深度神经网络的技术细节，比如严谨的数学描述和软件实现也都包含其中。这些对语音识别领域的专家和有一定基础的读者来说都将是极其珍贵的资料。

本书的独特之处还在于，它并没有局限于目前常应用于语音识别技术的深度神经网络上，还兼顾包含了深度学习中的生成模型，这种模型可以很自然地嵌入先验的领域知识和问题约束。作者在背景材料中充分证实了自 20 世纪 90 年代早期起，语音识别领域研究者提出的深度动态生成模型（dynamic generative models）的丰富性，同时又将其与最近快速发展的深度鉴别性模型在统一的框架下进行了比较。书中以循环神经网络和隐动态模型为例，对这两种截然不同的深度模型进行了全方位有见地的优劣比较。这为语音识别中的深度学习发展和其他信号及信息处理领域开启了一个新的激动人心的方向。该书还满怀历史情怀地对四代语音识别技术进行了分析。当然，以深度学习为主要内容的第四代技术是本书所详细阐述的，特别是 DNN 和深度生成模型的无缝结合，将使得知识扩展可以在一种最自然的方式下完成。

总的来说，该书可能成为语音识别领域工作者在第四代语音识别技术时代的重要参考书。全书不但巧妙地涵盖了一些基本概念，使你能够理解语音识别全貌，还对近两年兴盛起来的强大的深度学习方法进行了深入的细节介绍。读完本书，你将可以看清前沿的语音识别是如何构建在深度神经网络技术上的，可以满怀自信地去搭建识别能力达到甚至超越人类的语音识别系统。

Sadaoki Furui
芝加哥丰田技术研究所所长，东京理工学院教授

前言

以自然语言人机交互为主要目标的自动语音识别（ASR），在近几十年来一直是研究的热点。在 2000 年以前，有众多语音识别相关的核心技术涌现出来，例如：混合高斯模型（GMM）、隐马尔可夫模型（HMM）、梅尔倒谱系数（MFCC）及其差分、*n* 元词组语言模型（LM）、鉴别性训练以及多种自适应技术。这些技术极大地推进了 ASR 以及相关领域的发展。但是比较起来，在 2000 年到 2010 年间，虽然 GMM-HMM 序列鉴别性训练这种重要的技术被成功应用到实际系统中，但是在语音识别领域中无论是理论研究还是实际应用，进展都相对缓慢与平淡。

然而在过去的几年里，语音识别领域的研究热情又一次被点燃。由于移动设备对语音识别的需求与日俱增，并且众多新型语音应用，例如，语音搜索（VS）、短信听写（SMD）、虚拟语音助手（例如，苹果的 Siri、Google Now 以及微软的 Cortana）等在移动互联世界获得了成功，新一轮的研究热潮自然被带动起来。此外，由于计算能力的显著提升以及大数据的驱动，深度学习在大词汇连续语音识别下的成功应用也是同样重要的影响因素。比起此前最先进的识别技术——GMM-HMM 框架，深度学习在众多真实世界的大词汇连续语音识别任务中都使得识别的错误率降低了三分之一或更多，识别率也进入到真实用户可以接受的范围内。举例来说，绝大多数 SMD 系统的识别准确率都超过了 90%，甚至有些系统超过了 95%。

作为研究者，我们参与并见证了这许许多多令人兴奋的深度学习技术上的发展。考虑到近年来在学术领域与工业领域迸发的 ASR 研究热潮，我们认为是时候写一本书来总结语音识别领域的技术进展，尤其是近年来的最新进展。

最近 20 年，随着语音识别领域的不断发展，很多关于语音识别以及机器学习的优秀书籍相继问世，这里列举一部分：

- Deep Learning: Methods and Applications, by Li Deng and Dong Yu (June, 2014)
- Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods, by Joseph Keshet, Samy Bengio (Jan, 2009)
- Speech Recognition Over Digital Channels: Robustness and Standards, by Antonio Peinado and Jose Segura (Sept, 2006)
- Pattern Recognition in Speech and Language Processing, by Wu Chou and Biing-Hwang Juang (Feb, 2003)
- Speech Processing — A Dynamic and Optimization-Oriented Approach, by Li Deng and Doug O'Shaughnessy (June 2003)
- Spoken Language Processing: A Guide to Theory, Algorithm and System Development, by Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon (April 2001)
- Digital Speech Processing: Synthesis, and Recognition, Second Edition, by Sadaoki Furui (June, 2001)
- Speech Communications: Human and Machine, Second Edition, by Douglas O'Shaughnessy (June, 2000)
- Speech and Language Processing — An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, by Daniel Jurafsky and James Martin (April, 2000)
- Speech and Audio Signal Processing, by Ben Gold and Nelson Morgan (April, 2000)
- Statistical Methods for Speech Recognition, by Fred Jelinek (June, 1997)
- Fundamentals of Speech Recognition, by Lawrence Rabiner and Biing-Hwang Juang (April, 1993)
- Acoustical and Environmental Robustness in Automatic Speech Recognition, by Alex Acero (Nov, 1992)

然而，所有这些书或者是出版于 2009 年以前，也就是深度学习理论被提出之前，或者是像我们 2014 年出版的综述书籍，都没有特别关注深度学习技术在语音识别领域的应用。早期的书籍缺少 2010 年以后的深度学习新技术，而语音识别领域以及深度学习的研究者所需求的技术及数学细节更是没能涵盖其中。不同于以上书籍，本书除了涵盖必要的背景材料外，特别整理了近年来语音识别领域上深度学习以及鉴别性层次模型的相关研究。本书涵盖了一系列深度学习模型的理论基础及对其的理解，其中包括深度神经网络（DNN）、受限玻耳兹曼机（RBM）、降噪自动编码器、深度置信网络、循环神经网络（RNN）、长短时记忆（LSTM）RNN，以及各种将它们应用到

实际系统的技术，例如，DNN-HMM 混合系统、tandem 和瓶颈系统、多任务学习及迁移学习、序列鉴别性训练以及 DNN 自适应技术。本书更加细致地讨论了搭建真实世界实时语音识别系统时的注意事项、技巧、配置、深层模型的加速以及其他相关技术。为了更好地介绍基础背景，本书有两章讨论了 GMM 与 HMM 的相关内容。然而由于本书的主题是深度学习以及层次性建模，因而我们略过了 GMM-HMM 的技术细节。所以本书是上面罗列参考书籍的补充，而不是替代。我们相信本书将有益于语音处理及机器学习领域的在读研究生、研究者、实践者、工程师，以及科学家的学习研究工作。我们希望，本书在提供领域内相关技术的参考以外，能够激发更多新的想法与创新，进一步促进 ASR 的发展。

在本书的撰写过程中，Alex Acero、Geoffrey Zweig、Qiang Huo、Frank Seide、Jasha Droppo、Mike Seltzer 以及 Chin-Hui Lee 都提供了大量的支持与鼓励。同时，我们也要感谢 Springer 的编辑 Agata Oelschlaeger 和 Kiruthika Poomalai，他们的耐心和及时的帮助使得本书能够顺利出版。

俞　栋　　邓　力
美国华盛顿西雅图
2014 年 7 月

术语缩写

ADMM 乘子方向交替算法

AE-BN 瓶颈自动编码器

ALM 增广拉格朗日乘子

AM 声学模型

ANN 人工神经网络

ANN-HMM 人工神经网络–隐马尔可夫模型

ASGD 异步随机梯度下降

ASR 自动语音识别

BMMI 增强型最大互信息

BP 反向传播

BPTT 沿时反向传播

CD 对比散度

CD-DNN-HMM 上下文相关的深度神经网络–隐马尔可夫模型系统

CE 交叉熵

CHiME 多声源环境下的计算听觉

CN 计算型网络

CNN 卷积神经网络

CNTK 计算型神经网络工具包

CT 保守训练

DAG 有向无环图

DaT 设备感知训练

DBN 深度置信网络

DNN 深度神经网络

DNN-GMM-HMM 深度神经网络–混合高斯模型–隐马尔可夫模型

DNN-HMM 深度神经网络–隐马尔可夫模型

DP 动态规划

DPT 鉴别性预训练

EBW 扩展 Baum-Welch 算法

EM 期望最大化

F-smoothing 帧平滑

fDLR 特征空间鉴别性线性回归

fMLLR 特征空间最大似然线性回归

FSA 特征空间说话人自适应

GMM 混合高斯模型

GPGPU 通用图形处理单元

HDM 隐动态模型

HMM 隐马尔可夫模型

HTM 隐轨迹模型

IID 独立同分布

KL-HMM 基于 KL 散度的 HMM

KLD Kullback-Leibler 散度（KL 距离）

LBP 逐层的反向传播

LHN 线性隐含网络

LIN 线性输入网络

LM 语言模型

LON 线性输出网络

LSTM 长短时记忆单元

LVCSR 大词汇连续语音识别

LVSR 大词汇语音识别

MAP 最大后验

MBR 最小贝叶斯风险

MFCC 梅尔倒谱系数

MLP 多层感知器

MMI 最大互信息

MPE 最小音素错误

MSE 均方误差

MTL 多任务学习

NAT 噪声自适应训练

NaT 噪声感知训练

NCE 误差对比估计

NLL 负对数似然

oDLR 输出特征的鉴别性线性回归

PCA 主成分分析

PLP 感知线性预测

RBM 受限玻尔兹曼机

ReLU 整流线性单元

RKL 反向 KL 散度（KL 距离）

RNN 循环神经网络

ROVER 识别错误票选降低技术

RTF 实时率

SaT 说话人感知训练

SCARF 分段条件随机场

SGD 随机梯度下降

SHL-MDNN 共享隐层的多语言深度神经网络