

# 《中文信息学报》稿件排版格式

文章编号: 1003-0077 (2011) 00-0000-00

## 基于音系学模型的手语理解\*

姚登峰<sup>1,2</sup>, 江铭虎<sup>2</sup>, 阿布都克力木·阿布力孜<sup>2</sup>, 李晗静<sup>1</sup>, 哈里旦木·阿布都克里木<sup>3</sup>

(1.北京市信息服务工程重点实验室(北京联合大学), 北京 100101;

2.清华大学人文学院计算语言学实验室、心理学与认知科学研究中心, 北京 100084;

3.清华大学计算机科学与技术系智能技术与系统国家重点实验室, 北京 100084)

**摘要:** 我们试图模拟人脑处理手势信号的过程, 设计了一个混合的深层神经网络模型来解决基于音系学模型的手语理解问题, 即手语音韵信息到文本转换的问题。为此我们首先综合了手语语言学里同时性和序列性这两个观点的长处, 提出了一个手语音系学的改进模型。并针对难点设计了边感知边理解的算法, 直接从语言学的音韵特征推断汉语文本, 相比从视觉特征推断出汉语文本是一个很大的飞跃。实验验证了该认知计算技术的有效性, 为实现类人智能奠定了技术基础。

**关键词:** 音韵参数; 手语; 深度学习; 音系学模型

中图分类号: TP391

文献标识码: A

## Sign Language Understanding Based on Phonology Model

Dengfeng Yao<sup>1,2</sup>, Minghu Jiang<sup>2</sup>, Abudoukelimu·Abulizi<sup>2</sup>, Hanjing Li<sup>1</sup>,  
Halidanmu·Abudukelimu<sup>3</sup>

(1.Beijing Key Lab of Information Service Engineering, Beijing Union University, Beijing,  
100101, China;

2. Lab of Computational Linguistics, School of Humanities, Center for Psychology and Cognitive  
Science, Tsinghua University, Beijing, 100084, China;

3. State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for  
Information Science and Technology, Department of Computer Sci. and Tech., Tsinghua  
University, Beijing, China)

**Abstract:** we try to simulate the process of sign processing in the human brain, and design a hybrid neural network model to solve the sign language understanding based on phonological model, that is the problem of converting the phonological information of hand to Chinese text. We first integrate the advantages of the two perspectives of simultaneity and sequence in sign language, and propose an improved model of sign language phonology. The first-perception first-comprehension algorithm is designed for the difficulties, this algorithm is based on the cognitive mechanism of the brain, and it processes Chinese text directly from phonological features of the sign that can act as linguistic features. Compared with the traditional method that deduces Chinese text from graphic features, this algorithm represents tremendous progress in cognitive computing. Experimental results verify the feasibility of the intelligent cognitive technology, which lays a technical foundation to realize robot intelligence.

\*收稿日期: 定稿日期:

**基金项目:** 国家自然科学基金 (61433015; 91420202; 61602040); 国家社会科学基金 (14ZDB154); 教育部人文社会科学研究规划基金 (14YJC740104); 国家语委重点项目 (ZDI135-31); 北京市属高校高水平教师队伍建设支持计划高水平创新团队建设计划 (IDHT20170511); 北京市教委科技计划项目 (KM201711417006); 清华大学自主科研项目两岸清华大学专项 (20161080056)

**作者简介:** 姚登峰 (1979—), 男, 副教授, 主要研究领域为语言认知与计算、信息无障碍; 江铭虎 (1962—), 通讯作者, 男, 教授, 主要研究领域为语言认知与计算; 阿布都克力木·阿布力孜 (1983—), 男, 博士研究生, 主要研究领域为语言认知、认知神经科学。

**Key words:**Phonological Parameters; Sign Language; Deep Learning; Phonology Model

手语识别理解侧重于计算机图像处理,虽然大部分独立词的手语识别研究考虑手语表达的多模式特征,连续手语识别的研究考虑相邻手语词的上下文语义消歧,涉及手语语言学知识极为有限。正如语音识别需要语言计算理论一样,手语识别也需要手语计算的知识,是手势动作识别和手语含义识别理解的综合。长期以来手语识别理解缺乏手语语言学家的参与,手语识别理解并未取得较大的进展。本文深入研究了手语理解的问题,从语言学的角度,提出在音系学层面上进行手语理解,为此提出了基于音系学模型的手语理解算法。

## 1 基于语言学特征的手势理解综述

目前手势识别研究基本都参考了 Stokoe 理论,在 Stokoe 之前,手势被认为是不能分析的整体,而且没有内部结构。1960 年 Stokoe 提出手势(同时)由三个部分(参数)组成,包括手势位置(他称之为“tabula”或“tab”)、手形(“designator”或“dez”),以及运动(“signation”或“sig”)[1]。上世纪 70 年代初期语言学家给这三个参数增加了一个要素:方向,之后语言学者大都采用四个参数的说法。

Vogler 与 Metaxas 在 Stokoe 的双手手形和运动子单元模型的基础上提出了 PaHMMs [2]。此后手语识别文献使用 Stokoe 提出的语言学特征,主要是基于少量样本来学习新手势。

Lichtenauer 等人为未登录手势提出一种自动构建分类器的方法[3],具体是收集很多手势者的手势特征,与新手势的特征进行比较,从而为目标新手势构建分类器模型。这种方法依赖于很大的基础特征训练集(75 人打出的 120 个手势),并允许使用一次性学习来训练一个新手势分类器。Bowden 等人还提出使用一个训练样本就能正确分类新手势的手语识别系统[4],并给出了一个单一的训练例子。他们的方法使用了 2 级分类器,其中一级使用硬编码分类器来检测手形、布局、运动和位置。在应用动态分类手势前,二级使用了 ICA

(Independent Component Analysis)从一级得到的 34 维特征向量消除了噪声。在少量训练实例和缺乏语法知识的情况下获得了很好的效果。Kadir 等人扩展了前述工作,并基于 Boosting 级联弱分类器以躯干为中心的描述特征(将运动规格化为二维空间)来检测头部和手部,然后使用 2 级分类器,其中一级分类器生成语言学特征向量,二级分类器在马尔可夫链上使用 Viterbi 算法来获取最高识别概率[5]。Cooper 与 Bowden 延续了这项工作,使用了分类器集合来检测训练样本和分类样本的手势特征[6],将这些特征组合作为 2 级分类器的输入,使用了一阶马尔可夫假设。

这些文献表明基于 Stokoe 的语言学特征,其手语识别率等同于基于图形学特征的识别率。这显然有违于引进语言学特征的初衷,其根本原因在于 Stokoe 模型认为包括手形、运动、位置等在内的音系参数是同时出现的。也就是说手形、移动、位置、方向和非手动特征是同时表达出来的,这样一个手势就可以一下子表征出来。但是手势是存在序列性的,因为在有声语言中声音是线性连续的,所以手语也应与有声语言一样,是手势的序列性表达(sequential presentation)。音系结构通常通过形态变化来表现。如 Wendy Sandier 使用线性 CV(辅音元音)理论来证明序列性,因为手部一般在某个运动位置开始,进行运动,最后运动完成时停止在某个位置上,所以她认为美国手语只有两种基本音段,即运动和位置这两个音韵参数,分别对应有声语言里的元音和辅音,与前面说法有些类似[7][10]。同样还有 Baus 等人认为手语存在着与有声语言类似的“辅音-元音-辅音”(CVC)结构,具体就是“位置-运动-位置”结构,但是与语义有关的只有位置参数。至于方向、手形两个参数,虽然可能会对大脑语音加工有影响,但尚未发现其作用[8]。

我们认为手势可同时结合序列性和同时性,这就需要一个合适的音系学模型来表达手语的序列性和同时性。因为得到的手势音韵信息是一系列连续的音韵信息,如果不含分隔符号,

则切分音节时可能会造成歧义。大量跨手语语言的语料证明手语中存在着音节结构[9]。为此，我们先进行音韵参数感知加工，在此基础上提出改进的音系学模型，即增强运动-保持模型。

## 2 音韵参数感知加工

最近涌现出的 3D 体感设备为手势感知创造了可能，这种设备是利用 3D 体感摄像头来获取目标的图像信息，通常是目标的深度和红外图像信息，它借鉴了人眼的认知原理将传统的 2D 物体转换到 3D 空间。通过这种 3D 体感设备可以计算出位置、手形、方向、运动、非手动特征 5 个参数，这 5 个参数本身就是手语语言学里的音韵特征，相比较口语的语音识别，这些物理特征可用视觉直观感受到，感知阶段计算音韵特征的流程如图 1 所示，这是一个从粗粒度音韵特征到细粒度特征的计算过程，符合人脑对图像特征加工的认知机理。

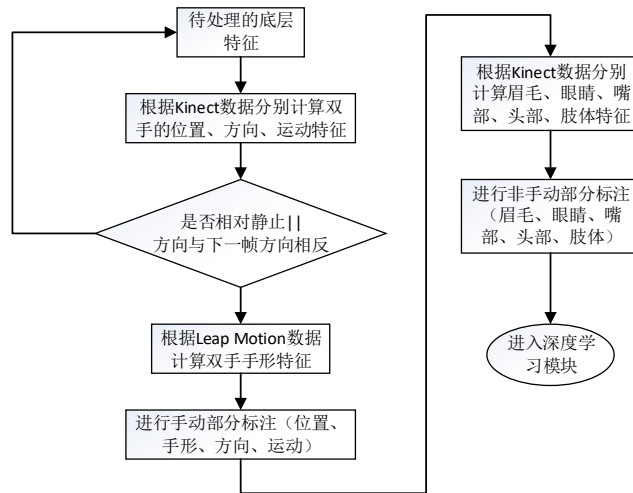


图 1 感知阶段音韵特征的计算流程图

手势的手形音韵参数，则由 Leap Motion 给出的每个手掌关节的 3D 坐标值计算得到，这里总共统计了 69 个手形。粗粒度的音韵参数如位置、方向、运动三个特征可由 Kinect 或 Leap Motion 计算得到，以 Kinect 为例，位置是根据 Kinect 给出的深度信息和 3D 坐标来返回两个参数，第一个参数是判断手部是否与身体接触，分别有接触、正常位置、远离身体三个值；第二个参数是判断手部在身体的位置，通常有头顶、太阳穴、耳部、头额等 28 个值。其中眉毛是根据 Kinect 给出的深度信息和 3D 坐标来判断眉毛是否上扬，分别有三个值，以此来判断手势者的情绪，由于眉毛变化通常伴随着眼睛、嘴部、头部和肢体的变化，为减少计算工作量，我们只将眉毛作为主要特征，其他作为次要特征。在判断主要特征眉毛上扬时，还需要验证次要特征，如眼睛是否变宽变大，头部和身体是否前倾，亦或肩部是否提起，在确认为疑问句后，再判断是否有双眼斜视（反问句没有双眼斜视）、头部是否摆动（一般疑问句没有头部摆动）等特征。

为了减少计算工作量，以 Leap Motion 为例，实际每秒只计算 3 个帧，即编号为 1、145、289 的帧，以帧为单位返回原始数据，以这些原始数据计算每一帧的返回参数为 (LA, LB, LC, LD, RA, RB, RC, RD, T)，其中具体参数值见图 2，这样手势“你”某一时刻返回的帧参数为 (0, 0, 0, 0, 1, 7, 1, 4, 64320)，即某一时刻该右手手势位于中性空间（不与胸部接触的空间），为数字 1 手形，正从后向前运动，手掌方向从左到右，该帧的时刻为 00: 01: 43: 20，其中 4 个 0 表示 Leap Motion 未检测到左手，返回左手的四个音韵参数均为 0。为了切分方便，M 音段的第一个音韵参数位置以中间帧的位置为准，第二个音韵参数手形以最后一帧的手形为准，第三个音韵参数运动可用当前帧 (F) 与上一帧 (F-1) 的差值表示当前

帧的变化量，具体以手掌中心坐标（Palm Position）的变化来表示，第四个音韵参数是手掌方向，需计算 M 音段最后帧的手掌由六个点组成的平面垂直向量（Palm Normal）来表示。

位置		手形		运动	
编号		编号		编号	
A1	中性空间	B1	拇指伸出,其余四指握拳	C1	手由后向前水平移动
A2	胸部	B2	手掌伸直,拇指弯曲贴在掌心,其余四指并齐	C2	手由前向后水平移动
A3	口部	B3	拇指弯曲,其余四指兵器弯向拇指成C形	C3	手前后移动
A4	头一侧或两侧空间	B4	手握拳,拇指搭在食指第二节上或者搭在中指第二节上。	C4	手左右水平移动
A5	额头	B5	食、中、无名、小三指伸直,分开不并紧,拇指和食指弯曲,拇指搭在食指上。	C5	手向身体一侧水平移动
A6	面前	B6	食、中二指伸直并紧,其余三指弯曲,拇指搭在无名指上。	C6	手一前一后排列
A7	脸颊	B7	食指伸直,其余四指握拳。	C7	手水平靠近另一手
A8	眼部	B8	食指伸出弯曲,其余四指握拳,拇指搭在中指上。	C8	手一前一后排列
A9	下巴	B9	食指伸直,中指伸直跟食指成直角,拇指跟中指交叉相搭,其余二指弯曲。	C9	手跨越另一手
A10	头部	B10	拇、食二指伸直分开,形成形,其余三指弯向掌心。	C10	手由一侧向身体正中做弧形移动
A11	肩部	B11	无名指、小指弯曲,拇指搭在无名指上,其余二指并齐,向下弯曲临空压在拇指上。	C11	手由一侧向另一侧做弧形移动
A12	头以上空间	B12	食、中、无名、小四指并齐弯曲,拇指跟食指、中指相抵成空拳。	C12	手拇指和食指套在另一手的拇指
A13	眼部	B13	拇指跟食指相抵成圆圈,其余三指伸直并紧。	C13	手指搭在另一手的手指上
A14	腰部	B14	拇指跟食指、中指相抵,其余二指弯向掌心。	C14	手由下向上做弧形移动
A15	耳部	B15	食、中、无名、小四指并齐弯曲,手指靠近手掌一节跟手掌成角度,拇指伸出。	C15	手由后向前做弧形移动
A16	颈部	B16	拇指跟中指、无名指相抵,成圆圈,食指和小指伸出。	C16	手由前向后做弧形移动
A17	牙齿	B17	手掌伸直,食、中、无名、小四指并齐。	C17	手一前一后向上移动
A18	腹部	B18	食指和中指伸直分开,成形,其余三指弯曲,拇指搭在无名指上。	C18	手一上一下不想接触
A19	太阳穴	B19	食、中、无名三指伸直分开,成形,其余二指弯曲相搭。	C19	手由上向斜下方移动
A20	上臂	B20	中指搭在食指上,成交叉形,其余三指弯向掌心,拇指搭在无名指上。	C20	手由下向斜上方移动
A21	小臂	B21	拇指和小指伸直,其余三指弯向掌心。	C21	手置于身体某部位
A22	腋下	B22	食指和小指伸直,其余三指弯曲,拇指搭在中指和无名指上。	C22	手敲打身体某部位
A23	肩下	B23	小指伸直,其余四指弯向掌心。	C23	手揪身上穿的衣服
A24	身体一侧或两侧	B24	拇指和食指张开稍曲不接触,其余三指弯向掌心。	C24	手抚摸身体某部位
A25	肘部	B25	五指张开稍曲。	C25	手在另一手上磨动
A26	腰部以下	B26	拇指伸出,食、中、无名和小指并齐与拇指呈)形。	C26	手自上而下移动
A27	上臂+肘部+小臂	B27	五指叠合在一起。	C27	手自上而下移动
A28	手	B28	拇指和小指伸直,其余三指并齐,五指并齐。	C28	手置于身体某部位

图 2 感知阶段音韵特征的具体值

### 3. 基于音韵参数的手势理解算法

这个阶段得到的标注文本就是手语的音韵信息，其任务是从这些音韵信息得到手语文本。根据这 5 个参数得到手势文本，本质上属于从拼音到汉字、从英语音标到英语单词的转换。因此从第一阶段自动标注好的音韵信息推断出手语文本类似于从连续的汉语拼音推断出汉语句子文本，或者从连续的英语音标推断出英语文本。注：这些连续汉语拼音或英语音标没有间隔符号，是连在一起的。这又是一个新的难题，但直接从这种语言学的音韵特征推断出汉语文本，至少要比直接从视觉手势特征推断汉语文本要前进了一大步。手语音韵信息到文本转换的流程如图 3 所示，这个过程与有声语言的传统标注技术相反，有声语言标注是先得到文本，再对文本进行拼音、词性等标注工作。

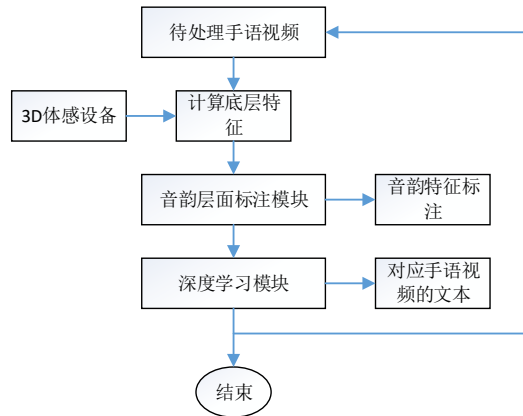


图 3 边感知边理解的算法流程图

#### 3.1 手势音韵信息——汉字文本转换的难点

有声语言的同音字现象很常见，以汉语为例，曾经有专家做了统计，若不考虑声调，新华字典里有 7536 个同音字，大约占 10%，平均每个读音就有 18.29 个汉字，如此高的比例给汉字消歧带来了很大的困难[11]。关于同音手势，目前未看到有关报道。本文对自建的中

国手语语料库做了统计，手语的同音手势占有所有手势的 46.86%，几乎占了一半。其中以词频统计，常用的同音手势占了 41.37%，不常见的同音手势占了 5.23%。因此手语与有声语言相比，同音情况更多、更复杂，其同音手势识别问题是音韵标注到文本转换问题的关键。

从手势音韵信息到汉语文本还存在特殊之处，文本采用的汉语是普通话的书写系统，具有表音和表意两方面的特点。而汉语拼音本身就是语音的书面符号，具有表音的作用，因此从汉语拼音得到汉字序列，虽然需要分析形、音、义等多方面信息，以及这些信息的综合判断，但毕竟存在着音的关系。而手势音韵信息是以手部为发音器官，从音系学角度得到的发音特征，它毕竟不同于国际音标或者西文的拼音文字，从手语音韵特征到口语的书面系统转换完全不存在任何关系。尤其是手语中分类词谓语句（手语独有的语言学现象，有声语言里没有）不是一个真正的手势，不受手语音系学里对称性和支配性等这些限制条件的影响[13]，这样包括分类词谓语句在内的很多手势其音韵结构较为复杂，没法用 1-2 个中文单词来标注。因此相对汉语口语，我们要完成的从手语音韵信息到汉语文本的转换任务是一个难度更大的消歧任务。把手语的音韵信息转化为口语文本，要弄清楚手语音韵学、句法学和形态学等独有的规律和特征。汉语是聋人的第二语言，因此与汉语语法相比，手语句子并不规范，其手语语法尚未规范化，再加上手语本身形态变化更为复杂，且每个手势复杂程度分布很不均匀，根据自建手语语料库统计，平均每个手势需要 2-6 个汉字来表述，但有些手势最多需要 62 个汉字才能表述其完整的意义。如前所述，在手语中同音手势的情况比有声语言更为常见，但是每个手势对应的候选词汇一般为 4-8 个，与汉语拼音基本相同。形态变化再加上候选词汇众多，将两者综合，使消歧难度比单纯的从拼音到汉字转换更大，因此手语本身的形态、语法是影响手语音韵信息到汉语文本的转换准确率的重要因素，也是实现从手势音韵信息到汉字文本转换的根本问题所在。

### 3.2 手语音系学的改进模型

大量跨手语语言的语料证明手语中存在着音节结构。为此，我们先引入 Liddell 和 Johnson 提出的运动-保持模型（Movement - Hold Model）[错误!未找到引用源。](#)，在此基础上进行改进。运动-保持模型认为打手势似乎是手在不停地运动，然而运动 (Movement) 和保持 (Hold) 两个音位音段 (phonological segments) 约各占一半时间，在语音上同等重要。将运动和保持分别表示为 M 和 H，这些音段都包含了完整的手部配置和语音特征（手形、运动、位置和方向）。

运动-保持模型基本提法是，手势由保持音段和运动音段构成，它们按序列生成。关于手形、位置、方向和非手动特征的信息通过每个单元一系列发音特征表现出来。这些特征类似于口语发声。保持音段定义为所有发音以稳定状态呈现的时间，而运动音段则定义为多个发音变换的时间，即一次至少一个参数发生变化，当然可能只有一个手形或位置参数的变化，但也有可能是手形和位置两个参数同时变化，这些变化就在运动音段内发生。比如手势“虫子”就只有手形变化，手势“到”只有位置变化，而手势“兴趣”在手势运动过程中既有手形变化，也有位置变化。手势变化在于主手的手掌方向，从主手掌心向下到主手掌心向上。这种说法与 Stokoe 认为手势参数同时生成的说法完全不同，但是与口语音段结构的说法一致。运动-保持模型解决了 Stokoe 模型的描述性问题。一些手势序列很重要，且要具有对比性，这套系统能够有效描述序列，还能提供足够多的手语描述细节，并清楚地描述和解释无数个发生在手语中的手势过程。

本文为了音节切分方便，对运动-保持模型（Movement - Hold Model）进行了改进，命名为增强运动-保持模型（Enhanced Movement - Hold Model），不考虑运动和保持两个音位音段约各占一半时间的划分，只考虑手势者在打出手势时，最开始手部是置于躯干两侧或者放平静止不动，手势的四个音韵参数（手形、运动、位置和方向）均没有变化，这时以 H 音段开始，接着手部到达手势初始位置时，为 M 音段+H 音段，即任何手势句子开始时，都是以

HMH 音段为开始，此后手势只要有一个参数发生变化，即可定义为 M 音段，运动停止时，即可定义为 H 音段。根据这种改进的增强运动-保持模型(Enhanced Movement - Hold Model)，在中国手语中单个手势的保持和运动组合可能有：HMH、MMH、HMHMMH、MHM、MHMMH、MMHH、HMHMHMHM 等 28 种组合。若不考虑书空（指在空气中书写汉字，是中国手语的一种构词方式）等复杂情况下，可能的组合可下降到 12 种。由此可见，并非所有手势都是保持-运动-保持（HMHMH）结构。但是，至少存在六种可能的手势结构，而 HM 并不包含在其中（见表 1）。在大多数情况下，HMHMH 型音节在中国手语中的分布非常广泛。并非所有的组合都能被语言结构所包含。不同音段结构反映了意义上的差异，如快（HMHMH）和速度（HMHMMMH）之间的音段结构差异。在手势“尊重”中，音段结构差异反映的是年龄差异或地域差异，即，上了年纪的手势者可能用 HMHMH 的变体来描述，而年轻人则用 HMHMHMHM 来描述。需要注意的是，手势的意义差异，或者手势者地域与年龄的差异，这些差异是手势组合方式的差异来源。运动-保持模型为描述这些差异给出了清晰而准确的思路。

表 1 中国手语中可能的音节结构组合

结构	中国手语
HMH	旅游、演讲
H	尝试、电脑 a
HMH	研究、哪儿
HMHMH	借 b、龙
HMMMMH	欢迎、翻译

注：a 表示做这些手势时手指要扭动，表示内部运动

B 运动-保持模型表示的“借”，描述成 HMHMHMH

我们在分析中国手语音段切分时，发现中国手语与汉语口语一样，连续手势序列并不是单个音节的简单组合，手语句子里每个手势组成部分以不同顺序组织，而且相互影响。因为受协同发音、韵律等因素的影响，手语也存在音变现象，从而导致连续手势序列与单独的手势音节有很大的不同。流畅手语是每秒 2-3 个手势，比较慢，每句最长不超过 12 个手势音节。此外手势者的表达手语的风格不同，并且手势者之间也存在音变的个体差异。本文以文献[13]为依据，目前已发现中国手语有四个音变，分别是运动增音（movement epenthesis）、保持缺失（hold deletion）、音位转换（metathesis）和同化（assimilation）。我们可以发现中国手语音节切分比汉语口语音节切分更为简单，至少某个手势过渡到下一个手势发生音变时，只是 MH 的前后序列发生了变化，并且这些变化仍能被体感设备感应到，没有超出运动-保持模型的范围。根据以上理论，我们可以根据图 4 得到以下例子：





没

图 4 手势句子“成长天安门没”

汉语句子：我自从出生以来，没去过天安门。

手语句子：成长（指从小到大）天-安-门没。

音韵特征序列：HMHMH M HMMMMH M HMH M HMH M HMH

实际数据序列：H(0, 0, 0, 0, 26, 25, 21, 5)M(0, 0, 0, 0, 26, 25, 5, 2)H(0, 0, 0, 0, 24, 25, 21, 2)M(0, 0, 0, 0, 24, 25, 27, 2)H(0, 0, 0, 0, 4, 25, 49, 2) M(0, 0, 0, 0, 4, 25, 19, 5) H(0, 0, 0, 0, 5, 7, 21, 5)M(0, 0, 0, 0, 5, 7, 11, 5)M(0, 0, 0, 0, 5, 7, 11, 5)M(0, 0, 0, 0, 5, 7, 11, 5)M(0, 0, 0, 0, 5, 7, 11, 5)H(0, 0, 0, 0, 5, 7, 21, 5) M(0, 0, 0, 0, 16, 49, 19, 2) H(0, 0, 0, 0, 2, 49, 21, 2)M(0, 0, 0, 0, 26, 49, 26, 2)H(0, 0, 0, 0, 18, 49, 21, 2) M(18, 49, 21, 3, 18, 49, 21, 3) H(4, 49, 21, 3, 4, 49, 21, 3)M(2, 49, 10, 3, 2, 49, 10, 3)H(2, 49, 21, 3, 2, 49, 21, 3) M(2, 12, 21, 6, 2, 12, 21, 5) H(2, 12, 21, 6, 2, 12, 21, 5)M(26, 25, 26, 6, 26, 25, 26, 5) H(26, 25, 21, 6, 26, 25, 21, 5)

### 3.3 手势理解与消歧

设  $S = \{s_1 s_2 \dots s_n\}$  为手势音韵特征序列，这样从音韵特征序列到汉语文本转换的任务就可以形式化描述如下：假设需要找到一个对应的汉语文本  $C = \{c_1 c_2 \dots c_n\}$ ，这样转换的目标就是使  $C$  在候选汉语文本中最符合手语语法和所在语境。其中  $c_n \in \text{char}\{s_j \dots s_{j+k}\}$ ，char 为音韵信息  $s_j \dots s_{j+k}$  对应的候选汉字序列之一。

消歧的重点在于手势音韵信息以及手语语法和语用信息的特征提取，我们考虑使用深层结构，深层结构具有较强的学习能力来获取特征，在图像识别等领域得到了广泛的应用。为了实现机器人的智能认知，也需要把原本人工完成的特征提取任务交给计算机来完成，通常这个工作需要由领域专家根据经验和运气来进行人工提取能够描述领域的最恰当的特征，但有时人工提取的特征并不一定能够代表领域特点，且需大量的耗时耗神。如能由计算机自动学习提取特征，那将大大减轻手工提取特征的工作量，也是未来实现智能认知的方向。

很多文献表明构造深层神经网络时需要注重模型的混合搭配，这样才能达到模拟大脑的工作机理，即在输入原始数据时，每处理一层，提取的概念特征更加抽象[14]。为此在构建底层网络时，我们采用了带噪声的自动编码器(Denoise Auto-encoder, DAE)，目的是通过非线性组合高维底层特征，以便学习得到低维抽象特征。我们希望利用自动编码器来提取未发现的特征，完成手势音韵信息和语法语用信息的特征提取。接着使用受限玻尔兹曼机(Restricted Boltzmann Machines, RBM)进一步自动学习更抽象、更有效的高层特征，最后采用BP算法更新整个网络的权重，对整个网络进行微调。最终用多元Logistic分类模型(Multi-class Logistic Classifier, MLC)进行分类，因为经过前两个网络后，数据已基本训练完成，没有必要再花更多的时间来训练。因此采用了更加简单的Logistic分类模型。该混合神经网络以每个待处理的手语音韵信息序列作为输入，依靠深度学习强大的无监督学习特征的能力[16-19]，在多层隐层中进行逐层训练学习，以获取音韵信息的抽象表示，最后将产生的特征通过分类器得到可能性最高的汉语文本，从而完成音韵信息到汉语文本的转换工作。因此我们提出的模型如图5所示。

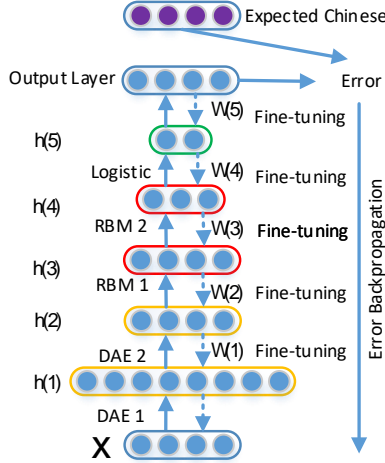


图 5 智能认知的基础——混合深层神经网络

只要设计的深层神经网络的结构合理，便可从感知阶段得到的音韵信息中提取出对汉语文本分类帮助较大的特征，从而解决手势音韵参数到汉语文本的转换问题。

如图 5 所示，输入层  $X$  输入感知阶段得到的手势音韵特征序列，将汉语文本作为这种混合深层神经网络的输出节点。基础模块第一层和第二层均采用了常规 DAE 模型[20]，其原型是自编码神经网络 (Auto-encoder, AE)，该网络通过一种尽可能复现输入信号的无监督学习算法完成网络的预训练。

由于感知阶段获取的手势音韵信息序列受 3D 体感设备的精确性所影响，体感设备硬件本身返回的参数值会有一些的几率出现错误，有可能多输入、少输入甚至错输入一些音韵信息，此外手势者在打手势的过程中可能会添加一些个性化的手势（非真正的中国手语），甚至手足舞蹈，使得手势输入随意性很高，对基于手势音韵信息的特征提取带来了更高的要求。为了增强 AE 模型的鲁棒性，更好地编码存在噪音的音韵信息，我们可以对输入的音韵信息  $x$  加入一定噪声  $q_D$ ，即用含有噪声的数据  $\bar{x}$  代替  $x$  作为输入，然后再将其输入到编码器中进行训练。因此恒等函数由  $g_{\theta'}(f_{\theta}(x)) \approx x$  变为  $g_{\theta'}(f_{\theta}(\bar{x})) \approx x$ ，学习得到最优参数  $\theta^*$  见公式(1)，这就是 Vincent 提出的 DAE 模型。与此 DAE 模型不同的是，本文采取的方法除了选取部分数据强制变为 0 以外，也随机挑选了一定比例的数据强制变为 1，目的是保证模型避免受到无关输入或者个性化音韵信息的影响，消除音韵信息的不规范性。

$$\theta^*, \theta'^* = \operatorname{argmin}_{\theta} \frac{1}{n} \sum_{n=1}^N L(x^i, g_{\theta'}(f_{\theta}(\bar{x}^i))) \quad (1)$$

在获取特征后，我们使用了深度信念网络(Deep Belief Network, DBN)实现去噪自编码器的输出数据的降维。这里采用了两层 RBM, RBM 是一个基于统计力学的概率图生成模型，可以自动从训练集里提取高阶抽象的特征，并且提供较好的神经网络初始权值，这样就可以把权值控制在对全局训练最有利的范围内，从而降低对学习目标过拟合的风险。预训练时，可采取对比散度算法变型进行近似求解，即通过最小化两个散度的不同来进行预训练[22]，以得到模型参数。具体是运行若干次 Gibbs 采样得到足够多的样本，然后通过计算平均值来求出归一化常数。由于 RBM 训练过程与维度无关，所以可以利用 RBM 对数据进行有效的投影，之后在最顶层加上逻辑斯蒂克模型作为分类器，对输出层的结果做排序，排序靠前的汉语文本即为该深层混合神经网络对手语音韵信息转换的预测结果。

最后是模型的微调，这里使用 BP 算法更新整个网络的权值，对混合神经网络的性能做调整。由于 BP 算法本身的缺陷，使用时需要注意训练过程要找到合理的初始值，否则这样

的网络将容易陷入局部极小值。

## 4 实验

### 4.1 实验准备

为了验证以上边感知边理解的算法，我们使用体感设备 Leap Motion 做了先行验证，本文硬件实验环境为：Intel 酷睿 i7-4770s@3.7GHZ 四核 CPU，16G 内存，Intel HD Graphics 4600 显卡。首先基于 Leap Motion，选择 Unity3D4.10 作为实现平台开发了一个中国手语手势音韵特征采集系统，可实现简单的手势识别功能，重点是采集手形、运动、位置、方向这四个参数，未采集非手动特征。

首先请三名聋生作为被试，以人民邮电出版社的《成语故事》（陈敏编）和人民文学出版社的《中国现代寓言故事》（安武林编）这两本书为故事来源，要求聋生阅读后，用自己的手语语言表达一遍，并要求尽量在水平方向打出手势，尽量减少握拳或垂直方向手势识别出现的误差。因此得到的龟兔赛跑、守株待兔、草木皆兵等故事视频，虽然其意思大致相同，但采集到的音韵特征序列大不相同。为了防止手指颤抖引起的误差，采用离散数据取样，即每隔一定时间做一次取样。由于 Leap Motion 没有视频录像功能，采集时同时采用普通摄像机进行拍摄，与 Leap Motion 进行同步拍摄，得到的视频和音韵特征序列将作为后期语料分析的基础，能够被 ELAN 软件阅读，参见图 6，其中第一行为经过深度网络输出后的文本，其他各行均为人工标注后的文本。其音韵信息未显示，与其他标注信息同存于 XML 文件。

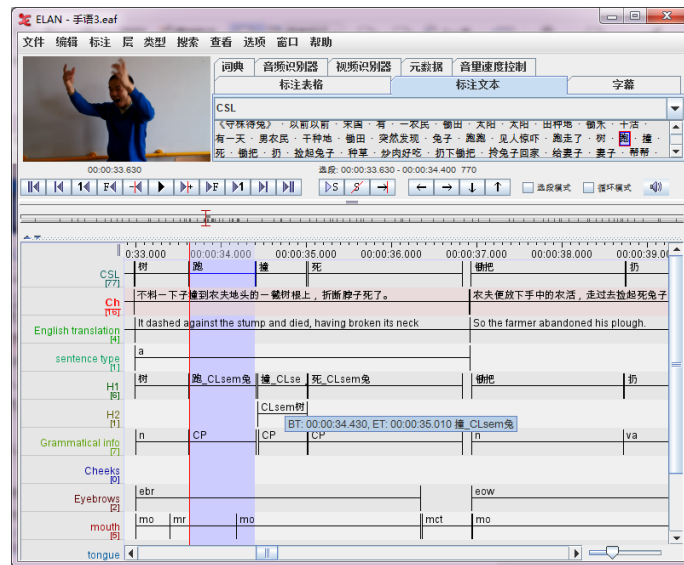


图 6 使用 ELAN 软件处理手语语料

拍摄视频时，对于每个手势者，要求打出手势的速度为 2-3 个手势/秒，尽量在 5 分钟之内表达完整的句子或段落意思。这样后期就可以切割成每段 5 分钟的视频文件，每两段视频文件之间没有交集。经 LEAP MOTION 同步采集后，经过整理可得到 400 个音韵特征标注语料。为了完成实验，我们将这些标注后的音韵特征文件分为 3 部分：300 个语料文件作为训练集，50 个语料文件作为验证集，剩下 50 个语料文件作为测试集。音韵特征覆盖常用的手势音节结构，从手语句子过渡到下一个句子的短暂停用“X”表示。由于对于视频而言，成语或寓言故事是相对独立的单元，因此我们在划分 3 个语料集合时保证不同集合的数据来自于不同的故事。

其中标注后的音韵特征文件为 txt 文本文件：

HUANGLEI\_01\_00003

H(0, 0, 0, 0, 26, 25, 21, 5)M(0, 0, 0, 0, 26, 25, 5, 2).....  
 HUANGLEI\_01\_00004  
 (0, 0, 0, 0, 18, 49, 21, 2)M(18, 49, 21, 3, 18, 49, 21, 3).....  
 HUANGLEI\_01\_00005  
 H(2, 12, 21, 6, 2, 12, 21, 5)M(26, 25, 26, 6, 26, 25, 26, 5) .....  
 .....

为了便于计算机处理，这些文件已处理成以下格式。

HUANGLEI\_01\_00003  
 (0, 0, 0, 0, 0, 26, 25, 21, 5)(1, 0, 0, 0, 0, 26, 25, 5, 2).....  
 HUANGLEI\_01\_00004  
 (0, 0, 0, 0, 0, 18, 49, 21, 2)(1, 18, 49, 21, 3, 18, 49, 21, 3).....  
 HUANGLEI\_01\_00005  
 (0, 2, 12, 21, 6, 2, 12, 21, 5)(1, 26, 25, 26, 6, 26, 25, 26, 5) .....  
 .....

训练用的汉语文本同样也为 txt 文本文件

HUANGLEI\_01\_00003           突然/兔/藏/窜 X 见/有/人/惊吓.....  
 HUANGLEI\_01\_00004           回/家/炒/吃/完/后 X.....  
 HUANGLEI\_01\_00005           全/人/听/农夫/笑/笑/他 X.....  
 .....

## 4.2 实验设置

预训练受限玻尔兹曼机时每次更新参数都运行一次 Gibbs 采样，每一层迭代次数为 30 次，微调时反向传播算法迭代的次数上限为 250 次，同时权值约束系数均为 0.1。DAE 层数设置为 2 层，且第二层设为 500，以便压缩数据，同时传送到第三层 RBM 网络，最后用逻辑斯特层输出手语句子的最大要求 12 个。对此，模型的参数定义如表 2。

表 2：模型参数列表

超参数	值	
微调学习率	0.1	
微调迭代次数	250	
DA 预训练学习率	0.002	
预训练迭代次数	30	
RBM 学习率	0.01	
RBM 随机比例	0.2	
L1 和 L2 正则化权重	0.1	
原始数据	672	
输出数据	12	
维数	第一层	2016
	第二层	500
	第三层	500
	第四层	300

## 4.3 实验结果

模型对比的结果如表 3 所示。

表 3：各模型正确率对比

	BP 神经网络	HMM	条件随机场	边感知边理解算法
手势者 A	58.37%	69.29%	69.74%	72.98%
手势者 B	61.39%	69.73%	70.12%	73.72%
手势者 C	60.78%	70.25%	71.46%	73.13%

从结果上看,与预想没有太大差异,因为BP神经网络对语言关联模式的处理能力不强,因此效果最差。HMM和条件随机场效果良好,说明这两个模型适合处理如不分段的音韵特征,比较适合用于自然语言处理。边感知边理解的算法由于用到了混合深度神经网络,综合了组成神经网络的优点,效果比这些模型要好些,因此把已有模型结合起来,可以博采每种模型的优点,尽力避免每种模型的缺点。HMM模型和条件随机场模型虽然也是常用的统计模型,但也存在局限性。HMM模型的特点是某一状态只能利用仅与其以前的状态有关的信息,后面的状态未充分利用,而条件随机场虽然克服了HMM模型的缺陷,但需要人为描述特征,并且训练是基于最大似然估计,未融入标注正确率的信息。因此使用单一的HMM和条件随机场很难进一步改善性能。而深度学习是模仿人脑认知,能够提取具有潜在复杂结构规则的手语视频等丰富结构数据的本质特征,这也是我们考虑引入深度神经网络作为助聋机器人智能认知基础的重要原因之一。

整体上,边感知边理解算法识别效果要比传统的统计模型要好,BP神经网络对于时序数据模式的处理能力较弱。同时与其他模型相比,条件随机场识别效果不明显,这说明提取的音韵特征对提升效果帮助不大,不是一个有效的特征。同时表明,随着特征维数的增大,HMM和条件随机场的识别正确率随之上升,说明对近距离的音韵参数处理得较好。图7分别显示了第一层隐节点数量为2016、2688、3360时的预测误差曲线,可以看到,随着节点数的增大,过拟合现象较严重。

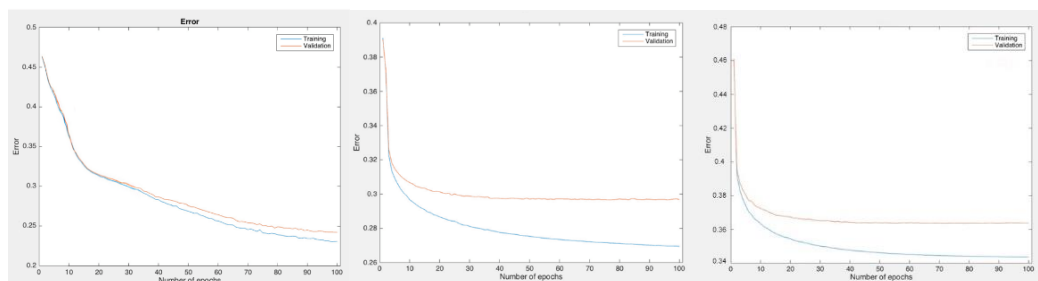


图7 误差预测曲线

由此表明隐层节点数量要适中,要尽可能的紧凑,即不能过少,达不到实验的精度要求,也不宜过多,否则会导致表达能力过强而产生过拟合现象,效果反而更糟。当然隐层节点数除了要考虑输入/输出层节点数,还需要考虑问题的复杂程度、样本的特征等因素。

#### 4.4 参数对实验结果的影响

训练好DAE后,用其各层之间的权值来初始化上一节中的深度网络,然后传到RBM时,再利用BP反向传播算法训练该深度网络。由此得到每个epoch对应的训练时间如图8所示。

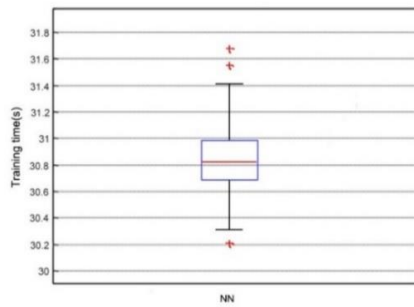


图 8 epoch 的训练时间

网络各层之间的参数对应的参数更新平均值随 epoch 变化趋势如图 9 所示，由此可以看出权值更新率也是个重要的参数，不能太小，否则会导致收敛过慢，也不能太大，否则会造成网络不稳定。

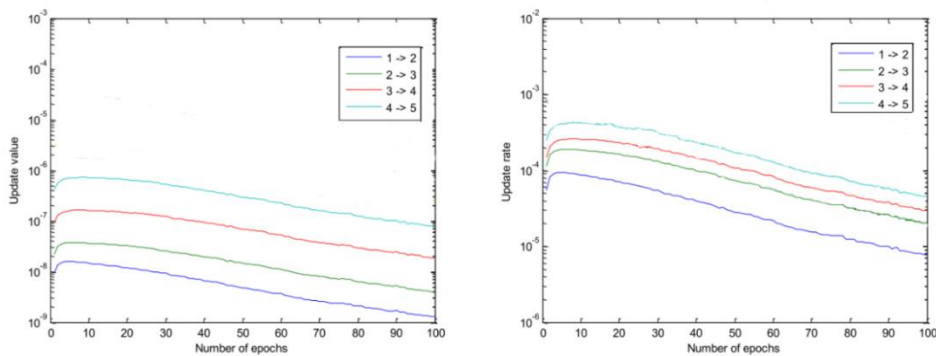


图 9 权值更新率变化曲线

由图可见，深度神经网络的梯度扩散问题有一定的改善。这说明混合神经网络采用去噪自编码器的无监督学习得到的各层参数作为深度网络的初始值，深度网络能够很好地迭代收敛，从而获得很好的分类效果。最终得到的分类效果如图 10 所示。

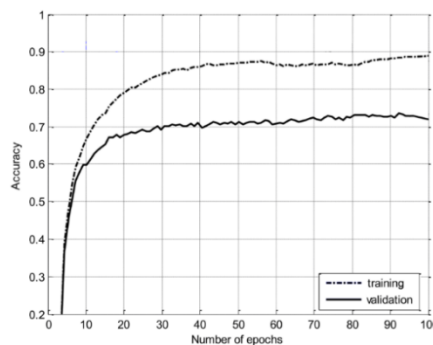


图 10 迭代次数对正确率的影响曲线

在训练和测试时，记录随着混合神经网络的训练时间的增加，正确率的变化规律。从图 10 可以看出，随着训练迭代次数的增长，训练样本和测试样本的识别率都会增长，最后趋于稳定，到达 80 次的时候，训练样本的正确率还会继续轻微的提高；由于训练过拟合的原

因,测试集正确率开始轻微的下降,这也是混合神经网络,包括去噪自编码器模型的一个特点,不是训练时间越长,测试结果就越好。

根据以上分析,我们认为这种边感知边理解的算法还有改进的空间,虽然效果高于其他对比的模型,但可能是因为手语相比口语,语法相对简单,很少见到长难句。手语与口语不同,汉语拼音转化为汉字时,存在着很多同音字,导致消歧困难,需要利用语境和各类知识。而手语也存在着同音现象,但手语可通过非手动特征来帮助消歧,并不需要过多的利用语境和世界知识。我们推测,手形、运动、位置这些参数在识别同音手势时并不起主要作用,而起决定作用的是非手动特征,它也恰恰是最复杂的特征。因为非手动特征包括面部表情和肢体动作,而面部表情又包括眉毛、嘴巴、脸颊等表情,当然引入非手动特征后,在这种情况下,中国手语理论上可组成无限个音韵特征,计算代价庞大,下一步需要解决这个问题。此外手势的复杂程度分布不均匀,若能将音韵信息的频率引入到监督信息中,可改进模型的识别准确率,因此改进神经网络的损失函数,使每一维的监督信息不再平等区分,这也是要进一步解决的问题。

## 5 小结

本文提出的先标注后文本的思路,是直接从语言学的角度得出汉语文本,因此无论是静态手势,还是动态手势,识别率都明显比基于计算机视觉的识别率要高。我们提出的技术验证了手语智能认知的可行性与准确性,下一步目标是提高识别速度,缩短响应时间。今后将继续进行手势识别训练的样本获取工作,扩充手语语料库的样本数,并将面部表情等非手动特征与手部动作结合起来,从而进行多感知机的语言认知计算领域深层次问题的研究。

## 参考文献

- [1] Stokoe, W.C.: Sign language structure: An outline of the visual communication systems of the american deaf. *Studies in Linguistics: Occasional Papers* 1960, 8: 3-37
- [2] Vogler, C., Metaxas, D.: Parallel hidden markov models for American Sign Language recognition// *Procs. of ICCV, Corfu, Greece* .1999,1:116-122.
- [3] Lichtenauer J, Hendriks E, Reinders M. Learning to recognize a sign from a single example//2008 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG' 08) . IEEE, 2008: 1-6.
- [4] Bowden, R., Windridge, D., Kadir, T., Zisserman, A., Brady, M.: A linguistic feature vector for the visual interpretation of sign language//*Computer Vision-ECCV 2004*. Springer Berlin Heidelberg, 2004: 390-401.
- [5] Kadir T, Bowden R, Ong E J, et al. Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition//*BMVC*. 2004: 1-10.
- [6] Cooper H, Bowden R. Large lexicon detection of sign language//*Human-Computer Interaction*. Springer Berlin Heidelberg, 2007: 88-97.
- [7] Sandler W. Phonological representation of the sign: Linearity and nonlinearity in American Sign Language. Walter de Gruyter, 1989.
- [8] Baus, C., Gutiérrez-Sigut, E., Quer, J., &Carreiras, M. Lexical access in Catalan signed language (LSC) production. *Cognition*, 2008, 108(3):856-865.
- [9] Jantunen, T., &Takkinen, R. Syllable structure in sign language phonology. na. In Brentari(ed.).*Sign Languages*.Cambridge:Cambridge University Press,2010.312-331
- [10] Perlmutter D M. Sonority and syllable structure in American Sign Language. *Linguistic inquiry*, 1992, 23(3): 407-442.
- [11] 吴军,王作英.一种基于语言理解的输入方法——智能拼音输入方法. *中文信息学报*, 1996, 10(2): 56-61.
- [12] Liddell, S. K., & Johnson, R. E. American Sign Language compound formation processes, lexicalization, and

phonological remnants. *Natural Language & Linguistic Theory*, 1986, 4(4): 445-513.

- [13] Valli, C., & Lucas, C. *Linguistics of American Sign Language: An introduction*. Gallaudet University Press. 2000.
- [14] Battison R. *Lexical Borrowing in American Sign Language*, Linstok Press, Silver Spring, MD. 1978.
- [15] Bengio, Y. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2009, 2(1): 1-127.
- [16] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 2007, 19: 153.
- [17] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., &Manzagol, P. A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 2010, 11: 3371-3408.
- [18] Bengio Y, Yao L, Alain G, et al. Generalized denoising auto-encoders as generative models//*Advances in Neural Information Processing Systems*. 2013: 899-907.
- [19] Salakhutdinov R, Hinton G. Semantic hashing. *International Journal of Approximate Reasoning*, 2009, 50(7): 969-978.
- [20] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders//*Proceedings of the 25th international conference on Machine learning*. ACM, 2008: 1096-1103.
- [21] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, 313(5786): 504-507.
- [22] Tieleman T. Training restricted Boltzmann machines using approximations to the likelihood gradient//*Proceedings of the 25th international conference on Machine learning*. ACM, 2008: 1064-1071.

通讯作者联系方式：江铭虎 通讯地址：北京市海淀区清华大学文北楼 306 室 清华大学人文学院计算语言学实验室（邮编 100084）。手机号 13520115507，邮箱 [jiang.mh@tsinghua.edu.cn](mailto:jiang.mh@tsinghua.edu.cn); [yaodengfeng@gmail.com](mailto:yaodengfeng@gmail.com)