以语音识别技术逆向分析汉语远场群体讨论中非母语者的交互策略

(Using Automatic Speech Recognition Technology to Reverse Analyze Communication Strategies between Non-Native Speakers in a Chinese Long Distance Group Discussion)

砂冈和子
(Sunaoka, Kazuko)
早稻田大学
(Waseda University)
ksunaoka@gmail.com

摘要：近几年，语音识别技术（Automatic Speech Recognition: ASR）的精度大幅提升，已突破了从技术走向实用的门槛。本文对非汉语母语者（NNS）与母语者（NS）的远场群体讨论语料用 ASR 技术进行识别精度。逆向验证了最新 ASR 对单一发言人、母语者、标准口语的识别精度非常高，已达到现场应用的水平。但不管对 NS 还是 NNS，对含有情感及多声源等干扰等语音，ASR 识别率都出现大幅度下降。因此很难应用到识别具有远场(far-field）,多通道(multi-channel）,多模态（multi-modal）特征的语音。相比之下，参加群体讨论会的 NNS 存在汉语口音不够标准，语句碎片化等问题，但充分利用多个信息通道、多种沟通模态来与 NS 进行互动共享信息。最后简介未来 ASR 技术的趋势，同时显示几种“ASR＋汉语教学”模式，从而探讨如何更好地与智能语言工具互补共存。

Abstract: In recent years, the accuracy of Automatic Speech Recognition (ASR) has greatly improved in terms of its practical applications. An ASR test on a real corpus of a multi-person long-distance group discussion between non-native speakers (NNS) and native speakers (NS) of Chinese was used to compare the accuracy of ASR with NNS speakers. It was found that the latest ASR has very high recognition accuracy for single speakers, native speakers, and the standard spoken language. ASR has now reached a new level in field applications. However, for both NS and NNS, the recognition rate of ASR significantly decreased while capturing emotional and multi-channel speech. Therefore, it is difficult to apply to far-field, multi-channel, or multi-modal speech. In contrast, NNS participants who made full use of multiple information channels and modalities were able to successfully communicate and interact, although their Chinese pronunciation was not standard and included fragmented statements. This paper also discusses trends in future ASR technology and introduces several

"ASR + Chinese teaching" methods to explore how they may better coexist with smart language tools.

关键词：语音识别技术，远场群体讨论，多通道，多模态，互动信息共享

Key words: Automatic Speech Recognition, Long Distance Group Discussion, Multi-channel, Multi-modal, Interaction by information sharing

1. 研究动机和目的

近几年，基于大规模数据库的深度神经网络（Deep Neural Networks，DNN）等声学模型在 ASR 研究上均获得了巨大成功(Graves et al., 2016；刘洋，2017)。谷歌、微软、苹果、亚马逊等全球智能语言技术大巨头研发的 ASR 识别率声称已堪比人类听力水平[1]。百度、科大讯飞、搜狗等中国公司对汉语普通话的语音识别率都已达到 97%，识别速度为每分钟 400 字之快(陈鹏，2017;戴礼荣等，2017）。ASR 是人机交互的基础，也是推动机器翻译，自然语言理解等技术发展的前提条件。语音输入法、语音助手、车载语音交互系统等 ASR 智能语言技术已渗入到人们生活的方方面面，并开始应用于会议演讲等实时互译的技术上(中村等人，2017；河原等人，2018）。日本政府研究机构开发的同声传译 Voice Tra，其日英文口译与笔译水平分别与 TOEIC 800 分到 900 相当[2]。

ASR 与机器翻译（MT）的发展既能低成本、高效率地消除不同语言之间的沟通障碍，又能为视听障碍者带来极大的方便。与此同时，自动翻译的普及给外语教学产业带来了强烈的冲击[3]。AI 语言技术会不会改变外语学习的方法？如有正面效果，它对第二语言学得有如何影响？实际上一些公司宣传的性能评价过高, 目前市面上的 ASR 错误率超过 15%甚至 30%[4]。它的难题主要在远场的精确识别以及对于口音、多人语音、多语种、大词汇等场景数据的获取上。当前的人机对话，预先定义好的特定域(domain)内才能实行，可扩展性仍然存在很大的问题（中村等人，2017; 篠田，2017）。

二语教学致力于在模拟真实的语言环境中，注重于物理、生理、心理方面的互动与情感，培养学生的语用能力和沟通能力(Clifford et al., 2013; 罗华珍等人，2017; 徐琦璐,2017;)。ASR 可作为一个很强的听写机，既可以增加自己可操作的教材资源，还可以辅助汉语发音练习。而 ASR 技术还不成熟，直接应用于会话学习或课堂教

[1] 《电子发烧友网》, http://www.elecfans.com/video/yinpinjishu/20161019441265.html [2016-10-19].
[2] VoiceTra. 7.0 (iOS 版). http://voicetra.nict.go.jp/ [2018-10-3].
[3] 《科技视界》, 2016 年 20 期. http://bianke.cnki.net/web/article/F085_1/KJSJ201620081.html.
《人工智能》, http://www.stdaily.com/rgzn/tuijianq/2017-09/04/content_574332.shtml [2017-09-04].
[4] 《瞭望》, https://www.iyiou.com/p/78283 [2018-08-04]

学活动弊大于利。如何设计“AI＋汉语教学”模式是我们面临的重要课题。本文基于这些研究动机和目的，使用远场群体讨论的语料，对此应用 ASR 技术的测验，从而了解 AI 与人类对语义理解之优劣之处。

以后分四个部分进行论述：首先将介绍本次研究的语料与分析框架;其次将对比 ASR 对 NS 与 NNS 两种语音的识别结果;接着浅析非母语者参加外语群体讨论时的交互策略与当前和未来 ASR 技术的趋势；最后提示几种“ASR＋汉语教学”模式作为本文的小结。

2. 分析方法与语料特征

2.1 分析框架

本文使用远场群体讨论课（后文称“讨论课”）的课堂录影材料，以信息共享互动理论（Theory of interaction by information sharing）作为分析框架(安西，2017)，用民俗学 (Ethnomethodology)的话语分析方法，对“讨论课”学生的交互行为进行观察。由于“讨论课”的语料具有多通道、多模态特征的自然口语。所以我们对收集的语料以文字和符号，尽可能地详细转写。一方面试用 ASR 对此小片语料进行听写测验转写成文字，从而比较 ASR 的精度与 NNS 学生交互策略的差异（详见第 4 节）。

信息共享互动理论基于认知神经科学和信息处理最新研究的成果，将构建一种包含所有生物与物体的交互机制与共享信息的模型（后述）。本文将 NNS 和 NS 之间交互方式用以民俗学话语分析方法进行细致描述（Goffman,1981；Sacks, Schegloff, & Jefferson ,1974; Schegloff, 1996），再用信息共享互动理论的交互概念给它试加标记，从而获取群体讨论成员交互过程的认知功能与它的机制。这样既有利于 NNS 所发挥的多种认知功能与互动策略，还可以互补印证与目前 ASR 功能的差异，帮助二语教学对智能语言工具更好地理解进行和取舍。

2.2 术语的定义

由于本文中所使用术语的含义与该术语在它领域的技术类定义有些许出入，特将这些术语在本文中的意思解释如下：

- 信息共享互动理论（Theory of interaction by information sharing）
- 通过认知神经科学和信息处理科学相结合，统一解释物体交互的机制，从而构建一种跨媒体的信息处理框架(Anzai, 1992; Anzai, 2013；安西, 2017; 潘煜等人, 2018；Tomasello et al., 2005)。该理论不仅对人人交互，还对人机交互的认知过程都有较强的理论和应用价值。该理论认为交互者的内部由叫做

GRAMES 的系统与机制构成[5]，并相互衔接和连贯进行信息处理。如目标导向机制（G：Goal-directed mechanism），奖励系统（R：Reward systems），注意力系统（A：Attention systems），动机机制（M：Motivational mechanisms），情绪机制（E：mechanisms for Emotion），社会信息处理机制（S：Social information processing mechanisms）等。本文将GRAMES的概念扩充解释为讨论者所运用的谈话策略**表 4**，初步探索对群体讨论语料作标记。

- 远场（far-field）：指的是发言人距离麦克风较远。本文用的是含有发言人之间的空间距离与网速等干扰因素的广义“远场”(Clark & Brennan， 1991)，以便验证 ASR 技术对远程群体讨论录音语料的识别精度。
- 多模态（multi-modal）：除了用以语音传递信息外，还通过感知觉、意识、记忆、注意、联想、动机和情绪等非语言模式进行交互。
- 多通道/多道（multi-channel）：在多数交互对象之间，通过动作、手势、表情、视线、姿势等进行交流。相对于单一发音人、单通道/单道（single-channel）的语料，对 ASR 难度较高。
- 智能语言工具：包括 ASR 与 MT 以及对话系统等使用 AI 技术开发的语音产品。

2.3 语料特征

笔者自 2011 年以来，通过互联网在北京、台北、东京、横滨、北九州五地之间实施远程汉语讨论课(砂冈,2016)。这堂课基于早稻田大学 Cross-Cultural Distance Learning（CCDL）的建构主义教学理念，针对 CEFR 进阶 B 级以上汉语水平的非母语学生，以培养 Intercultural Competence 能力为学习目标，通过让国内外学生同步互动的平台之上，展开非导向性对话（Non task oriented dialog）。一般每轮对话由一名主持人与一名指定发言人进行一对一问答，但通常由非指定听者参与讨论，有时旁听者之间以视线或姿势进行交互(上同)。

本文使用的是于 2016 年 5 月 19 日实施的一次主题为“介绍自己养宠物经历”（后文称“养宠物”）的对话片段。此次 4 校共 13 名学生参加，具体属性如**表 1** 所示。尽管言语能力占优势的 NS（8 名）和 NS-EC（2 名）掌握了话语领导地位，3 名 NNS 也积极参与讨论，在整个讨论时间 73 分钟之中，很少出现讨论中断等尴尬状况，如**表 6** 自 2B-1 至 2B-6 这一段话总共花 18 分钟，其中发言时间占 90%，而话轮空白（静场）只占 2%。

表 2 所示（A）朗读，（B）演讲等距麦克风单独、近讲、无噪音、中立情绪、单模态等语音类易于 ASR 识别。与此相比（C）现场直播，（D）自由对话等多通道、噪音大、带情绪的语音，为 ASR 带来难题。“讨论课”的语料属于（E）群体讨论语料，既有多人数（共有 13 名 Speaker）、超远场、强噪声等特点外，还含

[5] 基于当前脑神经科学研究的水平，与其他神经基础的结构相对清楚分离的功能称为“系统”，而还未分离的功能叫“机制”。由每项机制的头字母命名为 GRAMES（安西,2017）。

有不同汉语口音、以及二语学习者的非正规语音干扰。因为自然口语的讨论，所以还有大量副语言（如韵律、语气词）、非语音信息（如停顿、笑容）等因素（砂冈，2016；Sunaoka,K.2018 a）。第 4 节**图 3** 所示，目前 ASR 还未支持多模态、多通道信息的处理与理解。

表 1 参加 2016/5/19 讨论课学生的属性

单位	NS（MorF）	NS-EC	NNS	Total
WT 校	2（M1/F1）	1（F）	2（M1/F1）	5（M2/F3）
WB 校	4（M2/F2）			4（M2/F2）
B 校		1（M）	1（M）	2（M）
T 校	2（M2）			2（M）
Total	8（M5/F3）	2（M1/F1）	3（M2/F1）	13（M8/F5）

说明：WT：早稻田大学东京校园；WB：早稻田大学北九州校园；B：北京大学； T：台湾师范大学；M:Man；F：Female; NS:以汉语为母语者； NS：非汉语母语者； NS-EC： Ethnic Chinese 华裔学生
在论文引用时使用这些标号。

表 2 ASR 对语料难度类型(Feature comparison of speech types)

	语音类型	发言人数	通道	话速 mora/s word/min	距 话筒	噪音 水平	情绪强弱	模态	识别难度（%）
A	朗读	单人	单一	7.26(189)	近	安静	平坦	单	易
B	演讲	单人	单一	7.31(191)	近	较安静	较平坦	单/多	较易 約 90%
C	现场直播	单/多人	多道	8.51(222)	近	嘈杂	大起大落	单/多	较难 約 60-80%
D	自由对话	多人	多道	每人不一	较远	嘈杂	每人不一	多	较难
E	群体讨论	多人	多道	每人不一	远	嘈杂	每人不一	多模态	极难

（改编自 有木康雄，2003）

说明

1）Mora 是日语发音的基本单位,以“元音”或“辅音+元音”构成，用平假名表示一个音拍。日语词汇一般为汉字和假名，有时用片假名或罗马字混合书写。日文没有严格的正书法，

因此常用词汇平均由几个字构成就说法不一。话速(word/min)是笔者将原英文的词数换成日文字数（約 2.3 倍）后计算。

2）识别难度据（河原等，2018)。

2.4 测试方法

本研究使用 Google Translate 将语料从语音转换成文字。该平台具有识别精度高，支持多种语言，全球免费在线语音识别与同步翻译等特点[6]。因为随着 Google 不断改进其产品所用的内部语音识别技术，所以对同一样本进行了两次测试。一次为 2018 年 3 月 31 日，第二次是 2018 年 11 月 10 日。结果发现第二次识别精度显著提高。主要表现在 1）强化抗噪音干扰，增强了远场识别功能[7]。但如果多个声源在相近的方位还是难以识别**图 3**。2）情感语音识别能力提升（陈师哲等人，2018; Do et al., 2015） 3）同步翻译的结果比以前自然得多。感觉替换错误减少，而换误和漏误有所增加。

到第二次测试时期为止，Google Translate 主要支持单声道语音文件、单一话者韵律的识别，不支持同一音频文件中话者的转换[8]。因此我们先把音频文件转为单声道，并按照话者进行切分，然后把切分后的文件再输入到 ASR 系统上进行识别。识别操作在安静的室内进行，先将谷歌平台设置成需要的语言，点击图标开始收音，再将需要被识别的音频对着电脑片刻之后，谷歌便会在左侧框中输出识别结果**图 1**。另外，由于 Google Translate 平台每一次翻译的结果都有些许差异[9]，本研究采用同一天（2018 年 11 月 10 日）的多次识别结果中错误较少、翻译质量较好一次的识别结果进行分析比较。

3. ASR 语音识别结果

3.1 ASR 对母语者对话语音的识别结果

下面是将一名来自北京的 WT 校留学生（WT/NS1，汉族）与台湾的主持学生（T/NS1，对外华语教育专业研究生）的一段对话用 Google Translate 转换成文字信息的结果。同原发言相比，共 163 个字中只有 6 个单词被识别错，其识别精度竟达 96%[10]。测试用的样本即便是远程讨论课现场采录的 MP4 压缩格式，音频质量较差，对 Google Translate 并没有造成影响。不过同步自动翻译出了人脑不可能犯的错译，部分英译令人无法理解（后述）。

[6] 其他还有 Google Cloud Speech-to-Text 等更先进的技术，但需要付费，不适合教学用。至于科大讯飞、百度等中国制作的 ASR 精度更适于汉语口语的识别及合成（笪骏，2018），因为登记需要中国大陆手机号，所以不方便国外用户使用。免费的语音开放平台中国手机号

[7] 百度云 Far field speech recognition，https://cloud.baidu.com/product/speech/fsr [2018-10-02]

[8] 还可供选择视频转录模型，

[9] Google 在云端有不同数据中心处理，另外每次搜索算法各自不同，因而产生差异。

[10] 如果把 1A-2 段漏听的 7-8 个词算在一块，识别度不会讲到 90%以下。

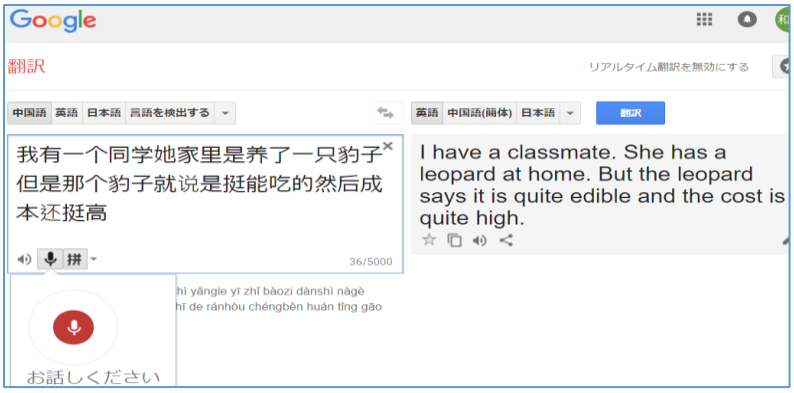

图 1 Google 对 NS 话语的识别

有关 6 个 ASR 听错的词，主要因为发言人把那些词说得较重而造成的。比如 1A-3 段，WT/NS1 将句中“大”和“养”的语调稍微重又长些**表 4**、1B-3，结果 ASR 分别听错为“打”“讲”等字，甚至发生同一个语音换成几种不同文字的情况**表 3**。其实“养”是当天讨论题目的核心动词，大家在讨论中都反复听和说过好多次，并在此段前面 1A-1 段里也出现。所以参加讨论的同学不会听错的。果然不少 NS 与 NS-EC 同学已经笑出来**表 4**、1B-c。

既然 WT/NS1 的发音都符合 ASR 的语音要求，如汉语母语者标准语音，手拿话筒**图 2** 近讲，约 200 多字/分的说话速度[11]等，为何 ASR 识别失败？原因在于 ASR 跨句交叉匹配功能远不如人类。尽管长短时记忆神经网络（Long Short Memory Networks:LSTM）及它的提升版为目前在 ASR 声学模型中广泛应用的算法，可以有效的对时序信号的长时相关性进行建模。理论上可以看到无穷远的上下文，长时间可以保存记忆[12]，实际上，程序运行速度和精度之间此升彼降的矛盾关系，成为制约 ASR 性能提升的瓶颈（中村，2017;篠田，2017）（第 4 节再述）。

另外 ASR 把两处“豹子”分别听错为“报纸”和“能报”（**表 3**，1B-2、**表 4**，1A-2、1A-5，WT/NS1 和 T/NS1 的发言），不仅与原发言内容相去甚远，整句的意思差强人意。ASR 识别错误的原因除了 WT/NS1 对这个词语气放得重些外，还有 T/NS1 说话里夹杂重叠和笑声**表 4**，1A-2，导致话速减慢（约 140 多字/分），造成 ASR 识别不清。后面再有 T/NS1 和 T/NS2 两名台湾学生同时说出“花豹”等词汇，结果发话重复（1B-2），造成 ASR 停掉了识别工作，出现漏听现象。尽管最新 ASR 会支持一个一音频文件中包含不同话者的识别，如果多个声源在相近的方位，ASR 还是难以识别。包括 T/NS1，两名台湾学生都是汉语教学专业。NS1 的华语稍带有南方口音，但其他地方的发言被识别得很清楚，说明他说的汉语足够标准。可知即便是 NS 的自然标准口语，ASR 还难于识别多通道的语料。

[11] 中国人普通话语速是 200 到 250 字左右/分，新闻播音速度在约 300 字左右/分（孟国 2006）。
[12] 《瞭望》采访科大讯飞鄢志杰，https://www.iyiou.com/p/78283 [2018-08-04]

ASR 同步汉英翻译输出了不可思议的译文**图 1**，**表 3** 1A-3，**表 4** 1B-3，这是因为 MT 对文章语境缺乏整体把控能力。此处没能抓好中文的主谓结构，结果英译失败。

- 但是[据]说 [它]的成长速度非常的快→[机器翻译]＊The leopard says it is quite edible and the cost is quite high.）
- 过了半年就已经比我那同学还要打[大]了→[机器翻译]＊After half a year, he has already played more than my classmate.

上面重点指出目前 ASR 在语音识别方面的盲点。智能语言工具还未能支持的背后，往往含有人类特有的交互方式。先将发言内容与非发言及互动行都转写出来，其后在第 4 节将综合分析群体讨论会上，NNS 如何与 NS 交互并共享信息。

表 3（会话片段 1A）Google ASR 识别结果

（下划线处是识别错误）

（ASR 转换后的文本一律无标点符号，去除了元发言的语气词和重复等口语成分。下划线红色是识别错处）

话轮记号	Speaker 记号	ASR 转换后的文本
1A-1	WT/NS1	我有一个同学她家里是养了一只豹子但是那个豹子就说是挺能吃的然后成本还挺高
1A-2	T/NS1	我想确认一下是报纸是就是一种大型的猫科动物有就（之后不识别）
1A-3	WT/NS1	是的他开始进的时候那个豹子还很小但是说他的成长速度非常的快然后过了半年就已经比我那同学还要打了
1A-4	T/NS1	你的同学现在还好吗
1A-5	WT/NS1	家里人呀我也同学还不错能报这是在笼子里的所以现在可能还没有问题呢
1A-6	T/NS1	那行谢谢谢同学分享

表 4（会话片段 1 B）人工转写

话轮记号	Speaker 记号	交互功能标记 (大约按出现顺序标记)	Time(sec) Total 65 sec.		发言转写（含交互行为的描述）
1B-1	WT/NS1	(A)(G)	9	12	*啊：：我有一个同学,然后他家里是养了一只豹子↑,但是那个豹子就是说很能吃的，然后成本还挺高.*
1B-a	交互行为	(A)(E)(S)			包括 T 校其他学校 NS 与 NS(EC) 表示惊讶或发笑。NNS 倒没有表情反应，恐怕没听懂“豹子”这种词汇
	静场		3		
1B-2	T/NS1	(A)(G)(M)(E)(S)	13	14	*ah…〈我想确认↑一下是..豹子↑是..就是那种..大型↑的..猫科动物..有..就是 hh..有耍师..那一种 hh 花豹的那一种 hh 吗.〉*

	T/NS2	(A)(G)(R)	重复		*（同时说）ng 是花豹*
1B-b	交互行为	(S)(M)(A)			13 个同学当中有 5-6 个 NS 与 NS(EC)有笑脸，但 NNS 还没有笑脸，一名 NNS 开始查电子词典。
	静场		1		
1B-3	WT/NS1	(G)(A)	12	15	*啊：是的，但是他开始养↑的时候那个豹子还很小，但是据说它的成长速度非常地快，然后过了半年就已经比我那同学还要大↑了.*
1B-c	交互行为	(A)(M)(S)			WT 校一名 NS 拿出自己的手机给旁边坐的两名 NNS 看，他们侧身观看之后才露出笑脸(图 3)
	静场		3		
1B-4	T/NS1	(M)(A)(E)(R)	3	8	*〈..你的..同学..现在..还..好 hh..吗.〉*
1B-d	交互行为	(S)(E)			几乎所有同学笑逐颜开，有的发出笑声
	静场		5		
1B-5	WT/NS1	(G)(E)(A)	8	9	*是他家里人养的，我同学还不错，豹子是在笼子里的，所以现在可能还没有问题的.*
1B-e	交互行为	(S)(E)(R)			所有同学发笑，幕后的教员也发出了笑声
	静场		1		
1B-6	T/NS1	(E)(S)(R)	6	*7*	*hh hh [好..谢谢..谢谢 hh..这位同学的分享] .*
1B-f	交互行为	(S)(M)			在大家的爆笑中主持人开始指定下一个发言人

3.2 ASR 对非母语者语音的识别结果

会话片段 2A 是一名非汉语学生（B 校 NNS1）与其他发言人的一轮对话。下面是用 Google Translate 转换成文字信息的结果。

表 5 （会话片段 2A）Google ASR 识别结果（下划线红色处是识别错误）

话轮记号	Speaker 记号	ASR 转换后的文本
2A-1	B/NNS1	我以前养过的苍雪的话跟家里的衣柜去宠物店然后一起商量来的但是我现在也现在在家里养的兔子的话我爸爸就你一个人去买的所以引起了妈妈的反感谢谢
2A-2	B/NNS1	其实我妈妈基本上讨厌宠物兔子的话比较一下规模比较大的优点就座椅沙发没有那么做所以基本上没有问题的但是就是因为那个爸爸去工作吧所以那个时候养要养的是妈妈吧所以就让买买醉的发自然后引起了很大的反感谢谢

2A-3	T/NS1	请问那只兔子还在吗
2A-5	B/NNS1	还在已经碎了

据此 ASR 输出的文字，很难理解这轮会话的内容（2A-1，2A-2，2A-5）。与下面对同一轮会话内容的人工转写结果**表 6** 相比较，得知 B/NNS1 发不准汉语发音造成 ASR 识别失败。一是他发[u]时,圆唇的程度不够（如“鼠、突、臭、五”）中，二是送气不够（如“突、臭”），另外后鼻音与四声不到家（如 “然、反、感”），导致 ASR 都不能正确识别，甚至输出的文字一团乱麻。

表 6 （会话片段 2B）人工转写

话轮记号	Speaker 记号	交互功能标记 (大约按出现顺序标记)	Time(sec) Total≒17.7min.		发言转写（含交互行为的描述）
2B-1	B/NNS1	(G) (S)	12		((确认自己手机上的发言稿))大家好,ah: :我,我在日本,我现在,养的是兔子,一个兔子. ah: :之前养过的是仓鼠.
2B-a			≒14 min.		此后 4 校 13 名学生都轮流发言，共有 26 个话轮
2B-2	B/NNS1	(S)(G)	23		好，我以前养过的仓鼠的话 ah：：跟家里人一起去 ah 宠物店=然后一起商量买的.但是 eng 我现在我现在在家里养的..ah：兔子的话..我爸爸突然一个人去买的..ah：所以 eng：引起了妈妈的反感..谢谢.
2B-b	交互行为	(S)(R)	≒2 min.		B NNS1 说完后和旁边的同学一起露出了笑颜。其他同学，除了主持人点头表示理解外，并没有笑。此后 3 校 5 名学生轮流发言，共有 12 个话轮.
2B-3	WT/NNS3	(S)(M)(G)(A)(R)(G)	13	16	((WT/NNS3 抢着话说))［不好意思］可以问一下北京同学的刚刚 ah：他说有愿意说引起了妈妈的反感 ah hh 我想问一下这个故事 hh 可以吗 hh hh.
2B-c	交互行为	(S)(A)(R)			WT 校所有同学笑容鼓励。主持人送视线同意，指定 WT/NNS1 发言。
	静场		3		
2B-4	B/NNS1	(S)(E)(A)（G）	43	45	((笑声应邀，比手划脚地))ah：aha ha 其实我妈妈基本上讨厌：宠物 eng.兔子的话，ah 比较比较规模比较大，而且有点臭，所以 ah：：仓鼠的话 ah：：没有那么臭.所以基本上是没有问题的.但是兔子的话，eng：：因为那个 ng：爸爸去工作吧，所以那个时候 ah 要养的是妈妈吧 hh.所以突然买买兔子的

					话 eng：突然买兔子然后引起了很大的反感，谢谢.
2B-d	交互行为	(S)(E)(R)			B 说完后满开笑脸。还有 3-4 名同学笑颜鼓励，但 S 主持人没有笑。
	静场		2		
2B-5	T/NS2	(A)(S)(G)(R)	3		*请问那只兔子还在吗*↑
2B-e	交互行为	(S)(E)(A)		5	旁坐的主持人 S 发出嗤笑声，其他同学也随着发出笑声或笑脸。
	静场		2		
2B-6	B/NNS1	(G)(E)(S)	3	3	*兔子还在，已经五岁了.*
2B-f	交互行为	(S)(E)(R)			所有同学都有笑容或发笑声

表 7 本文中语料文字转写及交互行为的标记

(3 sec.)	话语计时	↑	升调
,	话语继续	ah,hh,eng	重叠，笑声，吸呼气等副语言
.	话语结束	(( ))	笔者说明
:	表示拖音 (冒号越多表示拖音越长)	（G）	目标导向策略:如用长期记忆，语言功能等达成目标
..	停顿	（R）	奖励策略:如用动作，情感等表示鼓励
=	紧随话语	（A）	注意力策略:如感知联想，语言提醒等
--	话语截断	（M）	动机转移策略:调整目标优先级
［］	同步话语	（E）	情绪调节策略:如快乐，焦虑等
＜＞	符号内是语速明显较慢的话语	（S）	社会共存策略:如礼仪，同情，笑容等

图 2 NS 与 NNS 互动镜头(2B-3)

4. 非母语者交互能力与未来 ASR 技术

基于前节测验的结果，下面对 NNS 的多模态，多通道交互行为进行分析。然后整理当前 ASR 的功能范围及其未来应用技术的发展趋势。

4. 1 NNS 采取多模态交互策略

参加“讨论课”要求 NNS 学生 CEFR 的 B 级以上语言水平，不过大多数 NNS 离 B 级应有的独立交际能力还有一段距离，有时 NNS 汉语发音欠佳影响口语交际。ASR 更不善于识别非标准语音，结果输出让人费解的文字**表 5**。实际上，其他同学以记忆或类推与核实等手段，都理解 B/NNS1 想要说什么。对话句法理论指出，人类自然交流中，重复现象可以帮助构建话语的衔接和连贯。此时，感知、意识、记忆、注意、联想、动机和情绪等多模态信息为整体话语的语义连贯和深层理解提供充足的资源(Tannen, D. 2007; Susan M.et al., 2008）。**表 4** 与**表 6** 所示，多处可以发现 NNS 用多模态交互策略，试图补偿二语能力的差距与 NS 进行沟通。下面基于 GRAMES 改为谈话策略的标记，重新整理 NNS 与 NS 的多模态交互行为。

（表 6 会话片段，2B-3 到 2B-f，共约 69 秒） 划下线[__]指谈话策略，(S)(M)等指策略分类

B/NNS1 说完后，他好友 WT/NNS3 拿过话筒[(S)，(M)动机转移]，引起 B/NNS1 说过的一句[(G)目标导向]，向他要求澄清一遍 [(A)注意力，(R)(G)]。然后 WT 校所有同学满脸笑容鼓励(R)。主持人送视线同意，指定 WT/NNS 发言 [(S)(R)]。B/NNS1 就发出笑声应邀[(S)，(E)情绪调节]，比手划脚地[(S)]重述了前段的内容(2B-4)[(A)(G)]。此间其他同学们以笑颜鼓励（2B-c，2B-d）[(S)(E)(R)]。此后另外一位 NS 颇有风趣地插话提问(2B-5)[(A) (S)(G)(R)]，旁坐的主持人发出笑声，其他同学也随着发出笑声或笑脸[(S)(E)(A)]。对此 B/NNS1 斩钉截铁地答复(2B-6)[(G)(E)(S)]使得大家都笑起来（2B-f）[(S)(E)(R)]。

初步统计所用 GRAMES 策略的结果是;(S)社会共存策略 9 次＞(R)奖励策略 6 次＞(G)目标导向、(E)情绪调节、(A)注意力等策略，均 5 次＞(M)动机转移策略 1 次(次数不含重复出现的，下同)。其中 NNS 运行的策略从多到少排序后有，(S)7 次＞(E)5 次＞(R)4 次＞(G)(A)均 3 次＞(M)1 次。可见不仅是 NS 还是 NNS，运用社会共存策略的占多数，其次是奖励、目标导向、情绪调节、注意力等策略受 NS 的重用。而 NNS 较多用情绪调节策略的倾向。尽管随意选来的短时交互语料，这种结果值得我们重新思考群体讨论的组织框架与条件。

4 .2 NNS 运用社会共存策略

NNS 与 NS 共同维护课堂是“讨论课”的无言共识。各单位的教师、助手与技术人员在后方支持课程活动的顺利运行，保证大家平等参与讨论（砂冈，2016）。如上所示同学最多用**(S)**社会共存策略，可以认为跟这堂课的组织原则有关。课堂需要自我管理，使大家有意识和无意识地注重社会共存策略。其中被 NNS 采取的(S)策略有如，用社会礼仪打招呼（说谢谢）、笑颜笑声达成感情共鸣（DuBois，2014）、用手势或加强语气示意（比手划脚或斩钉截地答复）或主动帮助别人（拿过话筒），平衡分布话语权等行动。其他策略与(S)衔接，相互连贯加深交互（安西，2017）。可知 NNS 虽受外语能力的制约，对语言管理意识并不损于 NS。NNS

常用笑容礼仪等(S)策略，要缓解群体讨论的紧张气氛，成功扮演一个共生情感的中介人角色，与同侪合作共享信息(砂冈, 2018b)。

NNS 能够参与讨论离不开与 NS 的互动。其中与 NNS 接触经验丰富的 NS-E(Native Speaker-Experienced)，无意识地监测双方信息传输是否顺利，还可以具体提出支援方案等语言管理能力(栁田，2015; Sunaoka, 2018a)。如上面 WT/NNS1 经常注意全体会员有无举手要发言，将视线放到发言人的方向，使用体态语表示“了解”。另外一名 NS-E（W/NS2）主动为 NNS 提示生词，帮助他们语义理解，让他们能赶上会话进度有贡献**表 4**，自 1B-b 至 1B-d，e。譬如：

W/NS1 话中的“豹子”这个词汇，起初只有部分 NS 表示惊讶或发笑声，而 NNS 他们似乎没听懂，没有表情反应（1B-a）。此时主持人 T/NS1 向 W/NS1 确认是否真的“豹子”（1B-2）。这句话因发自于 T/NS1 的内心疑惑和惊奇的感情，富有情感，话速变慢。正因为这个语音特色造成 ASR 听别失败，而浓厚情感的音色与音调引起了大家对“豹子”这个生词的选择性注意（王萍丽等人, 2015)。显然 T/NS1 发言后一名 W/NNS3 开始查字典，但还是弄不明白。于是旁坐的 W/NS2 拿出自己的手机（看来帮他们查字典或网络检索）给旁边坐的两名 NNS 看（1B-c）**图 3**。他们侧身观看之后才露出笑脸。此后两名 NNS 都能赶上会话的进度，带着笑容参与讨论（1B-d，e），最后达成了全体成员对 W/NS1 发言内容的理解。NS-E 的无言社会共存行为促进了 NNS 交互与共享信息（Sunaoka, 2018a）。

4 .3 NNS 长时间运用情节记忆

参加网络视频交流活动，由空间距离和时间距离，导致身体动作（姿势、视线交集等）、感官知觉（如听觉、视觉、嗅觉、触觉）的使用都受到一定限制，从而提高参与者之间的交互成本(Clark & Brennan，1991)。正由此限制，群体讨论成员多用情节记忆（Episodic memory），注重推理和联想,以提高与同侪共享信息的效果。如自从 WT/NNS 第一次介绍自己养宠物经验后，第二次重述，竟花了 18 分钟**表 6**。中间还夹有几段其他同学的发言，之后因他好友 WT/NNS2 的澄清要求，WT/NNS1 有机会重复自己的发言，通过核实、重述、交叉匹配等手段，最终消解了大家的歧义。此间大家保持运用长时间情节记忆，尽管要花不少时间，加强互动共享信息。现在 ASR 既有长短时记忆功能（第 3 节已述），最新又出现了装有情节记忆能力的机器人[13]，但目前只能对物体的情节记忆，未能支持对抽象情节的记忆。智能语言工具还难以突破 NNS 学生长时间运用情节记忆的能力。

4.4 ASR 的现在与未来发展趋势

ASR 即使是人工智能目前落地最成功的技术，不过如上所示，ASR 与人脑的认知相比还存在诸多缺陷，但总是在朝着精度不断提高、功能更强大的方向发展。下面引自据最新文献和资料（中村等人,2017;篠田,2017;陈师哲等人，2018），总

[13] 麻省理工学院开发的「ComText」http://news.mit.edu/2017/robot-learns-to-follow-orders-like-alexa-0830. [2017-08-30]

结现阶段 ASR 的基本原理与功能范围，在此基础上介绍 ASR 及其应用技术的未来发展趋势。

ASR 整个系统由语音识别（听写功能），机器翻译（MT），语音合成（TTS）3 种技术组合而构成**图 3**，下同。其中统计匹配模式是整个系统的核心部分。近年来，基于海量语音、语言数据库，通过云端加强 DNN 训练的搜索算法模型，包括 TTS，ASR 与 MT 的精度亦获得了迅速提高。不过这些模式要求的数据都不是原始“自然口语”的语料，而是“标准声音信号”。由于目前 ASR 在嘈杂环境中获取特定声音较为困难，尤其在语言切换时，系统容易出现混乱。因此先去除副语言和非语言成分等“冗余信息”，之后对“声学波形”进行端点检测（除多余的静音和非说话声音）、降噪（去掉杂音与噪音）、特征提取等筛选处理（除掉口音，性别及年龄等个人语音特点），保留语音的“关键信息”，再按照一定规则对数据加以整理构成模式库，最后实行模式匹配。因此在应用 ASR 时，环境要安静、单人（Single Speaker）、说话要离麦克风近、发音要标准、不能持续对话、不能打断[14]。最近 ASR 对情感语音的识别精度有所提升（陈师哲等人，2018），但分开语音所含的韵律特征与说话者特征是语音工学的长年课题，至今尚未解决（篠田,2017）。

有关话者的姿势、视线交集等身体动作以及如视觉、嗅觉、触觉等非语言感官知觉信息的研究本属于开发仿生机器人（Human robot）的范畴（浅田，2010）。近几年 ASR 朝着同非语言技术并存与互补方向发展，如情感语音特征提取研究与图像识别技术的兼容不断取得进展（石黒,2015; 石田等人,2018），已开始应用到汉语的实用业务服务（潘忠德等人,2015; 韩伟等人,2016; 戴礼荣等人,2017; 李银河等人,2017）。但如何提取不同模态下（比如面部表情）最优的情感特征还没有定论。不同模型只能用分流、分层进行运算，计算复杂度极高，机器负载亦过大（上同）。

至于 ASR 的同步翻译功能，DNN 系统往往将词汇表限制为高频词，并将其他所有低频词视为未登录词（刘洋，2017）。词汇量的受限导致译文中出现原文单词漏译和重复翻译、少见的专有名词翻错或 unknown words 等现象（同上）。最新 ASR 提升了长时间记忆性能，但如上所示，还远不如人的记忆功能。尽管 DNN 对 MT 与 ASR 具有里程碑意义，但整个对话系统还是不完美的。因此目前对话系统只能面向旅游查询、订票、数据库检索等一个狭窄领域。包括情感语音识别，语言符号与情境、社会之间的语义协商能力，仍然是 ASR 的挑战课题（陈师哲等人, 2018; 刘振焘等人, 2017; 邵兵等人, 2016）。

[14] CSDN 博客 https://blog.csdn.net/ffmpeg4976/article/details/52348412 [2016-08-28]

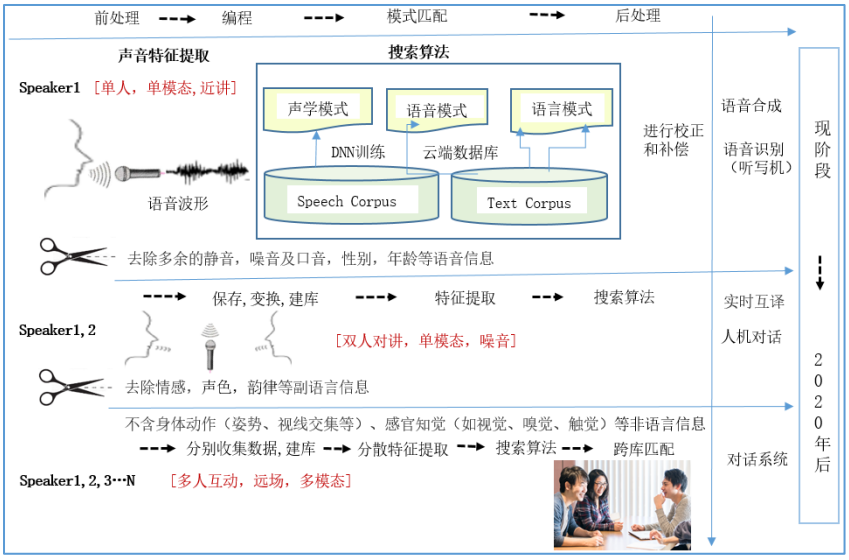

图 3 ASR 到对话系统的基本原理图（改编自中村哲，2016）

5. ASR 在汉语教学应用小结及展望

虽然人工智能语言技术还不太靠谱，但用起来省时省力，已成为一个能支持 NNS 二语习得的重要辅助工具之一。下面介绍几项 ASR＋汉语教学方案。

5.1 ASR＋汉语教学方案（由易到难的排列）

1) 中文听写机＋TTS（初中级）

ASR 可以作为一个很能干的“听写机”，它将逐字逐句听懂汉语并转化成书面中文字，既能省力又能增加自己可操作的教材。譬如为讲演或电影自动生成字幕（张雪梅等人，2012），口述听写成书面教材等，不仅对 L2 学生有用，还为听障人带来福音（河原等人，2018）。另外将 ASR 技术与(Text To Speech: TTS)相结合的教学模式更能提高二语教学的效果。TTS 是已达到成熟的一门技术，它精度比 ASR 还要高，并有如男生、女生、年长者、年轻者等可以选择有不同声音的语音朗读服务功能。如果用 TTS 听不同情感的朗读语音，可以提高汉语的听力能力，又能帮助学生自学能力。

2) 发音练习（初中级）

ASR 可作为一个辅助汉语发音练习的工具（笪骏，2015; Da,2018），如上述 NNS 说有几句发音不准确的话语，ARS 及时识别出了错误的汉语。借以 ARS 这种听写机的功能, 二语学生可以发现自己说汉语中的一些问题。还可以把教材文章朗读出来，用 ARS 来评估他/她朗读结果是否与原内容一致。缺点是 ARS 没有明确提示何处错、为何错。甚至还会出非类人或令人无法理解的识别错误。学生只好把输入用的文本和输出后的字符串相对照才知道自己哪里错。从语言教学的角度来讲，

这种靠感知和摸索的学习法，效率不佳。也可先让学生用 TTS 做复述练习（shadowing），然后再用 ASR 复述会减少二语学生的学习负担。

现在出现即时对汉语学习者的发音给出打分和正音反馈的产品。比如“尔雅中文 app”是一款以学生为主体进行发音练习，同时帮助老师监测学生的发音情况，为汉语教师提供客观数据反馈，以提高课堂语音教学效率，更科学的发音教学模式（魏巍等人，2018）。上面已述在发音时要注意环境安静、单人、标准发音、靠麦克风近讲、不能打断且不要太长，以免 ASR 识别失败。

3) 口语类机考（中高级）

ASR 技术已作为对英文学生进行口语类考试计算机考（机考）的实现手段。采用自动评分的方法，能够根据评估和反馈的结果，对口语考试实行进一步的优化，从而提高口语教学的质量，来满足学生口语学习的需要和口语能力的培养需求（陈卫兵,2015）。在完成人机对话模拟测试之后，教师应要对学生的测试结果进行评价总结，给予学生指导自己汉语能力缺陷的原因，以此弥补 AI 语言技术的局限性(上同)。

4) 对话练习（高级）

ASR 已开始应用到实用业务的问答系统（Question Answering）和信息抽取（Information Extraction）等领域中。但它对人机交互功能还有局限，因此对二语教学的应用目前限于较简单的指令控制，以及对口译课堂教学的部分应用（李霄垅等人, 2018）目前人机对话对二语习得的应用以英文教学占多数（陈卫兵,2015;杜建萍,2018;刘菁菁,2002），而中文教学的却少见。英文教学的例如 Microsoft Research 利用最新的人工智能聊天机器人，将华语为母语人士建立机器学习模型，可直接语音和软件沟通对话。日本一家私企开发的学英文 APP「Tera Talk」，通过与 AI 的模拟对话训练引导学生掌握口语表达技巧，低廉有效地自习英语，产品已推广到 136 个国家[15]。

随着 AI 语言技术的不断完善与成熟,基于云端和大数据的 ASR 引擎可以提供关于真实世界的信息反馈，还可以部分模拟真实世界的人际交互。如果未来 ARS 技术成功应用到互译系统上，它会成为支持交互式口语教学的重要辅助技术（朱坤鸿,2018）。

5.2 展望和课题

通过以上 ASR 与远程讨论课学生之间的语义协商的对比分析，得知 NNS 以跨段、多通道、多模态作为沟通渠道，与 NS 学生合作进行消解歧义。情感语音促成互动共鸣，自律协助行为课堂管理以及语义协商取得了很大的成效，使得全体最终

[15] https://itunes.apple.com/jp/app/terratalk/id1114037031?l=en&mt=8（iOS 版）
https://play.google.com/store/apps/details?id=jp.co.joyz.terratalk（Android 版）

达到了理解。与此相比，当前 AI 工具还严重缺乏社会性，无法主动选择目标，灵活转移注意力、调节情绪而达成目标。自我管理是外语教育的关键能力之一，它决定知识与技能的发挥效率[16]。人的语言无论是选择词汇，还是语气声调，都是自我管理的彰显，这才是“自然”语言的魅力所在。当前 ASR 还未支持副语言与非语言成分的识别和理解。期待未来 ASR 存有足够的发展空间**图 3**，与其他 AI 技术相结合，可以构建出更加复杂的应用，将成为一个能支持二语教学的主流工具。因此我们不要偏激地、激烈地否定或反对相关技术，也不要盲目憧憬它，而是理性地进行取舍。

当然只借助输入等待 AI 输出的结果，中间没有任何认知负荷，也很少与它互动合作协助等交互过程，可否益于二语习得是个最大的疑问。我们知道语言与认知的交互作用，如社会能力、情感、感知、运动、记忆功能、思维与解决问题等能力都为语言发展的重要认知资源。同时, 注意、监控、调整、纠错和反馈等一系列语言习得过程又是认知功能深化的过程(今井等人，2014)。因论述范围太大，这里暂不讨论，另有机会再进行探索。

“亚洲学生远程讨论课”在华语地区合作单位的支持下，已实施 16 年之久，并已保存了 300 多小时的现场录像（砂冈，2016）。本文只能用手工转写语料，观察分析都靠目视和听觉。今后期望利用视觉传感技术（Visual image sensing）以及人脸表情识别等技术，深层挖掘多模态信息（陈华斌等人,2017;李淑婧等人,2015; 松居，2018;佐藤，2017），将建构多模态二语群体讨论知识库，为汉语自然语言理解与教学研究提供数据。

致谢：本论文在第 10 届国际汉语电脑教学研讨会（TCLT10-2018，於台湾师范大学）以及第 11 届中文现代化国际研讨会（AMCL11-2018，於澳门科技大学）上发表后，再在此基础上进行修改并补充完成。在会上得到同仁指导，至此致谢！

参考文献

Anzai, Y. (1992). *Towards a new paradigm of human-robot-computer interaction*. Proceedings IEEE international workshop on robot and human communication, 11-17, Tokyo, Japan: IEEE.

Anzai, Y. (2013, March). *Human-robot interaction by information sharing*. Plenary lecture given at the 8th ACM/IEEE International Conference on Human-Robot Interaction, Tokyo, Japan.

Anzai, Y. (2017). Theory of interaction by information sharing. *Cognitive science. 24*(2), 234-260. [安西祐一郎. (2017). 情報共有によるインタラクションの理論. *認知科学, 24*(2), 234-260.]

[16]王文斌於第三届全国高等学校外语教育改革与发展高端论坛: http://heep.unipus.cn/news/content.php?NewsID=4322 [2018-03-27]

Arita, Y., Ogata, J., Fujimoto, M., & Tsukada, K. (2003). Sports live speech recognition using acoustic and language model adaptation: Application to highlight scene detection. *The Journal of the Institute of Electronics, Information and Communication Engineers (IEICE) Technical Report*, *102*(618), 33-40. [有木康雄, 緒方淳, 藤本雅清, & 塚田清志. (2003).音響・言語適応処理を用いたスポーツ実況中継音声の認識: ハイライトシーン検出への応用. *電子情報通信学会技術研究報告*, *102*(618), 33-40.]

Asada, M. ((2010). *What is a robot: Challenge the mystery of human brain and wisdom*. Tokyo, Japan: NHK Publishing, Inc. [浅田稔. ((2010). ロボットという思想 : 脳と知能の謎に挑む.東京: NHK 出版協会.]

Chen, H., Kong, M., Lv, N. (2017). Status and development of vision sensors on intelligentized robotic welding technologies. *Electric Welding Machine*, 47(3) , 1-7. [陈华斌,孔萌, & 吕娜. (2017).视觉传感技术在机器人智能化焊接中的研究现状. *电焊机*. 47(3), 1-7.]

Chen, P. (2017). Three waves of development in China's language industry. *Chinese Journal of Language Policy and Planning*, *5*(11), 20-28 [陈鹏. (2017). 当代中国语言产业发展的三次浪潮. *语言战略研究*, *5*(11), 20-28.]

Chen, S., Wang, S., & Jin. Q. (2018). Multimodal emotion recognition in multi-cultural conditions. *Journal of Software*. *29*(4), 1060-1070. Retrieved from http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=5412&journal_id=jos [陈师哲, 王帅, &金琴. (2018). 多文化场景下的多模态情感识别. *软件学报, 29*(4), 1060-1070. 参见 http://www.jos.org.cn/jos/ch/reader/create_pdf.aspx?file_no=5412&journal_id=jos]

Chen, W. (2015). Activity design of the seventh grade English oral teaching based on human-machine dialogue. *English Teachers*, 17, 59-62. [陈卫兵. (2015). 基于"人机对话"背景下七年级英语口语教学的活动设计.*英语教师*, 17, 59-62.]

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC, US: American Psychological Association.

Clifford, J., Merschel, L., & Munné, J. (2013). Surveying the landscape: What is the role of machine translation in language learning? *@tic. revista d'innovació educative*, 10, 108-122. Retrieved from https://dialnet.unirioja.es/descarga/articulo/4334892.pdf

Da, J. (2015). The Application of speech recognition technology in Chinese language learning: What can be learned from a pinyin lab session. *Journal of Technology and Chinese Language Teaching*, *6*(1), 16-24. Retrieved from http://www.tclt.us/journal/2015v6n1/da.pdf [笪骏. (2015). 语音识别技术在中文教学中的应用：一堂汉语拼音练习课的启示. *科技与中文教学*, *6*(1), 16-24. 参见 http://www.tclt.us/journal/2015v6n1/da.pdf]

Da, J. (2018, June). *Speech technology in Chinese language learning and teaching*. Workshop given at the 10th International Conference and Workshops on Technology and Chinese Language Teaching (TCLT 10), National Taiwan Normal University, Taipei.

Dai, L., Zhang S., & Huang Z. (2017). Deep learning for speech recognition: Review of state-of-the-arts technologies and prospects. *Journal of Data Acquisition and Processing*, 2, 221-231. [戴礼荣, 张仕良, 黄智颖. (2017). 基于深度学习的语音识别技术现状与展望.数据采集与处理. 2, 221-231.]

Do, Q. T., Toda, T., Neubig G., Sakti, S., & Nakamura S. (2017). Preserving word-level emphasis in speech-to-speech translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(3), 544-556. doi: 10.1109/TASLP.2016.2643280

Du, J. (2018). Practical research on improving middle school students' English listening ability. *Popular Science*. 7, 29. [杜建萍. (2018). 提高中学生英语听力能力的实践研究. *科学大众(科学教育)*, 7, 29.]

Goffman, E. (1981). *Forms of talk*. Philadelphia, PA: University of Pennsylvania Press.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwiska, A., … Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538, 471–476. Retrieved from https://www.nature.com/articles/nature20101

Han, W., Zhang, X., Bai, S., Zhang, R., & Ma, M. (2016). Deep learning for speech recognition, *Journal of Military Communications Technology*, 3.91-97 . [韩 伟, 张雄伟, 白崧廷, 张瑞昕, &马鸣. (2016). 深度学习理论及其应用专题讲座(四)第 7 讲深度学习在语音识别中的应用.军事通信技术, 3, 91-97]

Hori, T., Araki, S., Yoshioka T., Fujimoto, M., Watanabe, S., Oba, T., … Yamato, J. (2011). Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. Transactions on Audio, Speech, and Language Processing (IEEE). *Audio Speech Language Process*, *20*(2), 499–513.

Imai, M., & Saji, N. (2014). *Language and physicality*. Tokyo: Iwanami Shoten. [今井むつみ, 佐治伸郎. (2014). *言語と身体性*. 東京: 岩波書店.]

Ishida, M., Inoue, K., Nakamura, S., Takahashi, K., & Kawahara, T. (2018). An attentive listening system generating empathetic responses and promoting user utterances. *Journal of the Japanese Society for Artificial Intelligence SIG-SLUD*, *B5*(3), 7-12. [石田真也, 井上昂治, 中村静, 高梨克也, & 河原達也. (2018). 共感表出と発話促進のための聞き手応答を生成する傾聴対話システム. *人工知能学会研究会 SIG-SLUD, B5*(3), 7-12.]

Ishiguro, H. (2015). Studies on interactive robots. *Journal of the Japanese Society for Artificial Intelligence*, *30*(3), 377-382. [石黒浩. (2015). 人と関わるロボットの研究. *人工知能*, *30*(3), 377-382.]

Kawahara, T., & Akita, Y. (2018). Captioning lectures using automatic speech recognition for hearing-impaired people. *Acoustical Science and Technology*, *74*(3),1-8. [河原達也, &秋田祐哉. (2018). 聴覚障害者のための講演・講義の音声認識による字幕付与. *日本音響学会誌*, *70*(3), 1-8]

Li, S., Ji, P., Deng, J., Sun, B. & Q. Liu. (2015). Component-based facial expression recognition. *Application Research of Computers*, 3, 917-921,941. [李淑婧, 嵇朋朋, 邓健康, 孙玉宝, &刘青山. (2015). 基于面部结构的表情识别. *计算机应用研究*, 3, 917-921,941.]

Li, X., & Wang, M. (2018). Construction and research of the teaching model of using automatic speech recognition APP in simultaneous interpreting training course—A case study of voice note as an auxiliary tool. *Media in Foreign Language Instruction (TEFLE)*, 179, 12-18. [李霄垅, &王梦婕. (2018). 基于语音识别 APP 的同声传译能力培养教学模式建构与研究—以科大讯飞语记 APP 为例. *外语电化教学*, 179, 12-18.]

Li, Y., Li, X., Xu, N., Zhong, W., Zhao, X., Cheng, X., … Yuan, J. (2017). Research on speech emotion recognition classification algorithm. *Journal of Nanyang Normal University*, 6, 28-33. [李银河, 李雪晖, 徐楠, 钟文雅, 赵新仕, 程晓燕, … 袁键. (2017). 语音情感识别分类算法研究综述. *南阳师范学院学报*, 6, 28-33.]

Liu, Q. (2002). The man-machine method used in teaching speaking in college English. *Media in Foreign Language Instruction*, 1, 15-17. [刘菁菁. (2002). 人机对话在大学英语口语教学中的应用. *外语电化教学*, 1, 15-17.]

Liu, Y. (2017). Recent advances in neural machine translation, *Journal of Computer Research and Development*, *54*(6), 1144-1149. [刘洋. (2017. 神经机器翻译前沿进展. *计算机研究与发展*, *54*(6),1144-1149.]

Liu, Z., Xu, J., Wu, M., Cao, W., Chen, L., Ding, X., … Xie, Q. (2017). Review of Emotional Feature Extraction and Dimension Reduction Method for Speech Emotion Recognition. *Chinese Journal of Computers*, 41(12), 2833-2851. Retrieved from http://cjc.ict.ac.cn/online/onlinepaper/lzt-20181213153256.pdf [刘振焘, 徐建平, 吴敏, 曹卫华, 陈略峰, 丁学文, … 谢桥. (2017). 语音情感特征提取及其降维方法综述. *计算机学报*, *41*(12), 2833-2851. 参见 http://cjc.ict.ac.cn/online/onlinepaper/lzt-20181213153256.pdf]

Luo, H., Pan, Z., & Yi, Y. (2017). Current development and analysis on the prospects of artificial intelligence translation. *Electronics World*, 21, 21-23. [罗华珍, 潘正芹, &易永忠. (2017). 人工智能翻译的发展现状与前景分析. *电子世界*, 21, 21-23.]

Matsui, T. (2018). Learning analytics 4: Multimodal learning analytics. *Information Processing*, 59(9),810-814. [松居辰則. (2018). ラーニングアナリティクス 4. マルチモーダルラーニングアナリティクス. *情報処理*, *59*(9),810-814.]

Nakamura, S., Sakti, S., Neubig, G., Toda, T. (2017). *Introduction of speech-to-speech translation: It aims to automatic translation by computer*. Tokyo, Japan: Koronasha Press. [中村哲, Sakti, S., Neubig, G., &戸田智基. (2017). *音声言語の自動翻訳:コンピュータによる自動翻訳を目指して*. 東京: コロナ社]

Pan, Y., Wan, Y., & Chen, G. (2018). Neural information system research: Current status and prospects. *Journal of Management Science*, *21*(5), 1-21. [潘煜, 万岩, & 陈国青. (2018). 神经信息系统研究:现状与展望. *管理科学学报*, *21*(5), 1-21.]

Pan, Z., Cai, W., Zhu, J. & Cui, D. (2015). Application of speech recognition in the diagnosis of affective disorders. *Chinese Journal of Forensic Sciences*, 6, 85-89. [潘忠德, 蔡伟雄, 朱 杰, &崔东红. (2015). 情感障碍的语音识别研究进展. *中国司法鉴定*, 6, 85-89.]

Sacks, H., Schegloff, E. A, & Jefferson, G. (1974). Turn-taking system lectures on conversation (Nishizaka, A., Trans.). Kyoto: Sekaishisosha-Kyogakusha Co., Ltd. [H. サックス, E. A. シェグロフ, G. ジェファソン. 会話分析基本論集(西阪仰译.). (1974): 順番交替と修復の組織.京都: 世界思想社.]

Satoh, Y. (2017). Collaborative visual sensing for understanding group attention and behaviors. *Journal of the Japanese Society for Artificial Intelligence.* 32(5), 714-720. [佐藤 洋一. (2017). 集合視によるグループの注視・行動のセンシングと理解 (特集; 人と調和して協働する知的情報処理). *人工知能, 32*(5), 714-720.]

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. Ochs, E. A. Schegloff, & M. Thompson (Eds.), *Interaction and grammar* (pp. 3-35). Cambridge: Cambridge University Press.

Shao, B., & Du, P. (2016). The method of speech emotion recognition based on convolutional neural network, *Science and Technology Innovation Herald*, 6, 87-90. [邵兵, & 杜鹏飞. (2016). 基于卷积神经网络的语音情感识别方法. *科技创新导报*, 6, 87-90.]

Shinoda, K. (2017). *Speech recognition*. Tokyo: Kodansha Press. [篠田浩一. (2017). *音声認識.* 東京: 講談社.]

Sunaoka, K. (2016). Cooperative peer learning activity of online cross-cultural communication by sharing Emoji. *Journal of Technology and Chinese Language Teaching*, *7*(1), 43-55. Retrieved from http://www.tclt.us/journal/2016v7n1/sunaoka.pdf [砂冈和子. (2016). 移動通信終端促成遠程討論參加者間形成合作學習作用分析. *科技与中文教学, 7*(1), 43-55. 参见 http://www.tclt.us/journal/2016v7n1/sunaoka.pdf]

Sunaoka, K. (2018a). The interactive modes of non-native speakers in international Chinese language distance class discussions: The analysis of smiling as a facial cue. *Technology-Mediated Chinese language teaching; Innovation in Language Learning and Teaching*, 12, 24-34. London: Taylor Francis Group UK.

Sunaoka, K. (2018b). The language management awareness of L1 students in the international distance Chinese discussion class: A comparative analysis of the roles of NS and NNS in language contact situation. *Applied Linguistics Research on Chinese*, 7, 40-48. Beijing, China: The Commercial Press. [砂冈和子. (2018b). 国际远程汉语讨论课中母语学生的语言管理意识: 语言接触场面 NS 与 NNS 角色对比分析. *汉语应用语言学研究*, 7, 40-48. 北京: 商务印书馆.]

Susan, M., & Alison, M. (2008). Input, interaction, and output in second language acquisition. In B.V. Patten, & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 283-302). New York: Routledge.

Tannen, D. (2007). *Talking voices: Repetition, dialogue, and imagery in conversational discourse.* Cambridge: Cambridge University Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origin of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675–691.

Xu, Q. (2017). Research on the innovative professional interpreting training system against the backdrop of artificial intelligence. *Technology Enhanced Foreign Language Education*, 5, 87-92. [徐琦璐. (2017). 人工智能背景下的专业口译教学系统的创新研究. *外语电化教学*, 5, 87-92]

Yanagida, N. (2015). *Japanese native speakers' communication strategies in contact situations.* Tokyo: Koko Books, Ltd. [栁田直美. (2015). *接触场面における母语*

话者のコミュニケーション方略：情报やりとり方略の学习に着目して. 东京: ココ出版.]

Wang, P., & Li ,Y. (2015). Effects of negotiation of meaning and turn-taking structure: A case study of naturalistic interactions between a native speaker and a non-native speaker. *Chinese Teaching in The World*, *29*(3), 377-392 . [王萍丽 , & 李彦霖. (2015). 语义协商的效用与话步构成: 基于母语者和非母语者自然语言互动的个案研究. *世界汉语教学*, *29*(3), 377-392.]

Wei, W., & Zhang, J. (2018). An intelligent Chinese pronunciation teaching APP and the preliminary research of teaching experiment. *Journal of Technology and Chinese Language Teaching*, *9*(2), 83-97. [魏巍, &张劲松. (2018). 一款汉语智能语音教学 APP 及教学实验初步结果. *科技与中文教学*, *9*(2), 83-97.]

Zhang, X., & Wang, Y. (2012). On research on the effect of dynamic keyword captions on SLA. *Journal of Hebei University of Technology (Social Sciences Edition)*, *4*(1), 88-92. [张雪梅, &王怿旦. (2012). 动态关键词字幕对二语习得效果影响的研究. 河北工业大学学报(社会科学版), *4*(1), 88-92.]

Zhu, K. (2018). Research and implementation of multimodal teaching dialogue systems (Unpublished master's thesis). Beijing University of Posts and Telecommunications, Beijing, China. [朱坤鸿. (2018). 多模态教学对话系统研究与实现(未出版硕士论文). 北京邮电大学, 北京, 中国.]

