计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(1):1−10 [doi: 10.15888/j.cnki.csa.008323]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

语音识别研究综述①

马　晗, 唐柔冰, 张　义, 张巧灵

(浙江理工大学 信息学院, 杭州 310018)
通信作者: 张巧灵, E-mail: qlzhang@zstu.edu.cn

摘　要: 语音识别使声音变得“可读”, 让计算机能够“听懂”人类的语言并做出反应, 是人工智能实现人机交互的关键技术之一. 本文介绍了语音识别的发展历程, 阐述了语音识别的原理概念与基础框架, 分析了语音识别领域的研究热点和难点, 最后, 对语音识别技术进行了总结并就其未来研究进行了展望.

关键词: 语音识别; 声学模型; 语言模型; 人工智能

引用格式: 马晗,唐柔冰,张义,张巧灵.语音识别研究综述.计算机系统应用,2022,31(1):1–10. http://www.c-s-a.org.cn/1003-3254/8323.html

Survey on Speech Recognition

MA Han, TANG Rou-Bing, ZHANG Yi, ZHANG Qiao-Ling

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Speech recognition, which makes the voice readable and enables the computer to understand and respond to human language, is one of the key technologies for human-computer interaction in artificial intelligence. This study introduces the development of speech recognition, expounds the principles, concepts, and basic framework of speech recognition, and analyzes the research hotspots and difficulties in the related field. Finally, it summarizes the speech recognition technologies and presents an outlook on future research into this field.

Key words: speech recognition; acoustic model; language model; artificial intelligence

语言是人类最原始直接的一种交流方式, 通俗易懂、便于理解. 随着科技的发展, 语言交流不再只存在于人与人之间, 如何让机器“听懂”人类的语言并做出反应成为人工智能的重要课题, 语音智能交互技术应运而生. 作为其中重要一环的语音识别技术近年来不断发展, 走出了实验室, 随着人工智能进入人们的日常生活中. 当今市场上语音识别技术相关的软件、商品涉及人类生活的方方面面, 语音识别的实用性已经得到充分的印证. 如今语音识别技术已经成为人类社会智能化的关键一步, 能够极大提高人们生活的便捷度.

1　语音识别技术的发展历程

语音识别技术始于 20 世纪 50 年代, 贝尔实验室研发了 10 个孤立数字的语音识别系统, 此后, 语音识别相关研究大致经历了 3 个发展阶段. 第 1 阶段, 从 20 世纪 50 年代到 90 年代, 语音识别仍处于探索阶段. 这一阶段主要通过模板匹配—即将待识别的语音特征与训练中的模板进行匹配—进行语音识别. 典型的方法包括动态时间规整 (dynamic time warping, DTW) 技术和矢量量化 (vector quantification, VQ). DTW 依靠动态规划 (dynamic programming, DP) 技术解决了语音输入输出不定长的问题; VQ 则是对词库中的字、词等单元形成矢量量化的码本作为模板, 再用输入的语音特征矢量与模板进行匹配. 总体而言, 这一阶段主要实现了小词汇量、孤立词的语音识别. 20 世纪 80 年代至 21 世纪初为第 2 阶段, 这一阶段的语音识别主要以隐

① 基金项目: 国家自然科学基金 (61806178); 浙江省自然科学基金 (LY21F010015)
收稿时间: 2021-04-20; 修改时间: 2021-05-19; 采用时间: 2021-06-02; csa 在线出版时间: 2021-12-17

马尔科夫模型 (hidden Markov model, HMM) 为基础的概率统计模型为主, 识别的准确率和稳定性都得到极大提升. 该阶段的经典成果包括 1990年李开复等研发的 SPHINX 系统[1], 该系统以 GMM-HMM (Gaussian mixture model-hidden Markov model)为核心框架, 是有史以来第一个高性能的非特定人、大词汇量、连续语音识别系统. GMM-HMM 结构在相当长时间内一直占据语音识别系统的主流地位, 并且至今仍然是学习、理解语音识别技术的基石. 此外, 剑桥推出了以 HMM 为基础的语音识别工具包 HTK (hidden Markov model toolkit)[2]. 21 世纪至今是语音识别的第 3 阶段. 这一阶段的语音识别建立在深度学习基础上, 得益于神经网络对非线性模型和大数据的处理能力, 取得了大量成果. 2009 年 Mohamed 等[3] 提出深度置信网络 (deep belief network, DBN) 与 HMM 相结合的声学模型在小词汇量连续语音识别中取得成功. 2012 年深度神经网络与 HMM 相结合的声学模型 DNN-HMM 在大词汇量连续语音识别 (large vocabulary continuous speech recognition, LVCSR) 中取得成功[4], 掀起利用深度学习进行语音识别的浪潮. 此后, 以卷积神经网络 (convolutional neural network, CNN)、循环神经网络 (recurrent neural network, RNN) 等常见网络为基础的混合识别系统和端到端识别系统都获得了不错的识别结果和系统稳定性. 迄今为止, 以神经网络为基础的语音识别系统仍旧是国内外学者的研究热点.

我国的语音识别则起步于国家的“863 计划”和“973 计划”, 中科院声学所等研究所以及顶尖高校尝试实现长时语音的汉语识别工作, 如今中文语音识别技术已经达到了国际水准. 2015 年清华大学建立了第一个开源的中文语音数据库 THCHS-30[5] . 2016 年上海交通大学提出的非常深卷积网络 (very deep convolutional neural networks, VDCNN)[6] 提高了噪声语音识别的性能, 并在此基础上进一步提出了非常深卷积残差网络 (very deep convolutional residual network, VDCRN)[7]. 百度于 2014 年、2016 年依次推出了 DeepSpeech[8] 及其改进版本[9], 并在 2017 年提出 Cold Fusion[10] 以便于更好地利用语言学信息进行语音识别, 该系统以 LSTM-CTC (long short-term memory-connectionist temporal classification) 的端到端模型为基础, 在不同的噪声环境下实现了英语和普通话的语音识别. 2018 年科大讯飞提出的深度全序列卷积神经网络 (deep full-sequence convolution neural networks, DFCNN)[11] 直接对语音信号进行建模, 该模型采用的大量叠加卷积层能够储存更多历史信息, 获得了良好的识别效果. 同年, 阿里巴巴提出低帧率深度前馈记忆网络 (lower frame rate-deep feed forward sequential memory networks, LFR-DFSMN)[12], 将低帧率算法和 DFSMN 算法相结合, 使错误率降低了 20%, 解码速度却提升了近 3 倍.

总体而言, 当前主流语音识别技术主要在大词汇量连续语音数据集上, 基于深度神经网络进行模型构建和训练, 面向不同应用场景需求和数据特点对现有的神经网络不断改进, 相比于传统的统计方法取得了极大的性能提升.

2 语音识别基础

2.1 语音识别概念

语音识别是利用机器对语音信号进行识别和理解并将其转换成相应文本和命令的技术, 涉及到心理学、信号处理、统计学、数学和计算机等多门学科. 其本质是一种模式识别, 通过对未知语音和已知语音的比较, 匹配出最优的识别结果.

根据面向的应用场景不同, 语音识别存在许多不同的类型: 从对说话人的要求考虑可分为特定人和非特定人系统; 从识别内容考虑可分为孤立词识别和连续语音识别、命令及小词汇量识别和大词汇量识别、规范语言识别和口语识别; 从识别的速度考虑还可分为听写和自然语速的识别等[13] .

2.2 传统语音识别基本原理

通常, 语音识别过程大致分为两步: 第 1 步, 首先对语音信号提取特定的声学特征, 然后对声学特征进行“学习”或者说是“训练”, 即建立识别基本单元的声学模型和进行语言文法分析的语言模型; 第 2 步是“识别”, 根据识别系统的类型选择能够满足要求的识别方法, 采用语音分析方法分析出这种识别方法所要求的语音特征参数, 按照一定的准则和测度与系统模型进行比较, 通过判决得出识别结果.

设一段语音信号经过特征提取得到特征向量序列为 $X=[x_1, x_2, \cdots, x_N]$, 其中 x_i 是一帧的特征向量, $i=1, 2, \cdots, N$, N 为特征向量的数目. 该段语音对应的文本序列设为 $W=[w_1, w_2, \cdots, w_M]$, 其中 w_i 为基本组成单元, 如音素、单词、字符, $i=1, 2, \cdots, M$, M 为文本序列的维度. 从贝叶斯角度, 语音识别的目标就是从所有可能产生

特征向量 X 的文本序列中找到概率最大的 W^*, 可以用公式表示为式 (1) 优化问题:

$$W^* = \arg\max_W P(W|X) = \arg\max_W \frac{P(X|W)P(W)}{P(X)} \propto \arg\max_W P(X|W)P(W) \quad (1)$$

由式 (1) 可知, 要找到最可能的文本序列必须使两个概率 $P(X|W)$ 和 $P(W)$ 的乘积最大, 其中 $P(X|W)$ 为条件概率, 由声学模型决定; $P(W)$ 为先验概率, 由语言模型决定. 声学模型和语言模型对语音信号的表示越精准, 得到的语音系统效果越准确.

从语音识别系统的构成来讲, 一套完整的语音识别系统包括预处理、特征提取、声学模型、语言模型以及搜索算法等模块, 其结构示意图如图 1 所示. 其中较为重要的特征提取、声学模型和语言模型将在第 2.2 节中详细阐述.

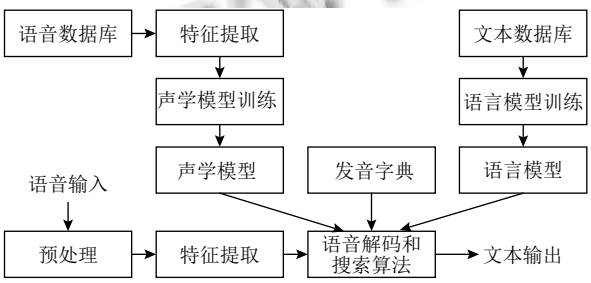

图 1 语音识别系统结构图

预处理包括预滤波、采样、模/数转换、预加重、分帧加窗、端点检测等操作. 其中, 信号分帧是将信号数字化后的语音信号分成短时信号作为识别的基本单位. 这主要是因为语音信号是非平稳信号, 且具有时变特性, 不易分析; 但其通常在短时间范围 (一般为 10–30 ms) 内其特性基本不变, 具有短时平稳性, 可以用来分析其特征参数.

搜索模块是指在训练好声学模型和语言模型后, 根据字典搜索最优路径, 即最可能的输出词序列. 传统的语音识别解码建立在加权有限状态转换器 (weighted finite state transducer, WFST) 所构成的动态网络上, 将 HMM 状态、词典和语法等结合起来. 目前端到端模型中主流的搜索算法为 Beam Search 等.

2.2.1 特征提取

通常, 在进行语音识别之前, 需要根据语音信号波形提取有效的声学特征. 特征提取的性能对后续语音识别系统的准确性极其关键, 因此需要具有一定的鲁棒性和区分性. 目前语音识别系统常用的声学特征有梅尔频率倒谱系数 (Mel-frequency cepstrum coefficient, MFCC)、感知线性预测系数 (perceptual linear predictive cepstrum coefficient, PLP)、线性预测倒谱系数 (linear prediction cepstral coefficient, LPCC)、梅尔滤波器组系数 (Mel filter bank, Fbank) 等.

MFCC 是最为经典的语音特征, 其提取过程如图 2 所示. MFCC 的提取模仿了人耳的听觉系统, 计算简单, 低频部分也有良好的频率分辨能力, 在噪声环境下具有一定的鲁棒性. 因此, 现阶段语音识别系统大多仍采用 MFCC 作为特征参数, 并取得了不错的识别效果.

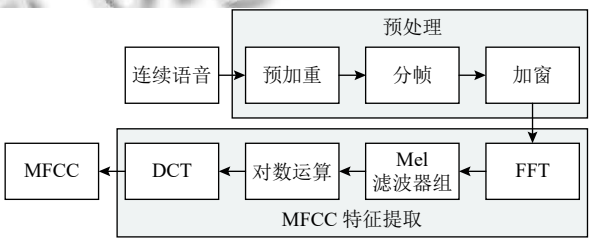

图 2 MFCC 的特征提取过程

2.2.2 声学模型

声学模型是对等式 (1) 中的 $P(X|W)$ 进行建模, 在语音特征与音素之间建立映射关系, 即给定模型后产生语音波形的概率, 其输入是语音信号经过特征提取后得到的特征向量序列. 声学模型整个语音识别系统中最重要的部分, 只有学好了发音, 才能顺利和发音词典、语言模型相结合得到较好的识别性能.

GMM-HMM 是最为常见的一种声学模型, 该模型利用 HMM 对时间序列的建模能力, 描述语音如何从一个短时平稳段过渡到下一个短时平稳段; 此外, HMM 的隐藏状态和观测状态的数目互不相干, 可以解决语音识别中输入输出不等长的问题. 该声学模型中的每个 HMM 都涉及到 3 个参数: 初始状态概率、状态转移概率和观测概率, 其中观测概率依赖于特征向量的概率分布, 采用高斯混合模型 GMM 进行建模.

GMM-HMM 声学模型在语音识别领域有很重要的地位, 其结构简单且区分度训练成熟, 训练速度也相对较快. 然而该模型中的 GMM 忽略时序信息, 每帧之间相对孤立, 对上下文信息利用并不充分. 且随着数据量的上升, GMM 需要优化的参数急剧增加, 这给声学模型带来了很大的计算负担, 浅层模型也难以学习非线性的特征变换.

深度学习的兴起为声学建模提供了新途径, 学者们用深度神经网络 (deep neural network, DNN) 代替 GMM 估计 HMM 的观测概率, 得到了 DNN-HMM 语音识别系统, 其结构如图 3 所示. DNN-HMM 采用 DNN 的每个输出节点来估计给定声学特征的条件下 HMM 某个状态的后验概率. DNN 模型的训练阶段大致分为两个步骤: 第 1 步是预训练, 利用无监督学习的算法训练受限波尔兹曼机 (restricted Boltzmann machine, RBM) , RBM 算法通过逐层训练并堆叠成深层置信网络 (deep belief networks, DBN); 第 2 步是区分性调整, 在 DBN 的最后一层上面增加一层 Softmax层, 将其用于初始化 DNN 的模型参数, 然后使用带标注的数据, 利用传统神经网络的学习算法 (如 BP 算法) 学习 DNN 的模型参数. 相比于 GMM-HMM, DNN-HMM 具有更好的泛化能力, 擅长举一反三, 帧与帧之间可以进行拼接输入, 特征参数也更加多样化, 且对所有状态只需训练一个神经网络. 文献 [4] 证实了神经网络在大词汇量语音识别领域的出色表现.

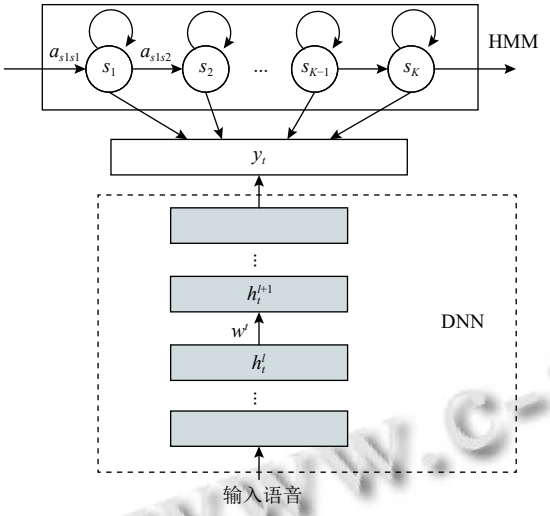

图 3 基于 DNN-HMM 的语音识别系统框架

通过将 DNN 取代 GMM 对 HMM 观测概率进行声学建模, DNN-HMM 相比 GMM-HMM 在语音识别性能方面有很大提升; 然而, DNN 对于时序信息的上下文建模能力以及灵活性等方面仍有欠缺. 针对这一问题, 对上下文信息利用能力更强的循环神经网络 RNN[14] 和卷积神经网络 CNN[15] 被引入声学建模中. 在 RNN 的网络结构中, 当前时刻的输出依赖记忆与当前时刻的输入, 这对于语音信号的上下文相关性建模非常有优势. 然而, RNN 存在因梯度消失和梯度爆炸而难以训练的问题, 于是研究人员引入门控机制, 得到梯度传播更加稳定的长短时记忆 (long short-term memory, LSTM) 网络. LSTM-RNN 对语音的上下文信息的利用率更高, 识别的准确率与鲁棒性也均有提升, 这些在文献 [16] 中能得到证实. CNN 的优势在于卷积的不变性和池化技术, 对上下文信息有建模能力, 对噪声具有鲁棒性, 并且可以减少计算量. 时延神经网络 (time delay neural network, TDNN) 是 CNN 对大词汇量连续语音识别的成功应用[17]. CLDNN (CNN-LSTM-DNN) 综合了三者的优点, 实验结果也证明了三者的结合得到了正向的收益[18].

总体而言, 近年来语音识别中对声学模型的研究仍集中在神经网络, 针对不同的应用场景和需求对上述经典网络结构进行综合和改进[19–21], 以期训练更复杂、更强大的声学模型.

2.2.3 语言模型

语言模型是用来预测字符 (词) 序列产生的概率, 判断一个语言序列是否为正常语句, 也就是解决如何计算等式 (1) 中的 $P(W)$. 传统的语言模型 n-gram[22] 是一种具有强马尔科夫独立性假设的模型, 它认为任意一个词出现的概率仅与前面有限的 n–1 个字出现的概率有关, 其公式表达如下:

$$P(W)=\prod_{i=1}^{m}P(w_i|w_1,w_2,\cdots,w_{i-1}) \propto \prod_{i=1}^{m}P(w_i|w_{i-n+1},\cdots,w_{i-1}) \tag{2}$$

$$P(w_i|w_{i-n+1},w_{i-n+2},\cdots,w_{i-1})=\frac{count(w_{i-n+1},w_{i-n+2},\cdots,w_{i-1},w_i)}{count(w_{i-n+1},w_{i-n+2},\cdots,w_{i-1})} \tag{3}$$

然而, 由于训练语料数据不足或者词组使用频率过低等常见因素, 测试集中可能会出现训练集中未出现过的词或某个子序列未在训练集中出现, 这将导致 n-gram 语言模型计算出的概率为零, 这种情况被称为未登录词 (out-of-vocabulary, OOV) 问题. 为缓解这个问题, 通常采用一些平滑技术, 常见的平滑处理有 Discounting、Interpolation 和 Backing-off 等. n-gram 模型的优势在于其参数易训练, 可解释性极强, 且完全包含了前 n–1 个词的全部信息, 能够节省解码时间; 但难以避免维数灾难的问题, 此外 n-gram 模型泛化能力弱, 容易出现 OOV 问题, 缺乏长期依赖.

随着深度学习的发展, 语言模型的研究也开始引

入深度神经网络. 从 n-gram 模型可以看出当前的词组出现依赖于前方的信息, 因此很适合用循环神经网络进行建模. Bengio 等将神经网络用于语言模型建模[23], 提出用词向量的概念, 用连续变量代替离散变量, 利用神经网络去建模当前词出现的概率与其前 n–1 个词之间的约束关系. 这种模型能够降低模型参数的数量, 具有一定的泛化能力, 能够较好地解决数据稀疏带来的问题, 但其对取得长距离信息仍束手无策. 为进一步解决问题, RNN 被用于语言模型建模[24]. RNNLM 中隐含层的循环能够获得更多上下文信息, 通过在整个训练集上优化交叉熵来训练模型, 使得网络能够尽可能建模出自然语言序列与后续词之间的内在联系. 其优势在于相同的网络结构和超参数可以处理任意长度的历史信息, 能够利用神经网络的表征学习能力, 极大程度避免了未登录问题; 但无法任意修改神经网络中的参数, 不利于新词的添加和修改, 且实时性不高.

语言模型的性能通常采用困惑度 (perplexity, PPL) 进行评价. PPL 定义为序列的概率几何平均数的倒数, 其公式定义如下:

$$P(w_i|w_{i-n+1},w_{i-n+2},\cdots,w_{i-1}) = \frac{count(w_{i-n+1},w_{i-n+2},\cdots,w_{i-1},w_i)}{count(w_{i-n+1},w_{i-n+2},\cdots,w_{i-1})} \tag{4}$$

PPL 越小表示在给定历史上出现下一个预测词的概率越高, 该模型的效果越好.

2.3 端到端语音识别

传统的语音识别由多个模块组成, 彼此独立训练, 但各个子模块的训练目标不一致, 容易产生误差累积, 使得子模块的最优解并不一定是全局最优解. 针对这个问题, 学者们提出了端到端的语音识别系统, 直接对等式 (1) 中的概率 $P(W|X)$ 进行建模, 将输入的语音波形 (或特征矢量序列) 直接转换成单词、字符序列. 端到端的语音识别将声学模型、语言模型、发音词典等模块被容纳至一个系统, 通过训练直接优化最终目标, 如词错误率 (word error rate, WER)、字错误率 (character error rate, CER), 极大地简化了整个建模过程. 目前端到端的语音识别方法主要有基于连接时序分类 (connectionist temporal classification, CTC)[25] 和基于注意力机制 (attention model)[26] 两类方法及其改进方法.

CTC 引入空白符号 (blank) 解决输入输出序列不等长的问题, 主要思想是最大化所有可能对应的序列概率之和, 无需考虑语音帧和字符的对齐关系, 只需要输入和输出就可以训练. CTC 实质是一种损失函数, 常与 LSTM 联合使用. 基于 CTC 的模型结构简单, 可读性较强, 但对发音词典和语言模型的依赖性较强, 且需要做独立性假设. RNN-Transducer 模型[27] 是对 CTC 的一种改进, 加入一个语言模型预测网络, 并和 CTC 网络通过一层全连接层得到新的输出, 这样解决了 CTC 输出需做条件独立性假设的问题, 能够对历史输出和历史语音特征进行信息累积, 更好地利用语言学信息提高识别准确率.

基于注意力机制的端到端模型最开始被用于机器翻译, 能够自动实现两种语言的不同长度单词序列之间的转换. 该模型主要由编码网络、解码网络和注意力子网络组成. 编码网络将语音特征序列经过深层神经网络映射成高维特征序列, 注意力网络分配权重系数, 解码网络负责输出预测的概率分布. 该模型不需要先验对齐信息, 也不用音素序列间的独立性假设, 不需要发音词典等人工知识, 可以真正实现端到端的建模. 2016 年谷歌提出了一个 Listen-Attend-Spell (LAS) 模型[28], 其结构框图如图 4 所示. LAS 模型真正实现了端到端, 所有组件联合训练, 也无独立性假设要求. 但 LAS 模型需要对整个输入序列之后进行识别, 因此实时性较差, 之后也有许多学者对该模型不断改进[29–31].

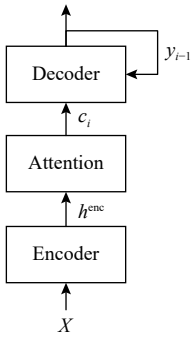

图 4 LAS 模型框架图

目前端到端的语音识别系统仍是语音识别领域的研究热点, 基于 CTC[32–34]、attention 机制[35] 以及两者结合的系统[36,37] 都取得了非常不错的成果. 其中 Transformer-Transducer 模型[38] 将 RNN-T 模型中的 RNN 替换为 Transformer 提升了计算效率, 还控制 attention 模块上下文时间片的宽度, 满足流式语音识别的需求. 2020 年谷歌提出的 ContextNet 模型[39], 采用 Squeeze-and-Excitation 模块获取全局信息, 并通过渐进降采样和模型缩放在减小模型参数和保持识别准确率

之间取得平衡. 在 Transformer 模型捕捉长距离交互的基础上加入了 CNN 擅长的局部提取特征得到 Conformer 模型[40], 实现以更少的参数达到更好的精度. 实际上端到端的语音识别系统在很多场景的识别效果已经超出传统结构下的识别系统, 但距其落地得到广泛商业应用仍有一段路要走.

3 语音识别的难点与热点

语音识别作为人机交互的关键技术一直是科技应用领域的研究热点. 目前, 语音识别技术从理论研究到产品的开发都已取得了很多的成果, 然而, 相关研究及应用落地仍然面临很大挑战, 具体可归纳为以下几方面:

鲁棒性语音识别: 目前, 理想条件下 (低噪声加近场) 的语音识别准确率已经达到一定程度. 然而, 在实际一些复杂语音环境下, 如声源远场等情景, 低信噪比、房间混响、回声干扰以及多声源信号干扰等因素, 使得语音识别任务面临很大挑战. 因此, 针对复杂环境研究鲁棒语音识别是目前语音识别领域的研究难点和热点. 当前, 针对复杂环境下的语音识别研究大致可以分为 4 个方向: (1) 在语音识别前端, 利用信号处理技术提高信号质量: 采用麦克风阵列技术采集远场声源信号, 然后通过声源定位[41]、回声消除[42]、声源分离或语音增强[43] 等提高语音信号质量. 例如, 文献 [44] 在基于深度学习的自适应声学回声消除 (acoustic echo cancellation, AEC) 中加入了背景关注模块以适应部署环境的变化, 以提高语音信号质量; 文献 [45] 以深度聚类为框架提出了结合频谱和空间信息的盲源分离方法; 文献 [46] 利用以基于生成式对抗网络 (generative adversial networks, GAN) 为基础框架的增强网络进行噪声抑制, 从而提高目标语音信号质量; (2) 寻找新的鲁棒性特征, 尽可能消除非目标语音信号的影响: 例如, 伽马通滤波器倒谱系数 (Gammatone frequency cepstrum coefficient, GFCC)[47] 等听觉特征参数更适合拟合人耳基底膜的选择性, 符合人耳听觉特征; 或者, 采用自动编码器[48]、迁移学习[49] 等多种方式提取更鲁棒的特征; (3) 模型的改进与自适应[50]: 上海交通大学提出的 VDCNN[6] 以及 VDCRN[7] 通过加深卷积层提升算法的鲁棒性, 文献 [51] 利用 GAN 中生成器与判别器的相互博弈和瓶颈特征构建声学模型, 文献 [52] 采用 teacher-student learning 的方式以干净语音训练的声学模型作为教师模型训练噪声环境下的学生模型; (4) 多模态数据融合[53]: 当在高噪声环境或多说话人造成语音重叠的情况下, 目标语音信号容易被噪声或其他非目标声源 (干扰信号)“淹没”, 这时仅凭拾音设备捕捉的“语音”信号往往无法获得良好的识别性能; 这时, 将语音信号和其他信号如声带的振动信号[54]、嘴部的图像信号[55] 等进行融合, 更好地提升识别系统的鲁棒性. 例如, 文献 [56] 以 RNN-T 为框架, 提出多模态注意力机制对音频和视频信息进行融合, 以提高识别性能; 文献 [57] 同样基于 RNN-T, 但利用 vision-to-phoneme model (V2P) 提取视觉特征, 连同音频特征以相同的帧频输入至编码器, 取得了良好的识别性能.

低资源语音识别: 这是对各种小语种语言识别研究的统称. 小语种不同于方言, 有独立完整的发音体系, 各异性较强但数据资源匮乏, 难以适应以汉语、英语为主的语音识别系统, 声学建模需要利用不充分的数据资源训练得到尽可能多的声学特征. 解决这一问题的基本思路可以概括为从主流语言的丰富资源中提取共性训练出可以公用的模型, 在此基础上训练小语种模型. 文献 [58] 为解决共享隐藏层中会学到不必要的特定信息这一问题, 提出了一个共享层和特有层平行的模型，它通过对抗性学习确保模型能够学习更多不同语种间的不变特征. 然而, 小语种种类繁多, 为了单独一种建立识别系统耗费过多资源并不划算, 因此现在主要研究多语种融合的语音识别系统[59,60].

语音的模糊性: 各种语言中都存在相似发音的词语, 不同的讲话者存在不同的发音习惯以及口音、方言等问题, 母语者和非母语者说同一种语言也存在不同的口音, 难以针对单独的口音构建模型. 针对多口音建模[61] 的问题, 现有的方法一般可以分为与口音无关和与口音相关两大类, 其中与口音无关的模型普遍表现更好一些. 文献 [62] 尝试通过特定口音模型的集合建立统一的多口音识别模型; 文献 [63] 通过多任务学习将声学模型和口音识别分类器联合; 文献 [64] 则基于 GAN 构建了预训练网络从声学特征中区分出不变的口音.

低计算资源: 精度高效果好的神经网络模型往往需要大量的计算资源且规模巨大, 但移动设备 (如手机、智能家居等) 计算能力和内存有限, 难以支撑, 因此需要对模型进行压缩及加速. 目前针对深度学习模型采用的压缩方法有网络剪枝、参数量化、知识蒸馏等. 文献 [65] 采用网络剪枝的方法构建了动态稀疏神

经网络 (dynamic sparsity neural networks, DSNN) , 提供不同稀疏级别的网络模型, 通过动态调整以适应不同资源和能量约束的多种硬件类型的能力. 文献 [66] 通过量化网络参数减少内存占用并加快计算速度. 知识蒸馏能够将复杂模型的知识迁入小模型, 已应用于对语音识别系统的语言模型[67]、声学模型[68] 和端到端模型[29,69,70] 等进行压缩. 文献 [71] 利用知识蒸馏将视听两模态的识别系统迁移至单听觉模型, 缩小了模型规模, 加快了训练速度, 却并不影响精度.

4 总结与展望

4.1 总结

本文主要对语音识别的发展、系统结构研究、热点及难点进行了阐述. 目前主流的语音识别方法大多基于深度神经网络. 这些方法大体分为两类: 一类是采用一定的神经网络取代传统语音识别方法中的个别模块, 如特征提取、声学模型或语言模型等; 另一类是基于神经网络实现端到端的语音识别. 相比于传统的识别方法, 基于深度神经网络的语音识别方法在性能上有了显著的提升. 在低噪音加近场等理想环境下, 当前的语音识别技术研究已经达到了商业需求. 然而, 在实际应用中存在各种复杂情况, 如声源远场、小语种识别、说话人口音、专业语言场景等, 这些情况使得复杂场景下的语音识别应用落地仍面临挑战. 此外, 尽管当前深度学习在语音识别的应用确实提高了识别率等性能, 但效果好的模型往往规模复杂且庞大、需要的数据资源较为冗余, 不适合用于移动设备 (如手机、智能穿戴设备等); 此外, 小语种、多口音、不同方言等的识别性能仍然差强人意. 总之, 当前语音识别领域已取得丰富的研究成果, 但仍有很长一段路要走.

4.2 展望

在未来很长一段时间内, 基于深度神经网络的语音识别仍是主流; 面向不同应用场景, 根据语音信号特点对现有神经网络结构进行改进仍是未来研究重点. 大体上, 未来语音识别领域的研究方向可大致归纳如下.

(1) 模型压缩与加速. 尽管当前深度学习在语音识别的应用确实提高了识别率等性能, 但效果好的模型往往规模复杂且庞大、需要的数据资源较为冗余, 不适合用于移动设备 (如手机、智能穿戴设备等), 因此对基于深度神经网络的语音识别系统进行网络模型压缩和加速, 将是未来语音识别的研究方向之一.

(2) 数据迁移. 在面对小样本数据或复杂问题时, 迁移学习是一种有效的方式. 在语音识别领域中, 采用迁移学习的方式对小语种、方言口音或含噪语音进行识别也是未来的研究方向之一.

(3) 多模态数据融合. 对于一些复杂的语音场景 (高噪声、混响、多源干扰等), 可以利用语音信号和其他信号 (如图像信号、振动信号等) 进行融合, 以提高语音识别性能, 也是未来研究研究方向之一.

(4) 多技术融合, 提高认知智能. 当前大多数语音识别算法只关注识别文字内容的正确性; 然而, 许多智能语音交互的应用 (如 QA 问答、多轮对话等) 还涉及到语义的理解. 因此, 将语音识别技术结合其他技术[72–75] 如自然语言处理 (natural language processing, NLP) 相结合以提升识别性能也是未来研究方向之一.

参考文献

1 Lee KF, Hon HW, Reddy R. An overview of the SPHINX speech recognition system. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1990, 38(1): 35–45.

2 Young SJ, Young S. The HTK hidden Markov model toolkit: Design and philosophy. 1994. https://www.researchgate.net/publication/263124034

3 Mohamed AR, Dahl G, Hinton G. Deep belief networks for phone recognition. Nips Workshop on Deep Learning for Speech Recognition and Related Applications. 2009, 1(9): 39.

4 Hinton G, Deng L, Yu D, *et al*. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82–97. [doi: 10.1109/MSP.2012.2205597]

5 Wang D, Zhang XW. THCHS-30: A free Chinese speech corpus. arXiv: 1512.01882, 2015.

6 Qian YM, Bi MX, Tan T, *et al*. Very deep convolutional neural networks for noise robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(12): 2263–2276. [doi: 10.1109/TASLP.2016.2602884]

7 Tan T, Qian YM, Hu H, *et al*. Adaptive very deep convolutional residual network for noise robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(8): 1393–1405. [doi: 10.1109/TASLP.2018.2825432]

8 Hannun A, Case C, Casper J, *et al*. DeepSpeech: Scaling up end-to-end speech recognition. arXiv: 1412.5567, 2014.

9 Amodei D, Ananthanarayanan S, Anubhai R, *et al*. Deep speech 2: End-to-end speech recognition in English and mandarin. Proceedings of the 33rd International Conference on Machine Learning. New York: ACM, 2016. 173–182.
10 Sriram A, Jun H, Satheesh S, *et al*. Cold fusion: Training Seq2Seq models together with language models. arXiv: 1708.06426v1, 2017.
11 Zhang WD, Zhang F, Chen W, *et al*. Fault state recognition of rolling bearing based fully convolutional network. Computing in Science & Engineering, 2019, 21(5): 55–63.
12 Zhang SL, Lei M, Yan ZJ, *et al*. Deep-FSMN for large vocabulary continuous speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5869–5873.
13 张学工. 模式识别. 3 版. 北京: 清华大学出版社, 2010.
14 Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 6645–6649.
15 Abdel-Hamid O, Mohamed AR, Jiang H, *et al*. Convolutional neural networks for speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533–1545. [doi: 10.1109/TASLP.2014.2339736]
16 Sak H, Senior A, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 15th Annual Conference of the International Speech Communication Association. Singapore, 2014.
17 Peddinti V, Povey D, Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts. Proceedings of Interspeech 2015, 2015. 3214–3218. [doi: 10.21437/Interspeech.2015-647]
18 Sainath TN, Vinyals O, Senior A, *et al*. Convolutional, long short-term memory, fully connected deep neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane: IEEE, 2015. 4580–4584.
19 Li J, Lavrukhin V, Ginsburg B, *et al*. Jasper: An end-to-end convolutional neural acoustic model. arXiv: 1904.03288v3, 2019.
20 Pundak G, Sainath TN. Highway-LSTM and recurrent highway networks for speech recognition. Proceedings of Interspeech2017.2017.1303–1307.[doi:10.21437/Interspeech.2017-429.]
21 Xiang HY, Ou ZJ. CRF-based single-stage acoustic modeling with CTC topology. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 5676–5680.
22 Bahl LR, Jelinek F, Mercer RL. A maximum likelihood approach to continuous speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983, PAMI-5(2): 179–190. [doi: 10.1109/TPAMI.1983.4767370]
23 Bengio Y, Ducharme R, Vincent P, *et al*. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137–1155.
24 Mikolov T, Karafiát M, Burget L, *et al*. Recurrent neural network based language model. Eleventh Annual Conference of the International Speech Communication Association. Makuhari: DBLP, 2010. 1045–1048.
25 Graves A, Fernández S, Gomez F, *et al*. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh: Association for Computing Machinery, 2006. 369–376.
26 Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015. arXiv: 1409.0473v6, 2015.
27 Graves A. Sequence transduction with recurrent neural networks. arXiv: 1211.3711, 2012.
28 Chan W, Jaitly N, Le Q, *et al*. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016. 4960–4964.
29 Pang RM, Sainath TN, Prabhavalkar R, *et al*. Compression of end-to-end models. Proceedings of Interspeech 2018. Hyderabad, 2018. 27–31.
30 Chiu CC, Sainath TN, Wu YH, *et al*. State-of-the-art speech recognition with sequence-to-sequence models. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 4774–4778.
31 Sun SN, Guo PC, Xie L, *et al*. Adversarial regularization for attention based end-to-end robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1826–1838. [doi: 10.1109/TASLP.2019.2933146]
32 Sak H, Senior A, Rao K, *et al*. Learning acoustic frame labeling for speech recognition with recurrent neural networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane: IEEE, 2015. 4280–4284.
33 Miao YJ, Gowayyed M, Metze F. EESEN: End-to-end

speech recognition using deep RNN models and WFST-based decoding. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). Scottsdale: IEEE, 2015. 167–174.

34 He YZ, Sainath TN, Prabhavalkar R, *et al*. Streaming end-to-end speech recognition for mobile devices. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6381–6385.

35 Dong LH, Xu S, Xu B. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5884–5888.

36 Moritz N, Hori T, Le Roux J. Triggered attention for end-to-end speech recognition. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 5666–5670.

37 Watanabe S, Hori T, Kim S, *et al*. Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240–1253. [doi: 10.1109/JSTSP.2017.2763455]

38 Zhang Q, Lu H, Sak H, *et al*. Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 7829–7833.

39 Han W, Zhang ZD, Zhang Y, *et al*. ContextNet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv: 2005.03191v3, 2020.

40 Gulati A, Qin J, Chiu CC, *et al*. Conformer: Convolution-augmented transformer for speech recognition. arXiv: 2005.08100v1, 2020.

41 Pertilä P, Parviainen M. Time difference of arrival estimation of speech signals using deep neural networks with integrated time-frequency masking. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 436–440.

42 Park J, Chang JH. State-space microphone array nonlinear acoustic echo cancellation using multi-microphone near-end speech covariance. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(10): 1520–1534. [doi: 10.1109/TASLP.2019.2923969]

43 Moore AH, Xue W, Naylor PA, *et al*. Noise covariance matrix estimation for rotating microphone arrays. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(3): 519–530. [doi: 10.1109/TASLP.2018.2882307]

44 Fazel A, El-Khamy M, Lee J. CAD-AEC: Context-aware deep acoustic echo cancellation. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 6919–6923.

45 Wang ZQ, Le Roux J, Hershey JR. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 1–5.

46 Ye SS, Hu XH, Xu XK. TDCGAN: Temporal dilated convolutional generative adversarial network for end-to-end speech enhancement. arXiv: 2008.07787, 2020.

47 Jiang Y, Liu RS, Bai Y. An auditory-based monaural feature for noisy and reverberant speech enhancement. 2017 International Conference on Computing Intelligence and Information System (CIIS). Nanjing: IEEE, 2017. 100–103.

48 Zhang HY, Liu CG, Inoue N, *et al*. Multi-task autoencoder for noise-robust speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5599–5603.

49 Meng Z, Li JY, Gong YF, *et al*. Adversarial teacher-student learning for unsupervised domain adaptation. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5949–5953.

50 Ganapathy S, Peddinti V. 3-D CNN models for far-field multi-channel speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5499–5503.

51 Liu B, Nie S, Zhang YP, *et al*. Boosting noise robustness of acoustic model via deep adversarial training. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 5034–5038.

52 Mošner L, Wu MH, Raju A, *et al*. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6475–6479.

53 Lee W, Seong JJ, Ozlu B, *et al*. Biosignal sensors and deep learning-based speech recognition: A review. Sensors, 2021, 21(4): 1399. [doi: 10.3390/s21041399]

54 Vijayan A, Mathai BM, Valsalan K, *et al*. Throat microphone speech recognition using MFCC. 2017 International Conference on Networks & Advances in Computational Technologies (NetACT). Thiruvananthapuram: IEEE, 2017. 392–395.

55 Pujari S, Sneha SK, Vinusha R, *et al*. A survey on deep learning based lip-reading techniques. 2021 3rd International

Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). Tirunelveli: IEEE, 2021. 1286–1293.

56 Zhou P, Yang WW, Chen W, *et al*. Modality attention for end-to-end audio-visual speech recognition. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2019. 6565–6569.

57 Makino T, Liao H, Assael Y, *et al*. Recurrent neural network transducer for audio-visual speech recognition. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Singapore: IEEE, 2019. 905–912.

58 Yi JY, Tao JH, Wen ZQ, *et al*. Adversarial multilingual training for low-resource speech recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 4899–4903.

59 Sahraeian R, Van Compernolle D. Cross-entropy training of DNN ensemble acoustic models for low-resource ASR. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 1991–2001. [doi: 10.1109/TASLP.2018.2851145]

60 Yi JY, Tao JH, Wen ZQ, *et al*. Language-adversarial transfer learning for low-resource speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(3): 621–630. [doi: 10.1109/TASLP.2018.2889606]

61 Yoo S, Song I, Bengio Y. A highly adaptive acoustic model for accurate multi-dialect speech recognition. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 5716–5720.

62 Elfeky M, Bastani M, Velez X, *et al*. Towards acoustic model unification across dialects. 2016 IEEE Spoken Language Technology Workshop (SLT). San Diego: IEEE, 2016. 624–628.

63 Kamper H, Niesler TR. Multi-accent speech recognition of Afrikaans, black and white varieties of South African English. 12th Annual Conference of the International Speech Communication Association. Florence: DBLP, 2011. 3189–3192.

64 Chen YC, Yang ZJ, Yeh CF, *et al*. Aipnet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition. 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 6979–6983.

65 Wu ZF, Zhao D, Liang Q, *et al*. Dynamic sparsity neural networks for automatic speech recognition. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 6014–6018.

66 Jacob B, Kligys S, Chen B, *et al*. Quantization and training of neural networks for efficient integer-arithmetic-only inference. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 2704–2713.

67 Shi YY, Hwang MY, Lei X, *et al*. Knowledge distillation for recurrent neural network language modeling with trust regularization. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 7230–7234.

68 Takashima R, Sheng L, Kawai H. Investigation of sequence-level knowledge distillation methods for CTC acoustic models. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6156–6160.

69 Kim HG, Na H, Lee H, *et al*. Knowledge distillation using output errors for self-attention end-to-end models. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6181–6185.

70 Zhang WY, Chang XK, Qian YM, *et al*. Improving end-to-end single-channel multi-talker speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1385–1394. [doi: 10.1109/TASLP.2020.2988423]

71 Pérez AF, Sanguineti V, Morerio P, *et al*. Audio-visual model distillation using acoustic images. Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Snowmass: IEEE, 2020. 2843–2852.

72 Kim S, Arora A, Le D, *et al*. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding. arXiv: 2104.02138, 2021.

73 Shenoy A, Bodapati S, Sunkara M, *et al*. "What 's the context?": Long context NLM adaptation for ASR rescoring in conversational agents. arXiv: 2104.11070, 2021.

74 Zhou ZY, Song XC, Botros R, *et al*. A neural network based ranking framework to improve ASR with NLU related knowledge deployed. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6450–6454.

75 Liu XY, Li MD, Chen LX, *et al*. ASR N-best fusion nets. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021. 7618–7622.