计算机科学
COMPUTER SCIENCE

智能语音技术端到端框架模型分析和趋势研究

李荪，曹峰

引用本文

李荪，曹峰. 智能语音技术端到端框架模型分析和趋势研究[J]. 计算机科学, 2022, 49(6A): 331-336.

LI Sun, CAO Feng. Analysis and Trend Research of End-to-End Framework Model of Intelligent Speech Technology[J]. Computer Science, 2022, 49(6A): 331-336.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210500180

智能语音技术端到端框架模型分析和趋势研究

李 荪 曹 峰
中国信息通信研究院 北京 100191
(lisun@caict.cn.cn)

摘 要 端到端(End-to-End)框架是一种基于深度神经网络可直接预测语音信号和目标语言字符的概率模型,从原始的数据输入到结果输出,中间的处理过程和神经网络成一体化,可脱离人类主观偏见,直接提取特征,从而充分挖掘数据信息,简化任务处理步骤。近几年,注意力机制的引入,辅助端到端架构实现了多模态间的相互映射,进一步提高了技术的整体性能。通过对近几年端到端技术在智能语音领域技术和应用的调研,端到端架构为语音模型算法提供了新的思想和方法,但也存在混合框架无法有效地平衡和兼顾单一技术特点,模型内部逻辑复杂使得人工介入调试困难、定制可扩展性减弱等问题。未来端到端一体化模型在语音领域应用方面还将有进一步的发展,一方面是前端到后端的模块端到端,忽略前端语音增强和后端语音识别中涉及多项输入的假设,将语音增强和声学建模一体化,另一方面是交互信息载体的端到端,聚焦于语音信号数据本身的信息提取和处理,使得人机交互更贴近真实人类语言的沟通方式。

关键词:端到端模型;智能语音;混合框架;人机交互
中图法分类号 TN912.34

Analysis and Trend Research of End-to-End Framework Model of Intelligent Speech Technology

LI Sun and CAO Feng
China Academy of Information and Communications Technology,Beijing 100191,China

Abstract The end-to-end framework is a probability model based on the depth neural network which can directly predict the speech signal and the target language character. From the original data input to the result output, the intermediate processing process and neural network are integrated, which can be separated from human subjective bias, directly extract the features, fully mine the data information, and simplify the task processing steps. In recent years, with the introduction of attention mechanism, the auxiliary end-to-end architecture realizes the mutual mapping between multimode, further improving the overall performance of the technology. Through the research on the technology and application of end-to-end technology in the field of intelligent speech in recent years, the end-to-end architecture provides a new idea and method for speech model algorithm, but there are also problems such as the mixed framework can not effectively balance and take into account the single technical characteristics, the complexity of the internal logic of the model makes it difficult for human intervention debugging, and the customization scalability is weakened. In the future, there will be further development in the application of the end-to-end integrated model in the field of speech. On the one hand, the front-end to back-end modules ignore the multiple input assumptions in front-end speech enhancement and back-end speech recognition to integrate speech enhancement and acoustic modeling. On the other hand, the end-to-end interactive information carrier focuses on the information extraction and processing of speech signal data itself the human-computer interaction is closer to the real human language communication.

Keywords End-to-end model, Intelligent voice, Hybrid framework, Human-computer interaction

1 引言

智能语音是实现人机语言的通信,包括语音识别技术(Automatic Speech Recognition,ASR)和语音合成技术(Text-ToSpeech,TTS)。智能语音需要将声学、语音识别、语义、搜索、内容等多种领域技术相融合,以实现自然的人机交互。语音作为信息入口方式已经出现在诸多人工智能(Artificial Intelligence,AI)产品形态中,智能手机、智能家居、智能音箱、智能可穿戴设备纷纷加入语音交互技术,利用语音的便利性提升产品用户体验,在搜索、导航、休闲娱乐等众多领域发挥了越来越重要的作用。总体来说,语音交互在手动输入不便的场景、远距离跨空间操作和多指令多意图执行中,相比传统键盘输入和视觉输出的人机交互模式,具有更高的信息传递效率,形成了以语音交互为主的感知交互状态。

随着人工智能技术的不断发展,智能语音技术性能得到了大幅度提升,传统非端到端技术需要首先对输入数据进行特征提取预处理,手工选取关键特征,将原始的数据信息进行降维处理,步骤繁琐且工程效率低。深度学习端到端架构为技术提供了新的发展思路,从原始的数据输入到结果输出,中间的处理过程和神经网络成一体化,相比传统的多步骤解决问题,端到端则是由输入端的数据直接得到输出端的结果,数据的分析处理呈黑盒状态。如图1所示,随着人工智能发展

通信作者:曹峰(caofeng@caict.ac.cn)

的三次浪潮，智能语音技术和模型研究有阶段性的成果，而语音识别技术的发展最为突出。

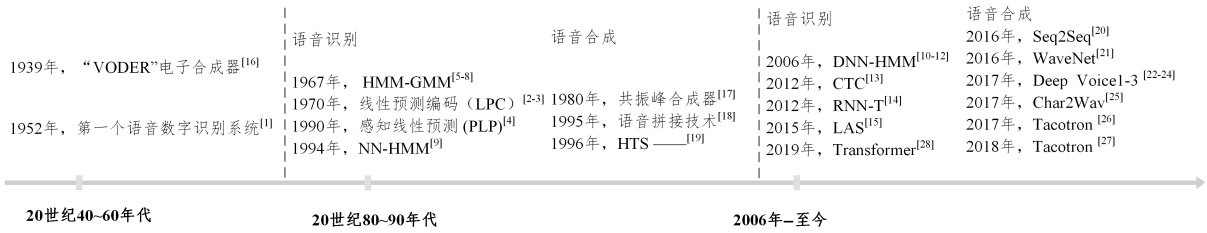

图 1 智能语音技术发展路线

Fig. 1 Development route of intelligent voice technology

为了解决人工干预和模型无法一体化的问题，近年来基于深度神经网络的直接预测语音信号和目标语言字符的概率模型——端到端架构（End-To-End Architecture），成为了研究者们重点探索的领域。

2 传统智能语音框架和技术

语音交互过程是采用语音方式处理和完成的，智能语音技术需要完成的主要任务是尽可能提取语音信号中的声音信息（特征内容和文本内容），同时将自然语言生成后的文本信息转化为语音信号，实现语音交互过程中的“听”输入和“说”输出，以及部分语音理解任务，如图 2 所示。

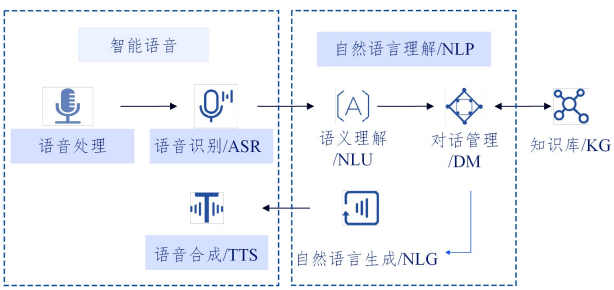

图 2 智能语音交互流程

Fig. 2 Intelligent voice interaction process

2.1 语音处理

在真实的语音交互场景中，目标声源通常离拾音设备较远，使得语音识别输入的目标语音信号衰减，同时还存在环境噪音和干扰信号，最终导致语音信号信噪比偏低，降低了语音识别准确率。通常来说，为了解决远场识别的问题，绝大部分的产品会采用前端语音处理技术来解决噪声、混响和回声等问题，如利用麦克风阵列多通道采集语音信号的目标声源方向和方位信息，对语音信号进行增强处理，消除干扰信号，继而使得后端语音识别采用近场识别方式。

近年来，深度学习正在逐渐代替传统的麦克风阵列数字信号处理技术，前端语音增强开始使用模型训练的方法来获取方向估计和波束生成（Beamforming，BF），并获得了性能上的提升。但是，由于前端增强模式是独立于后端的语音识别，二者的最终目的不一样，单一优化可能无法实现识别的最优化。而且，真实的使用环境非常复杂，通常前端是先定位声源方向再生成波束，继而对波束内的语音信号提高信噪比，因此声源的定位直接影响后续的处理环节，这就直接影响了第一次唤醒和识别的准确率[1,18]。

2.2 语音识别

语音识别，即对语音信号进行分析，将其转换为文字序列的过程。传统语音识别流程及框架如图 3 所示，首先通过信号处理模块提取解码器需要的特征向量，将声音转化为计算机可以识别的数字序列或向量，接着根据提取的特征序列在解码器中寻找最优解。其中，解码器中声学模型是将语音信号的观测特征与句子的语音建模单元联系起来，通过对大量的语音数据进行训练得出每一帧和状态所对应的概率，语言模型结合概率输出计算出概率最大的文字序列。发音字典则包含系统所能处理的单词的集合，为声学模型的建模单元和语言模型建模单元间的映射关系，组成一个搜索的状态空间用于解码器进行解码工作。最后，在声学模型、语言模型、发音词典共同组成的网络中解码出得分最高的序列，该序列即为识别出来的结果。

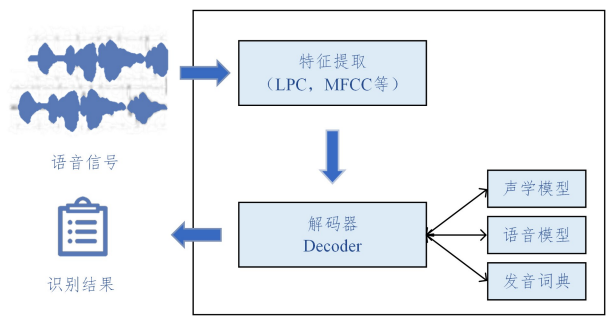

图 3 传统语音识别技术框架

Fig. 3 Framework of traditional ASR

目前，大多数的研究将声学模型和语言模型分开处理，且不同的语音识别系统主要体现为声学模型的差异性，而基于深度学习的端到端方法则将声学模型和语言模型一体化处理，为语音识别技术的发展提供了新的技术思路。

2.3 语音合成

语音合成将文本转化为语音信号，传统技术包括前端和后端两个模块，实现文本分析和语音生成，如图 4 所示。前端模块主要是对输入文本进行分析，提取后端模块所需要的语言学信息，对于中文合成系统而言，前端模块一般包含文本正则化、分词、词性预测、多音字消歧、韵律预测等子模块。后端模块根据前端分析结果，通过一定的方法生成语音波形，后端系统一般分为基于统计参数建模的语音合成（或称参数合成）以及基于单元挑选和波形拼接的语音合成（或称拼接合成）。

对于参数合成而言，该方法在训练阶段对语言声学特征、时长信息进行上下文相关建模，对声学特征参数做后处理，最终通过声码器恢复语音波形，该方法可以在语音库相对较小的情况下得到较为稳定的合成效果，但声码器可能会对音质造成损伤。对于拼接合成而言，通过模型计算挑选语音单元，再对选出的单元进行能量规整和波形拼接，该方法使用真实的语音片段最大限度地保留语音音质，但需要较大的音库，且具有领域限制性。

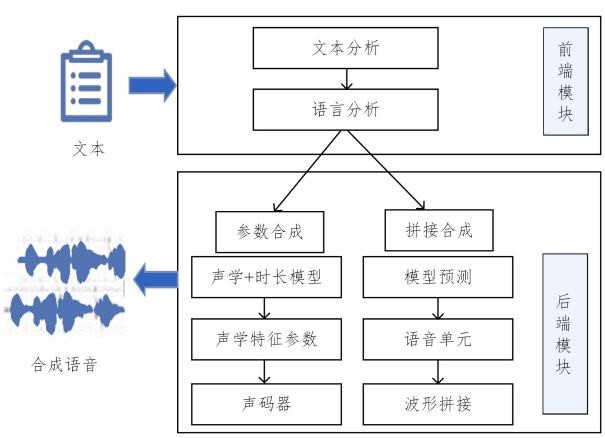

图 4　传统语音合成技术框架

Fig. 4　Framework of traditional TTS

3　端到端语音技术的主要方法

端到端的模型以单个系统的方式联合学习分离的组件，旨在直接实现语音文本的输入与解码识别，从而不需要繁杂的对齐工作与发音词典制作工作，具有节省大量前期准备的优势，真正做到语音和文本之间的转换。在端到端模型的训练过程中，将输入到输出的预测结果与真实结果进行比较，得到的误差在模型中每层传递，反向传播中每层将会依据误差进行微调，直至模型收敛到预期的效果，而中间所有的处理过程都包含在网络内部，不再是分为多个子模块来进行处理。

3.1　语音处理端到端

不考虑前端语音增强和后端语音识别中多涉及多项输入的假设，采用端到端的思想将语音增强和声学建模一体化，相关的研究结果已经证明，端到端一体化模型相比传统分离模型有性能上的提升。2017 年，谷歌首次提出语音增强和语音声学建模的一体化建模的概念，采用信号处理方法，将时域模型转换为频域模型结构 FCLP(Factored Complex Linear Projection)，经过空间和频域的滤波计算出多方向特征参数，然后直接将参数作为语音识别的输入，实现模型端到端一体化，该方法不仅达到了联合优化的目的，而且错误率降低了 16%。2019 年，百度进一步优化了语音增强和语音声学端到端一体化模型，基于复数的卷积神经网络不依赖任何信号处理方法，利用网络的多层次结果充分提取语音自身的特征参数，然后将增强信息传给后端识别模型，与基于数字信号处理的传统算法相比，识别率提升了 30%。

3.2　语音识别端到端

传统语音识别任务将处理流程分为多个子任务(声学模型、语言模型和发音词典)，而端到端的语音识别模型基于语音信号处理的特征信息作为输入，通过端到端神经网络直接产生对应的识别文本，大大简化了内部结构和训练过程，目前主要包括基于 CTC(Connectionist Temporal Classification)、LAS(Listen Attend and Spell)和 Attention 等方法。

3.2.1　典型方法的介绍

(1)基于 CTC 的方法。CTC 主要用于解决时序类数据的分类问题。传统语音识别的声学模型训练，需要得到每一帧数据的特征信息，并且对语音做反复多次迭代的对齐预处理。CTC 是一种损失函数，用于衡量输入的序列数据经过神经网络之后与真实的输出的差距，是一种完全的端到端的声学模型训练。CTC 只需输入序列到输出序列的结果，判断预测值与真实值是否接近，而不需要输出序列的每个结果在时间点与输入序列对其，大大节省了训练时间[13]。但是，基于 CTC 方法进行声学建模还存在一些问题，一方面无法结合语言模型进行联合建模，另一方面则是模型输出存在彼此依赖。

(2)基于 RNN-T(Recurrent Neural Network Transducer)的方法。针对 CTC 的一些问题，Alex Graves 提出基于 RNN-T 的方法对各个部分进行优化[14]。基于 RNN-T 的语音识别方法，使得语音识别模型具备端到端的联合优化，以及声学模型和语言模型联合建模的能力，利用联合网络将语言模型和声学模型状态通过拼接、相加等操作进行结合。同时，该方法具有单调性，能满足实时在线解码处理，更适用于在线识别任务，提高了应用场景的泛化能力。

(3)基于 LAS 的方法。LAS 是一种端到端的序列到序列(Sequence-to-Sequence，seq2seq)结构，直接从帧序列输出到概率转移矩阵，最早是由 Chan 等[15]提出的。LAS 模型中的所有模块都作为独立的端到端神经网络，再将这些单一的组件进行联合训练，区别于传统方法将模型分割为单独的模型。同时，LAS 模型不需要外部的人工设计参与，而是利用词典、转化器等标准化模块，再由完全的神经网络进行训练，使得训练更为简便。

(4)基于 Attention 机制的方法。与人类利用有限的注意力从大量信息中快速筛选出高价值信息类似，注意力模型更专注于让任务处理系统找到输入数据中显著的与当前输出相关的有用信息。与传统机器学习方法相比，注意力模型通过结构化地选取输入的子集来降低数据维度，极大地提高了信息处理的效率与准确性，减小了处理高维输入数据的计算负担。2018 年，谷歌推出基于注意力机制的端到端语音识别模型，该模型基于 seq2seq，结合注意力机制，通过单个模型实现语音序列到文本序列的不定长转换，词错率降到 5.6%，大小是传统模型的 1/18[33]。2019 年，卡内基梅隆大学和卡尔斯鲁厄理工学院的研究者提出了一种非常深的自注意力机制(Self-Attention)的网络，大幅提升了端到端语音识别系统的识别精度[34]。

3.2.2　模型比较分析

近年来，LSTM，CTC 和 LAS 等系列端到端语音识别证明端到端模型比传统的瀑布式模型有更好的表现。2019 年 Jia 等又提出了利用弱监督数据继续提升表现的方案，继续提升了端到端语音到文本序列转换模型的表现(表 1 列出了端到端语音识别模型的优点，表 2 列出了相关模型的结果)[29]。

表 1　端到端语音识别模型的比较

Table 1　Comparison of end-to-end ASR models

模型名称	特点
CTC	CTC 解决了 RNN 训练时目标标签与各输入帧的对齐问题 CTC 直接输出序列预测概率，无需外部后处理
RNN-T	RNN-T 集成了语言模型的声学模型，可以联合优化 RNN-T 可以考虑输出之间的依赖性
LAS	LAS 可以直接将语音话语转录成无需发音模型的字符 LAS 共同学习语音识别器的所有组件
Sequence to Sequence (+Attention)	模型可以考虑序列上下文信息 注意模型可以通过单个模型将语音序列转换为文本序列

表 2 端到端模型的识别错误率[29]

Table 2 Error rate of end-to-end models recognition[29]

模型名称	纯净数据		噪音数据	
	dict	vs	dict	vs
CTC	39.4	53.4	—	—
RNN-T	6.6	12.8	8.5	22.0
Seq2seq+Attention	6.3	11.2	8.1	19.7

注:dict 为开放式听写领域的测试集,vs 来自语音搜索查询领域的测试集

3.3 语音合成端到端

3.3.1 典型方法介绍

传统语音合成包含很多模型,如文本分析、声学模型和音频合成等,需要语言学相关背景知识和人工前期设计加工,使得语音合成步骤繁琐且困难。2017 年,谷歌提出了 Tacotron,这是一种端到端的生成文本到语音的模型,直接从字符合成语音。该模型基于注意序列到序列范式,以字符为输入输出原始谱图,采用多种技术提高了语音合成能力[30]。2018 年,Tacotron 2 简化了 Tacotron 模型,更新了注意力机制,提高了对准稳定性。端到端的语音合成系统简化了前端语言部分的处理,降低了对语言知识的高度依赖性,实现了直接输入文本或注音字符生成音频波形,且可以在不同的语种上进行复用。同时,利用深度神经网路,语音合成的自然度和表现力也得到了很大提升。

3.3.2 模型比较分析

以上研究证明,这些端到端模型能够促进智能语音技术的发展,具有巨大的潜在价值,Tacotron 和 Tacotron2 端到端语音合成模型的比较和实验结果如表 3—表 5[31]所列。

表 3 端到端语音合成模型的比较

Table 3 Comparison of end-to-end TTS models

模块	Tacotron	Tacotron 2
输入	Character	Character
编码	Prenet+CBHG	Convolution+LSTM
注意力机制	Soft alignment	Local sensitive attention
暂停预测	No	Yes
输出目标	Linearspectrogram	Mel spectrogram
声码器	Griffin-Lim	WaveNet

表 4 Tacotron 和传统方法实验结果的比较[30]

Table 4 Comparison of experimental results between Tacotron and traditional methods[30]

模型名称	MOS
Tacotron	3.82±0.085
Parametric	3.69±0.109
Concatenative	4.09±0.119

表 5 Tacotron 和 Tacotron 2 实验结果的比较[31]

Table 5 Comparison of experimental results between Tacotron and Tacotron 2[31]

模型名称	MOS
Tacotron(Griffin-Lim)	4.001±0.087
Tacotron 2(Mel+WaveNet)	4.526±0.066

综上所述,一方面,端到端架构直接提取特征,可充分挖掘数据信息。端到端结构利用纯机器学习的方法,直接训练数据到数据的神经网络,让网络自我发现数据中的统计信息,脱离人类的主观偏见,让数据资源发挥最大的效能。对于设计实现来说,由于无需数据特征提取的中间表示方式,因此简化了人工设计处理的步骤,端到端的架构使得整个流程变得更为简单。以语音识别为例,传统方法是通过语音信号频谱转换来提取人工设计特征,算法在音频片段中定位声音基础单位音位,然后将音位拼接形成独立的词,而在端到端语音识别模型中,输入为音频,而输出为文本,中间深度神经网络直接实现输入到输出的映射。但是,端到端的深度学习也面临着实际的问题。因为缺少可能有用的人工处理特征,模型中没有人类的先验知识作为有效输入,所以模型需要大量的数据进行学习和训练。

另一方面,端到端架构将独立技术融合,简化任务处理步骤。语音交互是单点技术融合、流程化级联处理的模式,从语音输入到语音输出,中间涉及到语音处理、文本理解生成和语音合成等多个步骤,而每个处理环节中也涉及到声学模型、文本模型等子模块。目前,就语音识别、语音合成等任务来说,已经实现文本和语音的端到端转换,不考虑音位、文本归一化处理,基本达到了单点技术的端到端[32]。同时,如果将语音交互看作任务,只关注语音输入和语音的输出,将中间多步骤处理过程黑盒化处理,不完全依赖文本作为信息载体,将语音理解和生成作为内容处理,实现与人类一样认知和感知的交互能力,其实才是完全达到端到端的语音交互。

4 端到端模型发展趋势分析

纵观以上对端到端架构的语音模型关键技术的对比分析可以发现,端到端架构为智能语音技术的发展提供了新的框架和思路,但是在实际的应用过程中还存在一些亟需解决的问题,例如技术的长期稳定可靠和效率要求。基于以上分析,针对基于深度学习端到端架构的语音模型关键技术,本文提出以下发展趋势。

(1)建立 End-To-End 混合框架模型

任何交互本身是呈整体化、流程化的无感认知和感知过程,为了进行技术实现和分析研究,将其拆分为多个独立环节,再通过接口设计和数据输入输出实现交互的目的。通常语音交互中涉及到各类的声学模型和语言模型,在端到端一体化的过程中也衍生了很多针对语音合成、语音识别的端到端模型,每个模型在性能和功能上表现出了不同的优点和缺点。在语音识别中,纯 CTC 解码通过预测每个帧的输出来识别语音,算法的实现基于每帧的解码保持彼此独立这一假设,因此缺乏解码过程中前后语音特征之间的联系,比较依赖语言模型的修正;纯 attention 解码过程则与输入语音的帧的顺序无关,每个解码单元是通过前一单元的解码结果与整体语音特征来生成当前的结果,解码过程忽略了语音的单调时序性。如果语音识别网络结构的编码结构之后,分别连接基于注意力机制和 CTC 的解码结构,则通过分配两种模型的权重再将反向梯度优化的结果相加,共同输出解码识别结果,可能对识别准确性有所提高[33]。对于语音合成,Tacotron2 不需要使用复杂特征工程的数据来训练模式,它集成了 WaveNet 和 Tacotron,使结果接近人类自然语言的质量[30]。虽然端到端模型方法的性能比以前有所提高,但它们将显示出不同的特性和功能。综合考虑各种融合方法的优缺点,建立混合框架模型将是端到端智能语音模型的一个重点研究方向。

(2)关注模型纠偏和可解释性问题

深度学习本身具有一定的局限性,端到端的序列学习将

算法模型中分离的计算模块逐渐整合成为从输入到输出的黑盒运算,这就带来了可解释性和纠偏问题。一方面,端到端使得模型本身架构和内部逻辑变得更为复杂,无法对计算过程的变量进行有效量化和人工介入调试。如果模型已经训练完成,对于某些发音和识别错误,就无法对问题进行精准定位和模型调优,可能还需要重新设计训练数据进行再次实验,这给工程应用带来了很大的困难和挑战。另一方面,定制可扩展性减弱,以语音合成为例,如果想实现合成发音的定制化能力,对韵律、声调、快慢等变量进行调节,由于端到端模型成为不可抽离的整体,其所有的变量信息经训练后都变成固定要素,很难进行人为干预和控制,整体产品的可复用性和灵活度就会大打折扣。此外,智能语音技术目前都还是以文本为中间载体,对语言结构的认知和语音特征抽离为算法模型的训练提供了基础支撑,而语音解析不应仅局限于文本单个字或者词,而应深入到语音信号内在所表达的含义和情感,这样才能更好地挖掘潜在表达的深层次信息,这就为模型的可解释性提出了新的挑战。

(3)聚焦类脑完全端到端交互研究

当前人机语音交互技术的实现主要是将任务拆分成源语言语音识别模型、文本到文本转换模型和目标语言语音合成模型串联而成。以上独立的技术各自都发展得很成熟,可以满足会议记录、车载导航等细分领域的应用,同时集成后的人机语音交互模型也纷纷被应用到智能机器人、智能家电等产品中。语音信息来源于发音人的思想,是人类交互交流、传递信息的主要工具,语音通过人类的感知和认知被转换为可理解的神经编码信息,人借助声音作为数据载体实现信息交互,以供大脑进行信息理解和交换。但是,自然人类语音交互,实际上是语音到语音的一体化信息交互,通过语音信号和脑神经信号的编解码将音频信息进行有效的转换,并不借助于文本模型作为中间理解的中间件,更类似于理想型的语音交互端到端模型。随着类脑感知、类脑学习、类脑控制等理论与方法逐步成熟,将语音直接量化为信息光谱图,神经声码器将输出频谱图转换为时域波形,完全模拟人类语音处理的过程,将语音内容、情感、声学特征都包含在波形信号中,从而不再考虑音素和词的影响,有效地规避了受到上下文限制的重音、音量、音调和速度等语音模糊性问题[39]。人机语音交互要实现等同于人类交流的水平,端到端架构为其提供了有效的手段和方法。

结束语　智能语音技术的发展从基于规则的建模方法到深度学习基于统计的模型架构,取得了各项性能上的进步,端到端架构通过一体化的神经网络模型实现多环节语音处理的归一化,由于其更贴近人类感知和认知的真实过程,因此为智能语音技术的下一步发展提供了新的思路。但是,端到端语音模型还存在模型复杂、纠偏困难以及可解释性差等问题,如何构建更加高效准确的端到端架构语音模型,将成为智能语音研究的重点。

参 考 文 献

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

[2] ATAL B S,HANAUER S L. Speech Analysis and Synthesis by Linear Prediction of the Speech Wave[J]. J. Acoust. Soc. Am.,1971,50(2):637-655,.

[3] ITAKURA F,SAITO S. A Statistical Method for Estimation of Speech Spectral Density and Format Frequencies[J]. Electronics and Communications in Japan,1970,53(A):36-43.

[4] HERMANSKY H. Perceptual Linear predictive(PLP) analysis of speech[J]. Journal of the Acoustical Society of America,1990,87(4):1738-1752.

[5] BAUM L E,EGON J A. An inequality with applications to statistical estimation for probabilistic functions of Markov process and to a model for ecology[J]. Bull. Amer. Meteorol. Soc. m,1967,73:360-363.

[6] BAUM L E,SHELL G R. Growth functions for transformations on manifolds[J]. Pacific Journal of Mathematics,1968,27(2):211-227.

[7] BAUM L E,PETRIE T,SOULES G,et al. A Maximization technique ccurring in statistical analysis of probabilistic functions of Markov chains[J]. Ann. Math. Stat.,1970,41(1):164-171.

[8] BAUM L E. An inequality and associated maximization techniques in statistical estimation for probabilistic functions of Markov processes[M]. Inequalities,1972,3:1-8.

[9] BOURLARD H A,MORGAN N. Connectionist Speech Recognition:A Hybrid Approach[M]. Springer US,1994.

[10] HINTON G,DENG L,YU D,et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition:The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine,2012,29(6):82-97.

[11] GRAVES A,FERNANDEZ S,GOMEZ F,et al. Connectionist Temporal Classification:Labelling Unsegmented Sequence Data with Recurrent Neural Networks[C]// ICML. Pittsburgh,USA,2006.

[12] GRAVES A. Supervised sequence labeling with recurrent neural networks[M]. vol. 385,Springer,2012.

[13] GRAVES A. Sequence transduction with recurrent neural networks[C]// ICML Representation Learning Worksop. 2012.

[14] GRAVES A. Sequence Transduction with Recurrent Neural Networks[J]. Computerence,2012,58(3):235-242.

[15] CHAN W,JAITLY N,LE Q,et al. Listen,attend and spell:A neural network for large vocabulary conversational speech recognition[C]// 2016 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP). IEEE,2016.

[16] ZHANG B,QUAN C Q,REN F J. Overview of speech synthesis methods and development[J]. Minicomputer System,2016,37(1):186-192.

[17] KLATT D H. Software for a cascade/parallel formant synthesizer[J]. The Journal of the Acoustical Society of America,1980,67(3):971-971.

[18] BLACK A W,CAMPBELLN. Optimising selection of units from speech databases for concatenative synthesis[J/OL]. 1996. https://www.researchgate.net/publication/2580972_Optimising_Selection_Of_Units_From_Speech_Databases_For_Concatenative_Synthesis.

[19] MASUKO T,TOKUDA K,KOBAYASHI T,et al. HMM-based speech synthesis with various voice characteristics[J]. The Jour-

nal of the Acoustical Society of America, 1996, 100(4): 2760-2760.

[20] WANG W F, XU S, XU B. First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention[C]// Proceedings Interspeech. 2016: 2243-2247.

[21] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. WaveNet: A Generative Model for Raw Audio[C]// Proceedings of 9th ISCA Speech Synthesis Workshop. Seoul: ISCA, 2016: 125.

[22] ARIK S O, CHRZANOWSKI M, COATES A, et al. Deep Voice: Real-time Neural Text-to-Speech[C]// Proceedings of the 34th International Conference on Machine Learning(ICML'17). Sydney: ACM, 2017: 195-204.

[23] GIBIANSKY A, ARIK S, DIAMOS G, et al. Deep voice 2: Multi-speaker neural text-to-speech[C]// Advances in Neural Information Processing Systems. 2017: 2962-2970.

[24] PING W, PENG K, GIBIANSKY A, et al. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning[J/OL]. 2017. https://arxiv.org/pdf/1710.07654.pdf.

[25] SOTELO J, MEHRI S, KUMAR K, et al. Char2wav: End-to-End speech synthesis[C]// Proceedings of the ICLR 2017 Workshop. Toulon: ICLR, 2017: 24-26.

[26] WANG Y, SKERRY-RYAN R, STANTON D, et al. Tacotron: Towards End-to-End Speech Synthesis[C]// Interspeech 2017 Stockholm. ISCA, 2017: 4006-4010.

[27] SHEN J, PAN R, WEISS R J, et al. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions[C]// Proceedings of 2018 International Conference on Acoustics, Speech, and Signal Processing. Calgary: IEEE, 2018: 4779-4783.

[28] JIA Y, JOHNSON M, MACHEREY W, et al. Leveraging Weakly Supervised Data to Improve End-to-end Speech-to-text Translation[C]// 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2019). IEEE, 2019.

[29] JIA Y, WEISS R J, BIADSYF, et al. Direct speech-to-speech translation with a sequence-to-sequence model[C]// Interspeech 2019. 2019.

[30] WANG Y, SKERRY-RYAN R, STANTON D, et al. Tacotron: Towards End-to-End Speech Synthesis[C]// Interspeech 2017. 2017.

[31] SHEN J, PANG R, WEISS R J, et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2018). IEEE, 2018.

[32] PRABHAVALKARR, RAO K, SAINATH T N, et al. A Comparison of Sequence-to-Sequence Models for Speech Recognition[C]// Interspeech 2017. 2017.

[33] WATANABE S, HORI T, KARITA S, et al. ESPnet: End-to-End Speech Processing Toolkit[C]// Interspeech 2018. 2018.

[34] PHAM N Q, NGUYEN T S, NIEHUES J, et al. Very Deep Self-Attention Networks for End-to-End Speech Recognition[J/OL]. 2019. https://arxiv.org/abs/1904.13377.

[35] YUAN Z, LYU Z, LI J, et al. An improved hybrid CTC-Attention model for speech recognition[J/OL]. 2018. https://arxiv.org/abs/1810.12020.

[36] CHAN W, JAITLY N, LE Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2016.

[37] XIAO X, WATANABE S, ERDOGAN H, et al. Deep beamforming networks for multi-channel speech recognition[C]// IEEE International Conference on Acoustics. IEEE, 2016.

[38] ANUMANCHIPALLI G K, CHARTIER J, CHANG E F. Speech synthesis from neural decoding of spoken sentences[J]. Nature, 2019, 568(7753): 493-498.

[39] CHIU C C, SAINATH T N, WU Y, et al. State-of-the-art Speech Recognition With Sequence-to-Sequence Models[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP 2018). 2017.

LI Sun, born in 1988, master, senior engineer. Her main research interests include artificial intelligence policy, standards, and industry research, covering machine learning, perceptual cognitive technology, and intelligent customer service.

CAO Feng, born in 1986, master, engineer. His main research interests include artificial intelligence evaluation standards, artificial intelligence engineering, robotic process automation and intelligent speech semantics.