语音合成

2019 年 1 月 13 日

语音合成是由文字生成声音的过程，通俗地说，就是让机器按人的指令发出声音。早在1769年，匈牙利发明家Wolfgang von Kempelen就建造了一台会说话的机器。这台机器用机械装置模拟人的发音机理，通过风箱驱动簧片产生声音。1845年奥地利发明家Joseph Faber发明了Euphonia，通过键盘可以发出声音。这些早期发声机器用机械装置模拟人的发音过程，清晰度较低，只能发一些简单音素和单词。

Figure 1: Saarland University大学于2007-2009年重现的Kempelen发声器。

计算机发明以后，语音合成技术开始快速发展。按时间顺序，语音合成方法可归纳为四种：**参数合成**、**拼接合成**、**统计模型合成**和**神经模型合成**。绝大多数合成方法都基于激励-响应（Source-Filter）模型，我们首先对该模型做简单介绍。

1 激励-响应模型

人类的发音可以描述为一个声门激励序列经过一个声道响应函数的滤波过程，称为激励-响应模型。基于这一模型，声音是由肺部气流冲击声门（主要是声带）产生振动，再通过口鼻所形成的声道共鸣产生。如果声带完全打开，则振动源是一个随机噪声，这时产生的声音是清辅音；如果声带拉紧，声门生成周期性振动，这时产生的声音是浊辅音和元音。在这一模型中，气流冲击声带称为声门激励，声道的共鸣称为声道响应，是对声门激励信号的滤波或调制。图2给出了这一语音生成过程。

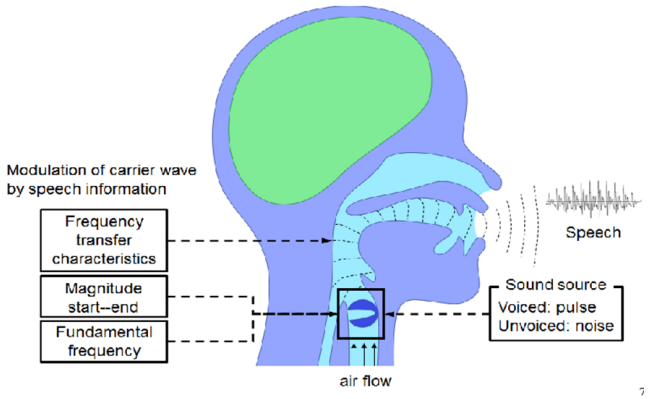

Figure 2: 人类发音的声门激励-声道响应模型。气流冲击声门产生混乱的或周期性的振动，这一振动通过口鼻形成的声道形成共鸣和调制。

我们以元音为例为观察上述声道-激励模型产生的语音信号的特点。基于该模型，语音信号是声门发出的基础周期信号经过声道的共鸣。这一过程有两个效果：一方面，原始周期信号在声道所形成的共振腔中反射叠加，会生成一系列倍频信号，在频域上形成周期性波动；另一方面，声道具有调制作用，对不同倍频信号产生不同的增益，形成频谱上的包络。图3给出语音信号的频谱特性，其中F_0是原始周期信号的频率，称为基频，F_1,F_2,F_3分别为频谱包络的极大值，称为第一、第二、第三共振峰。需要强调的是，F_0表征的是声门的激励特性，而各个共振峰反应的声道传递函数，二者具有完全不同的物理意义。

激励-响应模型是一个生成模型，给定基频和各个共振峰，即可计算出声门激励和声道的传递函数，将二者通过卷积即可以合成语音，另一方面，基于该模型，也可以实现语音信号的分解，将语音分解为声门信号和声道响应。这一分解事实上是一个解卷积过程。最著名的分解方法是基于线性预测（LPC）的分解，即假设语音信号具有线性可预测性，预测系数构成了声道特性，预测的

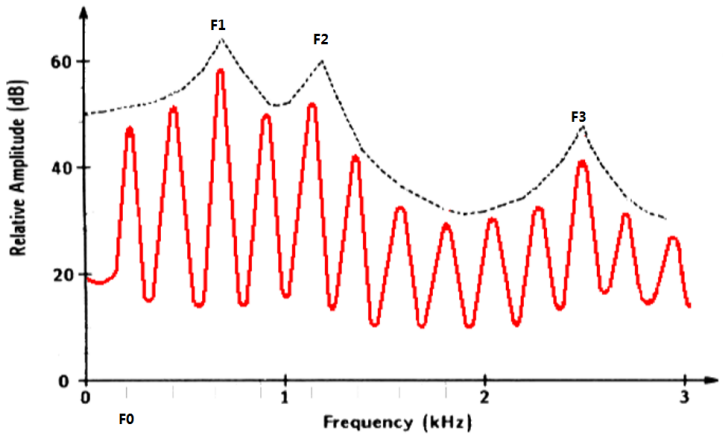

Figure 3: 语音信号是倍频后的声门激励和声道特性的卷积，在频域上表现为声道特性对不同频率幅值的调制。图中红色实线表示实际频谱，黑色虚线为频谱包络，代表声道的响应函数。

残差即对应声门激励。事实上，这一假设对应一个声道的全极点模型，LPC事实上是对该模型极点位置的估计。图4给出这一分解的示意图。

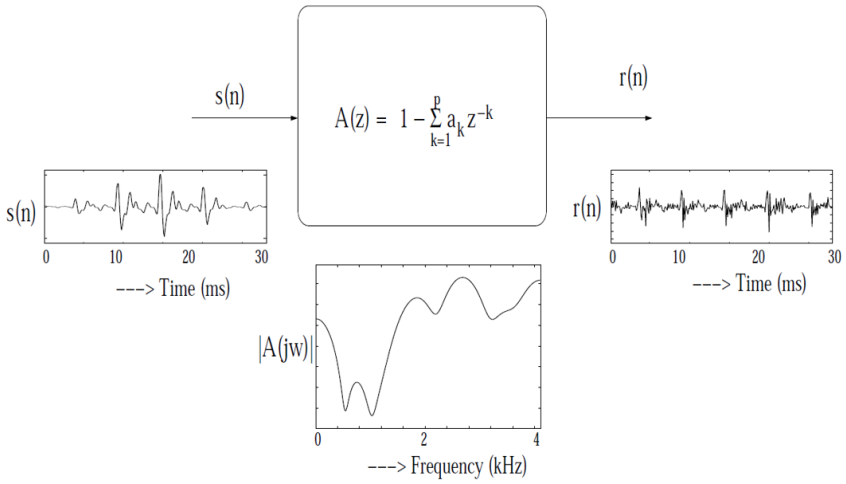

Figure 4: 基于线性回归模型（LPC）的语音信号分解。

我们已经看到，基于激励-响应模型，我们既可以对语音进行分解，也可以

基于分解得到的成分实现对声音的合成。对声音进行分解-合成的装置称为声码器，其中分解部分称为编码器，合成部分称为解码器。历史上第一个声码器由贝尔实验室于1930年发明，这一发明奠定了语音合成研究的基础。图5给出了一个声码器中的编-解码流程示意图，其中编码器将声音分解成基频和频谱包络两部分，前者代表声门特性，后者代表声道特性。这两部分信号经过通信信道传送到解码器，解码器利用这两路信息，基于激励-响应模型合成语音。

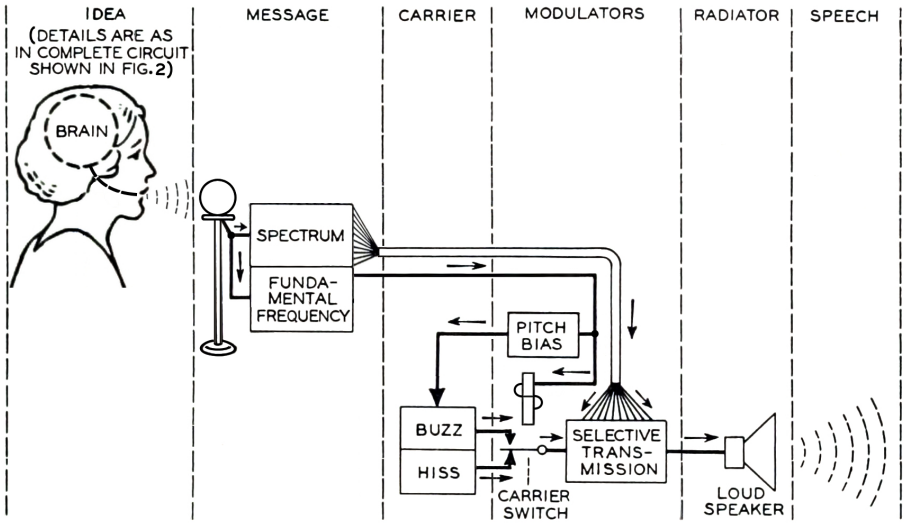

Figure 5: 声码器的基本流程。

2 参数合成

由前一节所介绍的声道-激励模型可知，语音信号可分解为声门激励和声道响应两部分，二者共同决定了发出声音的特性。因此，只要设计合适的激励信号和合适的声道响应函数，即可合成目标声音。

前面我们已经提到，声门激励与发音的类型有关，如辅音和元音，而声道特性直接决定发音的内容。图6给出不同元音在共振峰平面（F1-F2）上的分布情况，可以看到不同元音在这一平面上有显著区分，意味着我们只要指定F1和F2的值，即可生成对应的元音。在实际系统中，可以对现有发音库进行分析，统计不同音素（包括其上下文）的基础和各个共振峰的取值，在合成时利用这些取值即可合成出相应的声音。这一方法称为参数合成。

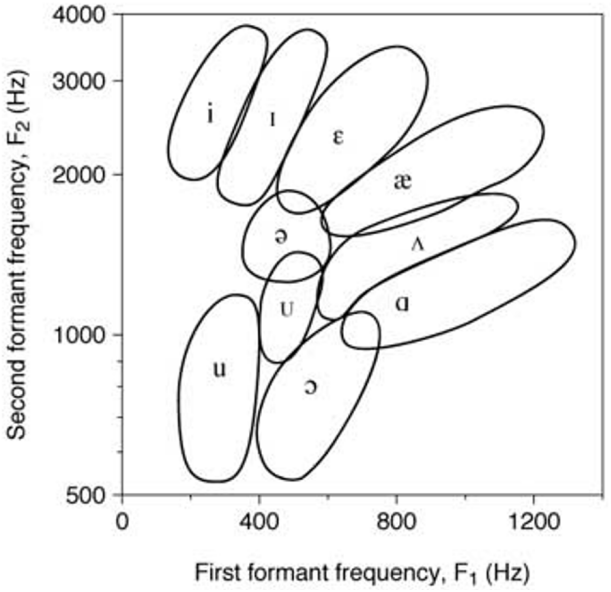

Figure 6: 不同元音在F1-F2平面上的分布。

DEC公司推出的DECTalk是这种合成方法的典型代表。参数合成的优点是计算量小，可调节性高，缺点是参数调节困难，生成的声音自然度较低。

DECtalk lets micros read messages over phones

BY PEGGY ZIENTARA
Senior Editor

The use of the telephone as a "universal terminal" came one step closer to reality when Digital Equipment Corporation (DEC) introduced a voice synthesizer unit that can "read" aloud ASCII text messages.

DECtalk, which the company says will be available in March at a price of $4000, allows you to call in to a personal computer from anywhere in the world by touch-tone telephone (including pay phones), to hear electronic mail messages stored in memory.

DECtalk can also "tell" you what you just typed into memory on the computer keyboard if you simply press a punctuation mark key and the Return key.

Business applications are expected to include access to company data bases by traveling sales representatives; access by lawyers to public and private computer files during trial breaks; direct customer transactions with banks and stock brokerage firms; direct catalog and retail sales; and access to transportation schedules, according to Ed Kramer, vice-president of corporate marketing for DEC.

Other applications will include aid for the speech-impaired (now in place at

via any touch-tone phone.

The DECtalk hardware, which fits easily under a 12-inch monitor, will feature eight different types of voices, including that of an old man, a deep-voiced man, a typical middle-aged man, two different women and a child. The unit attaches to a personal computer via an RS-232C serial interface.

DEC claims to have licked the traditional voice technology trade-offs of having either good voices with limited vocabularies or unlimited vocabularies with poor voices. DEC's unit has a "virtually unlimited vocabulary," Kramer said, and the speaking rate can be adjusted from 120 to 350 words per minute. Tone and modulation can also be regulated.

Heuristics — interpreting word context to improve pronunciation — and a computer model of the human vocal tract allow DECtalk to produce high-quality voices, in contrast with traditional stored digitized or synthesized voice technology, Kramer explained.

The unit accepts ASCII text, which it applies to an internal dictionary and a set of letter-to-sound rules indicating pronunciation and context interpretation to

Figure 7: 1984年InfoWorld报导DECTalk。DECTalk一直是著名科学家霍金的发音助手。

3 拼接合成

上个世纪九十年代，随着大规模语音库的积累，基于拼接的合成方法成为主流。这一方法将事先录好的语音切分成发音片段（一般为音素），在合成时从这些片段中选择合适的候选进行拼接组成句子 [1]。

拼接法需要处理两个主要问题。第一个问题是对上下文环境的处理，同一个音素在不同环境下发音会有所差别，因此需要根据环境选择合理的发音片段。这里的环境既包括前后音素，也可以包括词边界、词性、句子属性等。第二个问题是拼接连贯性问题。把两个发音片段拼在一起时，总会产生一些不连续性。一般采用在频域进行拼接和平滑的方法来增加连贯性。

早期的拼接合成多采用半音素（di-phone）作为发音片段。所谓Di-Phone，是两个相邻音素对中从一个音素的中心到另一个音素的中心对应的发音片段，如图8所示。半音素模型有若干好处，一是在进行语音标注时，标注音素的中心要比标注音素边界容易的多；二是半音素中包含了音素音的变换过程，有利于描述一定的上下文关系；三是音素中心一般比较稳定，在这里拼接有利于保持发音的连续性。

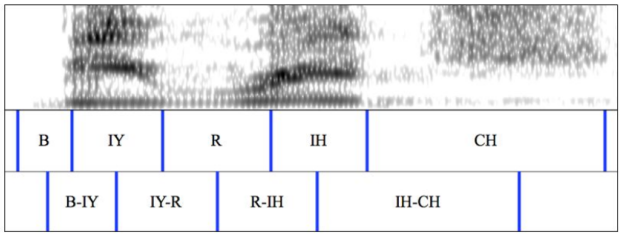

Figure 8: 语音流中的半音素（Di-Phone），其中B-IY，IY-R等都是半音素。

另一种拼接合成方法是基于单元选择（Unit Selection）的方法，如图 9所示。在这种方法里，每个发音单元是一个上下文相关音素，不同的音素可能有多个单元，每个单元拥有包括环境相关的各种标记，以及相应的基频、频谱包络等。在合成时，将目标句子的上下文环境选择合适的单元，将相应的基频和频谱包络拼接起来。在单元选择过程中，需要考虑两个准则，一是使得选择出的单元符合环境约束，二是使得选择出的单元在互相拼接时保持连贯性。和半音素方法相比，单元选择方法基于更大规模的数据库，对环境的建模更细致。理论上说，如果数据库规模足够大，我们总可以选出合适的单元，生成自然流畅的声音。

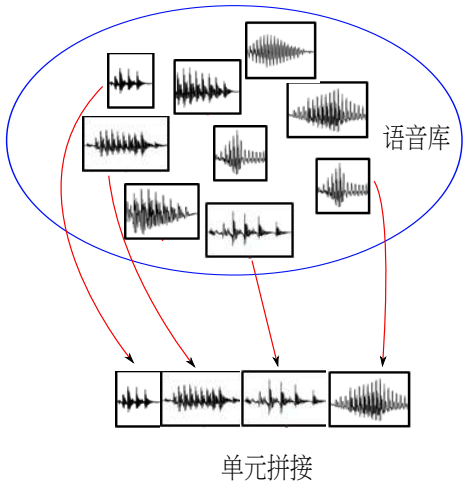

Figure 9: 基于单元选择的语音合成。合成时从语音库中选择合适的发音片段进行拼接。选择发音片段时需要考虑该片段是否符合上下文约束的发音要求，还需要考虑是否可以和其它片段实现自然拼接。图片设计基于HTS slides。[2]

4 统计模型合成

基于拼接的合成方法可以生成高质量的语音，但需要较大规模的语音库，录制成本较高，占用磁盘空间较大，在嵌入式设备上很难实用。另外，拼接方法无法灵活地改变声音的特性，在应用上有一定局限性。

为解决拼接方法的这一困难，人们提出统计模型方法。该方法为每个发音单元建立一个统计模型，在合成时仅利用这些模型进行生成。由于在合成时不需要语音库，基于统计模型的合成系统通常比较精简 [2]。特别重要的是，基于统计模型可以很容易实现对合成语音的控制，这对合成个性化的声音尤为重要。基于隐马尔可夫模型（HMM）的统计模型方法是这一时期的主流。这种方法对每个发音单元建立一个HMM模型，在合成时将句子中所有音素的HMM模型拼接起来形成一个组合模型，由该模型生成最匹配的语音。

具体而言，我们首先对语音信号进行切分，生成上下文相关音素，再对每个音素的长度、基频和频谱包络建立三个HMM模型。上下文环境信息一般包括前后各两个音素标识、词内位置、句内位置、词性、韵律信息等。长度模型是一个单一状态单个高斯的HMM，基频和频谱包络是多状态HMM。基频的状态概率分为两个空间，一个离散空间用来描述清辅音的零基频，一个连续空间用来描述浊辅音和元音的非零基频。频谱包络的HMM的状态概率模型是一个高斯混合模型（GMM），用来描述频谱包络在该状态的分布规律。

在实际合成时，首先需要基于长度模型选择每个音素的发音长度，之后基于基频模型和频谱包络模型生成每个音素的基频和频谱包络，最后基于生成的基频和频谱包络合成语音。在生成基频和频谱包络时，生成准则是使得生成基频和频谱包络在对应HMM中的输出概率最大化。如果不考虑语音帧之间的相关性时，则每个音素中每个状态在不同时刻的输出是不变的。显然，这一输出会在状态的分界面上产生跳跃，影响合成的质量。为解决这一问题，研究者提出基于动态特征建模的方法。基于这一模型，每个状态的输出需要考虑前后帧的相关性，这一约束使得生成的基频和频谱包络更加连续。

图 10给出对频谱包络中的某一维进行生成的过程，其中红色折线不考虑前后帧相关性时的最大概率输出；当考虑前后帧约束时，即生成更为连续的输出，如图中蓝色曲线所示。

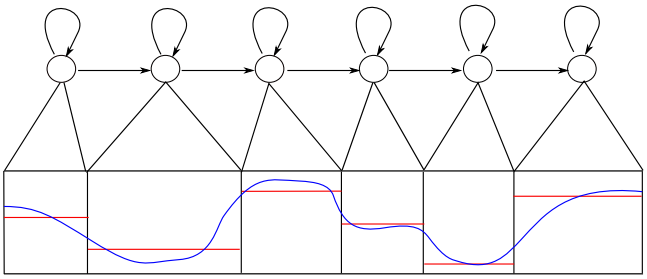

Figure 10: 基于HMM模型生成某一维频谱包络的过程。每个蓝色圆圈代表一个HMM状态，虚线箭头标出每个状态要生成的语音片段。下方两条水平平行线间的竖线表示每一帧的生成值。实际生成时，首先估计出每个音素中每个状态要生成的帧数，再对每个状态进行生成（图中红色折线）。考虑到前后帧的相关性，应对生成的语音帧进行平滑，即得到图中的蓝色生成曲线。图片设计基于HTS slides。

图11给出HMM合成系统的总体框架。在训练阶段，对数据库中的语音信号进行分析，分解成基频（对应声门）和频谱包络（对应声道响应），利用数据库中的标注信息，对每个上下文相关音素建立时长、基频和频谱包络的HMM模型。在合成阶段，首先对输入文本进行分析（一般称为前端处理），将文本转换成标注文件，基于该标注选择相应的音素，并将这些音素对应的基频HMM和频谱包络串连起来，生成句子的基频和频谱包络序列，最后送入声码器得到合成语音。

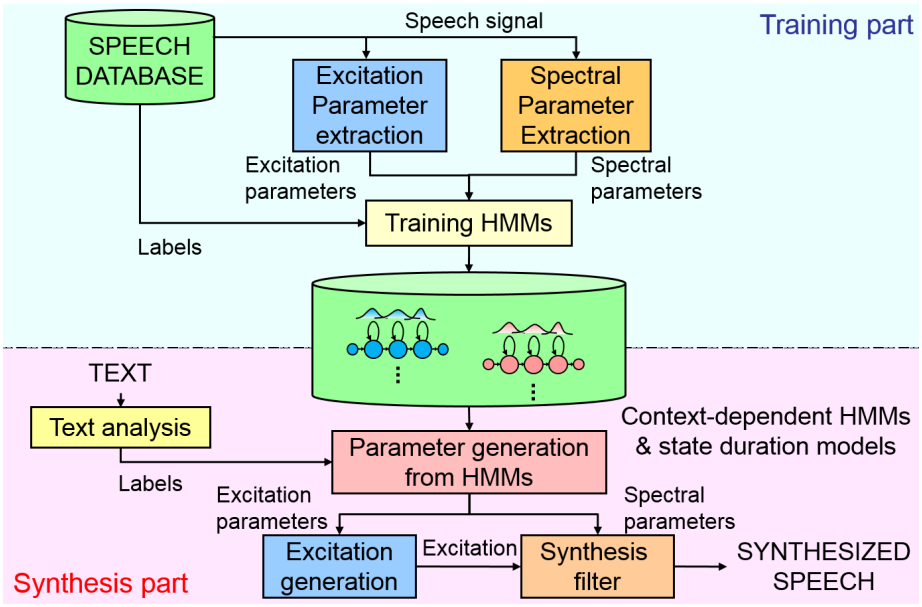

Figure 11: 基于HMM的语音合成系统框架图。图片来自HTS slides。

5 神经模型合成

在讨论语音识别时我们已经知到，HMM模型不能描述复杂的发音现象。当数据量增加时，基于HMM的语音合成系统的性能遇到了瓶颈。为此，研究者提出基于深度神经网络（DNN）的语音合成方法。这一工作由香港中文大学、微软、Google于2013年提出 [3, 4, 5, 6]。在这些模型中，研究者用神经网络取代HMM模型来预测每一帧语音的激励和调制信号，再通过声码器合成自然语音。

图12是Google的DNN合成系统示意图。该系统的前端处理部分和HMM系统是一样的，不同的是用DNN取代HMM来生成每个音素的的时长、基频和频谱包络。事实上，Google系统中生成的是这些量的均值和方差，基于这些统计量，可利用HMM系统类似的方法，通过考虑前后帧之间的相关性生成每一帧的实际参数，即图中的“Parameter Generation”部分。生成的参数最后还是要通过一个声码器来合成语音。

2014年以后，研究者对基于DNN的合成方法进行了一系列扩展，提出了基于RNN的合成系统。和DNN相比，RNN可以学习长时相关性，因此可以用来学习发音过程中前后音素之间的协同发音特性，从而得到更平滑自然的发音。图 13是微软发表的基于RNN的合成系统示意图 [7]。该模型首先生成每个音素

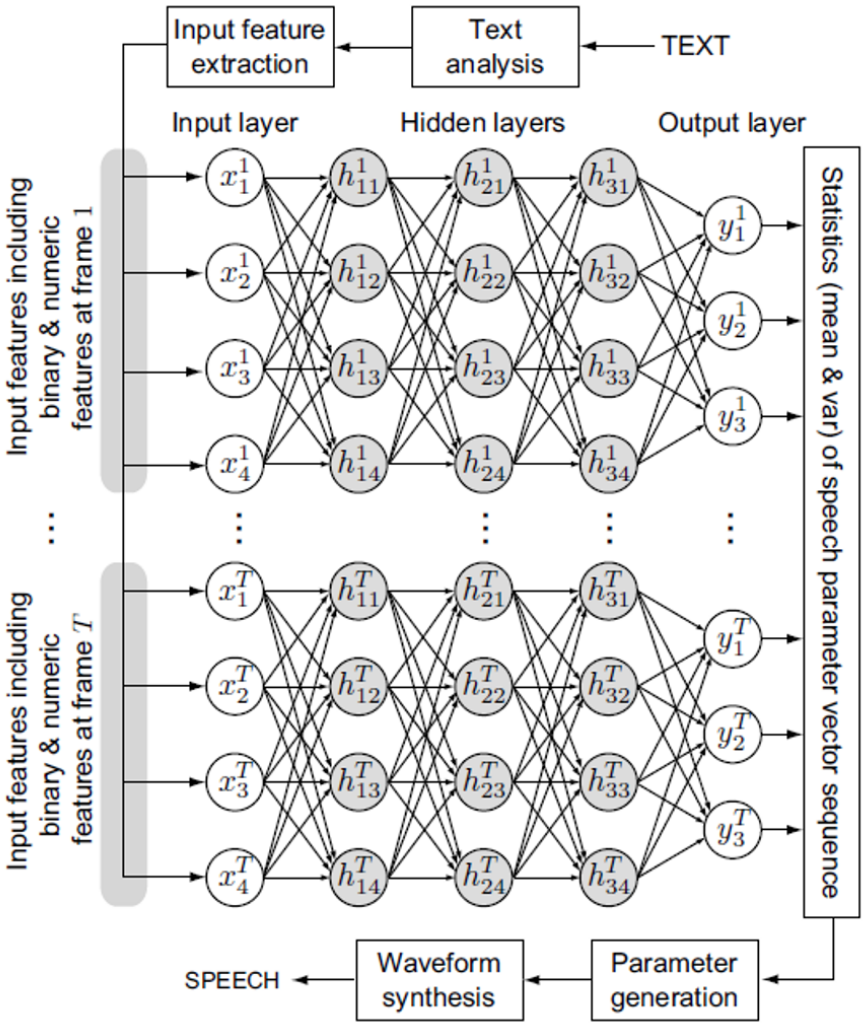

Figure 12: Google基于DNN的语音合成系统。该模型通过DNN来预测每个音素的时长、基频和频谱包络的均值和方差，基于此生成每帧的参数，最后送入声码器合成语音。图片设计源自 [6]。

的长度，然后基于一个双向RNN模型对每个语音帧直接预测基频和频谱包络参数。因此，该系统并不需要一个额外的参数生成器，由音素参数的统计量生成每帧的参数。

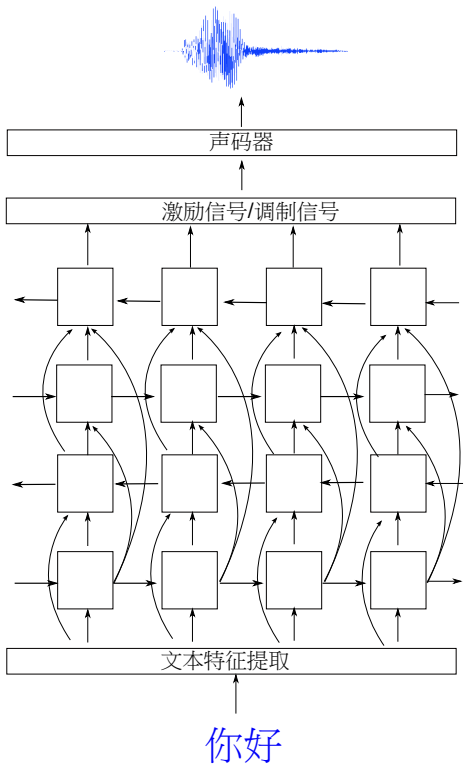

Figure 13: 微软基于RNN的语音合成过程。该模型基于多层双向RNN对输入句子进行映射，预测每一帧语音的激励和声道调制信号，最后通过声码器合成声音。图片设计源自 [7]。

6 基于注意力机制的合成系统

近两年来，基于神经模型的语音合成取得长足进展，其中基于注意力机制的语音合成系统最受瞩目 [8]。直观上，基于注意力机制的语音合成可类比人类的阅读-理解-复述过程。首先，用一个双向RNN将要发音的文本编码，这类似于阅读和理解；之后基于另一个RNN逐一生成每个语音帧，类似于对脑海中的记忆进行复述。在每一步生成时，系统基于注意力机制定位到要发音的文本，并利用该文本的信息指导生成过程。

图14是Google基于该思路设计的Tactoron系统。在该系统中，待合成的句子以字符串的形式输入到一个编码器，生成一个完整的句子编码。在合成时，基于一个RNN迭代合成每一帧参数，每合成一帧的时候，不仅将前一帧输出作为输入，同时基于Attention机制提取句子编码中相应的信息，使得生成的语音

与输入要求相一致。特别值得说明的是，这一方法的解码输出并不是基频和频谱包络，而是直接生成频谱。基于频谱可以直接合成原始语音，而不再需要一个声码器。一种常用的由频谱合成语音的算法是Griffin-Lim算法，该方法首先由频谱估计出相位，再将相信应用到频谱上，合成语音。Griffin-Lim算法简单方便，但性能不高。Google提出利用多层卷积网络由频谱直接合成语音，这一网络称为WaveNet[9]，如图15所示。

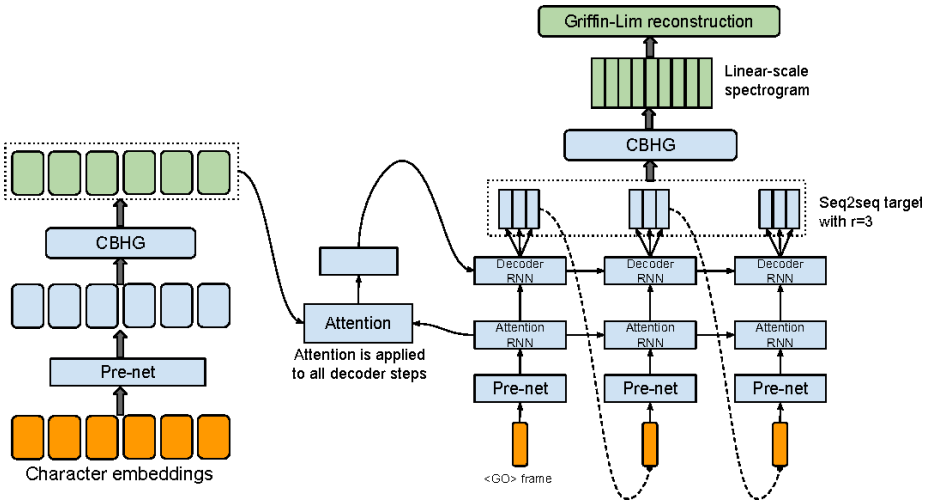

Figure 14: 基于注意力机制的Tacotron合成系统。图片源自 [8]。

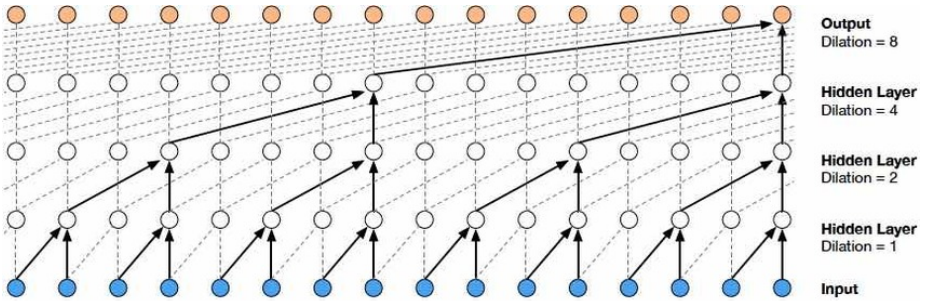

Figure 15: WaveNet结构。图片源自 [9]。

Tacotron已经成为语音合成系统的主流模型，可以合成非常自然的声音。人们对这一模型进行了一系列扩展，例如，DeepMind提出了一种并行WaveNet算法，可以使WaveNet的合成速度达到实时 [10]，Google将说话人信息表达成一个说话人向量，作为辅助输入，使得Tacotron系统可以合成不

同人的声音[11]。

7 小结

本节我们简要介绍了语音合成的基本方法。绝大部分合成方法基于激励-调制模型，该模型认为语音是由声门激励经过声道的调制过程，因此，只要设计合适的激励和声道响应函数，即可合成需要的声音。最早的语音合成系统基于共振峰参数生成声道响应函数，基于此合成语音。这一方法效率很高，但由于参数过于简单，质量不能保证。拼接法从大规模数据库中直接提取声门和声道参数，可以实现较逼真的合成，但在拼接点有不连续情况，且占用系统资源较多。统计模型法是一种对共振峰合成的回归，但利用了更多信息和更复杂的概率模型对声门和声道的参数进行预测。神经模型法用更复杂的神经网络代替HMM等统计模型，当训练数据量较大时可以突破统计模型在概率形式上的限制，因而合成更自然流畅的声音。近两年来，基于神经模型的合成系统进一步发展，特别是以Tacotron为代表的端对端系统，将待合成的句子直接映射到频谱，事实上已经摆脱了传统合成方法中对激励-调制模型的依赖，成为新一代合成系统。

References

[1] A. W. Black, "Perfect synthesis for all of the people all of the time," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis.* IEEE, 2002, pp. 167–170.

[2] K. Tokuda, K. Oura, K. Hashimoto, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, and A. Black, "The HMM-based speech synthesis system," *Online: http://hts. ics. nitech. ac. jp*, 2007.

[3] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *2013 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).* IEEE, 2013, pp. 8012–8016.

[4] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 3829–3833.

[5] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7962–7966.

[6] H. Zen and A. Senior, "Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis," in *2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3844–3848.

[7] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014, pp. 1964–1968.

[8] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint*, 2017.

[9] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[10] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[11] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.