志玲姐姐的导航声音是如何产生的？

开车走在路上，听着导航里志玲姐姐给你带路，是件非常愉快的事。这么甜美的导航声音是如何产生的呢？今天我们就聊一聊背后的语音合成（Speech Synthesis）技术。

让机器像人一样开口说话是人们很早就有的梦想，但真正的规研究还是在 Leonhard Euler 在 1750 年左右建立了声音的物理学原理之后。1769 年，Wolfgang von Kempelen，发明了一个模拟人类发音器官的发音器。如图 1 所示，这个发音器包括一个风箱来模拟人的声带，一个共鸣腔来模拟人的口唇。这种用机械模拟人发音的方法能产生和人类似的发音，是语音合成的初期探索[1]。

图 1：保存于德国 Saarland 大学的 Kempelen 发声器复制品。

1930 年， Bell 实验室发明了声码器（Vocoder），将人的声音分解成声带振动和口唇调制两部分，改变口唇部分的调制函数后，就可以合成出不同的声音。这种合成方式物理学基础明确，系统简单，在 80 年代很受欢迎。著名物理学家霍金的轮椅就是采用这种方式发声的。这种合成方式的缺点在于发音的机器味道很浓，流畅度也不够。用这种方法是无法生成志玲姐姐的声音的。

90 年代，人们采用更粗暴的方式来合成声音。研究者让播音员录制一个大规模声音库，然后从声音库中选出声音片段来，拼接成所要的句子。比如要合成“我想回家”，就在声音库里找到“我”、“想”、“回”、“家”这四个字对应的发音，再把他们拼成一句话。假如这个声音库是志玲姐姐录的，那我们就可以合成她的声音了。这种拼接法里最重要的事是选择合适的发音片段，因为同一个音节在不同环境下的真正发音是不太一样的，要选出最合适的发音片段并不容易。同时，为了拼出的声音更自然，质量更高，声音库自然是越大越好，因此需要大量录制工作。

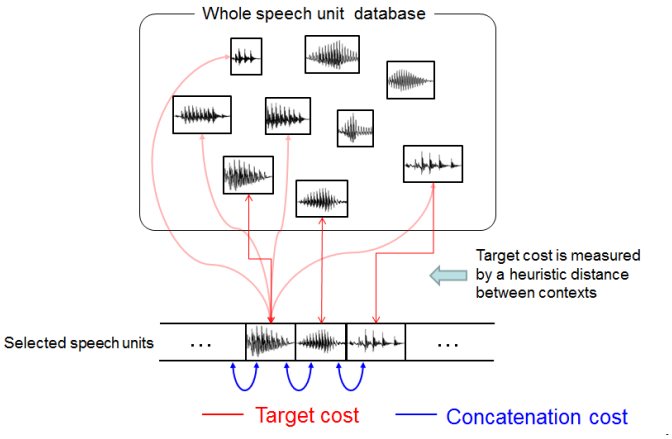

图 2：拼接法从数据库中选择声音片段进行合成 [2]

拼接法的一个缺点在于声音不容易改变。比如，我想换个人说话，就需要重新录制数据，如果换个情绪说话，还需要录制这个人在特定情绪下的声音，工作量太大了。研究者提出统计模型方法来解决这个问题。和拼接法不同，统计模型方法对每个发音构造一个统计模型，这样只要调整模型参数就可以得到新的发音，而这种参数调整只需要很少的数据。如果采用这种方法，只要请志玲姐姐读个几分钟就可以合成她的声音了。

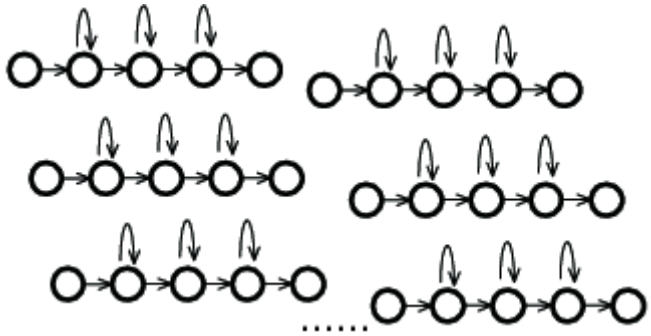

图 3：统计模型法将每个发音表示成一个概率模型

近年来，深度学习方法成为主流。和统计模型方法相比，深度神经网络对发音过程有更精细的刻画，因此可以合成非常自然逼真的声音。图 4 是 Google 发布的一个基于深度神经网络的合成模型，该模型将需要合成的句子通过一个序列到序列模型直接生成发音。因为在发音时对前后发音的相关性有细致的建模，这一模型可以生成很自然的发音。特别是，如果给这个模型输入一个表示发音人的向量，就可以随时随地改变发音的说话人特性了。如果用这个模型，志玲姐姐也许只要录几句话，就能帮我们导航了。不仅如此，基于深度学习，人们还可以控制发音的口音、情绪、语速、音调等各种参数，甚至造出虚拟人的声音。可以说，人们长久以来让机器开口说话的理想已经成为现实。

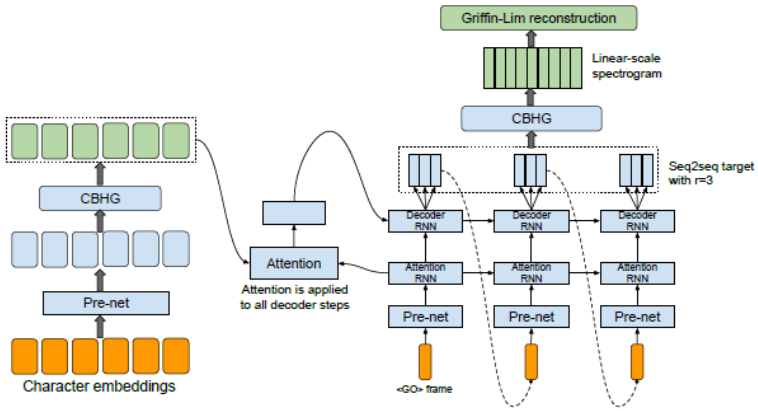

图 4：Google 发布的基于深度学习的 Tacotron 语音合成系统[3]

1. 王东，人工智能，清华大学出版社，2019.10.

2. HTS Slides, released by HTS Working Group, http://hts.sp.nitech.ac.jp/

3. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S. and Le, Q., 2017. Tacotron: Towards End-to-End Speech Synthesis. Proc. Interspeech 2017, pp.4006-4010.