机器如何实现听音辨人？

每个人的声音各有不同，就算是双胞胎，也具有各自不同的特性。我们都有这样的经验，接电话时，如果是自己熟悉的人，只要对方“喂”一声，就能判断出对方是谁。图 1 给出两个人发同一个字“绿”得到的声音频谱。可以看到，同样的声音由不同人发出来细节有很大差差。这些发音细节和说话人直接相关，因此通常被形象地称为“声纹”。

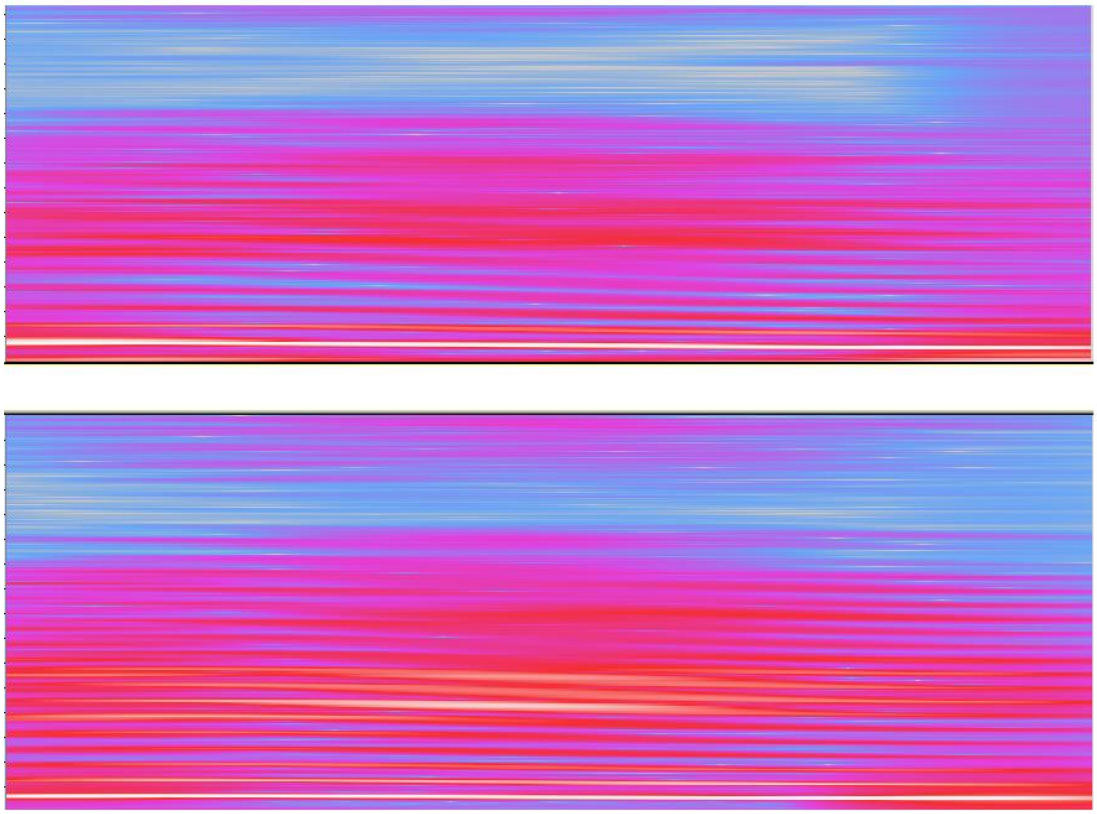

图 1：两个人发“绿”时的声音差异

利用不同人在发音上的个性化差异可以对身份进行验证，这一技术称为“声纹识别”。声纹识别有一些独特的优势，如验证方便，无需接触，隐私暴露较少等，还可以结合发音内容来确认验证人的真实意图。声纹识别有广泛应用，例如在手机银行转账时，加入声纹验证可极大降低账户被盗用风险；在智能加电中加入声纹认证可以让它表现的更智能（如可以按使用者偏好调节运行方式）；在刑侦破案中，声纹可以用来对嫌疑人进行筛查，或作为辅助证据确认嫌疑人。

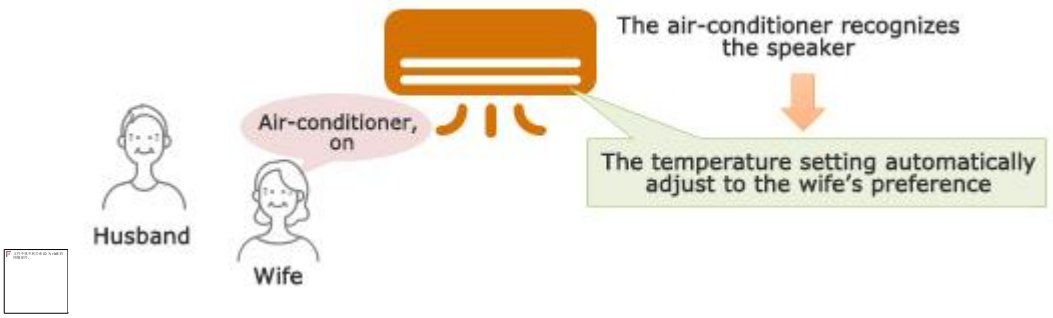

图 2：空调上加入声纹识别可以让空调认“主人”。[1]

如何实现声纹识别呢？传统方法一般采用特征提取-统计建模方案：首先提取和说话人相关的显著特征，再基于这些特征对每个人建立统计模型。近年来，研究界多采用深度学习方法。这种方法通过收集大规模人群的发音数据，通过训练一个深度神经网络来提取与说话人相关的显著特征。这种方法一般具有更好的抗干扰能力，在实际应用中表现出更优越的性能。图 3 给出基于深度学习方法得到的不同发音人的特征，其中每种颜色代表一个发音人，每个点代表一个句子。可以看到，这一方法可以实现对不同发音人的较好区分。

目前，声纹识别技术已经有一些商业化应用，但总体来说性能还有待加强，特别是复杂环境下（如远场、噪音、跨领域）的识别效果还有较大差距[3]。如果将声纹识别和其它生物认证技术（如人脸识别）进行结合，则有望显著提高认证的可靠性。

图 3：利用深度神经网络可以对说话人进行较好区分。[2]

[1] https://www.toshiba.co.jp/rdc/rd/detail_e/e2002_01.html
[2] Yunqi Cai, Lantian Li, Andrew Abel, Xiaoyan Zhu, Dong Wang, "Deep Normalization for Speaker Vectors", IEEE Transactions on Audio, Speech and Language Processing, 2020.
[3] Li L, Liu R, Kang J, et al. CN-Celeb: multi-genre speaker recognition[J]. arXiv preprint arXiv:2012.12468, 2020.