指数下降的学习率。目前最常用的是自适应的学习率，包括 Adagrad [73]、AdaDelta [74] 及 Adam [75] 等算法。

3.4.2 声纹识别中的神经网络

神经网络在声纹识别中的应用主要分为两种：间接用于声纹识别和直接用于声纹识别。前者指的是，在传统的声纹识别系统中，将一部分模块用神经网络来代替，但整体框架依然保持不变；而后者指的是，将整个声纹识别问题作为一个深度学习问题来处理，而不依赖于高斯混合模型、因子分析或概率线性判别等其他方法。

1. 串联深度特征

串联深度特征（tandem deep features）[76] 是上海交通大学于 2014 年提出的一种将神经网络间接用于声纹识别的技术。这种方法依然采用 3.3.3 节介绍的高斯混合模型-通用背景模型作为声纹识别的基本框架。与之不同的是，传统的高斯混合模型-通用背景模型的输入为 MFCC 或 PLP 等音频特征；而串联深度特征方法则利用多种不同类型的神经网络作为特征提取模块，提取出更高级的特征，与音频特征拼接后再作为高斯混合模型-通用背景模型的输入。

参考文献 [76] 中采用了三种神经网络来提取高级特征。第一种是受限玻尔兹曼机（restricted Boltzmann machine，RBM），这是一种非监督式的神经网络。通过这种网络获得的特征可能包括音素（phoneme）、说话人和信道等多方面的信息。第二种是用于将三连音素（triphone）进行分类的监督式前馈神经网络。第三种则是用于将说话人进行分类的监督式前馈神经网络。这三种神经网络中的后两种监督式神经网络的训练，都会用到 3.4.4 节将要介绍的交叉熵损失函数。

2. 基于神经网络的 i-vector

基于神经网络的 i-vector [77] 是另一项于 2014 年提出的将神经网络间接用于声纹识别的技术。这种方法的基本框架依然是 3.3.6 节介绍的 i-vector 方法。其差别在于，传统的 i-vector 方法基于高斯混合模型-通用背景模型；而参考文献 [77] 中描述的框架则将高斯混合模型-通用背景模型替换为用于进行语音识别的神经网络。

这里提到的用于语音识别的神经网络（ASR-DNN），非常类似于“串联深度特征”一节提到的参考文献 [76] 中的第二种网络。这种网络对音频特征帧进行分类，而分类的目标则为各种三连音素。高斯混合模型可以看作对音频特征的分布的一种概率描述，这里三连音素的个数对应着高斯混合模型的高斯分量的个数；类似地，用于语音识别