一种极短时说话人确认方法

背景

说话人确认是通过声音特征对说话人进行身份验证的方法。例如在银行应用中，用户先预留一段声音，当该用户再次使用银行服务时，系统提示输入验证语音。将验证语音与系统预留语音进行对比，即可判断该用户是否真实。说话人确认在现代信息社会中具有重要意义， 是识别行为人身份，保证社会安全的重要方法，具有广阔的市场前景。

技术综述

当前说话人确认方法以统计模型为主。当前最好的识别系统基于 i-vector 模型和 PLDA 模型。I-vector 模型对语音信号建立如下线性模型：

$$X = Tw + v$$

其中,X 为语音信号的 MFCC 特征，T 为一个低秩矩阵，w 为句子向量，即 i-vector，v 为高斯噪声。该模型事实上是一个概率 PCA 模型。实际应用中，一般将语音空间分成若干区域，对每个区域进行上述线性建模，但共享句子向量 w。因此，w 是该语子的一个低维向量表示。因为 w 中包含说话人、说话内容、信道等一系列信息，为提高对说话人的区分性，引入 PLDA 模型如下：

$$w = Hu + Kc + n$$

其中 u 是说话人向量，c 是表达向量，包括发音方式，信道等，n 是高斯噪声。通过 PLDA，将说话人特征和表达特征区分开，从而极大提高说话人识别的性能。

上述模型基于通用的 MFCC 特征，通过模型将说话人信息分离出来。该方法具有如下缺点：

1. 因为该方法基于信号的分布状态建模，因此需要较多的数据才能得到较好的结果。
2. 该方法要进行 i-vector 提取，PLDA 计算，计算量较大，且计算方式不易优化
3. 该方法受到信道、噪声、时变的影响大

本发明的提出一种基于深度神经网络的特征学习方法解决上述问题。

发明内容

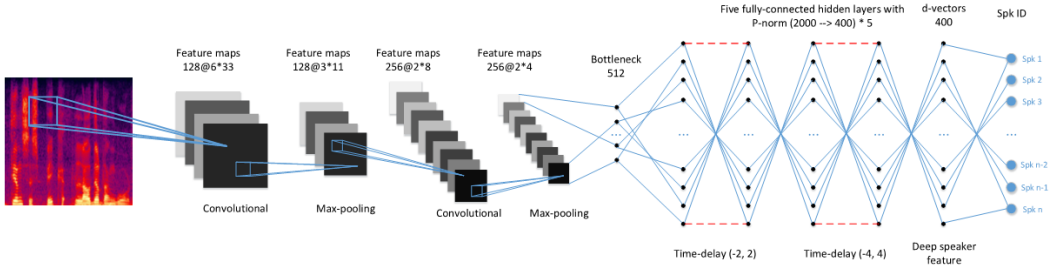
简述

当前基于 i-vector/PLDA 方法之所以具有上述各种缺陷，根本原因在于所用的语音特征（MFCC）是一种对说话人不具有区分性的特征，因此必须通过复杂的概率模型来实现对说话人因子的分离。本发明提出一种基于深度神经网络的特征学习方法。基于该方法，我们直接提取每帧信号中的说话人信息，基于该信息，可用非常简单的方法实现对发音人身份的判断。

说话人识别领域已经有基于深度神经网络的特征提取研究，然而，当前发表的方法都无法实现有效的特征提取，因此都需要后接一个统计模型，这事实上并不能避免 i-vector/PLDA 方法的缺陷。本发明所提出的模型结构极大提高了特征学习的质量，在帧级别（200 毫秒）即可实现判别精度 83%。

结构

本发明所述神经网络模型包括六部分：语音表示、卷积特征提取、降维、时延网络、说话人特征提取、说话分类。基于该原则的任何网络都可实现。下面是一个已经实现的网络结构如下。



（1） **语音表示：**我们采用传统的频谱作为对一段语音的时频表示。

（2） **卷积积特提取：**这一部分包括两个卷积层，每个卷积层后接一个最大池化层。第一个卷积层的卷积核为 128 个，每个卷积核的大小为 6x33；第一个池化层的池化窗口大小为 3x11。第二个卷积层的卷积核为 256 个，每个卷积核的大小为 2x8。第二个池化层的池化窗口大小为 2x4。

（3） **降维：**第二个池化层得到的 256 个特征平面被降维为 512 个神经单元。

（4） **时延网络：** 时延网络包括两个全连接层，每层通过时序拼接对上下文信息进行扩展。第一个时延层拼接前后各两帧信号，第二个时延层拼接前后各 4 帧信号。每个时延层后接一个 P-norm 层，以降低维度。每个 P-norm 层将延时层的大量输出降维为 400。

（5） **说话人特征提取：**将第二个时延层 P-norm 之后的 400 维输出经过一个线性变换得

到说话人特征。

（6） **说话人分类：**基于提取到的说话人特征，利用一个 softmax 分类模型对说话人进行分类。

训练

该网络训练时，将训练集中的所有说话人作为网络输出结点，以 xEntropy 作为训练目标函数，利用后向反馈算法进行学习。学习时，每一个语音帧是一个学习样本，学习可采用 NSGD 算法或任何深度神经网络学习方法。

验证

验证时，将验证语音通过所述神经网络前向计算，在特征提取端得到每一帧的说话人特征，基于该特征可利用任何统计模型进行对说话人进行判断。一个简单的方法是对该句中的所有特征取平均， 用简单的距离计算即可得到信任度。

发明优势

1. 性能更好。我们实验证明基于该特征，即使用最简单的加合方法，也能达到很好的识别效果。
2. 计算更快。本发明提出的发法在进行特征提取时仅包括神经网络计算，计算简单，且可通过各种软硬件加速方法实现更快计算。
3. 鲁棒性更强。通过大数据学习，得到的神经网络鲁棒性很强，因而可扩展到多种场景中。
4. 可应用于极短语音验证环境。该方法对极短语音特别有效，这在众多应用场景中非常重要。我们发现，即使用一个特征帧（包含上下文信息后，累计时长为 200 毫秒），也可以达到 83%的判断精度。