基于 X-vectors 的鲁棒的语种识别技术研究

于嘉威　解焱陆　张劲松

摘要 随着国际合作的日趋深入，语种识别技术在多语言语音处理系统中的作用越发重要。语音处理系统只有先识别出语音所属的语言种类，才能将不同语言的语法规则适用在该语音上，以完成系统的后续任务，因此如何有效提升语种识别系统的精确度，一直是研究重点。在这篇论文里，我们将 x-vectors 应用于语种识别任务，在这种结构里，包含跨时间信息的长时语种特征将会被网络结构里的时间池化层（temporal pooling layer）捕捉到。一旦抽取到 x-vectors，便可以像 i-vectors 那样采用同样的后端分类技术进行后续处理。在 2017 年的东方语种识别竞赛（AP17-OLR）中，x-vectors 的结果好于基线 i-vetors 系统，x-vectors 系统的等错误率（EER）为 5.13%，比 i-vectors 系统提高了 12.5%。最后，我们还实验了数据增强技术在 x-vectors 框架下的作用，发现使用数据增强技术（data augmentation）对系统的精确度有提升。

关键词 语种识别，深度神经网络，数据增强，I-vectors，X-vectors

Robust Language Identification using X-vectors

YU Jiawei　XIE Yanlu　ZHANG Jinsong

Abstract With the closer international cooperation, language recognition technology plays an increasingly important role in multilingual speech processing system. In order to complete the subsequent tasks, the speech processing system can only apply the grammar rules to the speech utterances by recognizing the language type initially. Hence how to effectively improve the accuracy of language recognition system has been the focus of researchers. In this paper, we apply x-vectors to the task of language identification. In x-vectors' framework, long-term language characteristics are captured in the network by a temporal pooling layer that aggregates information across time. Once extracted, x-vectors utilize the same classification technology developed for i-vectors. In the 2017 oriental language recognition (OLR) challenge, x-vectors' results shows improved performance compared to results of i-vectors baseline system, achieving 12.5% relative equal error rate (EER) improvement. In the post-evaluation analysis, we experiment on data augmentation with the x-vectors framework and find that the performance of language identification system is improved.

Key words Speaker identification, Deep neural networks, Data augmentation, I-vectors, X-vectors

1. 引言

现代的语种识别系统大多是基于 i-vectors[2]的，该方法围绕样本的高斯混合模型（Gaussian mixture model，GMM）[14]超向量与均值超向量之间的差异，将每个样本视为独立的个体，训练得到每个样本之间的全差异空间，然后得到样本之间差异的低维表示称之为 i-vectors,之后通过线性区分性分析(Linear discriminant analysis，LDA)或者类内协方差规整等技术对 i-vectors 进行类内类间差异补偿和降维，再采用 SVM 或者余弦距离来进行建模打分。

近几年, 使用深度神经网络（deep neural network, DNN）[7]来捕捉语种特征信息是目前十分活跃的研究领域：一方面从特征层面, 利用其强大的特征抽取能力,提取深度瓶颈特征(deep bottleneck feature, DBF)[21]；另一方面从模型域出发,提出了基于深度神经网络的总变化量因子分析（Total variability factor analysis，TV）建模策略[9]。此外,2014 年谷歌的研究人员将特征提取、特征变换和分类器融于一个神经网络模型中,提出了端到端的语种识别系统[10]。2016 年基于注意力的信号机制被引入到语种识别系统中[3]。

在语种识别领域，上面提到的基于深度神经网络的传统方法，对语种进行的分类通常是在帧层面的[10,4,5,6]，但在测试阶段，通常是在句子层面来计算打分，因此，在这篇论文里我们将应用在说话人识别领域里的 x-vectors 框架[15,16]适用在语种识别的任务里，这种框架结构会训练深度神经网络将一个变长的语音段映射到一个定长的嵌入层中，将语种在句子层面来进行分类，这样从中抽取出的包含更多时间信息的特征就称为 x-vectors。一旦抽取到 x-vectors，后端的处理便可使用在 i-vectors 上已经得到成熟运用的后端分类和打分技术。

后续论文的组织如下，第 2 节介绍 x-vectors 系统结构及数据增强技术的应用，第 3 节介绍实验的设置安排，第 4 节会给出实验结果与分析，最后第 5 节是结论与展望。

2. x-vectors 系统及数据增强技术

2.1 x-vectors 系统结构

x-vectors 系统最早是应用于说话人识别中的。X-vectors 系统包含一个前向深度神经网络，该神经网络将变长的语音段映射到一个定长的嵌入层，从该嵌入层中提取出的特征向量就称为 x-vectors。它和传统的说话人识别系统不同的地方就是加入了段（segment）层级的学习，而不单单是在帧层级。一旦提取到 x-vectors，它就将被后端的高斯分类器进行区分性的训练。

训练 x-vectors 的神经网络结构图如图 1 所示[15]。在时间池化层之前的结构是时延神经网络（Time delay neural network, TDNN）[12]，这一阶段是在帧层级上进行操作的，TDNN 的输入是一段语音，每次 TDNN 都取固定的帧数，这样一层一层向上传输，送入时间池化层，在这之后神经网络便在段层级上学习。时间池化层将 TDNN 的输出向量积累下来后，计算其均值与标准差作为该层的输出。时间池化层之后连接了两层全向连接层，x-vectors 就是在该层中提取出来的，最后加一个 softmax 层作为输出。

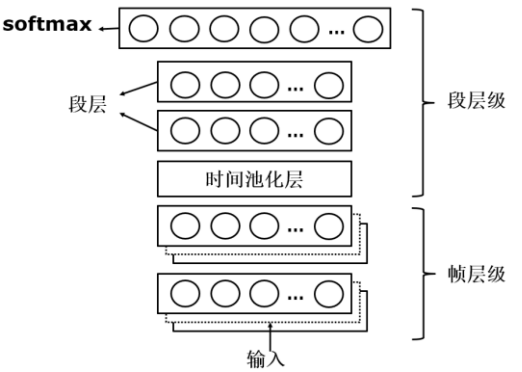

图 1：训练 x-vectors 的神经网络结构图

这样的网络结构可以帮助语种识别系统获得包含更多时间信息的特征，从而提高识别系统对语种进行分类的能力。

2.2 数据增强技术

在过去很长的一段时间里 i-vectors 语种识别系统的性能表现一直是优于神经网络识别系统，但随着互联网的发展以及硬件计算资源的提高进步，研究人员可以收集到更多的、高质量的数据，神经网络的威力便开始显现出来。[20]的研究表明，当有足够多的高质量的语音数据时，使用嵌入式（embeddings）神经网络的系统性能是好于 i-vectors 系统的。因此我们在这里使用了数据增强技术，以增加 x-vectors 神经网络的训练数据量和数据多样性，同时提升系统的鲁棒性。

3. 实验设计

3.1 语料简介

表 1: AP17-OLR 训练、测试集分布

AP17-OL10			AP17-训练集			AP17-测试集		
语种标识	详细描述	语音信道	话者个数	每个话者说的句子个数	句子总数	话者个数	每个话者说的句子个数	句子总数
ka-cn	哈萨克语	电话	86	50	4200	86	20	1800
ti-cn	西藏语		34	330	11100	34	50	1800
uy-id	维吾尔语		353	20	5800	353	5	1800
ct-cn	广东话		24	320	7559	6	300	1800
zh-cn	普通话		24	300	7198	6	300	1800
id-id	印尼语		24	320	7671	6	300	1800
ja-jp	日语		24	320	7662	6	300	1800
ru-ru	俄语		24	300	7190	6	300	1800
ko-kr	韩语		24	300	7196	6	300	1800
vi-vn	越南语		24	300	7200	6	300	1800

3.1.1 AP17-OLR

本实验语料取自 2017 年东方语种识别竞赛，该竞赛是由清华大学和海天瑞声联合举办[19]。语料由海天瑞声提供，共含有 10 种语言，语音的信道来源是传统的电话信道，频率为 16kHz。每一种语言的时长大约为 10 个小时，男女的性别比例为 1：1。训练集和测试集的划分如上面表 1 中所示。

3.1.2 数据增强语料

对语音数据加混响使用的语料取自 Ko 等人的 RIRs（room impulse responses）[8]；对原始语音数据加各种噪音的数据来自 MUSAN 语料库[17]，它包含超过 900 个噪声，42 小时的各种类型风格的音乐以及 60 个小时的 12 种不同语言的语音。为了满足语种识别任务的条件，这里我们仅使用噪音和音乐，而不使用它的语音数据。

3.2 前端特征提取

3.2.1 MFCC 特征

实验中，x-vectors 系统提取的是每一帧长为 25ms 的 23 维美尔频率倒谱系数（Mel Frequency cepstrum coefficient, MFCC）学特征，在送入 x-vectors 的深度神经网络之前，我们对特征使用了长度为 3s 的滑动窗口进行均值归一化处理，同时声学特征中的非语音特征帧都使用语音活性检测技术（Voice activity detection，VAD）[18]移除掉了。

3.2.2 VAD

VAD 是一项用于语音处理的技术，目的是检测语音信号是否存在，也可用于去除语音段中非语音的片段。

执行 VAD 的算法有很多种，对于在不同噪声环境下的语种识别任务，我们采用鲁棒性更好的频谱相减的方法以鉴别语音和非语音帧。频谱相减的优点是其执行计算的复杂度低，所以可以在较低的计算资源环境下工作。具体算法计算公式如下[18]：

$$\Omega \Rightarrow \begin{cases} T(i) > \max(T(j)_{j=1,2,\ldots,M}) - T_{SNR} \\ T(i) > T_{min} \end{cases} \quad (1)$$

其中$T(i)$表示第i帧语音信号经过噪声频谱相减后的数值。max()操作表示找到句子中所有帧频谱相减的最大值。T_{SNR}是刚刚max()函数找到的能量最大点处的降速值(downshift value)，T_{min}是判定某一帧语音是否是语音片段的最低阈值。未经处理的脉冲编码调制（Pulse-code modulation，pcm）数据的范围应该在-1 到 1 之间。在评估阶段，T_{SNR}和T_{min}分别取 46 分贝和-65 分贝。详细细节可以在[18]看到。

3.3 x-vectors 语种识别系统设置

3.3.1 深度神经网络架构

表 2: x-vectors 的深度神经网络结构（输入层是 F 维的特征，L 表示训练集语种的个数）[16]

神经网络层	网络层上下文	当前帧数
帧层 1	$[t-2, t+2]$	5
帧层 2	$\{t-2, t, t+2\}$	9
帧层 3	$\{t-3, t, t+3\}$	15
帧层 4	$\{t\}$	15
帧层 5	$\{t\}$	15
时间池化层	$[0, T)$	T
段层 1	$\{0\}$	T
段层 2	$\{0\}$	T
softmax 层	$\{0\}$	T

深度神经网络的设置如表 2 中所示，神经网络的输入是连续的 T 帧语音信号。神经网络的前五层处理的是帧层级的输入，这五层带有小的时间上下文信息被集中在当前帧 t 附近。例如，输入层，帧层 1，连接了 F 维的特征，包括 t-2，t-1，t，t+1，和 t+2 这些帧，共给出了 5 帧的上下文信息。这时，在下一层，帧层 2 结合了第一层在 t-2，t 和 t+2 处帧层 1 的输出，这里的时间上下文信息建立在上一层的基础上，所以帧层 2 可以包括 9 帧的上下文信息。这样的过程在接下来的每一层中都重复着，最后在帧层 5 总共可以看到 15 帧的上下文信息。

时间池化层将上面得到的信息打包成一个个语音片段（segment），这样后续的神经网络层便可以对该语音片段进行操作。这里数据池化层的输入是神经网络在前面几层处理后得到的一个 1500 维的向量序列，输出是这 1500 维向量的均值以及标准差。输出的这些数据串联在一起，并且通过段层 1 和 2，最后送入 softmax 层。激活函数使用的是修正线性单元（rectified linear units，ReLUs）。

3.3.2 数据训练

神经网络训练语种的分类使用了公式(2)所示的多层交叉熵目标函数[15]。这里的数据训练与以往的语种分类训练过程不同之处是，以往的训练是在帧层级上的，而 x-vectors 使用的是变长的语音特征段。

假设有 L 种语言在 N 个训练特征段里，其中$P(lang_l | x_{1:T}^{(n)})$是语种l在给定的矩阵 T 输入帧$x_1^{(n)}, x_2^{(n)}, \ldots\ldots, x_T^{(n)}$后的概率。如果语音特征段n的语种标签是l，则$d_{nl}$的值是1，否则其值为0。

$$\mathrm{E} = -\sum_{n=1}^{N}\sum_{l=1}^{L} d_{nl} log P(lang_l \mid x_{1:T}^{(n)}) \qquad (2)$$

训练的例子是从训练数据语音中选取 2-4 秒时长的语音段，同时训练的标签是给相应的语音段打上语种标签，网络会经过若干次迭代，训练使用随机梯度下降法。

3.3.3 数据增强

实验中我们对原始“干净”的语音数据，进行 4 种不同的数据增强策略，再加上原始的语音数据，增强后我们得到 5 倍于原数据的数据量，对一段语音的增强，我们有以下四种数据增强的方式可以选择：

速度扰动：对原语音数据进行 0.9 倍速度的加速，或 1.1 倍速度的减速。

音乐：这里我们选用 MUSAN 语料，这个语料是一个单个的音乐文件，不包含人声。我们从这个音乐文件里随机的切出一段适合原语音时长的音频文件加入到原始语音中（5-15 分贝的信噪比）。

噪声：同样选用 MUSAN 语料，把其中的噪声填进整个原始语音中（0-15 分贝的信噪比）。

混响：使用模拟房间脉冲响应（simulated RIRs）来对原始训练语音数据加入混响。

3.4 基线系统设置

3.4.1 i-vectors 系统

i-vectors 基线系统的输入特征采用的是 19 维的 MFCCs，然后对 MFCCs 特征做一阶、二阶差分，得到一个 60 维的特征向量。通用

背景模型（universal background mode，UBM）中包含 2048 个高斯混合单元，同时 i-vector 的维数是400。抽取到i-vectors特征向量后，后端使用 LDA 模型[13]，来提升得到的各语种信息之间的区分度，LDA 投影的维度是 9。

LDA 常被用于统计、机器学习和模式识别领域中，目的是在二分类或多分类问题中找到特征的线性组合关系。简单来说，LDA 的目标就是找到一个合理的投影矩阵 W，使得数据经过投影之后的类间(Between-class)协方差最大而类内(Within-class)协方差最小[11]。

通过LDA进行噪声补偿并降维得到的具有区分性的语音段表示后，便可以直接计算其余弦距离并得到识别结果。余弦距离的数学表达式如下:

$$\cos(\omega_{test,\mu_l}) = \frac{\omega_{test}^T \cdot \mu_l}{\|\omega_{test}^T\| \cdot \|\mu_l\|} \quad (3)$$

其中ω_{test}是待测语音段 i-vectors 投影后的矢量；μ_l是属于语种l的语音段 i-vectors 经过投影后的均值向量。在计算得到待测语音在所有语种上的余弦得分后，就可以对系统性能进行评价。

3.4.2 DNN 系统

我们设计了两个深度神经网络的基线系统，一个是传统的延时神经网络（TDNN），另一个是长短期记忆循环神经网络(LSTM-RNN)。这两个系统的主要区别是 TDNN 只能静态的捕捉输入点处的语种特征信息，而 LSTM-RNN 可以动态的捕捉到输入点处前后的语种特征信息，所以 LSTM-RNN 会比 TDNN 在结果预测的鲁棒性上表现更好，但 TDNN 的时间复杂度会更低，也更易于训练。

这两个基线系统都使用未经预处理过的 40 维的 MFCC 特征，延时神经网络采用一个对称的 4 帧滑动窗口来读取 MFCC 特征，长短期记忆循环神经网络(LSTM-RNN)采用对称的两帧时间窗来读取 MFCC。对于 TDNN 语种识别系统而言，有 6 个隐藏层，每一层有 2048 个节点，激活函数是 p 范数（p-norm）；LSTM 每一层设置的节点数为 1024。

4. 实验结果及分析

我们设置了两组实验，从两个方面来分析 x-vectors 语种识别系统在评价指标平均检测代价（C_{AVG}）和等错误率(equal error rate, EER)[1]上的差别，这两个方面分别是：在句子长度没有限制的情况下，使用整句的测试语句，比较基线系统和 x-vectors 系统之间的性能差别；第二个方面比较是否使用数据增强技术，x-vectors 语种识别系统的性能差别。

4.1 基线系统与 x-vectors 系统性能比较

这里我们比较了使用了数据增强技术的 x-vector 语种识别系统和基线系统 i-vectors、TDNN 以及 LSTM 在评价指标C_{AVG}和 EER 上的差别，实验结果如表 3 所示：

表 3: 基线系统和 x-vectors 系统性能比较

系统	C_{AVG}	%EER
MFCC+TDNN	0.1034	11.31
MFCC+LSTM	0.1154	12.76
MFCC+i-vector+LDA	0.0596	5.86
MFCC+x-vector+LDA+aug	**0.0523**	**5.13**

从表 3 中我们可以看到使用了数据增强技术的 x-vectors 系统性能表现最好，其C_{AVG}和 EER 值分别比基线系统中表现最好的 i-vectors 降低了 12.2%和 12.5%。更比同样基于神经网络的 TDNN 和 LSTM 的语种识别系统性能提升 2 倍多。

4.2 数据增强技术对系统性能的影响

实验中我们还测试了对神经网络的训练数据进行数据增强会对 x-vectors 系统产生怎样的影响。

表 4: 使用数据增强和没有使用数据增强的 x-vectors 系统性能比较

系统	C_{AVG}	%EER
x-vectors_aug	**0.0533**	**5.13**
x-vector_no_aug	0.0646	6.39

从表 4 中我们可以看到使用了数据增强技术的 x-vectors 系统（x-vectors_aug）表现要优于没有使用数据增强技术的 x-vectors 系统。这可能是因为数据增强技术使得神经网络的训练数据增多，并且提升了训练数据的

丰富性，而且使得 x-vectors 语种识别系统的鲁棒性更好，所以使用数据增强技术的 x-vectors 系统整体性能会好于没有使用数据增强技术的 x-vectors 系统。

5. 总结与展望

在这篇论文里我们将在说话人识别系统中使用的 x-vectors 系统框架适用于语种识别任务里，我们发现在 2017 年的东方语种识别竞赛任务中，x-vectors 系统的表现好于基线 i-vectors 系统。同时我们还试验了数据增强技术对 x-vectors 系统的影响，发现数据增强技术可以提升 x-vectors 系统的性能表现。

今后，可以探索其他影响语种识别系统性能的变量，如：语音的时长、噪音等等，观察不同的识别系统在这些变量影响下的表现，并找出合适的解决办法以提升语种识别系统的精确度。

6. 致谢

感谢解焱陆、张劲松老师在研究和写作过程中给予我的帮助和指导。

本研究受北京语言大学语言资源高精尖创新中心经费（KYR17005）及校级重大基础研究专项（16ZDJ03）（中央高校基本科研业务专项资金）资助。

7.参考文献

[1] Ambikairajah, E., Li, H., Wang, L., Yin, B., & Sethu, V. 2011. Language identification: A tutorial. *IEEE Circuits and Systems Magazine*, 11(2), 82-108.

[2] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. 2011. Front-end factor analysis for speaker verification. IEEE *Transactions on Audio, Speech, and Language Processing*, 19(4), 788-798.

[3] Geng, W., Wang, W., Zhao, Y., Cai, X., & Xu, B. 2016. End-to-End Language Identification Using Attention-Based Recurrent Neural Networks. In *INTERSPEECH* (pp. 2944-2948).

[4] Gonzalez-Dominguez, J., Lopez-Moreno, I., Sak, H., Gonzalez-Rodriguez, J., & Moreno, P. J. 2014. Automatic language identification using long short-term memory recurrent neural networks. *In Fifteenth Annual Conference of the International Speech Communication Association*.

[5] Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P. J., & Gonzalez-Rodriguez, J. 2015. Frame-by-frame language identification in short utterances using deep neural networks. Neural Networks, 64, 49-58.

[6] Garcia-Romero, D., & McCree, A. 2016. Stacked Long-Term TDNN for Spoken Language Recognition. In *INTERSPEECH* (pp. 3226-3230).

[7] Hinton, G. E., & Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science*, 313(5786), 504-507.

[8] Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. 2017. A study on data augmentation of reverberant speech for robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP)*, 2017 IEEE International Conference on. IEEE (pp. 5220-5224).

[9] Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014, May). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on (pp. 1695-1699). IEEE.

[10] Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez-Rodriguez, J., & Moreno, P. 2014. Automatic language identification using deep neural networks. In Acoustics, *Speech and Signal Processing (ICASSP)*, 2014 IEEE International Conference on (pp. 5337-5341). IEEE.

[11] Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K. R. 1999. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX*, 1999. Proceedings of the 1999 IEEE signal processing society workshop. (pp. 41-48). Ieee.

[12] Peddinti, V., Povey, D., & Khudanpur, S. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. *In Sixteenth Annual Conference of the International Speech Communication Association*.

[13] Prince, S. J., & Elder, J. H. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision*, 2007. ICCV 2007. IEEE 11th International Conference on (pp. 1-8). IEEE.

[14] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. 2000. Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 10(1-3), 19-41.

[15] Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. 2017. Deep neural network

embeddings for text-independent speaker verification. In Proc. *Interspeech* (pp. 999-1003).

[16] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. 2018. X-vectors: Robust DNN embeddings for speaker recognition. Submitted to *ICASSP*.

[17] Snyder, D., Chen, G., & Povey, D. 2015. Musan: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484.

[18] Sun, H., Ma, B., & Li, H. 2008. An efficient feature selection method for speaker recognition. In Chinese Spoken Language Processing, 2008. *ISCSLP'08. 6th International Symposium* on (pp. 1-4). IEEE.

[19] Tang, Z., Wang, D., Chen, Y., & Chen, Q. 2017. Ap17-olr challenge: data, plan, and baseline. 749-753.

[20] Variani, E., Lei, X., McDermott, E., Lopez-Moreno, I., & Gonzalez-Dominguez, J. 2014. Deep neural networks for small footprint text-dependent speaker verification. In *ICASSP* (Vol. 14, pp. 4052-4056).

[21] Yu, D., & Seltzer, M. L. 2011. Improved bottleneck features using pretrained deep neural networks. In *Twelfth annual conference of the international speech communication association*.

于嘉威 北京语言大学语言资源高精尖创新中心硕士研究生，语种识别、说话人识别。
E-mail: vyujiawei@gmail.com

解焱陆 北京语言大学语言资源高精尖创新中心副教授，博士，研究兴趣包括发音偏误检测，二语习得，语音信息处理，计算机辅助发音训练等。
E-mail: xieyanlu@blcu.edu.cn

张劲松 北京语言大学语言资源高精尖创新中心教授，博士，研究兴趣包括语音习得、韵律建模、 语音识别、实验语音学、计算机辅助发音教学。
E-mail: Jinsong.zhang@blcu.edu.cn