嵌入自联想神经网络的高斯混合模型说话人辨认

陈存宝 赵 力
(东南大学信息科学与工程学院 南京 210096)

摘 要：该文提出了一种嵌入自联想神经网络的高斯混合模型，它充分利用了神经网络和高斯混合模型各自的优点，以最大似然概率(ML)为准则，把它们作为一个整体来进行训练。训练过程中，高斯混合模型和神经网络的参数交替更新。由于神经网络起到了“数据整形”的作用，因而提高了类内数据的相似性。实验结果表明，采用该文提出的模型在各种信噪比情况下的识别率都比基线系统有所提高，最高能达到 19%。

关键词：说话人识别；高斯混合模型(GMM)；自联想神经网络(AANN)；嵌入
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2010)03-0528-05
DOI:10.3724/SP.J.1146.2008.00275

Speaker Identification Based on GMM with Embedded AANN

Chen Cun-bao Zhao Li
(*School of Information Science and Engineering*, *Southeast University*, *Nanjing* 210096, *China*)

Abstract: In this paper, a modified Gaussian Mixed Model (GMM) with an embedded Auto-Associate Neural Network (AANN) is proposed. It integrates the merits of GMM and AANN. GMM and AANN as a whole are trained by means of Maximum Likelihood (ML). In the process of training, the parameters of GMM and AANN are updated alternately. AANN reshapes the distribution of the data and improves the similarity of the data in one class. Experiments show that the proposed system improves accuracy rate against baseline GMM at all SNR, maximum to 19%.

Key words: Speaker identification; Gaussian Mixed Model (GMM); Auto-Associate Neural Network (AANN); Embedded

1 引言

自动说话人识别 (Automatic Speaker Recognition, ASR)[1−3] 是一个既有巨大吸引力而又有相当难度的课题。它主要包括说话人辨认(speaker identification)[4] 和说话人确认 (speaker verification)[5]两个范畴。前者是把待辨认语音判定为属于语音库中多个参考人之中的某一个；后者则是确认待测语音是否与所声明的那个人相符。另外对于说话人识别来说，不管是辨认还是证实，按照话语的文本都可以分为与文本有关(text-dependent)的说话人识别和与文本无关(text-independent)的说话人识别。本文主要讨论与文本无关的说话人辨认。

在说话人识别方法方面，高斯混合模型(GMM)方法已越来越受到人们重视[4−6]，它具有识别率高、训练简单的优点，已成为主流的识别方法。基于 GMM 超向量的支持向量机(SVM)和因子分析方法[7,8]则代表 GMM 方法的最新成果。由于 GMM 具有很好的数据分布表示能力，只要有足够多的项，足够多的训练数据，GMM 能够逼近任何分布模型。但是，GMM 只能逼近数据分布，它不能对数据作任何变换，因而对改变数据分布无能为力。另外，由于在选择 GMM 模型时，通常取各个高斯分布的方差矩阵为对角矩阵，这个假设也不是很合理，这要求高斯混合项足够多。

神经网络在说话人识别方面也占有重要的位置，多层感知器(MLP)、径向基(radial basis)网络、时延网络(TDNN)等已经成功应用于说话人识别[9,10]，并且取得很好的识别效果。神经网络对特征向量进行学习和变换，使变换得到的特征向量以某种方式逼近目标向量，逼近的准则通常是 MMSE。Yegnanarayana 等提出了利用 AANN 进行说话人识别的方法[11−13]，并且认为 AANN 的说话人识别效果可以和 GMM 的识别效果相媲美。

本文提出了一种结合 GMM 和 AANN 的说话人识别方法，把 AANN 嵌入到 GMM，通过 AANN

2008-03-17 收到，2010-01-06 改回
国家自然科学基金(60872073，60975017)和江苏省自然科学基金(BK2008291)资助课题
通信作者：陈存宝 chencunbao@gmail.com

学习特征向量间的差异，把特征向量集映射到能增大目标说话人似然概率的子空间，从而达到增大GMM 似然概率的目的。

本文如下组织：第 2 节介绍 GMM 模型以及AANN 用于说话人识别的情况；第 3 节提出嵌入自联想神经网络的高斯混合模型，并给出了训练算法；接着，实验数据、结果和相应的讨论在第 4 节中给出；最后，在第 5 节中进行了总结和讨论。

2　GMM 和 AANN 在说话人识别中的应用

一个 M 阶 GMM 的概率密度函数是由 M 个高斯概率密度函数加权求和得到的，可以用如下形式表示[4,5]：

$$p(\boldsymbol{x}_t|\boldsymbol{\lambda})=\sum_{i=1}^{M}p_i b_i(\boldsymbol{x}_t) \tag{1}$$

这里 $\boldsymbol{x}_t$ 是一 D 维随机向量，在说话人识别应用中，$\boldsymbol{x}_t$ 通常是能表征说话人的特征向量，例如梅尔倒谱(MFCC)或 LPC 倒谱(LPCC)；$b_i(\boldsymbol{x}_t)$，$i=1,2,\cdots,M$ 是成员密度函数；p_i，$i=1,2,\cdots,M$ 是混合权值。每个成员密度函数是一 D 维变量，均值矢量为 $\boldsymbol{u}_i$，协方差矩阵为 $\boldsymbol{\Sigma}_i$ 的高斯函数，形式如下：

$$b_i(\boldsymbol{x}_t)=\frac{1}{(2\pi)^{D/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{u}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_t-\boldsymbol{u}_i)\right\} \tag{2}$$

其中混合权值满足条件：$\sum_{i=1}^{M}p_i=1$。

完整的高斯混合模型由所有成员密度函数的均值矢量、协方差矩阵和混合权值参数组成。这些参数聚集一起表示为

$$\boldsymbol{\lambda}=\{(p_i,\boldsymbol{u}_i,\boldsymbol{\Sigma}_i),i=1,2,\cdots,M\} \tag{3}$$

一般说来，说话人识别训练和识别数据较少，所以实际使用 GMM 时，通常假设每个高斯混合密度的协方差矩阵是对角的，并且通过期望极大化(EM)算法来训练参数。

AANN 网络已成功应用于多种应用，如数据压缩，非线性主成分分析，数值逼近等等。Yegnanarayana 考虑了 GMM 的一些弱点，如混合项分布的选择，混合项数目的确定，他认为在一些分布场合下，AANN 比 GMM 应用得更好。他提出了利用 5 层 AANN (包括输入层)来克服 GMM 的弱点[11]，其神经元组成是 2L:4N:L:4N:2L，其中 L 表示线性，N 表示非线性。其中，第 3 层是压缩层，以便消除数据的冗余。Yegnanarayana 把误差曲面的概念引进 AANN，通过调整一个增益参数，可以获得想要的误差曲面。Yegnanarayana 把这个方法用于说话人确认，获得了与 GMM 接近的识别性能。

本文的设计是把 AANN 嵌入到 GMM 中，利用极大似然概率，使神经网络的学习朝着增大似然概率的方向进行。

3　嵌入 AANN 的 GMM 的模型及其算法

在综合 GMM 和 AANN 各自优点的基础上，本文提出在 GMM 中嵌入 AANN 的方案。AANN 起到增强说话人信息，抑制语义信息的作用。下面从训练和识别模型，训练方法，AANN 训练推导和收敛性证明几个方面来推导和说明本文的方法。

3.1　训练和辨认模型

图 1 是嵌入 AANN 的说话人辨认的训练和识别模型，它跟基线 GMM 模型在训练和识别方面都有所不同。

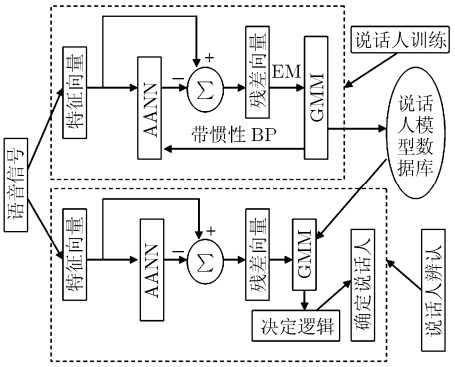

图 1 说话人训练和识别模型

训练时，特征向量先经过 AANN，获得的残差特征向量(即输入向量与 AANN 的输出之差)提供给 GMM，利用 EM 方法进行 GMM 模型训练；再以 GMM 似然概率为训练准则，利用带惯性的 BP 方法调整 AANN 的参数。这里 AANN 和 GMM 模型学习和训练的准则都是极大似然概率。模型通过学习，残差分布就有可能朝着增强目标说话人的似然概率的方向进行。

辨认时，由于 AANN 已经学习了特征空间的结构，所以输入特征向量，AANN 会把特征向量作适当的变换，然后提供给 GMM 模型。因为经过了学习，所以模型会起到增强目标模型的似然概率，降低非目标模型的似然概率。

3.2　训练方法

通过一个两阶段方法来训练这个模型，训练 AANN 的过程和训练 GMM 模型的过程交替进行。训练 GMM 时利用 EM 方法，训练 AANN 时利用带惯性的 BP 向后反演法。

训练过程描述如下：

(1)确定 GMM 模型和 AANN 结构；

(2)给定收敛条件和最大迭代次数；

(3)随机确定 AANN 和 GMM 模型参数；

(4)固定 AANN 参数，把特征向量输入 AANN，得到所有的残差向量；

(5)利用 EM 方法修正 GMM 模型各个高斯分布的权系数，均值向量和方差；

(6)固定修正后的 GMM 模型各个高斯分布的权系数，均值向量和方差，把残差代入，得到似然概率，利用带惯性的 BP 向后反演方法修正 AANN 参数；

(7)判断是否满足训练停止条件，是，停止训练，否，转(4)。

3.3 修正 AANN 参数的推导

由于采用两步迭代的方法求模型的参数，因此在迭代神经网络参数时，固定 GMM 的权系数，均值向量和方差阵，然后求使如下函数极大化的神经网络参数：

$$\boldsymbol{\omega}^* = \arg\max_{\boldsymbol{\omega}} \prod_{t=1}^{\mathrm{N}} p((\boldsymbol{x}_t - \boldsymbol{o}_t) \mid \boldsymbol{\lambda}) \tag{4}$$

$p(\boldsymbol{x} \mid \boldsymbol{\lambda})$ 见式(1)， $\boldsymbol{o}_t$ 为神经网络输出。

由于神经网络迭代时一般求极小值，而且和式比乘积更方便，所以对式(4)取对数后再取负，得到

$$\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}} \left(-\sum_{t=1}^{N} \ln p((\boldsymbol{x}_t - \boldsymbol{o}_t) \mid \boldsymbol{\lambda}) \right) \tag{5}$$

BP 方法[14]神经网络参数迭代过程一般如下(这里采用了单样本迭代法)：

$$\boldsymbol{\omega}_{ij}^k(m+1) = \boldsymbol{\omega}_{ij}^k(m) - \alpha \frac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{\omega}_{ij}^k} \Big|_{\boldsymbol{\omega}_{ij}^k = \boldsymbol{\omega}_{ij}^k(m)} \tag{6}$$

这里 α 为迭代步长， $F(\boldsymbol{x}) = -\ln p((\boldsymbol{x} - \boldsymbol{o}) \mid \boldsymbol{\lambda})$,为方便，省去了下标 t。

带惯性的 BP 方法[15]能够加速迭收敛过程，并能更好的处理局部极小值问题，带惯性的 BP 方法的公式如下：

$$\Delta\omega_{ij}^k(m+1) = \gamma \Delta\omega_{ij}^k(m) - (1-\gamma)\alpha \frac{\partial F(\boldsymbol{x})}{\partial \omega_{ij}^k} \Big|_{\omega_{ij}^k = \omega_{ij}^k(m)} \tag{7}$$

其中 $\Delta\omega_{ij}^k(m+1) = \omega_{ij}^k(m+1) - \omega_{ij}^k(m)$， m 为迭代次数，该值可根据实际情况选取， k 为神经网络的层序号， γ 为惯性系数，取 0 时即为式(6)。令

$$y_i^k = \sum_j \omega_{ij}^k o_j^{k-1} \tag{8}$$

$$o_i^k = f(y_i^k) \tag{9}$$

o_i^k 为 k 层 i 个神经元输入样本 x 时的输出， y_i^k 为 k 层 i 个神经元输入样本 $\boldsymbol{x}$ 时的输入， $f(y_i^k)$ 为激活函数。

要计算式(7)，需要求 $\dfrac{\partial F(\boldsymbol{x})}{\partial \boldsymbol{\omega}_{ij}^k}$：

$$\frac{\partial F(\boldsymbol{x})}{\partial \omega_{ij}^k} = \frac{\partial F(\boldsymbol{x})}{\partial y_i^k} \frac{\partial y_i^k}{\partial \omega_{ij}^k} \tag{10}$$

由式(8)知，

$$\frac{\partial y_i^k}{\partial \omega_{ij}^k} = o_j^{k-1} \tag{11}$$

由于输出层和隐含层的 $\dfrac{\partial F(\boldsymbol{x})}{\partial y_i^k}$ 计算方法不同，所以分别计算。先求输出层 $\dfrac{\partial F(\boldsymbol{x})}{\partial y_i^k}$：

$$\begin{aligned}\frac{\partial F(\boldsymbol{x})}{\partial y_i^k} &= -\frac{1}{p((\boldsymbol{x}-\boldsymbol{o}) \mid \boldsymbol{\lambda})} \frac{\partial p((\boldsymbol{x}-\boldsymbol{o}) \mid \boldsymbol{\lambda})}{\partial o_i^k} \frac{\partial o_i^k}{y_i^k} \\ &= -\frac{f'(y_i^k)}{p((\boldsymbol{x}-\boldsymbol{o}) \mid \boldsymbol{\lambda})} \sum_{n=1}^{M} p_n c_n \\ &\quad \cdot \left(\frac{a_n(\boldsymbol{x}-\boldsymbol{o}-\boldsymbol{u}_n)}{\boldsymbol{\sigma}_{n,i}^2} (x_i - o_i - u_{n,i}) \right)\end{aligned} \tag{12}$$

其中

$$a_n(\boldsymbol{x}-\boldsymbol{o}-\boldsymbol{u}_n) = \exp\left(-\frac{1}{2} (\boldsymbol{x}-\boldsymbol{o}-\boldsymbol{u}_n)^{\mathrm{T}} \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{x}-\boldsymbol{o}-\boldsymbol{u}_n) \right)$$

$$c_n = \frac{1}{(2\pi)^{D/2} \left| \boldsymbol{\Sigma}_n \right|^{1/2}}$$

再求隐含层 $\dfrac{\partial F(\boldsymbol{x})}{\partial y_i^k}$：

$$\begin{aligned}\frac{\partial F(\boldsymbol{x})}{\partial y_i^k} &= \sum_j \frac{\partial F(\boldsymbol{x})}{\partial y_j^{k+1}} \frac{\partial y_j^{k+1}}{\partial y_i^k} = \sum_j \frac{\partial F(\boldsymbol{x})}{\partial y_j^{k+1}} \frac{\partial \left(\sum_n \omega_{jn}^{k+1} o_n^k \right)}{\partial y_i^k} \\ &= f'(y_i^k) \sum_j \frac{\partial F(\boldsymbol{x})}{\partial y_i^{k+1}} \omega_{ji}^{k+1}\end{aligned} \tag{13}$$

由于向后反演，所以在计算 $\dfrac{\partial F(\boldsymbol{x})}{\partial y_i^k}$ 时 $\dfrac{\partial F(\boldsymbol{x})}{\partial y_i^{k+1}}$ 已知，代入式(13)即可求出 $\dfrac{\partial F(\boldsymbol{x})}{\partial y_i^k}$。

最后，简单证明该训练方法可以获得模型的一个局部极大点。由于 EM 训练过程中，每次求得的似然概率是单调非降的，如果在训练 AANN 时也选择步长因子使得每次迭代后的似然概率单调非降，那由高等数学上的一个定理，单调非降序列必有极限，由此可以证明训练过程可以收敛到局部极大点。

当然，为了避免过早收敛到局部极大点，在步长选择时，刚开始可以选择大一点，此时训练过程振荡反而有一定的好处，后面的迭代步长要控制好，不然可能训练过程会过长，甚至发散。

不管 GMM 还是 AANN，上面的方法理论上都是得到的局部极大点。因此，需要从多个初始值和

步长机制进行训练，以求得更好的模型参数。

4　实验结果

本文采用 NIST SRE 2006 年的 1conv4w-1conv4w 作为实验数据，在其中选了 107 个目标说话人，其中男性 63 名，女性 44 名。每个人选取大约 60–80 s 语音作为训练语音，其余语音作为测试语音，这样形成大约 23000 个测试。

为了测试噪声环境下的改进效果，选取的噪声数据是日本电子协会标准噪声数据库中的行驶中的汽车(2000cc 组，一般道路)内的噪声(平稳噪声)和展览会中的展示隔间内的噪声(非平稳噪声)。这些噪声被按一定的信噪比(SNR)叠加进 1conv4w-1conv4w 语音中，生成含噪语音。

语音预处理时，利用基于能量和过零率[1]的方法进行静音检测，去噪过程中使用了谱减法[16]，再通过 $f(Z)=1-0.97Z^{-1}$ 的滤波器进行预加重，进行长度 22.5 ms，窗移 10 ms 的分帧后，进行汉明加窗后，进行 20 阶 LPC 分析，然后从 20 阶 LPC 系数中求出 13 阶的倒谱系数和 13 阶 Δ 倒谱系数共 26 个参数作为说话人识别的特征参数。

采用 2L:4N:L:4N:2L 的 AANN，非线性激活函数采用 S 函数，神经网络的惯性系数 $\gamma=0.8$；无噪声时的混合项数见表 1，有噪声时 GMM 混合项数为 80，利用对角方差阵，最小 $\sigma=0.01$。由于本文的方法不影响后续的处理，所以只跟基线 GMM 作了比较。

表 1　无噪声时 1conv4w-1conv4w 识别率(%)比较

M	16	32	48	60	80
基线 GMM	85.1	90.3	93.2	95.4	96.4
AANN-GMM	90.3	93.9	96.1	97.3	97.7

与说话人确认不同，说话人辨认是根据提供的语音来判断是哪个目标说话人，所以说话人辨认一般采用正确识别率或错误识别率作为评判辨认效果的标准[4,6]，　$R=N_v/N_t$。其中，R 为正确识别率，N_v 为正确辨认的测试次数，N_t 总的测试次数。

实验结果见表 1，表 2 和表 3，表 1 显示了无噪音时 1conv4w-1conv4w 下改变混合项数的改进效果，表 2 和表 3 显示了在不同噪声、不同 SNR 条件下的改进效果。

从表 1 中我们确实看到嵌入 AANN 后，GMM 的识别效果确实有改进，并且混合项数 M 越少，改进效果越明显，这是由于类内子类较少时，神经网络的学习效果更好。

表 2　汽车内不同 SNR 时的识别率(%)比较

SNR	0	5	10	15	20	25	∞
基线 GMM	30.5	45.1	63.2	78.6	88.4	93.7	96.4
AANN-GMM	34.1	52.3	74.1	87.3	92.1	95.8	97.7

表 3　展示隔间内不同 SNR 时的识别率(%)比较

SNR	0	5	10	15	20	25	∞
基线 GMM	28.7	41.5	62.7	76.1	87.7	93.4	96.4
AANN-GMM	35.2	48.7	74.6	85.7	92.3	95.7	97.7

从表 2 和表 3 可以看出，当噪声过强时，本文方法的改进效果并不是太好，可能由于噪声太强，影响了神经网络的学习效果，但是当信噪比在一定范围时，改进效果较好。

5　结论

本文通过在 GMM 中嵌入一个 AANN，相当于对基线 GMM 同时进行特征域和模型域的变换。由于采用了极大似然概率指导 AANN 和 GMM 参数的学习，所以神经网络的学习结果起到了抑制语义特征，增强说话人特征的效果，特别在噪声环境下，还起到较好的消除噪声的效果。当神经网络的输出层权系数为 0 时，模型退化为基线 GMM。实验结果表明本文方法在无噪声时语音和噪声环境语音下的说话人辨认效果均比基线 GMM 都有所提高。

参 考 文 献

[1] 赵力. 语音信号处理. 北京：机械工业出版社, 2003: 236–253.
Zhao Li. Speech Signal Processing. Beijing: China Machine Press, 2003: 236–253.

[2] Campbell J P. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 1997, 85(9): 1437–1462.

[3] Bimbot F, Bonastre J F, and Fredouille C, *et al*.. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 2004, 2004(4): 430–451.

[4] Reynolds D A and Rose R C. Robust text-independent speaker identification using Gaussian mixture models. *IEEE Transactions on Speech Audio Processing*, 1995, 3(1): 72–83.

[5] Reynolds D A, Quatieri T, and Dunn R. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, 10(1): 19–41.

[6] Kwon S and Narayanan S. Robust speaker identification based on selective use of feature vectors. *Pattern Recognition Letters*, 2007, 28(1): 85–89.

[7] Campbell W M, Sturim D E, and Reynolds D A. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. Proceedings of ICASSP,

Toulouse, France, 2006: 97–100.

[8] Yin Shou-chun, Rose R, and Kenny P. A joint factor analysis approach to progressive model adaptation in text-independent speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2007, 15(7): 1999–2110.

[9] Mak M W, Allen W G, and Sexton G G. Speaker identification using multilayer perceptron and radial basis function networks. *Neurocomputing*, 1994, 6(1): 99–117.

[10] Bennani Y and Gallinari P. On the use of TDNN-extracted features information in talker identification. Proceedings of ICASSP, Toronto, Ont, Canada 1991: 385–388.

[11] Yegnanarayana B and Kishore S P. AANN: An alternative to GMM for pattern recognition. *Neural Networks*, 2002, 15(3): 459–469.

[12] Murty K S and Yegnanarayana B. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing Letters*, 2006, 13(1): 52–55.

[13] Kishore P and Sudhakar V. Significance of formants from difference spectrum for speaker identification. http://www.iiit.net/techreports/2007_23.2.0.pdf, 2007.2.

[14] Rumelhart D E, Hinton G E, and Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, 323(9): 533–536.

[15] Vogl T P, Mangis J K, and Aigler A K, *et al.*. Accelerating the convergence of the back-progagation method. *Biological Cybernetics*, 1988, 59(4): 256–264.

[16] Boll S F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoust, Speech & Signal Process*, 1979, 27(2): 113–120.

陈存宝： 男，1971 年生，博士生，研究方向为语音信号处理.

赵 力： 男，1958 年生，教授，博士生导师，研究方向为语音信号处理.