第 30 卷第 6 期　　声　学　学　报　　Vol. 30, No. 6
2005 年 11 月　　ACTA　ACUSTICA　　Nov., 2005

文本无关说话人识别的全特征矢量集模型及互信息评估方法*

俞一彪　　王朔中[1]
(苏州大学电子信息工程学院　苏州　215021)
(1　上海大学通信与信息工程学院　上海　200072)

2004 年 2 月 18 日收到
2004 年 11 月 10 日定稿

摘要　提出了一种文本无关说话人识别的全特征矢量集模型及互信息评估方法，该模型通过对一组说话人语音数据在特征空间进行聚类而形成，全面地反映了说话人语音的个性特征。对于说话人语音的似然度计算与判决，则提出了一种互信息评估方法，该算法综合分析距离空间和信息空间的似然度，并运用最大互信息判决准则进行识别判决。实验分析了线性预测倒谱系数 (LPCC) 和 Mel 频率倒谱系数 (MFCC) 两种情况下应用全特征矢量集模型和互信息评估算法的说话人识别性能，并与高斯混合模型进行了比较。结果表明：全特征矢量集模型和互信息评估算法能够充分反映说话人语音特征，并能够有效评估说话人语音特征相似程度，具有很好的识别性能，是有效的。

PACS 数：　43.60, 43.70

Text-independent speaker identification using complete feature corpus and mutual information evaluation

YU Yibiao　WANG Shuozhong[1]
(*School of Electronic Information Engineering, Soochow University* Suzhou 215021)
(1 *School of Communication and Information Engineering, Shanghai University* Shanghai 200072)

Received Feb. 18, 2004
Revised Nov. 10, 2004

Abstract　A complete feature corpus as speaker model and a evaluation algorithm of mutual information for text-independent speaker identification are proposed. The speaker model is trained by a clustering algorithm in feature vector space using speech samples with various representative pronunciation characteristics of the speaker. The evaluation algorithm is used to calculate the likelihood between input speech and the models in distance and information space, maximum mutual information decision rule is used to decide the identity of speaker. Experiments on performance analysis with comparison to GMM (Gaussian Mixture Model) method according to linear predictive cepstrum and Mel-fequency cepstrum parameters show the proposed model and evaluation algorithm is quite effective.

引言

作为一种基于生物特征的身份识别方法，说话人识别通过语音来识别说话人的身份，在电子银行、远程网络系统和数据库系统等的用户身份确认、电子侦听对话语音说话人身份的自动检测及其它各类安全系统的身份认证中有很大的应用价值，并具有其它生物特征身份识别方法所不具备的特点，即数据采集设备的非接触性和简易性。

说话人识别研究在国际[1,2]和国内[3,4]都取得了一定的成果，其核心内容包括说话人特征提取、说话人模型的建立以及似然度评估计算方法。目前，在文本无关说话人识别中常用的模型有 GMM 和 VQ 等[5−7]。GMM 高斯混合模型以多个正态分布逼近

* 国家自然科学基金 (60172016) 和江苏省高校自然科学基金 (04KJA510133) 资助项目

语音信号的实际统计分布，并运用 Bayes 分类器进行识别，这一模型通过EM算法较好地解决了语音信号统计分布特征的估计问题，但对时频动态变化的瞬时特征缺乏充分的建模能力，并且本身不能表达语音的时变特征，需要通过差分特征参数的引入来解决，因此，当说话人语音包含较快的动态瞬时变化特征时，识别性能会明显下降。VQ 模型采用矢量量化的方法形成每个说话人的语音码书，并运用矢量距离累加计算进行匹配，由于没有利用语音信号的统计特征，因此识别性能一般，鲁棒性不够好。

本文提出了一种说话人的全特征矢量集模型 CFC(Complete Feature vector Corpus) 和互信息评估算法 MIE(Mutual Information Evaluation)，通过 MIE 算法在特征距离空间和信息空间计算输入语音与 CFC 之间的似然度，并运用最大互信息判决准则 MMI(Maximum Mutual Information) 判别说话人的身份。实验对 CFC 模型和 MIE 算法在特征参数 LPCC 和 MFCC 两种情况下的识别性能进行了全面的分析，并与 GMM 模型在同样的测试环境与条件下进行了比较，结果显示，CFC-MIE 的识别性能较好，MFCC 特征参数比 LPCC 更能反映说话人特征。以下第 1 节将介绍全特征矢量集模型 CFC 的建立，第 2 节介绍互信息评估算法 MIE，第 3 节给出实验分析结果。

1　说话人的全特征矢量集模型

在文本无关的说话人识别中，说话人的语音模型应该能够充分反映说话人语音发音的个体特征，并实行语义特征的归一化。

全特征矢量集模型 CFC 的基本思想是通过对一组包含说话人各种语音发音个体特征的数据进行分析处理，提取相应的代表性特征矢量表示说话人语音模型，其训练过程如下。

训练语音信号由 N 段语音 $S_1(n)$, $S_2(n)$, $\cdots$, $S_N(n)$ 组成，其包含了说话人不同语音发音以及语音韵律的特点。预处理部分对这一组信号进行去噪处理，去除背景噪声，保留纯语音成份，并合并成一个完全由语音数据构成的训练语音信号 $S(n)$。进一步对 $S(n)$ 进行短时分析处理，提取特征参数，形成一原始特征矢量序列 $\{\boldsymbol{V}_i, \forall i\}$。聚类分析部分运用聚类分析算法对原始特征矢量进行聚类计算分析，提取代表性的特征矢量：$\{\boldsymbol{FV}_1, \boldsymbol{FV}_2, \cdots, \boldsymbol{FV}_M\}$ 作为说话人的全特征矢量集模型 CFC。聚类分析计算可以采用 K-Means 和 LBG 等多种算法来实现，本文采用的算法如下：

(1) 设置全特征矢量集 CFC 的大小 M，并以等间隔方式选取原始特征矢量序列 $\{\boldsymbol{V}_i, \forall i\}$ 中的 M 个特征矢量作为 CFC 的初始值：$\{\boldsymbol{FV}_1^\circ, \boldsymbol{FV}_2^\circ, \cdots, \boldsymbol{FV}_M^\circ\}$。

(2) 计算各原始特征矢量与 CFC 中各特征矢量之间的距离，并将原始特征矢量赋予与其距离最小的 CFC 特征矢量所在子集，即：

$$\boldsymbol{V}_i \Rightarrow FVS_q, \quad q = \mathrm{Arg}\min_k d(\boldsymbol{V}_i, \boldsymbol{FV}_k^\circ); \forall i.$$

(3) 对每个 CFC 特征矢量子集中的原始特征矢量在特征空间计算其均值，并将其作为新的 CFC 特征矢量，即：

$$\boldsymbol{FV}_k^n = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{V}_i, \quad \boldsymbol{V}_i \in FVS_k; \ \forall k$$

(4) 如果计算得出的新 CFC 特征矢量 $\boldsymbol{FV}_k^n, \forall k$ 与原 CFC 特征矢量 $\boldsymbol{FV}_k^\circ, \forall k$ 完全一致，则结束，并将该 CFC 作为说话人的全特征矢量集模型，否则继续。

(5) 将 $\boldsymbol{FV}_k^n, \forall k$ 替代原 CFC 特征矢量 $\boldsymbol{FV}_k^\circ, \forall k$，转 (2)。

由于用于训练的语音数据 $S_1(n)$, $S_2(n)$, $\cdots$, $S_N(n)$ 包含了说话人不同语音发音的声学与韵律特征，因而聚类训练所形成的特征矢量集 CFC 反映了说话人的全语音特征，并且，这样的特征矢量集并不包含语义与时序信息，实现了模型的语义归一化。

2　互信息评估算法与判决准则

目前，互信息理论在语音识别中的应用主要在一些参数的训练和距离尺度的理论描述 [8−10]。作者曾提出语音信号之间互信息的估计算法 [11]，其特点是用信息量来表示不同语音信号之间的似然度，计算中同时考虑了信号的时变特性与统计特性，在语音识别和基于文本的说话人识别中得到了很好的应用 [11,13]。但是，在文本无关的说话人识别中，需要计算的是输入语音与说话人模型之间的互信息，与语音识别和基于文本的说话人识别应用有很大的区别。第一，互信息反映的是说话人的特征信息，而不是语义信息，第二，互信息的计算应该考虑语音信号的语义归一化，尽可能地消除语义的影响，第三，互信息的计算不是针对两个语音信号，而是语音信号和说话人的语音模型。

2.1　基于互信息评估的说话人识别原理

设有 N 个说话人 SPK_1, SPK_2, $\cdots$, SPK_N，其对应的说话人语音模型采用全特征矢量集模型，分

别为 $\mathrm{CFC}_1, \mathrm{CFC}_2, \cdots, \mathrm{CFC}_N$ 。其中，每一个模型 $\mathrm{CFC}_k, \forall k$ 包含 M 个代表性特征矢量，即：

$$\mathrm{CFC}_k : \{\boldsymbol{FV}_1^k, \boldsymbol{FV}_2^k, \cdots, \boldsymbol{FV}_M^k\}$$

当某一说话人的测试语音输入时，经过预处理和特征提取，得到一特征矢量序列 XFV:$\{\boldsymbol{XFV}_1, \boldsymbol{XFV}_2, \cdots, \boldsymbol{XFV}_L\}$。XFV 和各说话人模型之间的互信息 $I(\mathrm{XFV}; \mathrm{CFC}_k)$ 反映了两者之间相互携带的信息量。根据互信息原理，信息量越大说明两者的相似程度越大。因此，可以根据最大互信息准则 MMI 来判决输入语音属于哪个说话人，即识别的说话人 d 应该满足下式：

$$d = \mathrm{Arg}\max_k \quad I(XFV; CFC_k).$$

如前所述，全特征矢量集模型 $CFC_k, \forall k$ 已经实行了语义的归一化，因此，在进行互信息计算中，只要通过适当的处理就可以实现语义的归一化。

2.2 互信息评估算法 MIE

依据互信息理论，输入语音信号XFV与说话人模型 CFC_k 之间的互信息可以由下式计算，其中，$H(XFV)$ 表示输入语音信号的熵，$H(XFV|CFC_k)$ 表示条件熵。

$$I(XFV; CFC_k) = H(XFV) - H(XFV|CFC_k).$$

上式中第 1 项与模型无关，在匹配判决中可以不加以考虑。第 2 项条件熵的计算需要运用条件概率来计算，但这样的条件概率无法从理论上推导得到，也无法通过大量样本的统计分析得出，必须通过一定的方法估计获得。本文提出了 MIE 算法来估计这一条件概率和互信息。

对输入语音信号 XFV 的每一个特征矢量 $\boldsymbol{XFV}_j$ $\forall j$，在距离空间求与说话人模型 CFC_k 的最佳匹配代表性特征矢量 $\boldsymbol{BFV}_j^k$，从而得到一个 XFV 与说话人模型 CFC_k 的最佳匹配代表性特征矢量序列，如下：

$$BFV_k : \{\boldsymbol{BFV}_1^k, \boldsymbol{BFV}_2^k, \cdots, \boldsymbol{BFV}_L^k\},$$
$$\boldsymbol{BFV}_j^k = \boldsymbol{FV}_o^k,$$
$$o = \arg\min_i \left\| \boldsymbol{XFV}_j - \boldsymbol{FV}_i^k \right\|.$$

对于特定的输入语音 XFV，由以上获取得的最佳匹配代表性特征矢量序列 BFV_k 代表说话人模型 CFC_k。求相互之间的互信息，计算公式如下：

$$I(XFV; CFC_k) \Rightarrow I(XFV; BFV_k) = H(XFV) - H(XFV|BFV_k).$$

求 XFV 与最佳匹配代表性特征矢量序列 BFV_k 的差，形成一个特征差矢量序列 DFV_k，如下：

$$DFV_k : XFV - BFV_k = \{\boldsymbol{DFV}_1^k, \boldsymbol{DFV}_2^k, \cdots, \boldsymbol{DFV}_L^k\} = \{\boldsymbol{XFV}_1 - \boldsymbol{BFV}_1^k, \boldsymbol{XFV}_2 - \boldsymbol{BFV}_2^k, \cdots, \boldsymbol{XFV}_L - \boldsymbol{BFV}_L^k\}.$$

特征差矢量序列 DFV_k 中的每个特征差矢量与输入信号特征矢量和最佳匹配代表性特征矢量之间存在关系：$\boldsymbol{DFV}^k = \boldsymbol{XFV} - \boldsymbol{BFV}^k$，或 $\boldsymbol{XFV} = \boldsymbol{BFV}^k + \boldsymbol{DFV}^k$。

根据概率统计理论，当 $\boldsymbol{XFV}$ 和 $\boldsymbol{BFV}^k$ 采用相同的特征参数，并且具有高斯正态分布统计特征时，特征差矢量 $\boldsymbol{DFV}^k$ 也是一个具有高斯正态分布统计特征的随机特征矢量，并且与 $\boldsymbol{BFV}^k$ 独立，其均值和方差可以根据 $\boldsymbol{XFV}$ 和 $\boldsymbol{BFV}^k$ 的均值与方差计算得到，也可以直接根据其样本数据通过最大似然估计方法估计得到。线性预测系数 LPC 以及 LPCC、MFCC 等参数具有近似的高斯正态分布统计特征，因此，当采用这些参数作为特征参数时，特征差矢量 $\boldsymbol{DFV}^k$ 的统计分布特征可以由概率密度函数 $N(\boldsymbol{m}_{dfv}^k, \boldsymbol{W}_{dfv}^k)$ 表示，输入语音信号 XFV 与说话人模型 CFC_k 之间的互信息可以计算如下：

$$\boldsymbol{XFV} = \boldsymbol{DFV}^k + \boldsymbol{BFV}^k,$$
$$p(\boldsymbol{XFV}|\boldsymbol{BFV}^k) = p(\boldsymbol{DFV}^k + \boldsymbol{BFV}^k|\boldsymbol{BFV}^k) = p(\boldsymbol{DFV}^k),$$
$$H(XFV|BFV_k) = -\iint p(\boldsymbol{BFV}^k) p(\boldsymbol{XFV}|\boldsymbol{BFV}^k) \log p(\boldsymbol{XFV}|\boldsymbol{BFV}^k) d\boldsymbol{XFV} d\boldsymbol{BFV}^k = -\int p(\boldsymbol{DFV}^k) \log p(\boldsymbol{DFV}^k) d\boldsymbol{DFV}^k = H(DFV_k),$$
$$H(DFV_k) = \frac{p}{2}\log(2\pi e) + \frac{1}{2}\log|\boldsymbol{W}_{dfv}^k|, \quad H(XFV) = \frac{p}{2}\log(2\pi e) + \frac{1}{2}\log|\boldsymbol{W}_{xfv}|,$$
$$I(XFV; BFV_k) = \frac{1}{2}\log\frac{|\boldsymbol{W}_{xfv}|}{|\boldsymbol{W}_{dfv}^k|}.$$

上式中，p 是特征矢量的维数，$\boldsymbol{W}_{xfv}$ 和 $\boldsymbol{W}_{dfv}^{k}$ 分别是输入语音信号特征矢量和特征差矢量的协方差矩阵，前者由于是与模型无关的，因此在实际的识别判决中并不需要计算。$\boldsymbol{W}_{dfv}^{k}$ 可以依据无偏估值理论由实际的特征差矢量分布数据估计获得，其估值计算公式如下：

$$\boldsymbol{W}_{dfv}^{k}=\frac{1}{L}\sum_{j=1}^{L}(\boldsymbol{DFV}_{j}^{k}-\boldsymbol{m}_{dfv}^{k})(\boldsymbol{DFV}_{j}^{k}-\boldsymbol{m}_{dfv}^{k})^{\mathrm{T}}=\frac{1}{L}\sum_{j=1}^{L}(\boldsymbol{XFV}_{j}-\boldsymbol{BFV}_{j}^{k}-\boldsymbol{m}_{xfv}+\boldsymbol{m}_{bfv}^{k})(\boldsymbol{XFV}_{j}-\boldsymbol{BFV}_{j}^{k}-\boldsymbol{m}_{xfv}+\boldsymbol{m}_{bfv}^{k})^{\mathrm{T}}$$

各特征矢量的均值 $\boldsymbol{m}_{dfv}^{k}$，$\boldsymbol{m}_{xfv}$，$\boldsymbol{m}_{bfv}^{k}$ 同样根据实际数据估计得到，但由于涉及统计估计的有效性，在实际应用中需要根据不同的情况作一定的变化。当输入信号的长度 L 较短时，各参数采用相同的均值效果好些，而当 L 较大时，采用不同的均值效果更好，并且，最佳匹配代表性特征矢量的均值 $\boldsymbol{m}_{bfv}^{k}$ 由相应的模型 CFC_k 直接估计效果更好。

2.3 最大互信息判决 MMI

输入语音与说话人模型之间的互信息 $I(XFV;CFC_k)$ 反映了两者的似然度，其值越大表示越相似。根据互信息评估算法MIE，互信息 $I(XFV;CFC_k)$ 可以由输入信号与最佳匹配代表性特征矢量序列之间的互信息 $I(XFV;BFV_k)$ 来替代，并且，最终由特征差矢量序列的协方差矩阵 $\boldsymbol{W}_{dfv}^{k}$ 与输入语音信号特征矢量序列的协方差矩阵 $\boldsymbol{W}_{xfv}$ 来决定。这样，最大互信息判决准则 MMI 如下：

$$d=\operatorname*{Arg\,max}_{k}\ I(XFV;BFV_k)=\operatorname*{Arg\,max}_{k}\ \frac{1}{2}\log\frac{|\boldsymbol{W}_{xfv}|}{|\boldsymbol{W}_{dfv}^{k}|}.$$

在对所有的说话人模型进行匹配判决过程中，输入语音信号特征矢量序列的协方差矩阵是不变的，所以对判决并没影响，可以在具体计算中不加考虑。另外，考虑到对数函数的单调递增特点，可以免除对数计算，最终 MMI 准则简化为如下：

$$d=\operatorname*{Arg\,min}_{k}|\boldsymbol{W}_{dfv}^{k}|.$$

即识别说话人所对应的模型具有最小的特征差矢量序列协方差矩阵值。

3 实验分析与比较

基于全特征矢量集模型 CFC 与互信息评估算法 MIE 的文本无关说话人识别，需要研究分析的问题包括：(1) 汉语说话人全特征矢量集模型 CFC 的大小，或 CFC 中需要多少代表性特征矢量才能充分表示说话人的语音特征，(2) 训练一个说话人的全特征矢量集模型 CFC 一般需要多少语音数据，(3) 互信息评估算法 MIE 的有效性或识别性能，(4) 不同特征参数，例如 LPCC、MFCC 情况下，CFC-MIE 的说话人识别性能，(5) 不同长度测试语音输入时，说话人识别性能的变化趋势，(6) 在相同训练语音数据、实验环境和条件下，CFC-MIE 与 GMM 的识别性能比较分析。

3.1 实验数据、环境与条件

语音数据选择 SD2002-D2 数据库，该数据库中包含了在普通实验室环境下通过计算机声音系统采集得到的 40 个说话人的 280 条语音片段，其中，男声 26 人，女声 14 人，每人分别输入 7 段语音，每段语音包括停顿间隙长度为 12 s。语音采样率为 11025 Hz，16 位量化，单声道输入。实验中，每说话人的前 4 段语音用于模型训练，后 3 段用于测试。

在模型训练和识别测试中，预处理部分首先消除输入语音信号的背景噪声，保留纯语音数据。短时分析采用 20 ms 矩形窗。特征参数 LPCC 和 MFCC 的计算采用 12 阶模型，其中，LPCC由12阶LPC线性预测系数推导得到，MFCC是基于Mel频率尺度的倒谱系数，通过计算 Mel 频率域均匀分布的 19 个三角滤波器组的 DFT 输出，并经 DCT 变换得到[12]，实验中选取第 1~12 个系数作为特征参数。

在本文的实验中，说话人模型CFC均采用每说话人的前 4 段语音信号进行训练，其纯语音成分的长度平均为 32 s。测试实验采用每个说话人的后 3 段语音。互信息评估中，如果没有特别声明，则特征差矢量序列的协方差矩阵计算一般采用相同的均值处理，以便消除当测试语音长度较短时 (小于 3 s) 均值估计误差带来的影响。

3.2 全特征矢量集大小分析

说话人的全特征矢量集模型CFC表达了说话人的语音特征，这种语音特征是语义和时序归一化的，因此体现了文本无关的特点。但是，对于汉语说话人来说，CFC 的大小应该如何规定呢？如果CFC太小，其代表性特征矢量就不能完备地反映说话人的所有语音特征，反之，如果很大，不仅计算量增大，而且，各说话人模型出现相似代表性特征矢量的概率就增加，模型之间的耦合度将变大。因此，选择合适的CFC大小对模型的准确性和识别性能有较大影响。

实验中分别对代表性特征矢量数为 100，200，

300, 400, 500 的五种不同大小的 CFC 的识别性能进行了分析，输入测试语音的长度分别为 1 s 和 2 s，其具体测试中的误识率如图 1 所示。

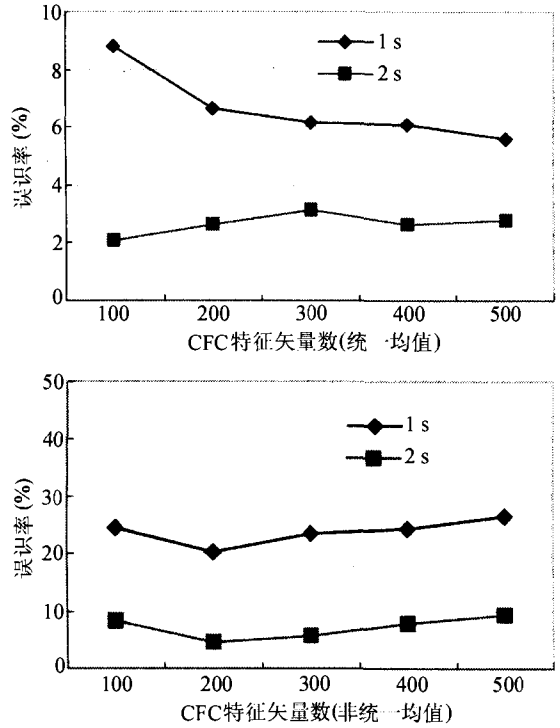

图 1 CFC 大小与识别性能的关系

从图 1 可以看到，如果采用相同均值时，当测试语音长度为 1 s 时，随着模型中特征矢量数从 100 增加到 200，误识率从 8.8% 下降到 6.63%，之后不断缓慢下降，但总体上较大；当测试语音长度为 2 s 时，误识率随模型大小的变化幅度不超过 1%，基本上是稳定的。如果采用非统一均值，两种测试语音长度下都在特征矢量数为 200 时得到最小误识率。因此，说话人全特征矢量集模型 CFC 选择 200 个代表性特征矢量较合适。

3.3 CFC-MIE 的识别性能分析

根据以往的说话人识别研究表明，线性预测倒谱系数 LPCC 以及 Mel 频率倒谱系数 MFCC 是说话人识别中比较有效的两种特征参数，实验分析了当采用以上两种参数作为特征参数时 CFC-MIE 的说话人识别性能，具体结果如图 2 和图 3 所示，其中 CFC 的大小为 200，特征矢量维数为 12 。

图 2 是在 12 阶 LPCC 特征参数情况下，采用统一均值 (CFC_MIE_LPCC) 和非统一均值 (CFC_MIE_LPCC_M) 得到的识别结果。可以看到，当测试语音时长分布为 1~3 s 时，由于无法从较短的语音信号中估计精确可靠的均值，采用统一均值的互信息计算方法识别性能更加优越。但当输入语音信号时长达到 4 s 以上时，此时采用非统一均值的互信息计算方法其识别性能比采用统一均值的方法要好，并在 5 s 时达到了 100%，这是因为当时长达到一定长度时，均值估计就比较精确可靠。从这一实验可以得到结论，在语音信号小于 4 s 时，互信息计算应该采用统一均值方法计算，而当超过 4 s 时，则应该采用非统一均值计算。同时，这一实验结果也说明了语音信号特征矢量的均值估计只有通过大于 4 s 的语音信号数据获得才是可靠、有意义的。

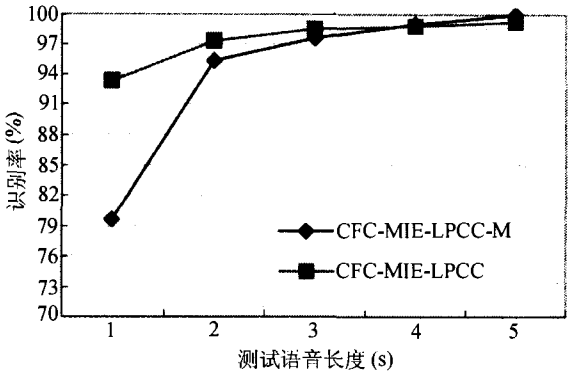

图 2 采用 LPCC 参数的识别性能

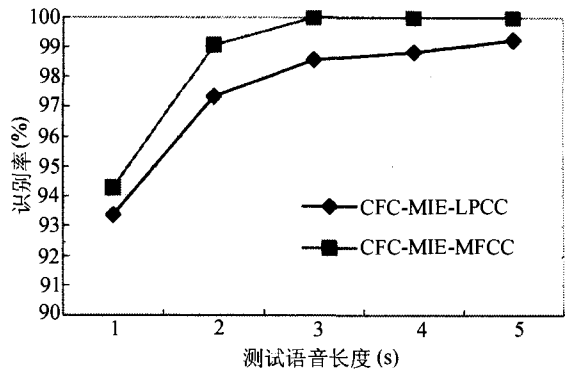

图 3 不同特征参数下识别性能比较

图 3 是采用统一均值情况下，分别采用 12 阶 MFCC 和 LPCC 特征参数时的识别性能比较。可以看出，总体上 CFC-MIE 具有很好的识别性能，并且，MFCC 作为特征参数比 LPCC 的性能更加优越。当输入测试语音为 1 s 时，尽管长度较短，但两种特征参数下识别率分别达到了 94.29% 和 93.37% 。当输入测试语音长度增加到 2 s 时，识别率分别提高到 99.07% 和 97.35%，并且，当输入测试语音长度增加到 3 s 时， MFCC 特征参数情况下的识别率达到了 100%，LPCC 特征参数情况下的识别率也达到了 98.58% 。

MFCC 参数无论在说话人识别应用中，还是在语音识别中都表现出比 LPCC 优越的识别性能，说明它在描述语音信号的语义和说话人个性特征两个方面都比 LPCC 更加有效。

3.4 CFC-MIE 与 GMM 的识别性能比较

高斯混合模型 GMM 是目前文本无关说话人识别中应用较广、性能较好的一种模型，该模型用若干高斯正态分布概率密度函数的混合加权来拟合实际说话人语音信号频谱的统计分布，并利用 Bayes 最大似然准则进行判决。

比较分析中，GMM 所采用的训练和测试语音数据、特征矢量的计算方法与维数、实验条件与环境都与 CFC-MIE 一致，其混合分量数为 16 。图 4 表示了不同测试语音长度情况下 GMM 与 CFC-MIE 采用统一均值进行互信息匹配计算时的识别性能比较，其中，特征参数为 MFCC, CFC 的大小为 200 。

图 4 显示，在各种测试语音长度下，基于全特征矢量集模型 CFC 与互信息评估算法 MIE 的识别性能要优于 GMM 模型的识别性能，并在测试语音长度到达 5 s 时趋向一致。

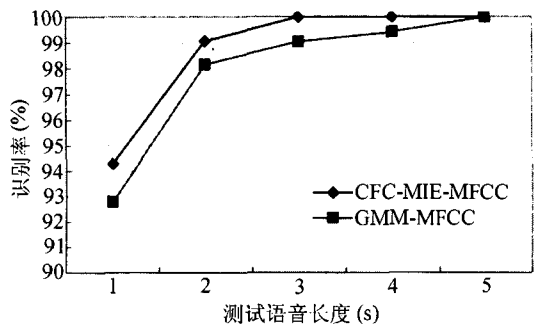

图 4　CFC-MIE 与 GMM 的识别性能比较

4　结论

本文描述了说话人的全特征矢量集模型 CFC 以及用于输入语音与说话人模型似然度计算的互信息评估算法 MIE，并通过实验分析了基于 CFC-MIE 的说话人识别性能。实验表明：(1) CFC-MIE 对文本无关的说话人识别是有效的，并具有很高的说话人识别性能，当特征参数为 MFCC，输入语音长度为 3 秒时，识别率达到了 100% 。(2) 相对 GMM 模型，CFC-MIE 的识别性能更加优越，识别率平均高 1%，并在输入语音长度达到 5 秒时趋于一致。(3) 对于 CFC-MIE，当特征参数采用 MFCC 时，其识别性能优于 LPCC，这与应用 GMM 等其它模型情况下得出的结论一致。(4) 对于汉语说话人，全特征矢量集模型 CFC 的大小取 200 较合适。

CFC 的训练速度很快，同样训练数据下，比 GMM 的 EM 训练算法快许多，并且，对于语音信号来说，代表性特征矢量初始值的选取既可以按等间隔方式从原始特征矢量中选取，也可以以随机的方式选取，不同的选取方式对识别性能的影响很小。实验中互信息评估都采用了相同均值的方式，其好处是可以避免输入测试语音较短时均值估计误差引起识别性能的下降，例如在输入测试语音长度为 1 s，特征参数采用 LPCC 时，采用相同均值和分别估计均值两种情况下的识别率分别为 93.37% 和 79.66%，说明这一处理是有效的。但是，当输入测试语音较长时，采用分别估计均值的方式或许更好些，例如，在输入语音长度为 5 s 的情况下，相同均值和分别估计均值两种情况下的识别率分别为 99.24% 和 100% 。

今后将考虑如何进一步提高短测试语音条件下的识别性能，并研究电信网络环境下基于 CFC-MIE 的文本无关说话人识别问题以及特征参数非线性变换与组合的问题。

参　考　文　献

1　Naik J. Speaker verification: A tutorial. *IEEE Commun. Mag.*, 1990; **28**(1): 42—48

2　Campbell J P. Speaker recognition: A tutorial. *IEEE Proc.*, 1997; **85**(9): 1436—1462

3　侯凤雷，王炳锡. 基于支持向量机的说话人辨认研究. 通信学报，2002; **23**(6): 61—67

4　岳喜才，伍晓宇等. 用神经阵列网络进行文本无关的说话人识别. 声学学报，2000; **25**(3): 230—234

5　Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian Mixture Speaker Models. *IEEE Speech and Audio*, 1995; **3**(1): 72—83

6　Tanprasert C, Achariyakulporn V. Comparative study of GMM, DTW and ANN on Thai speaker identification system. In: Proc. ICSLP, 2000 (Paper No.00718)

7　Chen C T, Chen C. Efficient genetic algorithm of codebook design for text-independent speaker recognition. *IEICE*, 2002, **E85-A**(11): 2529—2531

8　Lee Y -T. Information-theoretic distortion measures for speech recognition. *IEEE-ASSP*, 1991; **39**: 330—335

9　Okawa S, Kobayashi T, Shirai K. Automatic training of phoneme dictionary based on mutual information criterion. ICASSP, 1994: 241—244

10　Bahl L R, Brown P F. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. ICASSP, 1986: 49—52

11　俞一彪，赵鹤鸣等. 语音信号互信息估计的非线性搜索算法及识别应用. 信号处理，2002; **18**(2): 102—106

12　Shaughnessy D O. Speech communications-human and machine. IEEE Press, NJ., 2000: 378—383

13　俞一彪，王朔中. 基于互信息匹配模型的说话人识别. 声学学报，2004; **29**(5): 462—266