信道失配环境下鲁棒说话人识别①

冉国敬, 夏秀渝, 张凤仪

(四川大学 电子信息学院, 成都 610064)

摘 要: 目前说话人识别系统在理想环境下识别率已可达 90%以上, 但在实际通信环境下识别率却迅速下降.本文对信道失配环境下的鲁棒说话人识别进行研究. 首先建立了一个基于高斯混合模型(GMM)的说话人识别系统, 然后通过对实际通信信道的测试和分析, 提出了两种改进方法. 一是由实测数据建立了一个通用信道模型, 将干净语音经通用信道模型滤波后再作为训练语音训练说话人模型; 二是通过对比实测信道、理想低通信道及语音梅尔倒谱系数(MFCC)的特点, 提出合理舍去语音第一、二维特征参数的方法. 实验结果表明, 通过处理后, 系统在通信环境下的识别率提升了 20%左右, 与传统的倒谱均值减(CMS)方法相比,识别率提高了 9%-12%.

关键词: 说话人识别; 信道失配; 通用信道模型; 梅尔倒谱系数

Robust Speaker Recognition Under Channel Mismatch Environment

RAN Guo-Jing, XIA Xiu-Yu, ZHANG Feng-Yi

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610064, China)

Abstract: The recognition rate of the Speaker Recognition System under ideal condition can reach more than 90%, but the recognition rate will decrease rapidly under the traffic environment. In this paper, we discuss the robust speaker recognition under the channel mismatch environment. First, a speaker recognition system based on Gaussian mixture model(GMM) is set up. Then, two methods of improvement are put forward through testing and analysis of the actual communication channel. One is that we establish a general channel model to filter the clean speech which is regarded as the training speech. The other is that we put forward to abandon the first and second dimensional characteristic parameters through comparing the Mel-frequency cepstral coefficient of the measured channel, the ideal low-pass channel and the common speech. The experimental results show that the recognition rate under the channel mismatch environment is improved by 20% after processing and 9%-12% comparing with the traditional Cepstral Mean Subtraction(CMS).

Key words: speaker recognition; channel mismatch; general channel model; MFCC

说话人识别属于生物识别技术的一种, 以其独特的方便性、经济性和准确性等优势受到世人瞩目, 广泛应用于安全控制、保密部门身份验证、法庭鉴别等行业.

与文本无关的说话人识别技术是当前研究重点, 目前最常用的方法有: 基于 VQ 的方法[1], 基于 HMM 的方法[2], 基于 GMM 的方法[3-5]和基于人工神经网络的方法等. 目前基于 GMM 的说话人识别系统, 在安静环境下用高品质话筒采集语音, 对于几十名话者的识别率可达 90%以上. 但对于实际 GSM 网络传输的电话语音, 存在噪声的实际环境语音进行识别时性能显著恶化. 识别环境与训练环境失配导致的语音声学参数的变异是识别率下降的主要原因. 减小环境失配影响的方法主要有特征参数补偿和鲁棒特征提取[6-8]. MFCC 就是一种目前广泛采用具有鲁棒性的参数;实际应用中的特征参数补偿技术有: 谱减法(SS)倒谱均值减

① 收稿时间:2014-06-24;收到修改稿时间:2014-07-18

(CMS)、特征映射等. 其中谱减法主要用于消除环境中的加性干扰噪声, CMS 可用于消除线性信道干扰. 而特征映射需要首先训练一个通用背景模型(GMM-UBM), GMM-UBM 结构复杂而且收敛速度较慢.

本文采用MFCC参数[9], 基于GMM模型[10], 设计了一个 30 人的说话人识别系统, 在干净环境(训练和识别语音未受任何加性和卷积噪声影响)下该系统的识别率可以达到 93%, 然而在通信环境下(识别语音来自无线信道), 识别率则大幅度下降, 只能达到 58%. 对训练和识别语音都做了 CMS 处理之后, 系统识别率提升了 14%. 为了进一步提升系统的识别率, 我们在分析信道特性的基础上提出了两种改进方法, 一是由实测数据建立了一个通用信道模型, 将干净语音经通用信道模型滤波后再作为训练语音. 二是通过对比实际信道、理想低通信道及语音 MFCC 的特点, 提出合理舍去一二维特征参数的方法. 实验结果表明改进后系统的识别率在原来基础上可以提高 23%.

1　基于GMM的说话人识别系统介绍

GMM 说话人识别系统的基本原理如图 1 所示.

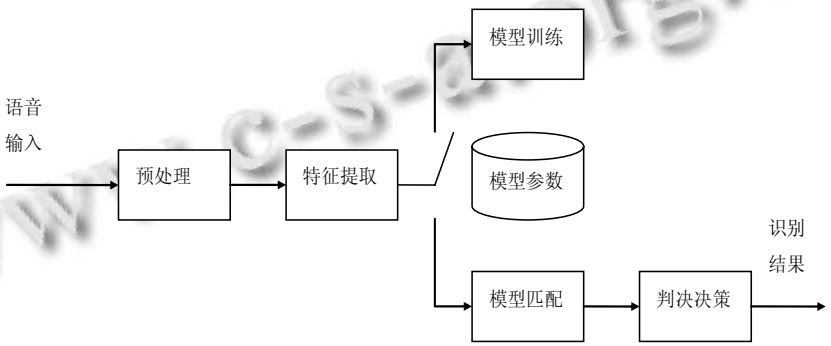

图 1　GMM 说话人识别基本原理图

首先, 对语音信号预处理, 包括语音信号的预加重, 分帧加窗. 预加重的目的是为了对语音的高频部分加重, 以去除口唇等辐射的影响. 通过传递函数为: $H(z)=1-az^{-1}$ 的一阶高通滤波器滤波实现预加重. 其中 a 为预加重系数, 通常, $0.9<a<1.0$, 本文 a 取 0.9375. 语音信号的分帧加窗, 主要是为了克服吉布斯现象, 使语音在短时($10ms-40ms$)内能够平滑过渡, 保持其连续性, 更加稳定. 本文采用 Hamming 窗, 其时域形式为:

$w(k)=0.54-0.64*\cos(2\pi\frac{k}{N-1})k=1,2,3...,N$, 其中 N 为窗长.

接着是提取语音信号的特征参数, 先对语音信号端点检测以提取语音的有用部分, 去除静音段. 因为浊音的能量较大, 清音和静音的能量较小, 但清音的短时过零率很大, 所以本文采用基于短时能量和过零率双门限判别法提取有声部分. 端点检测之后, 对语音帧进行短时傅里叶变换并计算其短时能量谱, 再用 S 个 Mel 带通滤波器组滤波, 最后对这 S 个滤波器的输出功率取对数和反离散余弦变换之后就得到 S 个 $MFCC$ 系数. 一般取 S 前 12---16 个.

在 GMM 训练阶段, 将提取的训练语音特征参数通过 GMM-EM 算法训练得到每个说话人的模型. GMM 作为一种概率密度函数的表达式. 它是多个单高斯分布的线性组合. 其中 M 阶 GMM 的概率密度函数如下:

$$P(o|\lambda)=\sum_{i=1}^{M}P(o,i|\lambda)=\sum_{i=1}^{M}c_iP(o|i,\lambda) \quad (1)$$

式中, λ 为GMM模型的参数集; o 为 K 维的声学特征矢量; i 为隐状态号, 也就是高斯分量的序号, M 阶 GMM 就有 M 个隐状态; c_i 为第 i 个分量的混合权值, 其值对应为隐状态 i 的先验概率. 式(1)中的 $P(o|i,\lambda)$ 为高斯混合分量, 是 $P(o|q=i,\lambda)$ 的简写形式, 对应隐状态 i 的观察概率密度函数, 一般用 K 维单高斯分布函数, 如下式所示:

$$P(o|i,\lambda)=N(o,\mu_i,\Sigma_i)=\frac{1}{(2\pi)^{K/2}|\Sigma_i|^{1/2}}\exp[-\frac{(o-\mu_i)^T\Sigma_i^{-1}(o-\mu_i)}{2}] \quad (2)$$

式中 μ_i 为均值矢量, Σ_i 为协方差矩阵, $i=1,2,...,M$.

分析 1 式可知 GMM 参数集 λ 可由各均值矢量、协方差矩阵及混合分量的权值组成, 表达成如下三元组的形式:

$$\lambda=\{c_i,\mu_i,\sum_i;(i=1,2,...,M)\} \tag{3}$$

只要拥有足够多的混合分量, 就可以逼近任意一种密度函数. 一般的语音特征都有平滑的概率密度函数, 因此有限个高斯密度函数(例如 16 个)就足以对语音特征的密度函数形成平滑逼近. 再适当的选择 GMM 的概率权重、均值和协方差, 就可以完成对一个概率密度函数的建模. 最大似然估计法是一种常用的估算 GMM 参数的方法, 最大似然估计的一个重要属性是对于足够多的训练特征矢量, 模型估计能收敛到真正的模型参数上. 然而, 求 GMM 的表达式并没有一个闭式解, 因而这是对一个“不完全数据”进行最大似然估计的问题. 针对这类问题, 其中一个解决方法就是使用 EM 算法, EM 算法会在迭代中改变 GMM 的参数估计, 在每次迭代中增加模型估计 λ 与观测特征矢量的匹配概率. 迭代公式如式(4)—式(6)所示.

$$c_i=\frac{1}{T}\sum_{t=1}^{T}P(q_t=i\mid o_t,\lambda) \tag{4}$$

$$\mu_i=\frac{\sum_{t=1}^{T}P(q_t=i\mid o_t,\lambda)o_t}{\sum_{t=1}^{T}P(q_t=i\mid o_t,\lambda)} \tag{5}$$

$$\sigma_{ik}^2=\frac{\sum_{t=1}^{T}P(q_t=i\mid o_t,\lambda)(o_{tk}-\mu_{tk})^2}{\sum_{t=1}^{T}P(q_t=i\mid o_t,\lambda)}\quad k=0,1,...,K-1 \tag{6}$$

每一次更新 λ, 都会对应一个 $P(O\mid\lambda)$, 经过多次迭代更新, 直到前后两次 $P(O\mid\lambda)$ 之间的差值小于一个门限值, 本文实验取 0. 0000001. 需要注意的是, 在实际说话人模型训练过程中, 由于语音数据自身的问题, 可能会出现个别协方差的值非常小的情况, 为了避免这种情况, 本文对每个协方差 σ_{ik}^2 的值进行查看, 并设置一个门限 0. 01, 如果这个协方差小于 0. 01, 那么就用 0. 01 取代替这个协方差值.

对于有 N 个人的说话人识别系统, 其中每个说话人用一个 GMM 模型来代表, 记为 $\lambda_1,\lambda_2\ldots\lambda_N$. 在识别阶段, 假设测试语音的声学特征矢量序列为 $O=\{o_1,o_2...o_T\}$, 则该人为第 n 个人的后验概率为:

$$p(\lambda_n\mid O)=\frac{p(O\mid\lambda_n)p(\lambda_n)}{p(O)}=\frac{p(O\mid\lambda_n)p(\lambda_n)}{\sum_{m=1}^{N}p(O\mid\lambda_m)p(\lambda_m)} \tag{7}$$

式中 $p(\lambda_n)$ 为第 n 个说话人的先验概率, $p(O)$ 为所有说话人条件下特征矢量集 O 的概率, $p(O\mid\lambda_n)$ 是第 n 个人产生特征矢量集 O 的条件概率. 识别结果由最大后验概率准则给出, 即:

$$n^*=\arg\max_{1\le n\le N}P(\lambda_n\mid O) \tag{8}$$

2 信道特性分析及鲁棒说话人识别研究

我们的仿真实验表明, 训练和测试语音来自相同环境时, 系统的识别率都较高. 但是, 当训练和测试语音来自不同环境时则识别性能明显降低. 由此得出训练环境和识别环境的不匹配是造成说话人识别率下降的一个重要原因.

改善训练和识别环境的匹配度, 可以很大程度上提高系统的识别率. 为了进一步提高训练环境和识别环境的不匹配情况下的说话人识别率, 我们对实际采用的信道进行了测试和分析, 并提出了两种简单有效的鲁棒说话人识别改进方法.

2.1 信道特性测试分析

通常, 语音通过信道后可以表示为:

$$x(n)=\{f(s(n)+v_1(n))\}*h(n)+v_2(n) \tag{9}$$

式中 $s(n)$ 为干净语音, $v_1(n)$ 为信道加性噪声, $h(n)$ 为信道函数, $v_2(n)$ 为背景加性噪声, f 为麦克风的非线性干扰函数. 我们录制语音的实验环境比较安静, 所以背景加性噪声的干扰可以近似排除. 由于电话麦克风的不同产生的非线性干扰比较复杂, 而且近年来通信行业的发展, 驻极体电话的非线性影响也较小, 所以本文主要针对无线信道的卷积效应对说话人识别的影响展开分析.

利用输入信号和输出信号之间的关系即可测出被测系统的冲击响应. 原理图如图 2.

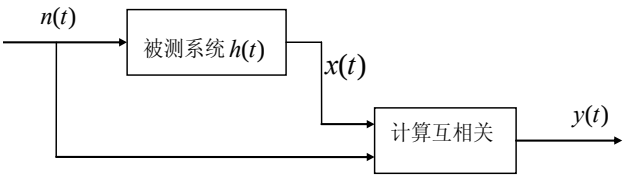

图 2 互相关法测系统冲击响应原理图、

设 $n(t)$ 为白噪声, 根据随机信号分析相应的理论, 计算系统输入 $n(t)$ 和输出 $x(t)$ 的互相关, 可间接求得系统冲击响应.

$$R_{nx}(\tau)=R_n(\tau)*h(\tau)=\frac{N_0}{2}\delta(\tau)*h(\tau)=\frac{N_0}{2}h(\tau) \tag{10}$$

测试时, 我们利用预先录制的一段白噪声信号,

在不同地点，不同时段，对不同通话中的电话播放，一共测得 15 段不同信道环境下的测试信号. 其中，白噪声开始前有大约 3 秒的静音段. 为了使输入输出信号帧间更好的对应，对电话原始输入和接收端的信号都设置了能量门限用来判断信号的起点，再按 256 个样点一段分段，每个对应小段利用公式 $H(j\omega)=N(j\omega)^{*}X(j\omega)$ 计算 $h(t)$ 的频率特性. 其中*表示复共轭，$H(j\omega)$、$N(j\omega)$、$X(j\omega)$ 分别为 $h(t)$、$n(t)$、$x(t)$ 的傅里叶变换. 最后对所有小段计算的 $H(j\omega)$ 求均值，并归一化得到该信道的频率特性. 如图 3 所示，a、b、c 为随机测得的 3 种无线信道幅频特性，其特性比较类似. d 为测得的 15 种信道的平均幅频特性，e 为其幅频特性的方差. f 为这 15 种信道的平均冲击响应.

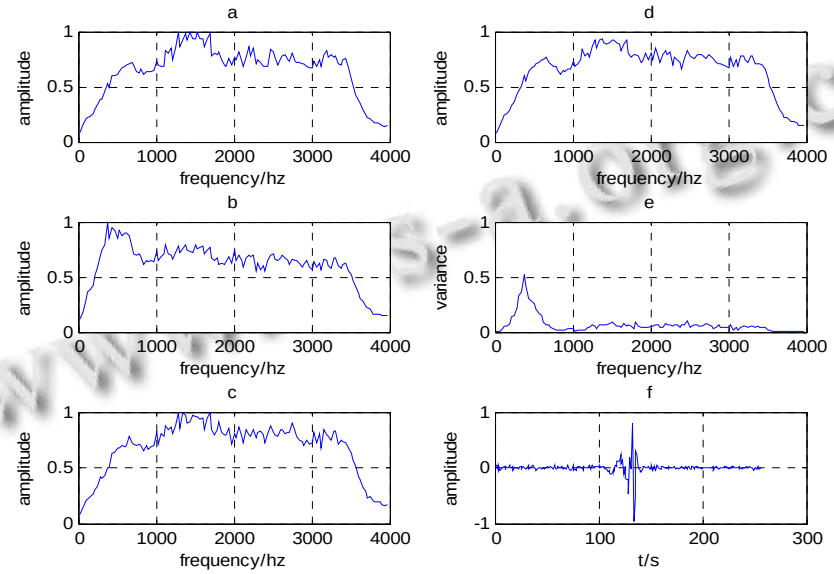

图 3　信道特性

对比图 3 中的 3 个随机信道幅频特性和 e 中的平均信道幅频特性，可知这些信道大同小异，整体走势近似. 分析 e 可知在低频部分，信道变化比较大，而高频分变化比较小. 根据实际信道特性的测试结果，我们提出了以下两种改善训练和测试语音特征参数匹配度的方法.

2.2 鲁棒性说话人识别方法研究

2. 2. 1 改进一

理论上，无线信道相当于一个带通滤波器，它会削弱 300HZ 以下和 3400HZ 以上的信号，语音信号通过信道后将产生信道失真. 所以我们在干净环境下训练的说话人系统在通信环境下识别率下降比较大. 若已知信道特性，让干净语音通过该信道后作为训练语音，则系统的识别性能可以大幅提高. 我们实验的结果是识别率可以达到 88%. 但实际应用中，每次通话的信道都不同，而且在不知输入信号的情况下想实时获得信道特性也比较困难.

不过从我们前期对不同信道的测试结果看，多次测量得到的信道特性比较类似，所以我们可以用前期已测信道特性的统计平均作为实际信道的通用模型，将干净语音通过该模型后作为训练语音. 具体原理如图 4.

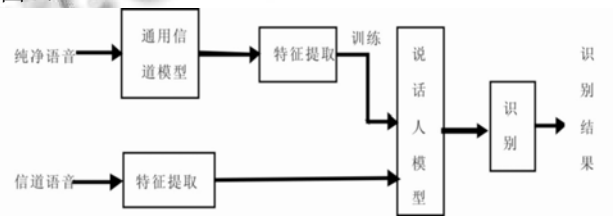

图 4　改进的失配环境说话人识别系统(1)

将纯净的语音信号通过实际测量得到的通用信道模型滤波，作为训练语音. 实验表明，这样处理后，系统的识别率有所提升.

2. 2. 2 改进二

由于识别系统采用的特征参数是 MFCC，所以我们还分析了理想低通信道、实测信道以及干净语音帧对应 MFCC 的特点，如图 5-7:

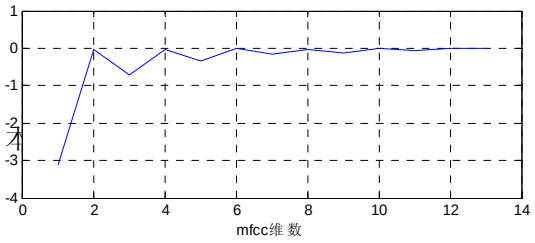

图 5 理想低通滤波器 Mfcc

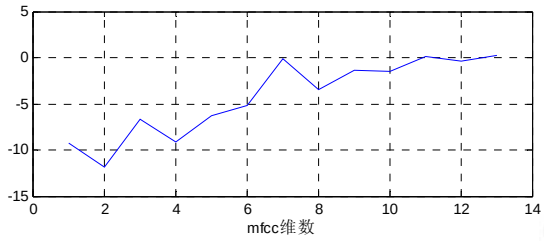

图 6 实测信道 Mfcc

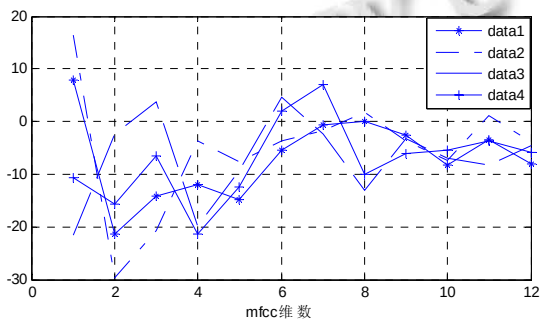

图 7 多帧语音的 Mfcc

图 5 到图 7 分别为理想低通滤波器的 13 阶 MFCC、实测 GSM 无线信道的 13 阶 MFCC 和一段语音中随机选取的 4 帧数据的 12 阶 MFCC. 其中 MFCC 参数的第一维是反映倒谱能量的, 都比较大. 有关研究表明[3], 去掉第一维参数后, 系统识别率会提高. 比较上面 3 张图的 MEL 倒谱系数, 可以发现, 理想低通滤波器的第一维系数很大, 其余的都在 0 附近波动; 实测信道的MEL 倒谱第一维和第二维都比较大; 而对于语音的 MEL 倒谱系数, 不存在明显的规律性. 信道冲击响应的 MFCC 的第二维系数较大, 说明它对语音第二维的影响也较大, 如果去掉第二维, 系统的识别效果会得到一定的改善. 改进后的识别系统原理图如图 8 所示.

提取语音 MFCC 参数后, 对参数进行特征补偿, 即去除第 1、2 维特征参数, 再进行后续的训练和识别. 实验表明, 这种方法是可行的.

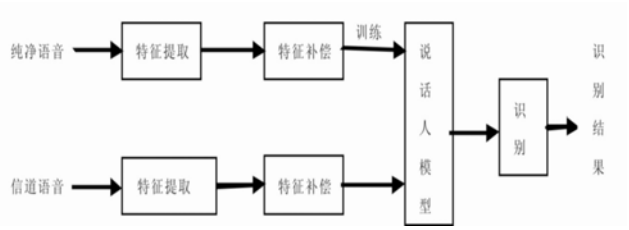

图 8 改进的失配环境说话人识别系统(2)

3 实验仿真

3.1 实验方法

本文实验语音数据来源于国际知名语音库 Timit, 随机选取其中 30 个说话人每人 10 句发音, 作为实验测试对象. Timit 数据库的语音本身为 16khz 采样率, 由于一般电话信道带宽小于 4kHz, 为了减小输入输出语音的不匹配程度, 我们将所有试验用语音都降采样为 8khz. 实验方法如图 9 所示, 两台电脑 C1 和 C2 放置于不同地点, 将用到的 300 句语音存放在 C1 中, 随机选择两部电话 A 和 B 分别放置于两电脑旁, 用 A 电话呼叫 B 电话, B 接听, 保持通话中. 将 C1 的播放功能打开, C2 的录音功能打开. 按顺序播放 A 旁边的电脑 C1 中的 300 句语音, 由电话 A 通过基站传给电话 B, 并用 B 旁边的电脑 C2 录下来.

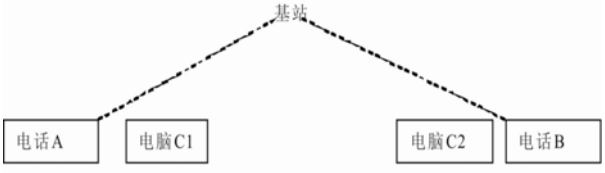

图 9 信道语音获得原理图

3.2 实验结果

本文采用 GMM-EM 算法训练得到各说话人的 GMM 模型. 文中将原始的不通过无线信道的语音称为干净语音, 将通过信道的语音称为信道语音, 后文将用干净和信道来代替这两种语音. 每一个说话人, 4 句用于训练, 6 句用于识别. 我们通过测量多部电话机, 一共采用了 300 句语音做实验. 我们对干净—干净, 干净—信道, 干净—信道(CMS), 信道—信道等多种情况进行了测试, 实验结果如表 1. 其中“干净—干净”表示采用干净语音训练, 并用干净语音识别的情况; “干净—信道”表示采用干净语音训练, 而用信道语音识别的情况, 表中分别列出了直接测试结果以及采用

本文改进方法一、二后的结果; “信道—信道”则表示训练语音和识别语音经过了相同信道的情况.

表 1 系统不同环境下识别率

	识别率(%)
干净——干净	93
干净——信道	58
干净——信道(CMS)	72
信道——信道	88
干净——信道(改进一)	84
干净——信道(改进二)	81

分析表 1 可知, 在干净环境下, 识别率能达到 93%, 在训练和测试语音为相同信道失真的情况下, 系统的识别率也能达到 88%, 说明在环境匹配的情况下, 系统的识别率还是比较理想的. 但是识别环境与训练环境失配情况下, 由于信道对识别语音的干扰. 系统识别率会迅速降低, 只能达到 58%.

采用传统的倒谱均值减方法(CMS), 识别率提升了 14%. 采用改进方法一, 将干净语音送入实际测量得到的通用信道模型滤波后, 再作为训练语音, 相应的识别率提高为 84%, 由此可知通过信道滤波技术能够使得语音的训练和识别环境有效的匹配. 但由于真实信道环境复杂多变, 已测的通用信道模型不能很好的拟合每个具体的信道, 但在一定程度削弱了信道的影响. 采用改进方法二, 将语音的 MFCC 第二维参数去掉, 识别率也提高到 81%, 正是由于信道对于语音 MFCC 参数的第二维影响较大, 所以去掉第二维后系统的识别率提升比较明显. 两种不同的处理方法都是为了尽量减小训练环境和识别环境不同造成的语音参数失配. 系统的识别率在原来干净——信道的情况下都提高了 20%左右, 与传统的 CMS 方法相比, 识别率提升了 9%-12%.

4 小结

本文针对实际通信环境下说话人系统识别率迅速下降的问题. 通过对实际通信信道的测试和分析, 提出了两种改进方法. 一是由实测数据建立了一个通用信道模型, 将干净语音经通用信道模型滤波后再作为训练语音. 二是通过对比实际信道、理想低通信道及语音 MFCC 的特点, 提出合理舍去一二维特征参数的方法. 实验结果验证了这两种方法的可行性, 与传统的 CMS 方法相比, 识别率也有所提高. 通信环境下识别比较困难的主要原因在于通信线路有其独特的信噪比和频响, 以及通信语音会受到瞬间干扰和非线性影响. 针对通信环境下语音的特点, 下一步我们准备开展有关短语音说话人识别方面的研究, 希望在拥有有限长度的语音数据的条件下, 尽可能快的识别出说话人.

参考文献

1 王伟,邓辉文.基于 MFCC 参数和 VQ 的说话人识别系统.仪器仪表学报,2006,27(Z3):2253–2255.

2 张永亮,张先庭,鲁宇明.基于 FMFCC 和 HMM 的说话人识别.计算机仿真,2010,27(5):352–354,358.

3 辛全超,吴萍.基于 GMM 的说话人识别研究与实践.计算机与数字工程,2009,37(6):11–15.

4 陈芬菲.基于 GMM 的说话人识别系统.微处理机,2006,27(4):76–77,79.

5 曹洁,潘鹏.基于 GMM 的说话人识别技术研究.计算机工程与应用,2011,47(11):114–117.

6 李轶杰,郭武,戴礼荣.话者识别的信道补偿.小型微型计算机系统,2008,29(12):2344–2347.

7 吴海洋,杨飞然,周琳,吴镇扬.矢量泰勒级数特征补偿的说话人识别.声学学报,2013,38(1):105–112.

8 梁春燕,张翔,杨琳,张建平,颜永红.最小方差无失真响应感知倒谱系数在说话人识别中的应用.声学学报,2012,6:673–678.

9 Quatieri TF. Discrete-time speech signal processing: Principles and practice. Pearson Education India, 2002.

10 Reynolds DA, Quatieri TF, Dunn RB. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 2000, 10(1): 19–41.