DOI:10.11992/tis.201511008
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160315.1248.018.html

一种语音特征提取中 Mel 倒谱系数的后处理算法

张毅[1],谢延义[2],罗元[3],席兵[3]
(1.重庆邮电大学 先进制造工程学院,重庆 400065; 2. 重庆邮电大学 自动化学院,重庆 400065; 3. 重庆邮电大学 光电工程学院,重庆 400065)

摘 要:为提高语音识别系统的鲁棒性,本文以 Mel 频率倒谱系数(MFCC)为基础,结合均值消减法、方差归一化、时间序列滤波法和加权自回归移动平均滤波法,提出了一种后处理算法,本文将该算法命名为 MVDA 后处理法,所得语音特征参数简称 MVDA。本文首先从理论上推导了 MVDA 后处理法可以去除加性噪声和卷积噪声的干扰,接着针对 MVDA 与 MFCC 做了对比试验,并分析了含噪语音与语音信号的欧氏距离变化,证明 MVDA 后处理法的每一步均有效降低了噪声的干扰,且得出了 MVDA 在不同噪声环境中均更优的结论。这种简洁的语音特征不仅可以达到许多复杂语音特征处理方法的效果,而且有效减少了自动语音识别系统的计算量。

关键词:后处理;语音特征;语音识别;噪声;鲁棒性
中图分类号:TP391.4 **文献标志码**:A **文章编号**:1673-4785(2016)02-0208-07

中文引用格式:张毅,谢延义,罗元,等. 一种语音特征提取中 Mel 倒谱系数的后处理算法[J]. 智能系统学报, 2016, 11(2): 208-215.
英文引用格式:ZHANG Yi, XIE Yanyi, LUO Yuan, et al. Postprocessing method of MFCC in speech feature extraction[J]. CAAI transactions on intelligent systems, 2016, 11(2): 208-215.

Postprocessing method of MFCC in speech feature extraction

ZHANG Yi[1], XIE Yanyi[2], LUO Yuan[3], XI Bing[3]
(1. Institute of Advanced Manufacturing Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2. College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 3. College of Opto Electronic Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: To improve the robustness of automatic speech recognition systems, a new speech feature postprocessing method based on the Mel-frequency Cepstral Coefficient (MFCC) is proposed, which is named the MVDA postprocessing method. The postprocessed feature parameters are named MVDAs. This technique combines mean subtraction, variance normalization, time sequence fltering, and autoregressive moving average flters. Experiments were conducted to compare MVDA and MFCC. Changes in the Euclidean distance of the speech with noise and the speech signal were analyzed, proving that every step of MVDA postprocessing could effectively reduce the noise interference. Thus, all MVDAs in different noise environments were superior. This simple feature does not only achieve the effect of many complex speech feature processing methods but also effectively reduces the computational complexity of automatic speech recognition systems.

Keywords: postprocessing; phonetic feature; speech recognition; noise; robustness

为了提高语音识别系统的鲁棒性,谱减法、卡尔曼滤波[1-2]和麦克风阵列[3]等语音增强技术得到应用和推广。语音特征的失真造成声学空间的变形,

收稿日期:2015-11-06. **网络出版日期**:2016-03-15.
基金项目:重庆市科委前沿技术专项重点项目(cstc2015jcyjBX0066).
通信作者:谢延义. E-mail:811719530@qq.com.

对此声学模型可以相应地调整,以弥补训练和测试语音之间的差异,这种调整通常被称为噪声模型补偿技术[4-5]。由于语音去噪的复杂性,甚至小词汇的自动语音识别系统都采用了相对复杂的处理方法[6]。这些复杂的处理方法往往会造成较大的计算量和不必要的时延,降低自动语音识别系统的灵活性[7]。

因此本文综合考虑自动语音识别系统的鲁棒性和灵敏性,有针对性地提出了一种简洁的语音信号后处理方法——MVDA 后处理法。同时,也改善了传统的 MFCC 特征提取方法中采用三角滤波器组带来的相邻频带之间的频谱能量相互泄露,且不利于反映共振特性的问题,为整个语音识别系统的优化提供了基础。实验表明,MVDA 后处理法在不同的噪声环境中的鲁棒性和灵敏性都要高于传统的 MFCC 特征提取法。

1　噪声分类和 MFCC

自动语音识别系统的鲁棒性取决于噪声、语音特征和语音信号处理方法。本节首先定义了日常声学环境中常见的噪声类型,对噪声的分类有利于本文更加清晰地分析特征失真,并且有利于描述 MVDA 后处理法。

1.1 噪声的分类

通常处理的噪声分为加性噪声和卷积噪声。加性噪声可以描述为

$$x(t) = s(t) + n(t) \tag{1}$$

式中:$\{s(t)\}$ 是语音信号,$\{n(t)\}$ 是加性噪声,$\{x(t)\}$ 是含噪语音。卷积噪声可以描述为

$$x(t) = s(t) * n(t) \tag{2}$$

式中:$*$ 是卷积符号,$h(t)$ 是环境导致的卷积噪声。此处假设环境是稳定的,在实际环境中,两种类型的噪声同时存在。因此加噪语音可以描述为

$$x(t) = s(t) * h(t) + n(t) \tag{3}$$

式(3)可以被看成是一般噪声情况下,表明语音成分与噪声成分的一种方法,可以简化为

$$x(t) = F(s(t)) \tag{4}$$

式中:F 指非线性时变环境下语音信号的映射。由于语音信号具有短时连续性,经过分帧加窗之后,语音信号在短时内接近线性时不变。

另一种失真产生于噪声环境下样本采集过程中的 Lombard 效应[8],如延长元音的持续时间和频谱向高频率倾斜,从而改变了语音信号本身。因此,噪声环境下的 $\{s(t)\}$ 本身存在失真,这种失真可以看做是式(4)的一个特例。

1.2　基础语音特征 MFCC

本文以 Mel 频率倒谱系数为基础,提出了新的语音特征提取法。MFCC 的分析基于人的听觉机理,即根据人的听觉实验结果来分析语音的频谱,期望获得更好的语音特性。MFCC 分析依据的听觉机理有两个:1)人的主观感知频域的划定并不是线性的;2)人耳听觉的临界带原理。

一帧语音信号的 MFCC 参数可以表示为 $C \triangleq (C[1]\cdots C[D])^{\mathrm{T}}$,这里 D 表示倒谱系数的维数。MFCC 的定义如下:

$$C = G\ln Q \tag{5}$$

这里 $Q \triangleq (Q[1]\cdots Q[J])^{\mathrm{T}}$ 表示每帧的谱线能量经过梅尔三角滤波器处理后的梅尔能量谱。$\boldsymbol{G}$ 是代表离散余弦变换的 $I * J$ 阶矩阵,表示为

$$G_{ij} = \sqrt{\frac{2}{J}}\cos\left(\frac{\pi i}{J}(j-0.5)\right)$$
$$i = 1,2,\cdots,I, \quad j = 1,2,\cdots,J \tag{6}$$

MFCC 特征提取方法采用三角滤波器组处理,同时也带来的相邻频带之间频谱能量的相互泄露。

2　MVDA 后处理法步骤

MVDA 后处理法在 MFCC 特征提取法的基础上,融合了均值消减、方差归一化、时间序列滤波和加权自回归移动平均滤波法,图 1 为 MVDA 后处理法基本步骤。

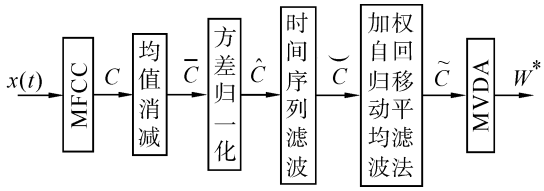


图 1　MVDA 后处理法

Fig.1　Postprocessing of MVDA

MVDA 的提出是为了解决 MFCC 特征参数的加性和卷积噪声的问题,均值消减和方差归一化在语音处理中已经得到了相对广泛的应用[9-12]。本文提出了结合时间序列滤波和加权自回归移动平均滤波法在频域的应用,可以获得相较于单独使用均值消减和方差归一化更好的效果。

本文用 $C^{(\tau)}$ 表示第 τ 帧语音的特征,则均值消减表示为

$$\bar{C}^{(\tau)} = C^{(\tau)} - \mu \tag{7}$$

式中：μ 是根据样本数据估计的均值项。方差归一化法表示为

$$\hat{C}^{(\tau)}[d] = (\sigma^2[d])^{-1/2}\bar{C}^{(\tau)}[d] \tag{8}$$

式中：$\hat{C}$ 是均值消减和方差归一化之后的特征，$\sigma^2[d]$ 是特征向量第 d 维的估计方差。本文的时间序列滤波法表示为

$$\breve{C}^{(\tau)} = \frac{\sum_{k=1}^{w} k^2\hat{C}^{(\tau+k)} - \sum_{k=1}^{w}(k-1)^2\hat{C}^{(\tau-k)}}{(4k-2)\sum_{k=1}^{w}k^2} \tag{9}$$

式中：$\breve{C}$ 是均值消减、方差归一化和时间序列滤波之后的特征，k 代表时间序列的宽度，w 为其最大宽度。本文的加权自回归移动平均滤波法表示为

$$\tilde{C}^{(\tau)} = [\tilde{C}^{(\tau-m)} + \cdots + (m-1)\tilde{C}^{(\tau-1)} + m\tilde{C}^{(\tau)}]/m^2 + [(m-1)\breve{C}^{(\tau+1)} + \cdots + \breve{C}^{(\tau+m)}]/m^2 \tag{10}$$

式中：$\breve{C}$ 是 MVDA 滤波之后的特征，m 代表加权自回归移动平均滤波法深度，特殊情况 $m = 1$ 表示没有加权自回归移动平均滤波处理，综合考虑算法的复杂度和准确度，一般取 $m = 3$。

均值 μ 和方差 σ^2 的估计可以采用多种方法。在方差估计法[13]中，均值和方差根据一整段对话语音估计。如果环境是静态的，则这种估计是相对稳定的。而根据在线估计法[14]，均值和方差可以不依赖将来的特征观察值，根据当前样本估计，这种策略时延低，适用于灵敏度要求高的系统。介于这两种策略之间的是语句估计法。本文中的所有结果都基于语句估计，其定义为

$$\mu = \frac{1}{T}\sum_{\tau=1}^{T} C^{(\tau)}$$

$$\sigma^2[d] = \frac{1}{T}\sum_{\tau=1}^{T}(C^{(\tau)}[d] - \mu[d])^2$$

式中：T 为给定语句中的帧数。注意在语句归一化法中，结果可能被语音前后的空白和噪声影响[15]，本文的研究假设在计算均值和方差统计之前，已经对语音进行了合理的分割。

3 噪声影响与 MVDA 滤波法分析

关于频域加性和卷积噪声，本文均作了详细的分析。本节从理论上推导 MVDA 滤波法，分析均值消减、方差归一化、时间序列滤波和加权自回归移动平均滤波法的去噪效果，并分析在滤波前后噪声对语音特征的影响。

3.1 均值消减

本文首先分析卷积噪声对语音特征造成的失真，并且得出均值消减可以有效去除卷积噪声。分析表明，频域均值消减导致参数在时不变卷积噪声下是稳定的。

卷积噪声在频域内表现为乘法运算，因此 $\{x(t)\}$、$\{s(t)\}$ 和 $\{h(t)\}$ 的功率谱可以表示为

$$P_x[k] = P_s[k]P_h[k]$$

式中：$P_x[k] = |X[k]^2|$，$X[k]$ 为语音信号 $x[n]$ 的离散傅里叶变换。根据式(5)，x 的第 i 维参数为

$$C_x[i] = \sum_{j=1}^{J} G_{ij}\ln\Big(\sum_{k=0}^{N-1} F_{jk}P_x[k]\Big)$$

式中：F_{jk} 表示第 j 个 Mel 特征滤波器的第 k 条谱线。

一般情况下，C_x 和 C_s 并不是简单的通过 h 关联，因为对数的参数求和不能被因式分解。如果假设 P_h 是相对平滑的，每一个 Mel 滤波器频带内卷积噪声的变化很小。

$$\sum_{k=0}^{N-1} F_{jk}P_x[k] = \sum_{k=0}^{N-1} F_{jk}P_s[k]P_h[k] \approx P_h[k_j]\sum_{k=0}^{N-1} F_{jk}P_s[k]$$

式中：$P_h[k_j]$ 为 $\{h(t)\}$ 在第 j 维滤波器中的能量谱。

$$C_x[i] = \sum_{j=1}^{J} G_{ij}\ln\Big(P_h[k_j]\sum_{k=0}^{N-1} F_{jk}P_s[k]\Big) = \sum_{j=1}^{J} G_{ij}\Big(\ln P_h[k_j] + \log\Big[\sum_{k=0}^{N-1} F_{jk}P_s[k]\Big]\Big) = \sum_{j=1}^{J} G_{ij}(\ln P_h[k_j] + \ln Q_s[j]) = B_h[i] + C_s[i]$$

式中 $B_h[i] \triangleq \sum_{j=1}^{J} G_{ij}\ln P_h[k_j]$。

上述假设不排除在 P_h 在 Mel 频域滤波器的不同频带内产生变化，而只要求其在每个频带内的变化足够小，该假设要求设计良好的传输设备通带。然而在多噪声环境中，从声源到接收者的多路径反射可能导致峰谷的频率响应[16]，不满足上述假设。因此第 i 维噪声和语音信号 MFCC 的差别与 $\{h(t)\}$，而与 $\{s(t)\}$ 无关。也就是说，卷积噪声增加特征的偏置取决于瞬时的信道特性数值。如果进一步假设噪声是稳态的，对于 MFCC，有

$$\bar{C}_x^{(\tau)}[i] = C_x^{(\tau)}[i] - \mu_x[i] = C_x^{(\tau)}[i] + B_h[i] - (\mu_s[i] + B_h[i]) = C_x^{(\tau)}[i] - \mu_s[i] = \bar{C}_s^{(\tau)}[i], i = 0,1,\cdots,I$$

因此在稳态噪声和相对平滑的卷积噪声环境下,均值消减特征不会改变。从而在语句结构中,如果环境噪声是卷积类型并且在语句内是稳态的、平滑的,均值消减法是有效的。对均值消减的上述特性均建立在卷积噪声的基础上。对于加性噪声的分析将在后面三级滤波中进行分析。

3.2　方差归一化

加性噪声不同于卷积噪声,在经过频域变换之后语音与加性噪声更加难以区分,为了更加方便地分析加性噪声环境下的语音信号,我们将含噪语音定义为

$$x(t;\gamma) = s(t) + n(t;\gamma) = s(t) + \gamma n_0(t)$$

式中:加性噪声 $n(t;\gamma) \triangleq \gamma n_o(t)$ 中的 γ 变量表示噪声的强度。本文首先分析加性噪声,然后分析语音信号。$n(t;\gamma)$ 和 $n_o(t)$ 在 Mel 频域的对数特征表示为

$$\ln Q_{n(\gamma)}[j] = \ln\left(\sum_{k=0}^{N-1} F_{jk}(\gamma^2 P_{n_o}[k])\right) = 2\ln|\gamma| + \ln\left(\sum_{k=0}^{N-1} F_{jk} P_{n_o}[k]\right) = 2\ln|\gamma| + \ln Q_{n_o}[j]$$

式中:$Q_{n(\gamma)}$ 和 Q_{n_o} 分别是 $n(t;\gamma)$ 和 $n_o(t)$ 的 Mel 频率谱表示,Mel 倒谱系数可以表示为

$$C_{n(\gamma)}[i] = \sum_{j=1}^{J} G_{ij}(2\ln|\gamma| + \ln Q_{n_o}[j]) = C_{n_o}[i]$$

式中:$C_{n(\gamma)}$ 和 C_{n_o} 分别是 $n(t;\gamma)$ 和 $n_o(t)$ 的倒谱,MFCC 并没有衰减。含噪语音的功率谱为

$$P_{x(\gamma)}[k] = |S[k]^2| + 2\gamma|S[k]N_o[k]| + \gamma^2|N_o[k]^2| = P_s[k] + 2\gamma|S[k]N_o[k]| + \gamma^2 P_{n_o}[k]$$

式中:$P_{x(\gamma)}$、P_s 和 P_{n_o} 分别表示 $x(t;\gamma)$、$s(t)$ 和 $n_o(t)$ 的功率谱。由于 Mel 分级是线性运算,因此

$$Q_{x(\gamma)}[j] = Q_s[j] + 2\gamma Q_1[j] + \gamma^2 Q_{n_o}[j]$$

式中:$Q_1[j] \triangleq \sum_{k=0}^{N-1} F_{jk}|S[k]N_o[k]|$,$Q_{x(\gamma)}$、$Q_s$ 和 Q_{n_o} 分别代表 $x(t;\gamma)$、$s(t)$ 和 $n_o(t)$ 的功率谱。Mel 特征频谱的失真由两部分构成:一部分取决于噪声和语音信号,并且与 γ 成正比。另一部分只取决于噪声,并且与 γ^2 成正比。根据式(5):

$$C_{x(\gamma)}[i] = \sum_{j=1}^{J} G_{ij}\ln Q_{x(\gamma)}[j] = \sum_{j=1}^{J} G_{ij}\ln(Q_s[j] + 2\gamma Q_1[j] + \gamma^2 Q_{n_o}[j]) = C_s[i] + \sum_{j=1}^{J} G_{ij}\ln\left(1 + 2\gamma\frac{Q_1[j]}{Q_s[j]} + \gamma^2\frac{Q_{n_o}[j]}{Q_s[j]}\right) = C_s[i] + \delta C_{x(\gamma)}[i],\quad i = 1,2,\cdots,I$$

语音失真为

$$\delta C_{x(\gamma)}[i] \triangleq \sum_{j=1}^{J} G_{ij}\ln\left(1 + 2\gamma\frac{Q_1[j]}{Q_s[j]} + \right)\gamma^2\frac{Q_{n_o}[j]}{Q_s[j]}$$

因此失真与语音信号 $s(t)$ 和噪声 $n(t;\gamma)$ 相关。一般强度的加性噪声影响与语音信号、噪声类型和噪声强度有着复杂的关系,因此加性噪声的滤波相对困难。当存在噪声语音数据样本时,可以考虑设计潜在的非线性变换来减小语音信号的失真。

均值消减法的使用无法弥补 ($C_{e2} \neq C_s$) 造成的失真。处理含噪语音的方法有两种,一种是直接使用含噪语音样本,另一种是非线性变换去噪,直接使用含噪语音必须与测试语音噪声匹配。

加性噪声造成的语音信号失真不仅仅取决于噪声的加性增益,而与语音信号和噪声均相关,因此很难去除加性噪声。在低噪声环境下这种关联并不明显。高噪声环境下,在去除噪声增益项之后,本文应用了方差归一化法以弥补语音信号特征的衰减。由于存在 γ^{-1} 的增益,在使用方差归一化法后,也无法得到零加性噪声的语音信号,因此处理后的语音特征很难满足要求。

3.3　时间序列滤波和加权自回归移动平均滤波

本文首先分析了没有假设 γ 的语音信号失真。以此为依据建立了方差归一化法,并基于该方法的不足,分析低噪声 $|\gamma| \ll 1$ 和高噪声 $|\gamma| \gg 1$,这两种噪声情况都可以通过近似来简化。

1)低加性噪声

当 $|\gamma| \ll 1$ 时,失真可以简化为

$$\delta C_{x(\gamma)}[i] \approx \sum_{j=1}^{J} G_{ij}\ln\left(1 + 2\gamma\frac{Q_1[j]}{Q_s[j]}\right) \approx 2\gamma C_{e1}[i]$$

式中:$C_{e1}[i] \triangleq \sum_{j=1}^{J} G_{ij}(Q_1[j]/Q_s[j])$,并且 $\ln(1+x) \approx x$。

2)高加性噪声

当 $|\gamma| \gg 1$ 时,失真可简化为

$$Q_{x(\gamma)}[i] \approx \sum_{j=1}^{J} G_{ij}\ln\left(\gamma^2\left(Q_{n_o}[j] + \frac{2}{\gamma}Q_1[j]\right)\right)$$

并且失真之后的 MFCC 特征近似为

$$C_{x(\gamma)}[i] \approx \sum_{j=1}^{J} G_{ij}\left(\gamma^2\left(Q_{n_o}[j] + \frac{2}{\gamma}Q_1[j]\right)\right) \approx \sum_{j=1}^{J} G_{ij}\ln\left(\gamma^2 Q_{n_o}\left(1 + \frac{2}{\gamma}\frac{Q_1[j]}{Q_{n_o}[j]}\right)\right) \approx$$

$$\sum_{j=1}^{J} G_{ij}\left(2\ln|\gamma| + \ln Q_{n_o}[j] + \frac{2}{\gamma}\frac{Q_1[j]}{Q_{n_o}[j]}\right) \approx C_{n_o}[i] + \frac{2}{\gamma}C_{e2}[i], \quad i = 1,2,\cdots,I$$

式中：$C_{e2}[i] \triangleq \sum_{j=1}^{J} G_{ij}(Q_1[j]/Q_{n_o}[j])$。其倒谱主要与噪声 $n_o(t)$ 相关，并且通过 C_{e2} 与语音强度成反比，倒谱特征的失真不只是偏置。由此，低噪声 $|\gamma| \ll 1$ 和高噪声 $|\gamma| \gg 1$ 时的噪声均反映了信号的不稳定性，因此强调语音动态特性和低频特性，将有助于加性噪声的去除。

人耳对语音的动态特征更为敏感，这种动态特性可以通过时间序列滤波实现。时间序列滤波之后的语音信号更接近真实语音信号。时间序列滤波器在语音信号静态特性的基础上，又兼顾了语音信号的动态特性，其使用达到了预期的目的。

由于人类的声音频率的结构性限制，发声时声道系统结构的改变有限，人类语音的重要信息主要是在低频段[17]。由于 MFCC 反映声道系统的特性，本文假设语音低频特征包含的信息更多。均值消减和方差归一化方法可以弥补能谱的下降，但却不能解决谱型平滑的问题。而加权自回归移动平均滤波由于强调了语音低频段的作用，并弱化了高频的影响。

4 实验设计及分析

实验数据库为用 cooledit 软件建立语音样本库。数据库规模为 100 人(50 男 50 女)，考虑时间的遍历性，同一段指令要求在不同的时间录制 10 遍。语音采样率 16 kHz，单声道，Windows PCM 编码格式，采样精度 16 位。噪声添加使用 Noise-92 库中的 pink、volvo、destroyerengine(DE)、和 white 噪声，根据随机时间偏移与纯净语音信号混合，形成 -5~20 dB 范围内不同信噪比的数据库。

本文语音信号分帧采用交叠分段的法，每帧 170 个采样点，叠加步长为 15 个采样点，对信号进行特征提取得 MFCC，设定特征维数为 25。再以 MFCC 为基础，获得 MVDA 语音特征。

图 2~9 是语音“12345”在噪声环境下，MVDA 特征向量的第一维和第 D 维特征。通过对比发现干净语音和不同信噪比的含噪语音的差异。均值消减和方差归化法使语音信号和含噪信号在同平均水平(均值消减)和总体规模(方差归一化法)的差异减小，然而差别依然明显。本文进一步使用了时间序列滤波和加权自回归移动平均滤波，差异进一步减小。

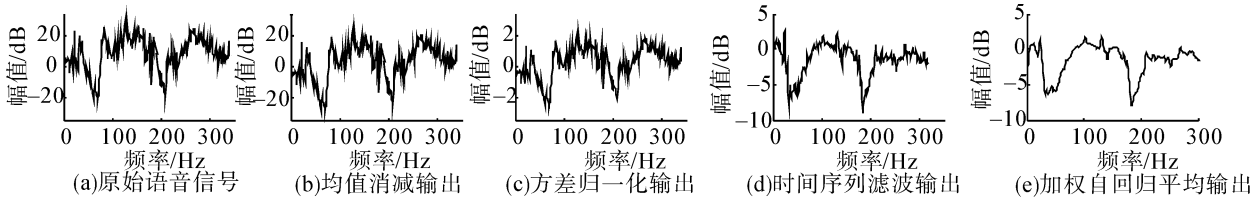


图 2 语音特征 $C[1]$ 噪声为 20 dB 时，MVDA 后处理输出

Fig.2 The MVDA postprocessing output of voice features $C[1]$ with noise of 20 dB

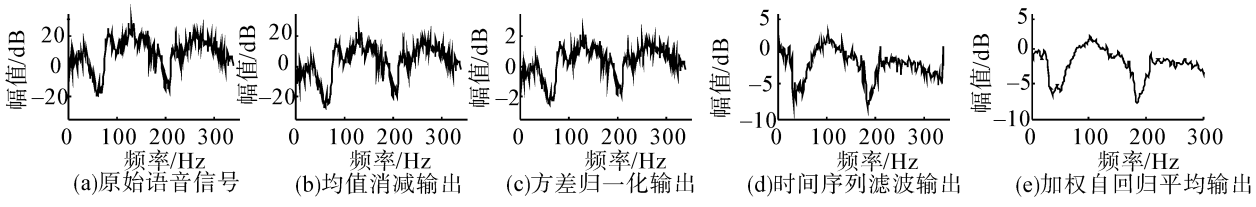


图 3 语音特征 $C[1]$ 噪声为 10 dB 时，MVDA 后处理输出

Fig.3 The MVDA postprocessing output of voice features $C[1]$ with noise of 10 dB

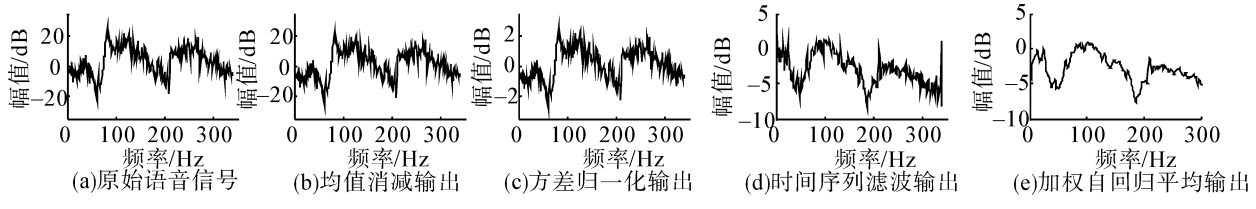


图 4 语音特征 $C[1]$ 噪声为 0 dB 时，MVDA 后处理输出

Fig.4 he MVDA postprocessing output of voice features $C[1]$ with noise of 0 dB

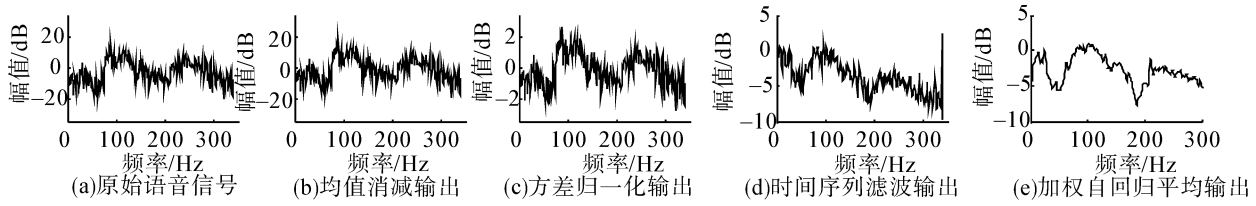


图 5　语音特征 $C[1]$ 噪声为 −5 dB 时,MVDA 后处理输出

Fig.5　The MVDA postprocessing output of voice features $C[1]$ with noise of −5 dB

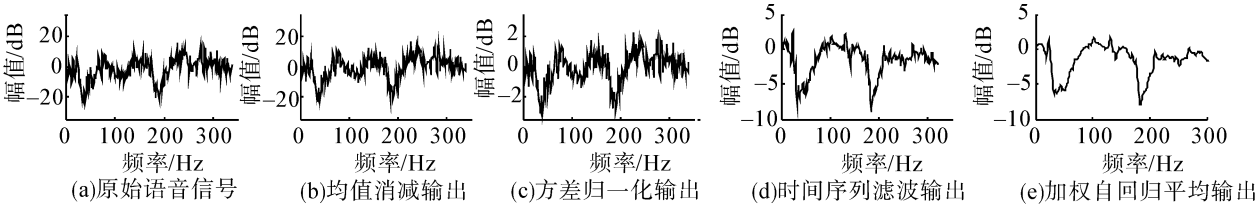


图 6　语音特征 $C[D]$ 噪声为 20 dB 时,MVDA 后处理输出

Fig.6　The MVDA postprocessing output of voice features $C[D]$ with noise of 20 dB

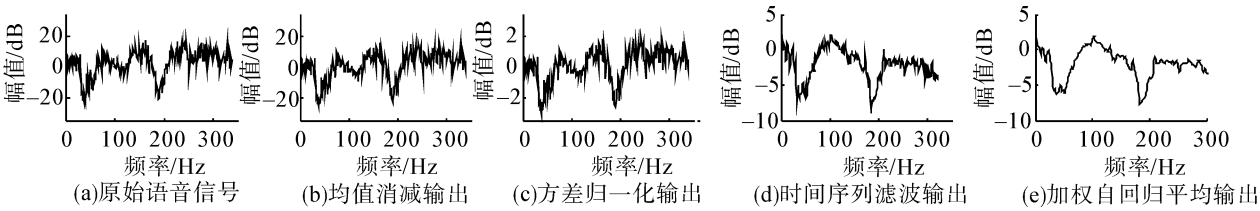


图 7　语音特征 $C[D]$ 噪声为 10 dB 时,MVDA 后处理输出

Fig.7　The MVDA postprocessing output of voice features $C[D]$ with noise of 10 dB

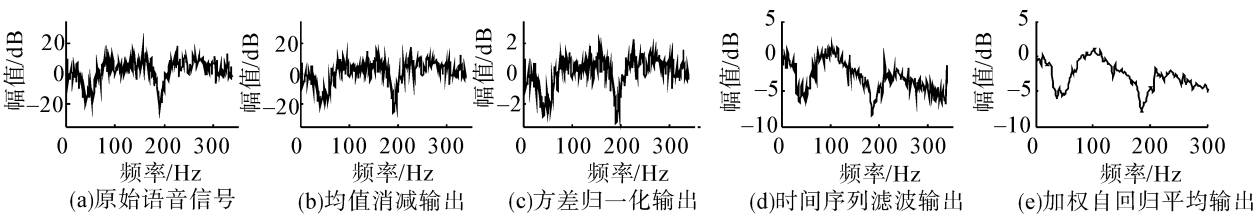


图 8　语音特征 $C[D]$ 噪声为 0 dB 时,MVDA 后处理输出

Fig.8　The MVDA postprocessing output of voice features $C[D]$ with noise of 0 dB

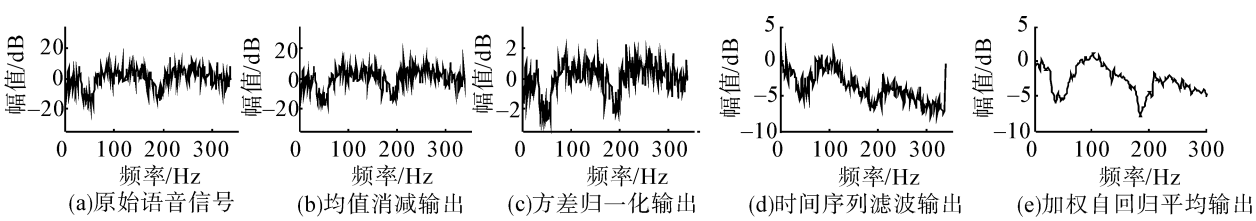


图 9　语音特征 $C[D]$ 噪声为−5 dB 时,MVDA 后处理输出

Fig.9　The MVDA postprocessing output of voice features $C[D]$ with noise of −5 dB

然而使用视觉检查推断语音识别处理方法的不确定性总是存在的。为了便于比较,本文计算了语音信号特征和带噪语音信号特征的欧式距离,具体数值见表 1。可以分析得出,含噪语音特征和无噪语音信号特征的欧式距离均与噪声强度正相关。均值消减和方差归一化减小了含噪语音特征与无噪语音信号特征的欧式距离。最终,时间序列滤波和加权自回归移动平均滤波进一步减小了欧式距离。根据表 1,加权自回归移动平均滤波处理后的带噪语音更加接近真实的语音信号。

表 1　含噪语音 MVDA 参数与语音信号的欧氏距离

Table 1　The compasison of training beteeen MVDA and MFCC

参数	20/dB	10/dB	0/dB	−10/dB
均值消减	939	1 356	1 845	1 956
方差归一化	129	196	259	346
时间序列滤波	78	112	136	203
加权自回归移动平均	61	69	72	76

将 MVDA 与 MFCC 特征在自动语音识别系统下进行语音识别实验对比,实验结果如图 4 。可以得出,信噪比较高时,MFCC 特征与 MVDA 特征的识别率基本相同,但随着信噪比降低,MVDA 语音特征的效果更加显著。

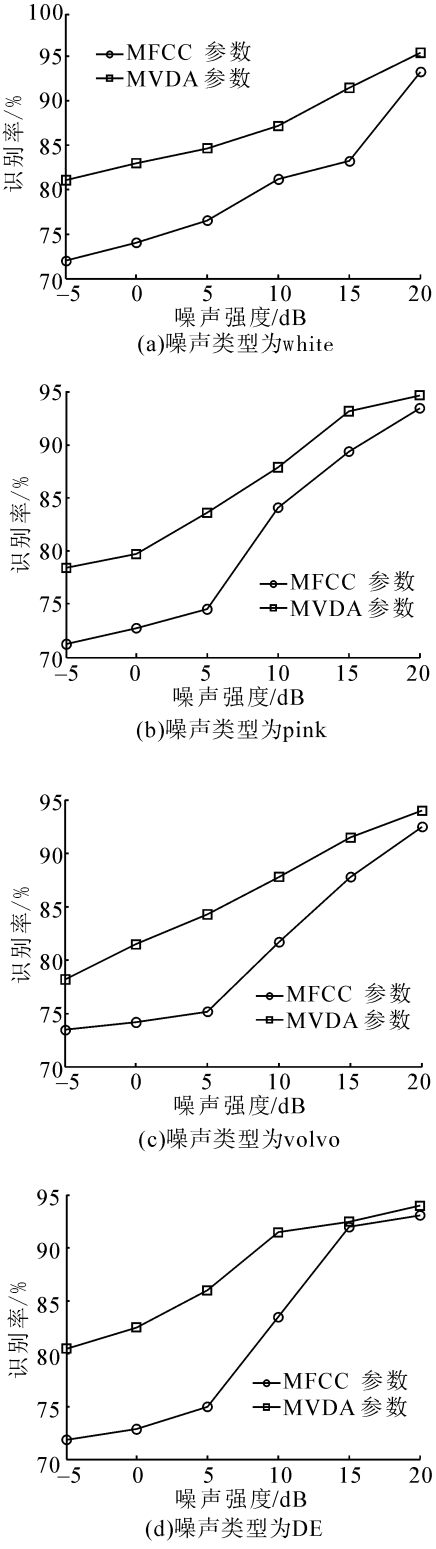


图 10 自动语音识别结果对比图

Fig.10 Comparison of automatic speech recognition results

5 结束语

本文的分析主要基于加性噪声和卷积噪声环境下 MFCC 特征参数的失真,针对这一问题提出了 MVDA 语音特征提取法。分析得出实验效果与语音基本特征、滤波器的类型均相关。在使用 MVDA 滤波法后,相较于 MFCC 语音特征,自动语音识别系统在不同性噪比环境下的识别率提高了 2.7% ~ 15.0%。MVDA 特征提取可以达到很多复杂去噪算法的效果,却可以减少系统对计算能力的要求,减小系统的时延。因此,MVDA 后处理法可以在更小的计算代价下提高系统的鲁棒性,具有较高的实际应用价值。

参考文献:

[1] PALIWAL K K, BASU A. A speech enhancement method based on Kalman fltering[C]//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Dallas, USA, 1997: 177-180.

[2] GIBSON J D, KOO B, GRAY S D. Filtering of Colored Noise for Speech Enhancement and Coding[J]. IEEE Transactions on Signal Processing, 1991, 39(8): 1732-1742.

[3] ZELINSKI R. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms[C]//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. New York, USA, 1998: 2578-2581.

[4] MYLLYMAKI M, VIRTANEN T. Non-stationary noise model compensation in voice activity detection[C]//Proceedings of IEEE International Conference on Signal Processing Conference. Glasgow, Scotland, 2009: 2186-2190.

[5] RAMFREZ J, SEGURA J C, BENFTEZ C, et al. Efficient voice activity detection algorithms using long-term speech information[J]. Speech communication, 2004, 42(3/4): 271-287.

[6] CHOWDHURY M, SELOUANI S A, O'SHAUGHNESSY D. A soft computing approach to improve the robustness of online ASR in previously unseen highly non-stationary acoustic environments[C]//Proceedings of the 11th IEEE International Conference on Information Science, Signal Processing and their Applications. Montreal, Canada, 2012: 522-527.

[7] GUPTA H A, RAJU A, ALWAN A. Non-linear dimension reduction of Gabor features for noise-robust ASR[C]//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Florence, Italy, 2014: 1715-1719.

[8] HANSEN J H L, VARADARAJAN V. Analysis and com-

pensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition[J]. IEEE transactions on audio, speech, and language processing, 2009, 17(2): 366-378.

[9] COOK G, ROBINSON T. Transcribing broadcast news with the 1997 abbot system[C]//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Seattle, USA, 1998: 917-920.

[10] KIM D S, LEE S Y, KIL R M. Auditory processing of speech signals for robust speech recognition in real-world noisy environments[J]. IEEE transactions on speech and audio processing, 1999, 7(1): 55-69.

[11] HAIN T, WOODLAND P C, EVERMANN G, et al. New features in the CU-HTK system for transcription of conversational telephone speech[C]//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City, UT, 2001(1): 57-60.

[12] LIN S H, CHEN B, YEH Y M. Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition[J]. IEEE transactions on audio, speech, and language processing, 2009, 17(1): 84-94.

[13] MORTIA S, UNOKI M, LU Xugang, et al. Robust voice activity detection based on concept of modulation transfer function in noisy reverberant environments[C]//Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP). Singapore, 2014: 108-112.

[14] CHANG J E, BAI J Y, ZENG Fangang. Unintelligible low frequency sound enhances simulated cochlear implant speech recognition in noise[J]. IEEe transactions on biomedical engineering, 2006, 53(12): 2598-2601.

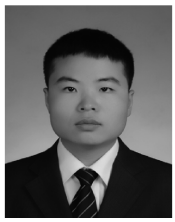
[15] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE transactions on acoustics, speech, and signal processing, 1999, 27(2): 113-120.

[16] MAMMONE R J, ZHANG Xiaoyu, RAMACHANDRAN R P. Robust speaker recognition: a feature-based approach[J]. IEEE signal processing magazine, 1996, 13(5): 58-71.

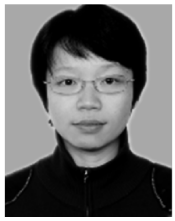
[17] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE transactions on acoustics, speech, and signal processing, 1999, 27(2): 113-120.

作者简介:

张毅,男,1966 年生,教授,博士生导师。主要研究方向机器人及应用、数据融合、信息无障碍技术。任重庆邮电大学国家信息无障碍工程研发中心主任,智能系统及机器人实验室主任,发表学术论文多篇。



谢延义,男,1989 年生,硕士研究生,主要研究方向为语音识别与智能机器人。



罗元,女,1972 年生,教授,博士,主要研究方向为信号与信息处理、数字图像处理。