

文章编号:1001-9081(2016)12-3369-05

DOI:10.11772/j.issn.1001-9081.2016.12.3369

用于自动语音识别系统的切换语音功率谱估计算法

刘金刚*, 周翊, 马永保, 刘宏清

(重庆邮电大学 通信与信息工程学院, 重庆 400065)

(* 通信作者电子邮箱 jg_liu@outlook.com)

摘要:针对语音识别系统在噪声环境下不能保持很好鲁棒性的问题,提出了一种切换语音功率谱估计算法。该算法假设语音的幅度服从 Chi 分布,提出了一种改进的基于最小均方误差(MMSE)的语音功率谱估计算法。然后,结合语音存在的概率(SPP),推导出改进的基于语音存在概率的 MMSE 估计器。接下来,将改进的 MMSE 估计器与传统的维纳滤波器结合。在噪声干扰比较大时,使用改进的 MMSE 估计器来估计纯净语音的功率谱,当噪声干扰较小时,改用传统的维纳滤波器以减少计算量,最终得到用于识别系统的切换语音功率谱估计算法。实验结果表明,所提算法相比传统的瑞利分布下的 MMSE 估计器在各种噪声的情况下识别率平均提高在 8 个百分点左右,在去除噪声干扰、提高识别系统鲁棒性的同时,减小了语音识别系统的功耗。

关键词:自动语音识别系统;鲁棒性;最小均方误差;语音存在概率;功率谱估计;维纳滤波器

中图分类号: TN912.35 **文献标志码:**A

Estimation algorithm of switching speech power spectrum for automatic speech recognition system

LIU Jingang*, ZHOU Yi, MA Yongbao, LIU Hongqing

(School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: In order to solve the poor robust problem of Automatic Speech Recognition (ASR) system in noisy environment, a new estimation algorithm of switching speech power spectrum was proposed. Firstly, based on the assumption of the speech spectral amplitude was better modelled for a Chi distribution, a modified estimation algorithm of speech power spectrum based on Minimum Mean Square Error (MMSE) was proposed. Then incorporating the Speech Presence Probability (SPP), a new MMSE estimator based on SPP was obtained. Next, the new approach and the conventional Wiener filter were combined to develop a switch algorithm. With the heavy noise environment, the modified MMSE estimator was used to estimate the clean speech power spectrum; otherwise, the Wiener filter was employed to reduce calculating amount. The final estimation algorithm of switching speech power spectrum for ASR system was obtained. The experimental results show that, compared with the traditional MMSE estimator with Rayleigh prior, the recognition accurate of the proposed algorithm was averagely improved by 8 percentage points in various noise environments. The proposed algorithm can improve the robustness of the ASR system by removing the noise, and reduce the computational cost.

Key words: Automatic Speech Recognition (ASR) system; robustness; Minimum Mean Square Error (MMSE); Speech Presence Probability (SPP); estimation of speech power spectrum; Wiener filter

0 引言

近年来,语音识别系统广泛应用于智能设备、车载系统和互联网等领域。在安静环境下,语音识别系统的识别率可高达 95%~99%,但在实际应用中,环境噪声会导致语音识别系统的识别率大大降低。文献[1]中提到了诸多用于提高语音识别系统鲁棒性的语音增强算法。其中,基于最小均方误差的短时谱估计语音增强算法具有复杂度不高、易于实时实现以及产生的音乐噪声小等特点,通常被用于语音识别系统中作为噪声抑制模块以改善语音的质量,从而提高识别系统的鲁棒性。基于文献[2]的研究,文献[3]发展出的最优改进

对数幅度谱(Optimal Modified Log Spectral Amplitude, OMLSA)估计算法被用于提高识别系统的鲁棒性^[4],并取得了一定的效果。但 OMLSA 算法是对语音的幅度谱进行估计和处理,而并非直接对用于识别的特征进行增强,因此该方法是次优的^[4]。文献[5~6]提出了一种最优的特征增强的估计算法,但其只能对梅尔倒谱系数(Mel-Frequency Cepstral Coefficient, MFCC)特征参数进行增强,而不适用于线性预测倒谱系数(Linear Predictive Cepstral Coefficient, LPCC)和感知线性预测(Perceptual Linear Prediction, PLP)系数等特征。

如文献[7]所述,在一些应用中,语音功率谱估计算法有时会比语音幅度谱估计算法取得更好的效果。因此,本文改

收稿日期:2016-05-25;修回日期:2016-07-12。

基金项目:国家自然科学基金资助项目(61501072);重庆市科委自然科学基金资助项目(cstc2015jcyjA40027)。

作者简介:刘金刚(1991—),男,山东诸城人,硕士研究生,主要研究方向:语音信号处理、语音增强;周翊(1974—),男,四川成都人,教授,博士,主要研究方向:自适应滤波、语音信号处理;马永保(1991—),男,甘肃武威人,硕士研究生,主要研究方向:语音信号处理、语音增强;刘宏清(1980—),男,黑龙江佳木斯人,教授,博士,主要研究方向:稀疏信号处理,阵列信号处理。

进了一种短时语音功率谱估计算法以提高语音识别系统的鲁棒性，并结合语音存在的概率进一步提高算法的性能。改进之处在于本文在推导结合了 SPP 的 MMSE 语音功率谱估计算法时，假设语音的幅度谱是服从 Chi 分布而非传统的瑞利分布的，该特性也在文献[8-9]中得以验证。此外，考虑到语音识别系统应用于移动终端设备时对功耗的限制，本文提出了一个切换的短时功率谱估计算法，以降低识别系统的总功耗。切换算法根据每一帧语音的最大似然先验信噪比和语音活动性检测（Voice Activity Detection, VAD）技术以判断使用两种的语音功率谱估计算法之一。在信噪比较低的时段，识别系统采用改进的 MMSE 语音功率谱估计算法估计纯净语音的功率谱；而在高信噪比时，识别系统使用复杂度较低的维纳语音功率谱估计器，以减小语音增强算法的运算量。因此新算法在去除噪声干扰以提高识别系统鲁棒性的同时，能节约语音识别系统的功耗。

1 模型假设和变量表示

带噪语音模型为： $y(n) = x(n) + d(n)$ 。其中： $y(n)$ 是带噪语音信号， $x(n)$ 是纯净语音信号， $d(n)$ 是加性噪声信号。对信号分帧后再进行傅里叶变换后得到： $Y_k(i) = X_k(i) + D_k(i)$ ，其中 $Y_k(i)$ ， $X_k(i)$ 和 $D_k(i)$ 分别代表信号 $y(n)$ 、 $x(n)$ 和 $d(n)$ 的第 i 帧中第 k 个频点，假设它们是零均值而且相互独立的随机变量。为了简化表示，可省去表示帧的变量 i ，例如第 i 帧带噪语音 $Y_k(i) = X_k(i) + D_k(i)$ 可表示为： $Y_k = X_k + D_k$ 。 $\xi_k = \lambda_x(k)/\lambda_d(k)$ 是第 i 帧中第 k 个频点的先验信噪比，而 $\gamma_k = |Y_k|^2/\lambda_d(k)$ 为对应频点的后验信噪比^[2]，其中 $|Y_k|^2$ 、 $\lambda_x(k)$ 和 $\lambda_d(k)$ 分别为带噪语音信号、纯净语音信号和噪声信号的方差。 $Y_k = R_k e^{j\varphi}$ 和 $X_k = A_k e^{j\theta}$ 分别是 Y_k 和 X_k 的极坐标表示形式， φ 和 θ 分别是对应的相位。

假设语音幅度谱的观测值服从 Chi 分布：

$$p(A_k) = \frac{2\beta^a}{\Gamma(a)} A_k^{2a-1} \exp(-\beta A_k^2) \quad (1)$$

其中： $2a$ 是 Chi 分布的自由度， $\beta = a/\lambda_x(k)$ 。Chi 分布在 $a = 1$ 时可以简化为一个瑞利分布。文献[10]中提到，在 $a < 1$ 时，Chi 分布可以很好地接近语音幅度谱的观察值。接下来，假设加性噪声 $d(n)$ 的离散傅里叶变换（Discrete Fourier Transform, DFT）系数服从复高斯分布，并且每一帧的每一个频点之间都相互独立^[2]，于是可得到带噪语音在已知纯净语音的幅度和相位条件下的后验概率密度函数，

$$p(Y_k | A_k, \theta) = \frac{1}{\pi \lambda_d} \exp(-|Y_k - A_k \exp(j\theta)|^2 / \lambda_d) = \frac{1}{\pi \lambda_d} \exp(-|R_k^2 + A_k^2 - 2A_k \operatorname{Re}\{\exp(-j\theta) Y_k\}|^2 / \lambda_d) \quad (2)$$

其中， $\operatorname{Re}\{\cdot\}$ 是取实部操作。

2 改进的 MMSE 语音功率谱估计算法

改进的 MMSE 语音功率谱估计算法假设了语音的幅度谱服从 Chi 分布，并结合了语音存在的概率。本章将分别介绍在语音幅度谱服从 Chi 分布假设下，MMSE 语音功率谱估计算法和结合了语音存在概率的 MMSE 语音功率谱估计算法的推导过程。

2.1 MMSE 语音功率谱估计

定义 $e_k = A_k^2$ 为每一帧语音的功率谱。接下来对 Chi 分

布下的语音的功率谱进行估计进行推导。首先，通过用贝叶斯定理计算语音幅度谱的后验概率密度函数 $p(A_k | Y_k)$ ：

$$p(A_k | Y_k) = \frac{p(Y_k | A_k) p(A_k)}{\int_0^\infty p(Y_k | A_k) p(A_k) dA_k} \quad (3)$$

其中， $p(Y_k | A_k)$ 由式(4)计算得到：

$$p(Y_k | A_k) = \int_0^{2\pi} p(Y_k | A_k, \theta) p(\theta) d\theta \quad (4)$$

其中 θ 是 $X(k)$ 的相位，是在 $[0, 2\pi]$ 内服从均匀分布的随机变量。将把式(2)代入(4)可得：

$$p(Y_k | A_k) = \frac{1}{\pi \lambda_d} \exp(-(R_k^2 + A_k^2)/\lambda_d) I_0(2A_k \sqrt{\nu_k/\lambda_d}) \quad (5)$$

式(5)中的积分通过文献[11]求解得到：

$$p(Y_k | A_k) = \frac{1}{\pi \lambda_d} \exp(-(R_k^2 + A_k^2)/\lambda_d) I_0(2A_k \sqrt{\nu_k/\lambda_d}) \quad (6)$$

其中： $\nu_k = \frac{\xi_k}{\xi_k + 1} \gamma_k$ ， $\lambda_k = \frac{\lambda_x}{\xi_k + 1}$ ， $I_0(\cdot)$ 是修正的第一类零阶贝塞尔函数。将式(6)和(1)代入到式(3)中，可得在已知带噪语音条件下的语音幅度谱的后验概率密度函数 $p(A_k | Y_k)$ ：

$$p(A_k | Y_k) = \frac{A_k^{2a-1} e^{-(\beta + 1/\lambda_d) A_k^2} I_0(2A_k \sqrt{\nu_k/\lambda_d})}{\int_0^\infty A_k^{2a-1} e^{-(\beta + 1/\lambda_d) A_k^2} I_0(2A_k \sqrt{\nu_k/\lambda_d}) dA_k} \quad (7)$$

其中，式(7)分母中的积分可通过文献[11]求得。代入式(7)可得到：

$$p(A_k | Y_k) = \frac{A_k^{2a-1} e^{-(\beta + 1/\lambda_d) A_k^2} I_0(2A_k \sqrt{\nu_k/\lambda_d})}{\Gamma(a) \Phi(a, 1; \frac{\nu_k}{\lambda_k (\beta + 1/\lambda_d)}) / 2(\beta + 1/\lambda_d)^a} \quad (8)$$

其中： $\Gamma(\cdot)$ 表示伽马函数； $\Phi(\cdot)$ 是合流超几何函数。

根据文献[12]，在已知带噪语音频谱的条件下，语音功率谱的后验概率密度函数 $p(e_k | Y_k)$ 可表示为：

$$p(e_k | Y_k) = p(A_k | Y_k) \cdot |\frac{dA_k}{de_k}| = \frac{p(A_k | Y_k)}{2 \sqrt{e_k}} \quad (9)$$

将式(8)代入(9)，把 $\sqrt{e_k}$ 用 A_k 替换，得到在 Chi 分布下语音功率谱的后验概率密度函数 $p(e_k | Y_k)$ 的表达式：

$$p(e_k | Y_k) = \frac{e_k^{a-1} e^{-(\beta + 1/\lambda_d) e_k} I_0(2e_k \sqrt{\nu_k/\lambda_d})}{\Gamma(a) / (\beta + 1/\lambda_d)^a \Phi(a, 1; \frac{\nu_k}{\lambda_k (\beta + 1/\lambda_d)})} \quad (10)$$

根据 MMSE 准则，纯净语音的功率谱估计可以表示为：

$$\hat{e}_k = E[e_k | Y_k] = \int_0^\infty e_k p(e_k | Y_k) de_k \quad (11)$$

最后，将(10)代入到(11)，然后使用文献[11]计算其中的积分，最终得到在 Chi 分布下，纯净语音功率谱的估计的闭合表达式：

$$\hat{e}_k = \frac{z \lambda_k}{\nu_k} \exp(z) \frac{\Gamma(a+1) \Phi(-a, 1; -z)}{\Gamma(a) \Phi(a, 1; z)} \quad (12)$$

其中， $z = \nu_k (a + \xi_k) / (1 + \xi_k)$ 。

2.2 MMSE 语音功率谱估计结合语音存在的概率

为了进一步提高在 Chi 分布下 MMSE 语音功率谱估计的

性能,计算出的语音功率谱 \hat{e}_k 与语音存在的概率结合得到改进的 MMSE 语音功率谱估计 \hat{e}_k^{app} :

$$\hat{e}_k^{\text{app}} = e_k \mid_{\xi_k = \xi'_k} p(H_1^k \mid Y_k) \quad (13)$$

其中, H_1^k 表示在第 k 个频点处存在语音, $p(H_1^k \mid Y_k)$ 代表在第 k 个频点语音存在的后验概率, 定义^[12]如下所示:

$$p(H_1^k \mid Y_k) = A_k / (1 + A_k) \quad (14)$$

其中, A_k 的定义如下所示:

$$A_k = \frac{1 - q_k}{q_k} \frac{p(Y_k \mid H_1^k)}{p(Y_k \mid H_0^k)} \quad (15)$$

其中: $p(Y_k \mid H_1^k)$ 和 $p(Y_k \mid H_0^k)$ 分别代表在语音存在和不存在的条件下 Y_k 的概率密度函数; q_k 表示在频点 k 处语音存在的概率, 则在频点 k 处语音不存在的概率即 $1 - q_k$ 。当语音不存在时, 即只存在噪声, 而已知噪声的 DFT 系数服从复高斯分布, 这样可以得到 $p(Y_k \mid H_0^k)$:

$$p(Y_k \mid H_0^k) = p(Y_k = D_k) = \frac{1}{\pi \lambda_d} \exp(-R_k^2 / \lambda_d) \quad (16)$$

而当语音存在时, 噪声也存在, 所以 $p(Y_k \mid H_1^k)$ 表示为:

$$p(H_1^k \mid Y_k) = p(Y_k = X_k + D_k) = p(Y_k = X_k) * p(Y_k = D_k) \quad (17)$$

式中: * 表示卷积操作, 将式(11)代入(17), 然后根据文献[13], 可以得到语音信号离散傅里叶变换系数分布的概率密度函数为:

$$p(Y_k = X_k) = \frac{1}{2\pi} \frac{2\beta^a}{\Gamma(a)} R_k^{2a-2} \exp(-\beta R_k^2) \quad (18)$$

接下来将 $\beta = a/\lambda_x$ 、式(16)和(18)一起代入(17)中。用极坐标的表示形式, 然后再做一次变量替换可以解出其中的复数卷积, 结果如下:

$$A_k = \frac{1 - q}{q} \left(\frac{1}{1 + \xi'_k/a} \right) \Phi \left(a, 1, \frac{\gamma_k}{1 + a/\xi'_k} \right) \quad (19)$$

当 $a = 1$, 语音幅度谱便服从瑞利分布, 式(19)将被简化为文献[2]中的形式: $A_k = \frac{1 - q}{q} \frac{\exp(\nu_k)}{1 + \xi'_k}$ 。

最后, 将式(12)、(14)代入(13), 得到在 Chi 分布下, 结合了语音存在概率的语音功率谱估计的改进的闭合表达式:

$$\hat{e}_k^{\text{app}} = \frac{z \lambda_k \exp(z)}{\nu_k} \frac{\Gamma(a+1) \Phi(-a, 1; -z)}{\Gamma(a) \Phi(a, 1; z)} \frac{A_k}{1 + A_k} \quad (20)$$

其中: $A_k = \frac{1 - q}{q} \left(\frac{1}{1 + \xi'_k/a} \right) \Phi \left(a, 1, \frac{\gamma_k}{1 + a/\xi'_k} \right)$, $\xi'_k = \frac{\xi_k}{1 - q}$,

$$z = \frac{\nu_k (a + \xi'_k)}{(1 + \xi'_k)}.$$

3 切换的语音功率谱估计算法

语音识别系统常常应用于如像手机、平板等移动终端设备中, 所以降低谱估计语音增强算法的复杂度从而节约功耗就很有必要。因此, 将改进的 MMSE 功率谱估计算法与传统的维纳语音功率谱估计器结合, 设计一种切换的语音功率谱估计算法。在低信噪比阶段用改进的 MMSE 功率谱估计算法尽可能抑制噪声; 而在高信噪比噪声环境中, 采用计算量较小的维纳滤波算法, 在去噪的同时节省系统的计算量。切换算法的流程介绍如下。

1) 计算每一帧的最大似然先验信噪比:

$$\xi_{\text{frame}}^{\text{ml}}(i) = \sum_{k=1}^{\text{len}} \lambda_x(k) / \sum_{k=1}^{\text{len}} \lambda_d(k) \quad (21)$$

其中, len 是每一帧语音的帧长。

2) 作 VAD 判决, 得到每一帧的 VAD 判决结果:

$$VAD(i) = \begin{cases} 1, & \xi_{\text{frame}}^{\text{ml}}(i) \geq A_{\text{thr}} \\ 0, & \text{其他} \end{cases} \quad (22)$$

其中, A_{thr} 是用来作 VAD 判决的阈值。

3) 切换相应的算法。首先, 令 $VAD(i) = 1$ 且 $VAD(i-N, i-(N-1), \dots, i-1) = 0$, 保证切换发生在语音的暂停的时刻, 这里 N 是静音段持续的帧数。然后将每一帧的最大似然先验信噪比与阈值 T_{switch} 作比较: 如果每一帧的最大似然先验信噪比大于阈值 T_{switch} , 就说明噪声干扰较小, 使用传统的维纳功率谱估计算法, 反之则使用本文改进的 MMSE 功率谱估计算法。

4 算法性能的仿真

4.1 MMSE 语音功率谱估计

分别对传统的瑞利分布下结合了语音存在概率的 MMSE (MMSE-Raleigh-Speech Presence Uncertainty, MMSE_Raleigh_SPU) 谱幅度估计算法、结合了倒谱平滑的最优的对数谱幅度 (Optimal Modified Log Spectral Amplitude-Temporal Cepstrum Smoothing, OMLSA-TCS) 估计算法、改进的功率谱幅度估计算法以及本文提出的切换的功率谱估计算法从频域分段信噪比^[4]、ITU-T P. 862 的语音质量感知评估 (Perceptual Evaluation of Speech Quality, PESQ) 得分^[15]以及 CMU 的 Pocketsphinx 语音识别系统^[16]的识别率三个方面进行仿真对比。其中频域加权分段信噪比的公式如下所示:

$$f_{\text{SNR}} = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^K B_j \lg \left[\frac{F^2(m, j)}{(F(m, j) - \hat{F}(m, j))^2} \right]}{\sum_{j=1}^K B_j} \quad (23)$$

其中: B_j 是第 j 个频带的权重; K 是频带的个数; M 是总帧数; $F(m, j)$ 是第 m 帧纯净信号的第 j 个频带的滤波带幅度; $\hat{F}(m, j)$ 是增强后的信号的滤波带幅度。采用频带的分段信噪比优于时域的分段信噪比, 这是因为增加了对数谱上下不同频带加不同的权重。此外, 可以基于感知去对频带进行划分, 这样更符合人耳的听觉特性, 因此更合理并具有灵活性。

语音识别系统的识别率用 PA ^[1] 表示, 计算公式如下:

$$PA = (N - D - S - I) / N \quad (24)$$

其中: N 是输入语音对应的文本个数, 用来作参考; D 、 S 和 I 分别代表识别结果不完整、词被替换和插入了其他词的情况的识别结果数量。只有识别出的结果跟参考的文本完全一致的时候才认为是识别正确的情况。

为了测试功率谱估计算法对语音识别系统抗噪声性能的提升效果, 我们录制了 400 句纯净的语音, 采样率为 16 kHz。分别从 NOISE92 噪声库^[17]中选取具有代表性的白噪声、粉红噪声、工厂噪声和 babble 噪声并且按照不同的信噪比 0 dB、5 dB、10 dB 和 15 dB 得到带噪语音。对 400 句语音预处理时, 帧长取 512 个点, 帧间叠加为 50%, 使用 512 个点的汉宁窗。本实验将改进的 MMSE 语音功率谱估计算法闭合表达式中的自由度设置为: $a = 0.1$ 。这是由于文献[10]中提到在 $a \in$

[0.05, 0.2] 时, Chi 分布能够更好地逼近语音幅度谱的真实分布。而且,本文通过实验已证明当 $a = 0.1$ 时,本文改进的算法有更高的识别率提升。此处不再作进一步的实验讨论。

本文改进的功率谱幅度算法将同以下三种算法对比:传统的 MMSE 功率谱估计算法 MMSERaleigh,也结合了语音存在的概率,但对语音幅度谱分布的假设是传统的瑞利分布;经典的结合倒谱平滑的最优的对数幅度谱估计算法(OMLSA-TCS),它是最优的 MMSE 对数幅度谱估计算法,并结合了基于倒谱平滑(Temporal Cepstrum Smoothing, TCS)的先验信噪比估计算法^[18];此外本文提出的改进的功率谱估计算法还将与切换的改进功率谱估计算法进行对比。TCS 先验信噪比算法将用于计算先验信噪比,分别用文献[3]和文献[19]中方法计算语音存在概率和噪声的功率谱。

通过 OMlsa-TCS 算法、传统的 MMSE Raleigh-SPU 功率谱估计算法、本文改进的 MMSE 功率谱算法以及切换的改进功率谱估计算法增强后语音的频域加权分段信噪比如图 1 所示。从图 1 中清楚地看出:本文改进的算法比其他三种算法有更高的频域加权分段信噪比。相比于传统的 MMSE Raleigh-SPU 功率谱估计算法,本文改进的 MMSE 功率谱估计算法有比较明显的信噪比提升。在不同类型和不同信噪比的带噪语音下,平均提升了约 2.15 dB。与 OMlsa-TCS 算法相比,本文改进的算法除了在 babble 噪声下略高 0.4 dB 之外,其他三种噪声下,本文的算法均取得了约 1.22 dB 的信噪比优势。

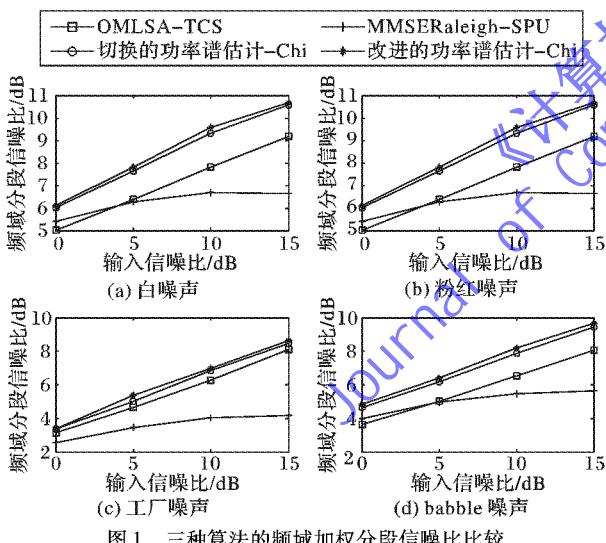


图 1 三种算法的频域加权分段信噪比比较

图 2 为上述四种算法对 400 句带噪语音增强后的 PESQ 得分比较图。总体上,本文改进的算法取得了较高的 PESQ 得分,其中,本文改进的 MMSE 功率谱估计算法取得了比传统的 MMSE Raleigh-SPU 功率谱估计算法更高的 PESQ 得分,平均提高了 0.58。在白噪声、粉红噪声和工厂噪声下,本文改进的方法的 PESQ 得分比 OMlsa-TCS 算法平均提高 0.2;在 babble 噪声下,两种算法有相近的 PESQ 得分,其中,本文改进的算法略高 0.07。

图 3 是四种算法增强后语音的识别率对比图。从图中可以看出,通过四种算法增强后语音的识别率都取得了比原始带噪语音更高值,从而说明这四种语音增强算法都可以提高语音识别系统的抗噪声性能。与两种传统的算法相比,本文改进的算法在获得较高频域分段信噪比和 PESQ 得分的同时

也获得了较高的识别率。与 OMlsa-TCS 算法相比,本文改进的算法除了在 babble 噪声下,平均略高 1.19 个百分点之外,在其他噪声下,均有比较明显的提升,分别比传统的 MMSE Raleigh-SPU 功率谱估计算法和 OMlsa-TCS 算法高 13 个百分点和 18 个百分点。此外,取得最低频域分段信噪比和 PESQ 得分的 MMSE Raleigh 功率谱估计算法,在低信噪比下的识别率超过了语音幅度谱算法 OMlsa-TCS,大约高了 5.3 个百分点,这也验证了文献[7]中提到的功率谱的估计算法有时比幅度谱的估计算法有更好的效果。

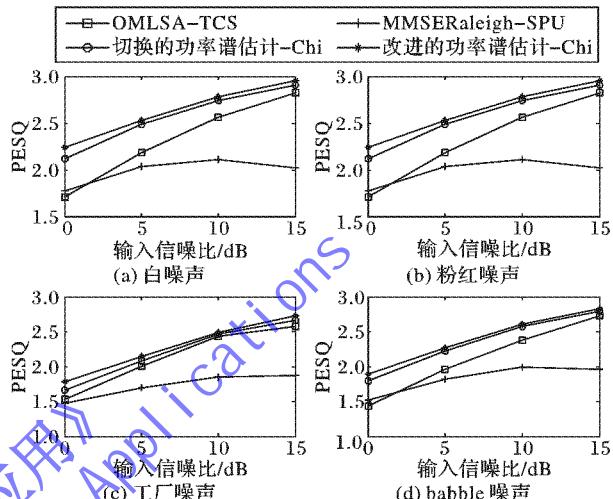


图 2 三种算法的 PESQ 得分比较

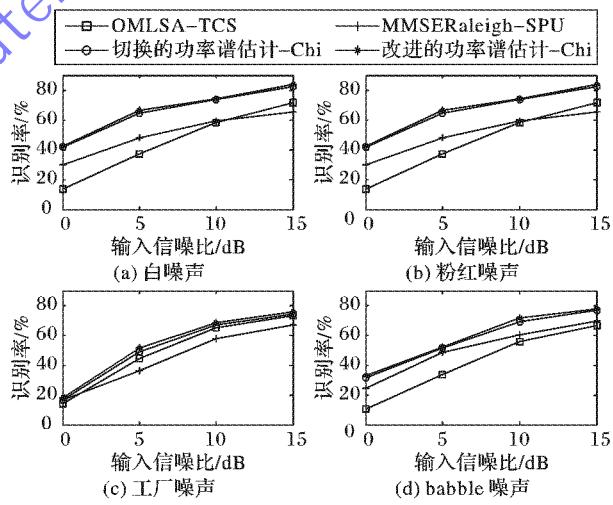


图 3 三种算法的识别率比较

4.2 切换的 MMSE 语音功率谱估计

由图 1~3 可以发现,虽然切换的改进功率谱估计算法,相比改进的 MMSE 功率谱幅度算法性能有一定的下降,但是,从频域分段信噪比、PESQ 和识别率的对比图来看。仍然优于传统的 MMSE 功率谱幅度估计算法和 OMlsa-TCS 算法。图 4 是用于语音识别系统中的切换算法的仿真图。其中,图 4(a)是根据每一帧语音的最大似然先验信噪比的 VAD 算法的判决结果。仿真的语音信号由两段带噪语音组成,前半句和后半句语音信号分别叠加了 0 dB 和 15 dB 的 babble 噪声。从图 4(a)中可以明显地看出,无论是在信噪比较低的前半句语音还是信噪比较高的后半句语音中,VAD 算法基本可以准确地判决出语音段和静音段。图 4(b)中画出了切换算法的切换点“*”,其大致在语音信号的 2.7 s 的位置。切换

算法根据“*”切换不同的算法。在切换点之前的 0~2.7 s 的语音信噪比较低,采用改进的 MMSE 功率谱估计算法对其进行噪声抑制;而对于信噪比较高的后半句语音信号,切换算法在切换点“*”处自动切换使用传统的维纳滤波增强算法对其进行噪声消除。

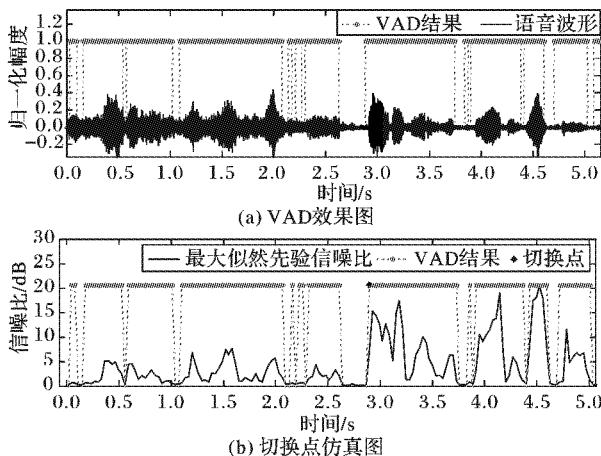


图 4 用于语音识别系统的切换算法仿真图

图 5 是切换的谱估计算法和 OMLSA-TCS 算法增强后语音的波形对比图。其中,图 5 (a) 为前半段和后半段分别叠加了 0 dB 和 15 dB babble 噪声的原始带噪语音波形图,图 5 (b) 和图 5 (c) 分别为此带噪语音通过切换的谱估计算法和改进的 MMSE 功率谱估计算法增强后语音的波形图。从图 5 中可以很清楚地看出:通过上述两种算法处理后语音的波形很接近。而且从图 1~3 中可以看出采用切换的功率谱幅度估计算法相对于整段语音全用改进的功率谱的估计方法的性能无明显的下降,而由于切换算法在高信噪比的后半段语音使用了传统的维纳滤波增强算法,其计算量比改进的 MMSE 功率谱估计算法小很多。从而表明:切换谱估计算法在保证了噪声抑制效果的同时减小了算法的计算量,降低了识别系统的功耗。

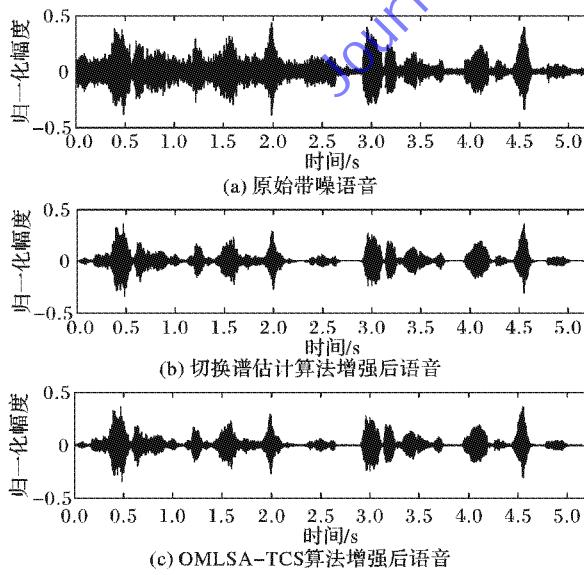


图 5 切换算法增强后语音的波形图

5 结语

本文提出了一种用于提高识别系统抗噪声鲁棒性的切换

语音功率谱估计算法。切换算法根据每一帧语音的平均 ML 先验信噪比和 VAD 判决结果,在改进的 MMSE 语音功率谱估计算法和传统的维纳语音功率谱估计算法之间切换。改进的 MMSE 语音功率谱估计算法根据语音 Chi 分布的先验假设,并结合语音存在的概率,然后由 MMSE 准则推导而得到。仿真实验表明切换算法能准确地切换功率谱估计算法,同时改进的切换算法也表现出了比传统的 MMSE 功率谱估计和幅度谱估计算法 OMLSA-TCS 算法更高的识别系统的识别率,并且维纳语音功率谱估计算法在高信噪比条件下,表现出了跟仅采用改进的 MMSE 功率谱估计算法相近的性能。所以本文提出的切换语音功率谱估计算法比在尽可能提高识别系统鲁棒性的同时,降低了识别系统的功耗。

参考文献:

- [1] VIRTANEN T, SINGH R, RAJ B. Techniques for Noise Robustness in Automatic Speech Recognition [M]. New York: Wiley & Sons, 2012: 228~231.
- [2] EPHRAIM Y, MALAH D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator [J]. IEEE Transactions on Acoustics Speech and Signal Processing, 1985, 33(2): 443~445.
- [3] COHEN I. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator [J]. IEEE Signal Processing Letters, 2002, 9(4): 113~116.
- [4] ASTUDILLO R F, ORGLMEISTER R. Computing MMSE estimates and residual uncertainty directly in the feature domain of ASR using STFT domain speech distortion models [J]. IEEE Transactions on Acoustics Speech and Signal Processing, 2013, 21(5): 1023~1034.
- [5] JENSEN J, TAN Z H. Minimum mean-square error estimation of Mel-frequency cepstral features theoretically consistent approach [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2015, 23(1): 186~197.
- [6] INDREBO K M, POVINELLI R J, JOHNSON M T. Minimum mean-squared error estimation of Mel-frequency cepstral coefficients using a novel distortion model [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2008, 16(8): 1654~1661.
- [7] LOIZOU P C. Speech Enhancement: Theory and Practice [M]. Boca Raton, FL: CRC Press, 2007: 119~122.
- [8] DAT T H, TAKEDA K, ITAKURA F. Generalized Gamma modeling of speech and its online estimation for speech enhancement [C]// Proceedings of the 2005 IEEE International Conference on Acoustics Speech and Signal Processing. Piscataway, NJ: IEEE, 2005, 4: 181~184.
- [9] LOTTER T, VARY P. Noise reduction by joint maximum a posteriori spectral amplitude and phase estimation with super-Gaussian speech modelling [C]// Proceedings of the 2004 European Conference on Signal Processing. Piscataway, NJ: IEEE, 2004: 1457~1460.
- [10] ERKELENS J S, HENDRIKS R C, HEUSDENS R, et al. Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors [J]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(6): 1741~1752.
- [11] GRADSHTEYN I S, RYZHIK I M. Table of Integrals, Series, and Products [M]. 7th ed. Cambridge, Massachusetts: Academic Press, 2007: 346~353, 699~711.

(下转第 3384 页)

- [7] LI S Y, CHENG Y C. Deterministic fuzzy time series model for forecasting enrollments [J]. Computers & Mathematics with Applications, 2007, 53(12): 1904 – 1920.
- [8] CHANG P C, LIU C H, FAN C Y, et al. Data clustering and fuzzy neural network for sales forecasting in printed circuit board industry [C]// CIDM 2007: Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining. Piscataway, NJ: IEEE, 2007: 107 – 113..
- [9] WONG W, BAI E, CHU A W C. Adaptive time-variant models for fuzzy-time-series forecasting [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2010, 40(6): 1531 – 1542.
- [10] ZHANG G P. Time series forecasting using a hybrid ARIMA and neural network model [J]. Neurocomputing, 2003, 50(1): 159 – 175.
- [11] SUN Z L, CHOI T M, AU K F, et al. Sales forecasting using extreme learning machine with applications in fashion retailing [J]. Decision Support Systems, 2008, 46(1): 411 – 419.
- [12] SUN Z L, AU K F, CHOI T M. A neuro-fuzzy inference system through integration of fuzzy logic and extreme learning machines [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2007, 37(5): 1321 – 1331.
- [13] WONG W K, GUO Z X. A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm [J]. International Journal of Production Economics, 2010, 128(2): 614 – 624.
- [14] HSU C C, CHEN C Y. Applications of improved grey prediction model for power demand forecasting [J]. Energy Conversion and Management, 2003, 44(14): 2241 – 2249.
- [15] 谢乃明, 刘思峰. 离散 GM(1,1) 模型与灰色预测模型建模机理 [J]. 系统工程理论与实践, 2005, 25 (1): 93 – 99. (XIE N M, et al. Discrete GM (1, 1) and mechanism of grey forecasting model [J]. Systems Engineering — Theory & Practice, 2005, 25 (1): 93 – 99.)
- [16] 任柏青. 基于关系数据库的领域本体构建方法的研究与实践 [D]. 北京: 北京邮电大学, 2009: 34 – 72. (REN B Q. Research and practice on a mechanism of generating ontology based on relational database [D]. Beijing: Beijing University of Posts and Telecommunications, 2009: 34 – 72.)
- [17] 王珊, 张延松. 面向并发 OLAP 的数据库查询处理方法: 中国, 201210113665.4 [P]. 2012-04-17. (WANG S, ZHANG Y S. Query processing method for paralleled OLAP database: China, 201210113665.4 [P]. 2012-04-17.)
- [18] 张平平, 伍俊良, 胡兴凯, 等. Gerschgorin 圆盘的分离 [J]. 西南师范大学学报(自然科学版), 2011, 36(3): 1 – 3. (ZHANG P P, WU J L, HU X K, et al. On the separation of Gerschgorin circular discs [J]. Journal of Southwest China Normal University (Natural Science Edition), 2011, 36(3): 1 – 3.)
- [19] WANG C, FAN J. Medical relation extraction with manifold models [C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2014: 828 – 838.
- [20] 姚富强. 通信抗干扰工程与实践 [M]. 北京: 电子工业出版社, 2008: 7 – 8. (YAO F Q. Communication Anti-jamming Engineering and Practice [M]. Beijing: Publishing House of Electronics Industry, 2008: 7 – 8.)

Background

LIU Weixiao, born in 1990, M. S. candidate. Her research interests include big data processing, data mining, machine learning, fashion sales forecasting.

(上接第 3373 页)

- [12] STARK A, PALIWAL K. MMSE estimation of log-filterbank energies for robust speech recognition [J]. Speech Communication, 2011, 53(3): 403 – 416.
- [13] FODOR B, FINGSCHEIDT T. MMSE speech enhancement under speech presence uncertainty assuming (generalized) Gamma speech priors throughout [C]// Proceedings of the 2012 IEEE International Conference on Acoustics Speech and Signal Processing. Piscataway, NJ: IEEE, 2012: 4033 – 4036.
- [14] TRIBOLET J M, NOLL P, MCDERMOTT B, et al. A study of complexity and quality of speech waveform coders [C]// Proceedings of the 1978 IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway, NJ: IEEE, 1978, 3: 586 – 590.
- [15] RIX A W, BEERENDS J G, HOLLMER M P, et al. Perceptual evaluation of speech quality (PESQ) — a new method for speech quality assessment of telephone networks and codecs [C]// Proceedings of the 2001 IEEE International Conference on Acoustics Speech and Signal Processing. Washington, DC: IEEE Computer Society, 2001, 2: 749 – 752.
- [16] Carnegie Mellon University. Carnegie Mellon University sphinx [EB/OL]. [2016-04-14]. <http://cmusphinx.sourceforge.net/>.
- [17] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems [J]. Speech Communication, 1993, 12(93): 247 – 251.
- [18] BREITHAUPT C, GERKMANN T, MARTIN R. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing [C]// Proceedings of the 2008 IEEE International Conference on Acoustics Speech and Signal Processing. Piscataway, NJ: IEEE, 2008: 4897 – 4900.
- [19] HENDRIKS R C, HEUSDENS R, JENSEN J. MMSE based noise PSD tracking with low complexity [C]// Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing. Piscataway, NJ: IEEE, 2010: 4266 – 4269.

Background

This work is partially supported by the National Natural Science Foundation of China (61501072), the Natural Science Foundation of Chongqing Science and Technology Commission (cstc2015jcyjA40027).

LIU Jingang, born in 1991, M. S. candidate. His research interests include speech signal processing, speech enhancement.

ZHOU Yi, born in 1974, Ph. D, professor. His research interests include adaptive filtering, speech signal processing.

MA Yongbao, born in 1989, M. S. candidate. His research interests include speech signal processing, speech enhancement.

LIU Hongqing, born in 1980, Ph. D, professor. His research interests include sparse signal processing, array signal processing.