文章编号:1004-9037(2014)02-0232-06

基于自适应超高斯混合模型的语音增强算法

赵改华 周 彬 张雄伟

(解放军理工大学指挥信息系统学院,南京,210007)

摘要:语音信号的频谱结构复杂性决定了其短时谱分布不能用单一的概率密度函数(Probability density function, PDF)准确描述。据此,提出了一种采用超高斯混合模型对语音信号幅度谱建模以实现语音增强的新方法。首先,采用超高斯混合模型对语音信号幅度谱的先验分布进行建模,相对于传统的单一模型,该模型能更好地描述语音信号的多类特性;然后,在增强过程中自适应更新混合分量的 PDF 及其权重,从而克服了传统模型难以跟踪语音信号分布动态变化的缺点。仿真结果表明与传统的短时谱估计算法相比,该算法的噪声抑制性能有较大的提升,增强语音的主观感知质量也有明显改善。

关键词:语音增强;超高斯混合模型;自适应

中图分类号:TP912.3 **文献标志码**:A

Speech Enhancement Algorithm Based on Adapted Super-Gaussian Mixture Model

Zhao Gaihua, *Zhou Bin*, *Zhang Xiongwei*

(College of Command Information Systems, PLA University of Science and Technology, Nanjing, 210007, China)

Abstract: The observation of speech spectral structure shows that the statistics of speech signal cannot be well determined by a simple probability density function (PDF). Therefore, a speech enhancement algorithm is presented based on the super-Gaussian mixture model. Firstly, the super Gaussian mixture model is employed to model the speech spectral amplitude, which is more flexible in capturing the statistical behavior of speech signals than the conventional simple speech model. Where after, PDF and weight of the mixture components are further adapted, which can overcome the disadvantage that the traditional simple speech model cannot well track the dynamic characteristics of the speech signal. The simulation results show that the proposed algorithm achieves better noise suppression and lower speech distortion compared with the conventional short-time spectral estimation algorithms.

Key words: speech enhancement; super-Gaussian mixture model; adaptation

引 言

在语音通信过程中,语音信号不可避免地会受到噪声的干扰,影响通信质量和语音信号的后续处理,语音增强技术是从带噪语音中尽可能提取原始纯净语音的重要手段,在提高语音可懂度、改善语音通信质量等方面有重要的应用。在众多的增强技术中基于统计模型的短时谱估计法以其复杂度低和相对有效的特点,长期以来受到了广大研究者的广泛关注。

基于统计模型的短时谱估计语音增强算法,主要是在不同的语音和噪声先验分布模型假设条件下,依据一定的准则,如:最小均方误差(Minimum mean square error, MMSE)、最大后验概率(Maximum a prosteriori, MAP)、最大似然值(Maximum likelihood, ML),对语音信号的短时谱进行最优估计。经典的基于统计模型的短时谱估计法是由 Ephraim 和 Malah 提出的基于高斯模型的最小均方误差短时幅度谱(Short-time spectral amplitude-minimum mean square error, STSA-MMSE)[1]估计算法。对语音信号统计模型的深入

收稿日期:2013-01-29;**修订日期**:2013-04-10

研究表明，超高斯模型更符合语音信号的实际分布[2]，据此，研究者提出了许多改进算法，例如文献[3～6]提出的基于超高斯语音模型的短时谱估计算法，包括基于超高斯模型的复频谱 MMSE 估计算法、基于超高斯模型的幅度谱 MAP 估计算法、基于 Gamma 模型的 DCT 域 MMSE 估计算法和基于超高斯模型的对数谱 MMSE 估计算法，相对于基于高斯模型的增强算法增强效果有所提升。

上述增强算法都假设语音信号幅度谱服从单一的分布函数，而事实上，由于语音信号产生的复杂性及其非平稳性，其分布很难用单一的函数准确描述[7]。因此，研究者提出了一些利用较为复杂的混合模型为语音信号建模的新方法，如文献[8]提出的基于高斯混合模型复频谱 MMSE 估计算法，文献[9]提出的基于瑞利混合模型的幅度谱 MMSE 估计算法，近年来，研究者提出了一些高斯混合模型的改进模型来进一步提高增强效果，例如：文献[10]提出的基于高斯尺度混合模型的对数谱估计算法，文献[11]提出的基于超高斯混合模型的幅度谱 MMSE 估计算法。相对于采用单一模型的增强算法，增强效果有较大提高。然而，这些混合模型对每帧语音信号建模时所用的混合分量及其权重都是固定的，而事实上，语音信号幅度谱的实际分布是动态变化的，每个混合分量与当前语音信号的相似度也是变化的，因此，固定的权重并不合理。同时，有些混合分量与当前语音信号相差较大的，将其引入混合模型不利于逼近当前语音信号的实际分布[12]。

针对上述算法存在的问题，本文提出了一种基于超高斯混合模型的语音增强算法。首先，采用 EM 算法将语音信号分为多个分量；然后，在增强过程中选择与当前帧相似度较大的部分混合分量，并利用初始增强语音更新选中混合分量的概率密度函数（Probability density function，PDF）；其次，估计对应每个选中混合分量的幅度谱最小均方误差估计式，并依据混合分量与当前帧的相似度更新对应的子类增强语音的权重；最终的增强语音由子类增强语音的加权和获得。

1　传统短时谱估计算法

假设 $s(n)$ 表示纯净语音信号，$x(n)$ 表示加性噪声信号，那么时域带噪语音信号可表示为 $y(n)=s(n)+x(n)$，对时域带噪语音信号进行分帧、加窗和 STFT 变换，得到带噪语音信号在频域内的表示为

$$Y(k,l)=S(k,l)+X(k,l) \tag{1}$$

式中：$l(l=0,1,2,\cdots)$ 表示帧序号；$k(k=0,1,\cdots)$ 表示频带序号，用幅度和相位表示为

$$R_k^l\exp(\mathrm{j}\alpha_k)=A_k^l\exp(\mathrm{j}\beta_k)+D_k^l\exp(\mathrm{j}\eta_k) \tag{2}$$

式中：R_k^l, A_k^l, D_k^l 分别表示带噪语音、纯净语音、噪声的幅度谱；$\alpha_k, \beta_k, \eta_k$ 表示对应的相位。假设各帧各频带之间相互独立，为了简化表示，下面将序号 k, l 省略。

一般地，假设噪声复频域系数的实部和虚部分别服从高斯分布，则噪声幅度谱系数服从瑞利分布[8]，表示为

$$f_D(d)=(2d/\sigma_d^2)\cdot\exp(-d^2/\sigma_d^2) \tag{3}$$

式中 σ_d^2 表示噪声系数方差。假设语音复频域系数的实部和虚部也分别服从高斯分布，则语音幅度谱系数服从瑞利分布，表示为

$$f_A(a)=(2a/\sigma_a^2)\cdot\exp(-a^2/\sigma_a^2) \tag{4}$$

式中 σ_a^2 表示语音系数方差，纯净语音的 MMSE 估计式为[1]

$$\hat{A}=E[A\mid R]=\Gamma(1.5)\frac{\sqrt{\varphi}}{\gamma}\exp\left(-\frac{\varphi}{2}\right)\cdot\left[(1+\varphi)I_0\left(\frac{\varphi}{2}\right)+\varphi I_1\left(\frac{\varphi}{2}\right)\right]R \tag{5}$$

式中：$\xi=\sigma_a^2/\sigma_d^2$ 代表先验信噪比；$\gamma=R^2/\sigma_d^2$ 代表后验信噪比；$\varphi=\gamma\cdot\xi/(1+\xi)$；$I_0(\cdot)$，$I_1(\cdot)$ 分别表示第零类和第一类修正贝塞尔函数。由于语音信号的非稳定性和复杂性，单一的某一种分布函数并不适用于所有的语音信号，因此，许多研究者提出了许多更为复杂的语音信号模型[8-9]，本文采用较为复杂的超高斯混合模型对语音信号幅度谱建模，并适时更新模型参数跟踪语音信号的变化。

2　本文算法

如前文所述，语音信号的复杂性和非稳定性决定了用单一的函数描述语音信号幅度谱的分布是不准确的，据此，本文提出了基于自适应超高斯混合模型的语音增强算法，不仅可以更好地逼近当前语音信号的实际分布，而且可以跟踪语音信号幅度谱分布随帧移的动态变化。算法可分为 3 个模块：训练模块、预处理模块、增强模块，如图 1 所示。

训练模块：采用超高斯混合模型，对大量的纯净语音进行训练，将语音信号分为多个超高斯分量，这些分量的参数将在增强模块适时更新。预处理模块：采用传统的 STSA-MMSE[1] 算法获得初始增强语音 $\overline{A}^{\mathrm{pri}}$。增强模块：首先，选择与 $\overline{A}^{\mathrm{pri}}$ 似然值较大的 I 个混合分量，然后，基于所选混合分量

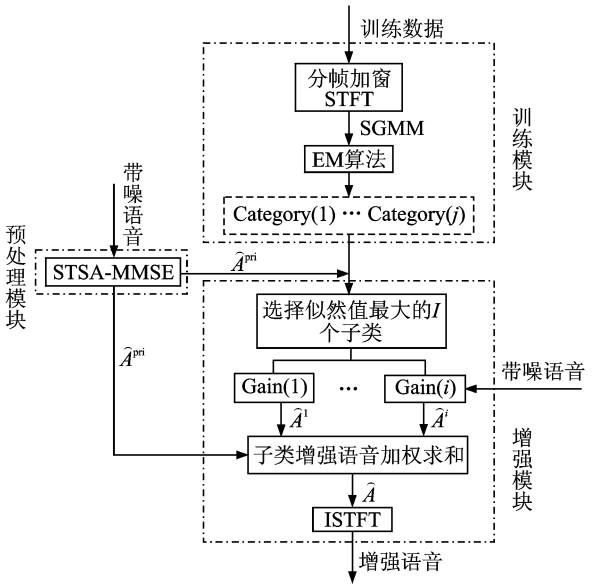

图 1 算法流程图

Fig. 1 Flowchart of the proposed algorithm

的 PDF,计算对应的幅度谱最小均方误差估计:$(\overline{A}^1,\cdots,\overline{A}^I)$,最终的增强语音由各子类增强语音的加权和确定。

2.1 训练模块

训练模块作用是采用超高斯混合模型,将语音信号分为多个分量,并确定每个分量的初始 PDF 及权重。每个分量采用文献[4]推导的语音信号幅度谱的超高斯分布建模为

$$f_c(a)=\frac{\mu^{\nu+1}}{\Gamma(\nu+1)}\frac{a^{\nu}}{\sigma_c^{\nu+1}}\exp\left(-\mu\frac{a}{\sigma_c}\right) \tag{6}$$

通过不同的参数(μ,ν)取值,式(6)可以非常精确地逼近对应的 Gamma 和 Laplace 分布。大量实验数据表明语音信号幅度谱的实际分布介于 Gamma 分布和 Laplace 分布之间,本文采用能够较为准确地逼近语音信号幅度谱实际分布的参数组:$(\mu=1.74,\nu=0.126)$[4]。语音信号幅度谱的超高斯混合分布表示为 C 个子分量的加权和,表示为

$$f_A(a)=\sum_{c=1}^{C}\omega_c f_c(a) \tag{7}$$

式中 ω_c 表示每个分量的权重,且满足限制条件

$$\begin{cases}\sum_{c=1}^{C}\omega_c=1\\ \int_0^{\infty}f_A(a)\mathrm{d}a=1\end{cases} \tag{8}$$

在限制条件下,采用 EM 参数估计法,对大量语音信号进行训练,获取参数组 $\lambda_c=(\omega_c;\sigma_c^2)$。

2.2 增强模块

增强模块是整个算法的核心部分。首先,选择与初始增强语音似然值最大的 I 个分量并更新其 PDF;然后,利用更新之后的 PDF 计算对应选中分量的最小均方误差估计式,并利用初始增强语音与对应分量的似然值计算子类增强语音的权重;子类增强语音加权求和即得到最终的增强语音。

2.2.1 子类的选择及更新

本文研究观察到,并不是每个子类的 PDF 都需要更新,而只需要更新与初始增强语音 $\overline{A}^{\mathrm{pri}}$ 似然值 $P(\overline{A}^{\mathrm{pri}}\mid\lambda_c)$ 较大的部分混合分量。理论上这样做也是合理的,与初始增强语音的似然值较小的混合分量,在很大概率上与实际语音相差较大,若每次都用当前的语音对应的初始增强语音更新全部的分量,则会导致所有的分量向着同一个模型收敛,不利于跟踪语音信号的动态特性。在试验中观察到选取 4 个似然值较大的分量,既能够达到较好的效果,而且不会导致混合分量同化。似然值的计算如下

$$P(\overline{A}^{\mathrm{pri}}\mid\lambda_c)=\omega_c\cdot f_c(\overline{A}^{\mathrm{pri}})=\omega_c\cdot\frac{\mu^{\nu+1}}{\Gamma(\nu+1)}\frac{(\overline{A}^{\mathrm{pri}})^{\nu}}{\sigma_c^{\nu+1}}\exp\left(-\mu\frac{\overline{A}^{\mathrm{pri}}}{\sigma_c}\right) \tag{9}$$

根据式(9),选择似然值较大的 4 个分量,用于对当前语音帧的估计。利用当前帧的初始增强语音 $\overline{A}^{\mathrm{pri}}$ 更新所选分量概率密度函数的方差为

$$\sigma_c^2=\eta\sigma_c^2+(1-\eta)(\overline{A}^{\mathrm{pri}})^2 \tag{10}$$

式中 η 表示模型更新的速度,在试验中观察到 $\eta=0.95$ 较为合适。

2.2.2 子类增强语音及其权重的估计

假设噪声服从高斯分布,则语音信号的幅度谱的 MMSE 估计式表示为[1]

$$\hat{A}=E\{A\mid R\}=\frac{\int_0^{\infty}a\cdot\exp\left(-\frac{a^2}{\sigma_d^2}\right)I_0\left(\frac{2ar}{\sigma_d^2}\right)\cdot f_A(a)\mathrm{d}a}{\int_0^{\infty}\exp\left(-\frac{a^2}{\sigma_d^2}\right)I_0\left(\frac{2ar}{\sigma_d^2}\right)\cdot f_A(a)\mathrm{d}a} \tag{11}$$

式(11)中零阶贝塞尔函数可以近似表示为[13]

$$I_0(x)=\frac{1}{\sqrt{2\pi x}}\cdot\exp(x) \tag{12}$$

假设语音信号的幅度谱服从超高斯分布,将近似式(12,6)代入式(11),根据文献[13]中的公式 3.462.1,求积分获得 I 类增强语音的幅度谱 MMSE 估计式为

$$\overline{A}^i=\frac{R}{\sqrt{2\gamma}}\cdot\frac{\exp(\mu^2/8\xi_i-\mu\sqrt{\gamma}/2\sqrt{\xi_i})}{\exp(\mu^2/8\xi_i-\mu\sqrt{\gamma}/2\sqrt{\xi_i})}\cdot\frac{(\Gamma(\nu+1.5)/\xi_i^{(\nu+1)/2})\cdot D_{-\nu-1.5}(\mu/\sqrt{2\xi_i}-\sqrt{2\gamma})}{(\Gamma(\nu+0.5)/\xi_i^{(\nu+1)/2})\cdot D_{-\nu-0.5}(\mu/\sqrt{2\xi_i}-\sqrt{2\gamma})} \tag{13}$$

式中：$\gamma=R^2/\sigma_d^2$ 表示后验信噪比，ξ_i 与先验信噪比成线性关系，表示为：$\xi_i=\sigma_i^2\cdot\xi$。I 类增强语音的权重 c_i 估计如下

$$c_i=h(P(\overline{A}^{\mathrm{pri}}\mid\lambda_i))\Big/\sum_{i=1}^{I}h(P(\overline{A}^{\mathrm{pri}}\mid\lambda_i)) \tag{14}$$

$$h(x)=\begin{cases}0.1 & \lg x\leqslant 0.1\\ \lg x & \text{其他}\end{cases} \tag{15}$$

权重的计算不将似然值 $P(\overline{A}^{\mathrm{pri}}\mid\lambda_i)$ 直接用于式(14)，而是采用其对数形式，是因为在试验中观察到，如果不对似然值作对数处理则会出现其中一个分量的权重趋近于 1，而其他分量的权值都趋近于 0，这是由于初始估计语音存在偏差造成的，显然这对最终增强语音估计的准确性是不利的。为避免这种现象，本文利用似然值 $P(\overline{A}^{\mathrm{pri}}\mid\lambda_i)$ 的对数进行权值计算。最终的增强语音由 I 类增强语音的加权和获得

$$\overline{A}=\sum_{i=1}^{I}c_i\overline{A}^{i} \tag{16}$$

式中：$\overline{A}^{i}$ 表示第 I 类增强语音；c_i 表示第 I 类增强语音的权重。

3　实验仿真

仿真实验在 MATLAB 环境下进行，将本文提出的增强算法与以下 2 种算法进行比较，包括：文献[6]提出的基于超高斯模型的 MMSE 对数谱估计法；文献[9]提出的基于瑞利混合模型的 MMSE 幅度谱估计法。为简化表示，这两种算法分别表示为 Super-gauss，RMM。本文提出的算法表示为：SGMM。

采用标准语音库 timit 中的纯净语音对超高斯混合模型进行训练。原始噪声信号从标准噪声库 Noisex92 中选取，包括高斯白噪声、汽车噪声，并下采样为 8 kHz。纯净语音信号为标准语音库 timit 中的标准汉语语音信号，采用 8 kHz 采样，时间长度约为 8 s，男女声各 8 句。利用 MATLAB 对噪声信号和纯净语音进行混和，信噪比分别定为 0，5，10 dB。噪声估计采用统计最小量跟踪算法[14]，先验信噪比计算采用面向判决的方法[1]为

$$\xi(m,k)=\alpha\cdot\frac{\overline{A}^2(m-1,k)}{\sigma_d^2(m,k)}+(1-\alpha)\cdot\max[(\gamma(m,k)-1),0] \tag{17}$$

式中 $\alpha=0.98$，采用增强后和增强前语音分段信噪比提高量来衡量不同短时谱估计算法的噪声抑制性能，分段信噪比定义为

$$\mathrm{SSNR}=\frac{1}{T}\sum_{l=0}^{T-1}10\lg\left(\frac{\sum_{n=0}^{R-1}s^2(lR+n)}{\sum_{n=0}^{R-1}(s(lR+n)-\hat{s}(lR+n))^2}\right) \tag{18}$$

式中：$\hat{s}(n)$ 和 $s(n)$ 分别为增强后语音和纯净语音的时域信号；T 表示语音信号的帧数。

表 1 给出了在不同噪声和信噪比条件下 3 种算法的分段信噪比的提高量。从表中可以看出，相较于单一成分的 Super-gauss 短时谱估计算法，由于采用了多种成分加权叠加的方式来更为精确地逼近语音信号的实际分布，RMM 算法和本文提出的 SGMM 算法在抑制噪声方面有更为显著的效果。

表 1　3 种算法的分段信噪比提高量　dB

Table 1　Improved segmental SNR of three algorithms

噪声	输入信噪比	Super-gauss	RMM	SGMM
高斯白噪声	0	12.348	12.778	13.215
	5	10.405	10.649	11.463
	10	9.130	9.375	10.021
	15	7.594	8.0876	8.904
汽车噪声	0	10.695	10.901	11.437
	5	8.590	8.875	9.597
	10	7.654	7.973	8.649
	15	6.494	6.869	7.087

采用对数频谱距离 LSD 衡量增强语音的失真度，对数谱距离定义为

$$\mathrm{LSD}=\frac{1}{T}\sum_{l=0}^{T-1}\sqrt{\sum_{k=0}^{L-1}\left(10\lg\frac{|S(k,l)|^2}{|\hat{S}(k,l)|^2}\right)^2} \tag{19}$$

式中：$\hat{S}(k,l)$ 和 $S(k,l)$ 分别为 $\hat{s}(n)$ 和 $s(n)$ 通过短时傅里叶变换后的第 k 帧的第 l 个频谱分量；T 表示语音信号的帧数。

图 2 给出了在不同噪声和信噪比条件下的 LSD 改进曲线。LSD 是一种语音信号失真测度，测度值越小表明语音信号谱失真越小，语音质量越接近原始语音。从图中可以看出，相较于固定模型的 Super-gauss 和 RMM 两种谱估计算法，本文提出的基于自适应超高斯混合模型的谱估计算法能够更好地描述原始语音的分布，增强语音的失真度更小。

采用客观质量评估方法 PESQ 衡量增强语音的质量。表 2 给出了在不同噪声和信噪比条件下

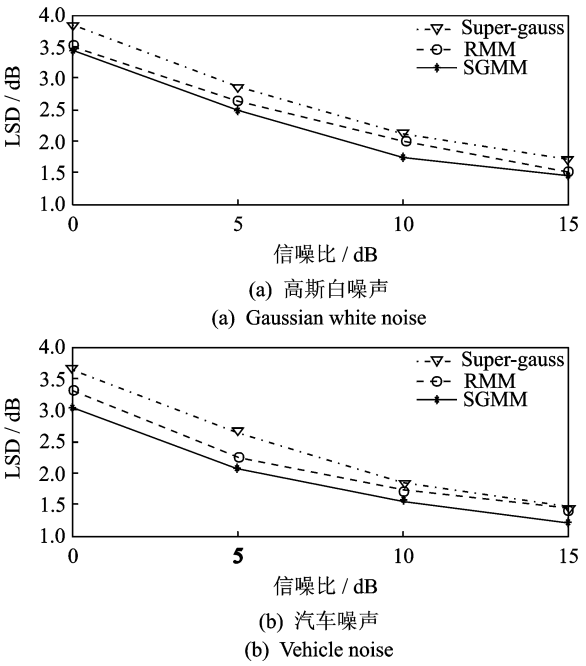

图 2 对数频谱距离改进曲线

Fig. 2 Improved LSD curve

的 PESQ 评估结果，从表 2 中可以看出，本文算法的增强语音的 PESQ 得分都明显高于其他两种谱估计算法，说明其具有更好的感知质量，主观测试也验证了这一结论。

表 2 四种算法 PESQ 评估得分

Table 2 PESQ scores of four algorithms

噪声	输入信噪比/dB	Super-gauss	RMM	SGMM
高斯白噪声	0	2.09	2.12	2.13
	5	2.40	2.46	2.52
	10	2.59	2.68	2.74
	15	2.82	2.88	2.94
汽车噪声	2.20	2.33	2.47	2.20
	2.61	2.64	2.87	2.61
	2.85	2.89	2.96	2.85
	3.05	3.09	3.16	3.05

由于在增强阶段，对于每帧语音信号都要重新选择混合分量并更新其权重，因此增强效果的提升是以计算量的增加为代价的。这也是许多类似自适应增强算法共同存在的情况[15]。

4 结束语

本文提出了一种新的基于自适应超高斯混合模型的语音增强算法，不仅将混合模型应用于超高斯幅度谱分布，且自适应更新模型参数，相对于传统的信号模型，本文提出的自适应超高斯混合模型能够更好地逼近语音信号的实际分布。仿真结果也验证了本文提出算法的优越性，不仅提高了噪声抑制性能，而且增强语音的失真度也有所下降。在下一步的工作中将针对噪声信号的非平稳性对噪声模型进行优化，有望提高算法的增强效果。

参考文献：

[1] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator[J]. IEEE Trans Acoust Speech, Signal Process, 1984,32(6):1109-1121.

[2] Gazor S, Zhang W. Speech probability distribution [J]. IEEE Signal Process Lett, 2003,10(7):2042-207.

[3] Martin R. Speech enhancement based on minimum mean-square error estimation and super Gaussian priors[J]. IEEE Trans Speech Audio Process, 2005,13(5):845-856.

[4] Lotter T, Vary P. Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model[J]. Eurasip J Signal Process, 2005,(7):1110-1126.

[5] 邹霞，陈亮，张雄伟. 一种基于 Gamma 语音模型的语音增强算法[J]. 通信学报，2006,27(10):118-123.
Zou Xia, Cheng Liang, Zhang Xiongwei. Speech enhancement with Gamma speech modeling[J]. Journal on Communications, 2006,27(10):118-123.

[6] Hendriks R C. Heusdens R, Jensen J. Log-spectral magnitude MMSE estimators under super-gaussian densities[J]. Inter Speech, 2009,10(6):1319-1322.

[7] Ephraim Y. A Bayesian estimation approach for speech enhancement using hidden Markov models [J]. IEEE Trans Acoust Speech, Signal Process, 1992, 40(4):725-735.

[8] Ding Guohong, Wang Xia, Cao Yang, et al. Speech enhancement based on speech spectral complex Gaussian mixture model[C]//IEEE Int Conf Acoustic, Speech, Signal Process (ICASSP). Philadephia, USA: IEEE, 2005:165-168.

[9] Erkelens J S, Jensen J, Heusdens R. Speech enhancement based on Rayleigh mixture modeling of speech spectral amplitude distributions[C]//European Signal Proc Conf (EUSIPCO). Poznan, Poland: [s. n.], 2007:65-69.

[10] Hao Jiucang, Lee Te-Won. Speech enhancement using Gaussian scale mixture models[J]. IEEE Trans

on ASLP, 2010,18(6):1127-1136.

[11] Wang Haiyan, Zhao Xiaohui, Gu Haijun. Speech enhancement using super gauss mixture model of speech spectral amplitude[J]. The Journal of China University of Posts and Telecommunications, 2011, 18(1):13-18.

[12] Jancovic P, Zou X, Köküer M. Speech enhancement based on sparse code shrinkage employing multiple speech models[J]. Speech Communication, 2012, 54:108-118.

[13] Gradshteyn I S, Ryzhik Z M. Table of integrals, series, and products[M]. New York: Academic, 1980.

[14] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics [J]. IEEE Transactions on Speech and Audio Processing, 2001,9(5):504:512.

[15] 曹斌芳，李建奇. 基于自适应仿生小波变换的语音增强方法[J]. 数据采集与处理，2010,25(6):741-745.
Cao Binfang, Li Jianqi. Speech enhancement method based on adaptive bionic wavelet transform[J]. Journal of Data Acquisition and Processing, 2010,25(6): 741-745.

作者简介：赵改华(1987-)，女，硕士研究生，研究方向：信息与信号处理、语音增强，E-mail: zhaogaihua. happy@163.com；周彬(1986-)，男，博士研究生，研究方向：信息与信号处理、语音编码、语音增强；张雄伟(1965-)，男，教授，博士生导师，研究方向：多媒体信息处理、智能计算机、压缩感知。