基于最大后验相位估计的多带谱减语音增强算法

李 真* 吴文锦 张 勤 任 慧
(中国传媒大学 北京 100024)

摘 要：传统语音增强算法中因为谱减法算法简单易于实现而得到广泛研究，谱减法的原理是将带噪语音幅度与估计的噪声幅度进行相减，并叠加带噪语音相位，进而重构增强语音谱。该方法在低信噪比下因为没有进行相位估计，会存在较大的估计误差，并且因为对噪声估计的不准确，会产生 “音乐噪声”。基于谱减法的缺点该文提出一种基于最大后验相位估计的多带谱减法，其中多带谱减法可减少“音乐噪声”的影响，最大后验方法估计纯净语音相位，可以减少在低信噪比时的估计误差。实验结果表明该方法在低信噪比时取得了较好的增强效果。

关键词：语音增强；最大后验相位估计；多带谱减；低信噪比
中图分类号： TN912 **文献标识码**： A **文章编号**：1009-5896(2017)09-2282-05
DOI: 10.11999/JEIT161381

Multi-band Spectral Subtraction of Speech Enhancement Based on Maximum Posteriori Phase Estimation

LI Zhen WU Wenjin ZHANG Qin REN Hui
(*Communication University of China*, *Beijing* 100024, *China*)

Abstract: The spectral subtraction speech enhancement is extensively used due to its simplicity and easy to implement. The principle of this method is to subtract the estimated magnitude of the noise from the magnitude of the noisy signal, but the phase of the noisy signal is unchanged. This conventional method produces the estimating error because it exploits the noisy phase, especially in low SNR, and it produces “musical noise” because of the inaccuracy of the noise estimation. This paper proposes a multi-band spectral subtraction algorithm based on maximum posteriori phase estimation. Experimental results show that the proposed method can get better performance than the conventional method especially in low SNR.

Key words: Speech enhancement; Maximum posteriori phase estimation; Multi-band spectral subtraction; Low SNR

1 引言

语音增强是语音信号处理技术的一个重要分支，语音在传输过程中不可避免受到周围环境或传输媒介引入的噪声的影响，在多数情况下，语音信号总是含有噪声成分。语音增强的目的是研究如何最大限度地从带噪语音信号中去除噪声的影响，得到增强的语音信号，提高受损语音信号的质量和可懂度。

本文主要研究只有一路语音信号可以利用的单通道语音增强算法。目前常用的单通道语音增强算法有基于短时谱(short time spectrum)估计的语音增强算法，基于信号子空间的语音增强算法。近年来随着信号处理技术的发展，一些新兴的数字信号处理技术也应用到了语音增强算法中，2016 年，Wang 等人[1]将压缩感知算法应用到语音增强领域，2008 年 Wójcicki 等人[2]提出改变带噪语音的相位谱，而不改变带噪语音的幅度谱的相位补偿(PSC)语音增强方法。近几年基于相位估计的语音增强算法引起了一些研究者的重视，Mowlaee 等人[3,4]提出了基于相位分解和信噪比信息的谐波相位估计方法。其中基于短时谱估计的算法由于具有使用信噪比范围大、算法高效简单、易于实时处理等特点而得到广泛应用。基于短时谱估计(STSA)的语音增强算法基本上可分为谱减法[5]、维纳滤波算法[6]、最小均方误差估计算法(MMSE)[7]。

在这些常用语音增强算法中，谱减法因其计算简单易于实现得到广泛应用，但它也存在着很多问题，如引入“音乐噪声”，以及谱估计误差的存在。文献[8]提出了多带谱减法，能减少音乐噪声的影响。

收稿日期：2016-12-21；改回日期：2017-04-25；网络出版：2017-05-26
*通信作者：李真 lizhen@cuc.edu.cn
基金项目：“十二五”国家科技支撑计划重大项目(2012BAH38F00)
Foundation Item: The National Science and Technology Planning Project (2012BAH38F00)

对于幅度谱估计的误差问题，也有其他方法对谱减法进行改进，但是这些方法没有对相位进行改变，使用的是带噪语音信号的相位，在低信噪比(<6 dB)时，带噪相位会导致语音信号变得粗糙，并达到可能被听觉所感知的程度，进而降低语音质量[9]。基于此本文提出一种基于相位估计的多带谱减法，能有效改善低信噪比条件下因为采用带噪语音相位进行增强所产生的估计误差问题。

2　谱减语音增强算法及存在问题

假设带噪语音信号可写为

$$y(n)=s(n)+d(n) \tag{1}$$

其中，$y(n)$是带噪语音信号，$s(n)$是纯净语音信号，$d(n)$是平稳加性噪声信号，$s(n)$与$d(n)$相互独立。对$y(n)$进行分帧加窗，并进行短时傅里叶变换，得到

$$Y(n,k)=S(n,k)+D(n,k) \tag{2}$$

其中，$Y(n,k),S(n,k),D(n,k)$分别是含噪语音、纯净语音和噪声信号的短时傅里叶变换。

谱减法算式为

$$\left|\widehat{S}(\omega)\right|=\left(\left|Y(\omega)\right|-\left|\widehat{D}(\omega)\right|\right)\mathrm{e}^{\mathrm{j}\varphi_y(\omega)} \tag{3}$$

由于噪声的频谱特性，噪声不会对语音的整个频谱都具有同等的影响，有些频率上的影响会比别的频率更加严重，基于此文献[8]提出了一种基于多带方式的谱减法，将语音频谱划分为N个不同的子带，谱减法在每个子带独立进行。第i个子带的纯净语音信号谱的估计如式(4)：

$$\left|\widehat{S}_i(\omega_k)\right|^2=\left|\overline{Y}_i(\omega_k)\right|^2-\alpha_i\cdot\delta_i\left|\widehat{D}(\omega_k)\right|^2,\ b_i\le\omega_k\le e_i \tag{4}$$

α_i为第i个子带的过减因子，δ_i为子带减法因子，它可以按子带独立设置以满足不同的噪声抑制要求。这两个参数的设置方法可参考文献[8]。多带谱减法可有效降低谱减法的“音乐噪声”问题。

图 1 所示是带噪语音、纯净语音和噪声 3 个信号的极坐标关系图，从图中可以看出估计的纯净语音信号和目的纯净语音信号之间有很大不同，从而产生了误差ΔE。图 2 所示，如果纯净语音信号幅度保持不变，并且纯净语音和噪声之间相位差也保持不变，随着噪声的增加，即信噪比的降低，误差从$\Delta E1$增加到$\Delta E2$[10]。所以传统谱减法的误差是因为纯净语音和噪声之间存在的相位差而产生的，并且误差会随着信噪比的下降而进一步增加。

3　基于后验相位估计的多带谱减法

谱减法的估计误差问题是因为在进行信号综合时采用带噪语音的相位，因而造成了低信噪比的估计误差较大。为了进一步提高低信噪比时谱减法增强后的语音质量，本文采用最大后验相位估计方法估计纯净语音相位[11]，并将其引入多带谱减法。

3.1 最大后验相位估计

假设带噪语音可表示为谐波信号和噪声的和，

$$y(n)=\sum_{h=1}^{H_l}A_h\cos\left(h\omega_0 n+\theta_h+d(n)\right) \tag{5}$$

根据似然函数

$$p(y\mid\theta)=\frac{1}{2\pi\sigma_d^2}\exp\left\{-\frac{1}{2\sigma_d^2}\sum_{n=0}^{N-1}\left(y(n)-A\cos\left(h\omega_0 n+\theta\right)\right)^2\right\} \tag{6}$$

其中，σ_d^2是噪声方差。假设相位服从冯•米塞斯分布。冯•米塞斯分布也被称作循环正态分布。是已知循环均值μ_c，浓度k的最大熵分布，被广泛用在语音分析与合成中。

$$\theta_h\sim VM\left(\mu_h,k_h\right),\ p(\theta_h)=\frac{\exp\left(k_h\cos\left(\theta_h-\mu_h\right)\right)}{2\pi I_0\left(k_h\right)} \tag{7}$$

其中，$I_0(\bullet)$是零阶贝塞尔函数，定义$x=[x(0)\ x(1)\ \cdots\ x(N_W-1)]$，$\theta=[\theta_1\ \theta_2\ \cdots\ \theta_H]$，最大后验相位估计是对式(8)求解

$$\hat{\theta}=\arg\max_{\theta}\frac{p(y\mid\theta)p(\theta)}{p(y)}=\arg\max_{\theta}\ p(y\mid\theta)p(\theta) \tag{8}$$

将式(6)，式(7)代入式(8)并去除常数部分，得到

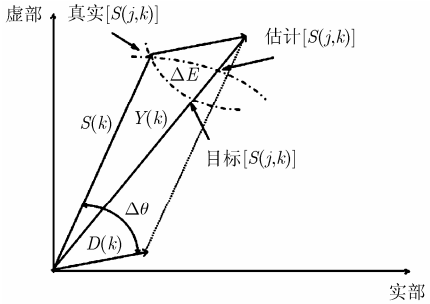

图 1 信号间的极坐标图

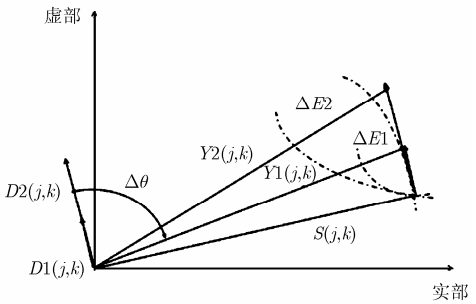

图 2 信噪比下降时相位差影响

$$L(\theta_h)=-\frac{1}{2\sigma_d^2}\sum_{n=0}^{N_W-1}\left(y(n)-\sum_{h=1}^{H}A_h\cos(h\omega_0 n+\theta_h)\right)^2+\sum_{h=1}^{H}k_h\cos(\theta_h-\mu_h) \tag{9}$$

求取 $L(\theta)$ 对 θ 的导数，并使其值为 0，可得到最大后验相位估计。参考文献[12]假设当 $h\omega_0$ 不在 0 或 π 附近，并且 N 足够大时，

$$\sum_{n=0}^{N_W-1}\sin(h\omega_0 n+\theta)\cos(h\omega_0 n+\theta)\approx 0$$

则得到

$$\begin{aligned}\frac{dL(\theta_h)}{d\theta_h}=&-\frac{1}{\sigma_d^2}\sum_{n=0}^{N_W-1}A_h y(n)\sin(h\omega_0 n+\theta_h)-k_h\sin(\theta_h-\mu_h)\\=&\cos(\theta_h)\left(\frac{A_h}{\sigma_d^2}\sum_{n=0}^{N_W-1}y(n)\sin(h\omega_0 n)+k_h\sin(\mu_h)\right)\\&+\sin(\theta_h)\left(\frac{A_h}{\sigma_d^2}\sum_{n=0}^{N_W-1}y(n)\cos(h\omega_0 n)+k_h\cos(\mu_h)\right)\\=&0\end{aligned} \tag{10}$$

用 θ_{MAP} 表示经过最大后验求得的相位估计值，即式(10)的解，则

$$\theta_{\mathrm{MAP}}=\tan^{-1}\frac{-\frac{A}{\sigma_d^2}\sum_{n=0}^{N_W-1}y(n)\sin(h\omega_0 n)+k_h\sin(\mu_h)}{\frac{A}{\sigma_d^2}\sum_{n=0}^{N_W-1}y(n)\cos(h\omega_0 n)+k_h\cos(\mu_h)} \tag{11}$$

从式(11)可看出最大后验相位估计是冯•米塞斯参数 μ_h, k_h，数据长度 N_W，以及信噪比 A/σ_d^2 的函数，冯•米塞斯分布的参数估计可参考文献[11]。

3.2 基于最大后验相位估计的多带谱减语音增强算法

基于前面所述传统谱减法存在“音乐噪声”及算法误差的问题，本文提出采用多带谱减法去除“音乐噪声”，采用估计纯净语音相位与多带谱减法估计幅度进行综合的方法减小低信噪比时的估计误差，算法结构如图 3 所示。

4 实验仿真与分析

根据图 3 所示最大后验相位估计多带谱减算法的结构，通过 MATLAB 实现，输入语音信号采用 NOIZEUS 语料库中的语音，采样率 8 kHz，实验首先生成不同信噪比(0 dB, 5 dB, 10 dB, 15 dB)下添加 white, street, car, restaurant 4 种不同噪声的语音文件。实验采用汉宁窗，帧长 32 ms，帧间重叠 24 ms，带噪语音经过分帧、加窗，傅里叶变换，再将整个频带分为 6 个不重叠的子带，根据式(4)估计纯净语音的幅度谱。再按式(11)纯净语音相位的最大后验估计，最后将幅度谱和相位谱结合得到纯净语音信号谱，进行傅里叶逆变化，并重叠相加后得到增强后的语音信号。

语音增强后的客观语音质量评价采用感知语音质量评估测度 PESQ(Perceptual Evaluation of Speech Quality)得分进行评测，表 1，表 2，表 3，表 4 给出了采用传统谱减法、多带谱减法及本文方法在不同噪声情况下增强后的 PESQ 结果，由其结果可以看出，相比于没有采用相位估计的谱减法，本文算法在低信噪比(<6 dB)时拥有最强的效果。但是随着信噪比提高该方法的优越性在 White 噪声和 car 噪声情况下不再明显。

语音质量的提高不一定带来可懂度的改善，这是由于在抑制声学噪声的同时可能产生纯净语音信号的失真。基于此图 4 给出了对不同方法增强后的语音可懂度评价结果用 STOI(Short-Time Objective Intelligibility measure)[13]值比较，从图 4

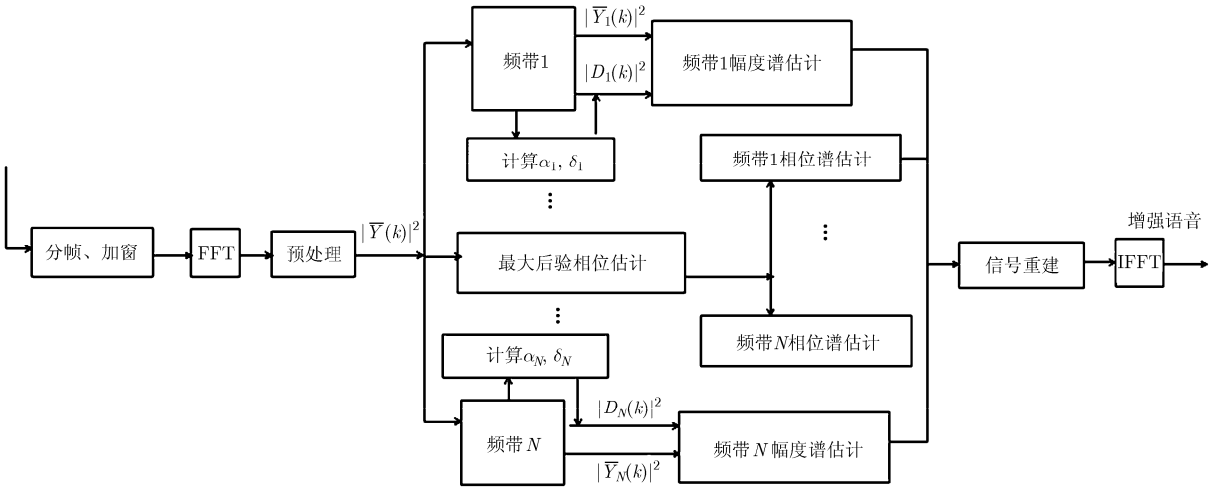

图 3 基于最大后验相位估计的多带谱减语音增强算法结构

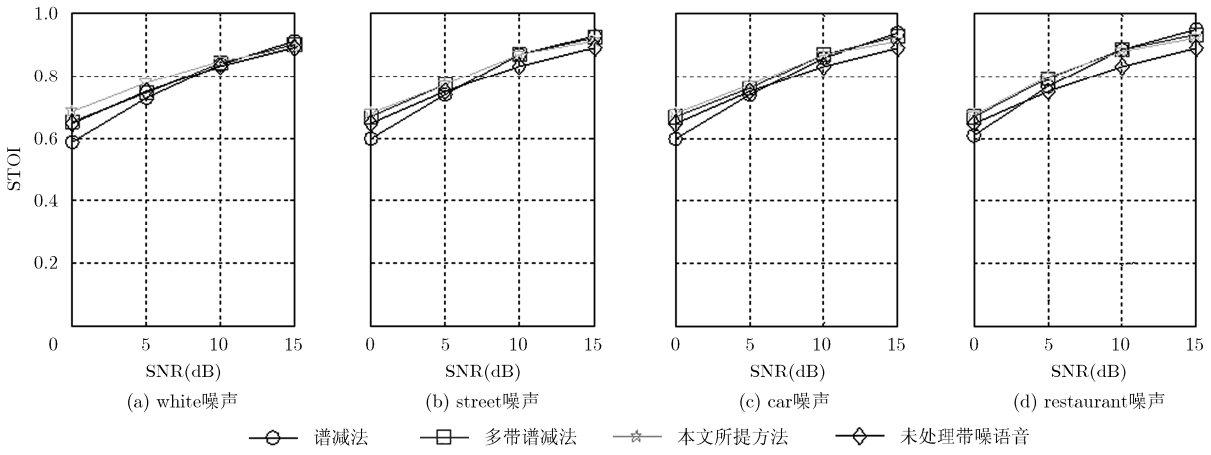

图 4 不同谱减增强方法对不同噪声语音增强后可懂度比较

表 1 谱减法、多带谱减法和本文方法对带 White 噪声语音信号增强后的平均 PESQ 得分

信噪比(dB)	纯净语音	带噪语音	谱减法	多带谱减法	本文方法
0	4.50	1.59	1.76	1.66	1.90
5	4.50	1.83	2.22	2.04	2.26
10	4.50	2.14	2.64	2.54	2.60
15	4.50	2.47	3.06	2.90	2.86

表 2 谱减法、多带谱减法和本文方法对 street 噪声语音信号增强后的平均 PESQ 得分

信噪比(dB)	纯净语音	带噪语音	谱减法	多带谱减法	本文方法
0	4.50	1.59	1.67	1.82	1.91
5	4.50	1.83	2.09	2.19	2.24
10	4.50	2.14	2.50	2.56	2.59
15	4.50	2.47	2.82	2.84	2.85

表 3 谱减法、多带谱减法和本文方法对 car 噪声语音信号增强后的平均 PESQ 得分

信噪比(dB)	纯净语音	带噪语音	谱减法	多带谱减法	本文方法
0	4.50	1.59	1.72	1.79	1.91
5	4.50	1.83	2.09	2.14	2.24
10	4.50	2.14	2.53	2.57	2.59
15	4.50	2.47	2.96	2.91	2.85

可以看出在不同噪声情况下，在低信噪比时，本文算法的语音可懂度均略高于其他方法。

为了评价相位改进对语音增强效果的影响，文献[9]提出了相位偏差的概念，式(12)定义了相位偏差是带噪语音相位和纯净语音相位的差。

表 4 谱减法、多带谱减法和本文方法对 restaurant 噪声语音信号增强后的平均 PESQ 得分

信噪比(dB)	纯净语音	带噪语音	谱减法	多带谱减法	本文方法
0	4.50	1.59	1.66	1.85	1.87
5	4.50	1.83	2.01	2.19	2.24
10	4.50	2.14	2.45	2.57	2.58
15	4.50	2.47	2.80	2.85	2.87

$$\varphi_{\mathrm{dev}} = \varphi_y(k,l) - \varphi_s(k,l) \tag{12}$$

文献[14]进一步提出一个相位偏差的度量指标 PD，

$$d_{\mathrm{PD}} = \frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K}\left(\cos\left(\varphi_{\mathrm{dev}}(k,l) - \cos\widehat{\varphi}_{\mathrm{dev}}(k,l)\right)\right)^2 \tag{13}$$

纯净语音的相位偏差 PD 值为 0, PD 值越小，语音质量越高。图 5 分别给出了不同噪声情况下在不同信噪比下通过不同增强方法的相位偏差值对比，从图中可以看出本文所提方法在低信噪比(<6 dB)时的相位偏差最小。

5 结论

本文基于传统谱减法所存在的“音乐噪声”和估计误差问题提出了基于最大后验相位估计的多带谱减语音增强方法，对受加性噪声干扰的带噪语音进行语音增强。在假设相位服从冯•米塞斯分布的前提下对纯净语音进行最大后验相位估计，与多带谱减法进行的幅度谱估计相结合，得到估计的纯净语音谱，进一步进行逆傅里叶变换得到估计的纯净语音。实验结果表明本文方法在低信噪比条件下的语音感知质量 PESQ 有明显提高，解决了传统谱减法所存在的问题。

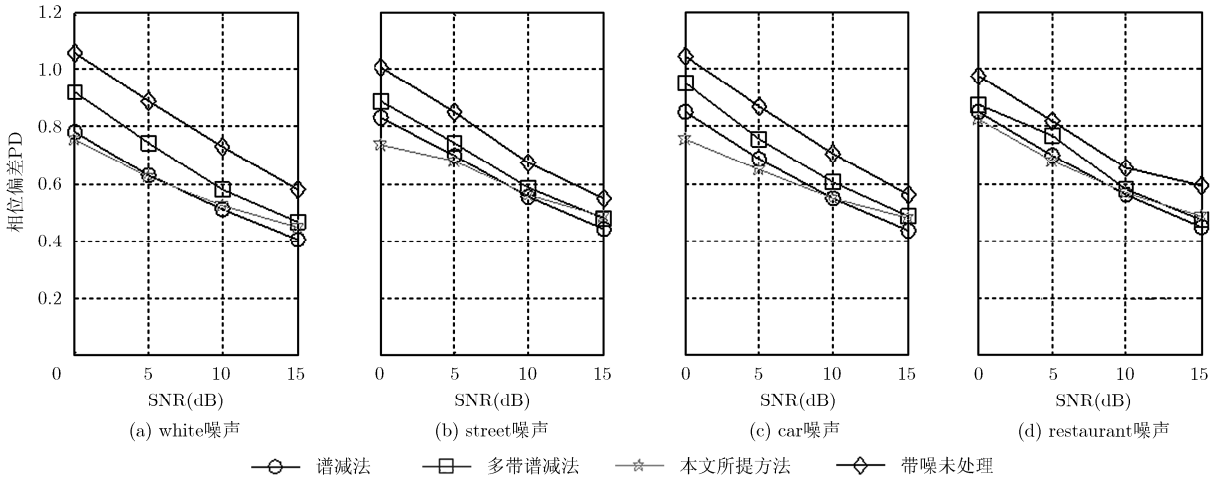

图 5 不同谱减增强方法对不同噪声语音增强后相位偏差比较

参 考 文 献

[1] WANG Jiaching, LIN Changhong, WANG Shufan, *et al*. Compressive Sensing-based speech enhancement[J]. *IEEE Transactions on Audio*, *Speech and Language Processing*, 2016, 24(11): 2122–2131. doi: 10.1109/TASLP.2016.2598306.

[2] WÓJCICKI K, MILACIC M, STARK A, *et al*. Exploiting conjugate symmetry of the short-time fourier spectrum for speech enhancement[J]. *IEEE Signal Processing Letters*, 2008, 15: 461–464. doi: 10.1109/LSP.2008.923579.

[3] MOWLAEE P and KULMER J. Harmonic phase estimation in single-channel speech enhancement using phase decomposition and SNR information[J]. *IEEE Transactions on Audio*, *Speech and Language Processing*, 2015, 23(9): 1521–1532. doi: 10.1109/TASLP.2015.2439038.

[4] KULMER J and MOWLAEE P. Phase estimation in single channel speech enhancement using phase decomposition[J]. *IEEE Signal Processing Letters*, 2015, 22(5): 598–602. doi: 10.1109/LSP.2014.2365040.

[5] BOLLS F. Suppression of acoustic noise in speech using spectral subtraction [J]. *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, 1979, 27(2): 113–120. doi: 10.1109/TASSP.1979.1163209.

[6] WIENER N. The Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications[M]. Cambridge: Massachusetts, MIT, 1949: 81–101.

[7] EPHRAIM Y and MALAH D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator[J]. *IEEE Transactions on Acoustics*, *Speech*, *and Signal Processing*, 1984, 32(6): 1109–1121. doi: 10.1109/TASSP.1984.1164453.

[8] KAMATH S and LOIZOU P C. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise[C]. IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, 2002: IV-4164-IV-4164.

[9] VARY P. Noise Suppression by spectral magnitude estimation: Mechanism and theoretical limits[J]. *Signal Processing*, 1985, 8(4): 387–400. doi: 10.1016/0165-1684(85)90002-7.

[10] SAMUI S. Improved single channel phase-aware speech enhancement technique for low signal-to-noise ration signal[J]. *IET Signal Processing*, 2016, 10(6): 641–650. doi: 10.1049/iet-spr.2015.0182.

[11] KULMER J and MOWLAEE P. Harmonic phase estimation in single-channel speech enhancement using Von Mises distribution and prior SNR[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 2015: 5063–5067.

[12] KAY S M. Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory[M]. New Jersey: Prentice Hall PTR, 1993: 164–172.

[13] TAAL C H, HENDRIKS R C, HEUSDENS R, *et al*. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. *IEEE Transactions on Audio*, *Speech and Languages*, 2011, 19(7): 2125–2136. doi: 10.1109/TASL.2011.2114881.

[14] GAICH A and MOWLAEE P. On speech quality estimation of phase-aware single-channel speech enhancement[C]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), Brisbane, Australia, 2015: 216–220.

李 真： 女，1978 年生，讲师，研究方向为语音数字信号处理.

吴文锦： 女，1993 年生，硕士生，研究方向为信号与信息处理.

张 勤： 男，1956 年生，教授，研究方向为宽带信息网络.

任 慧： 男，1966 年生，教授，研究方向为自动监控.