基于谱优势与非线性变换混合的立体声声学回声消除方法

杨鹤飞 郑成诗 李晓东*
(中国科学院声学研究所 北京 100190)
(中国科学院上海高等研究院 上海 201210)

摘 要：立体声声学回声控制系统中，两通道信号间的强相关性导致自适应算法的解不唯一，滤波器失调较大。为了解决此问题，并尽可能地保留语音质量，该文基于心理声学上的谱优势效应提出一种新型的混合立体声声学回声消除方法。根据谱优势效应，在3个低次谐频处注入能量较弱的正弦信号，以减弱低频相关性。同时对非线性变换法进行改进，以用于中高频去相关处理。与传统方法的多方面性能测试对比结果表明：所提方法能有效改善失调性能并提高收敛速度，且具有较小的语音失真。

关键词：语音信号处理；立体声声学回声消除；去相关；谱优势；非线性变换
中图分类号： TN912.3 **文献标识码**： A **文章编号**：1009-5896(2015)02-0373-07
DOI: 10.11999/JEIT140274

A Stereo Acoustic Echo Cancellation Method Based on the Hybrid of Spectral Dominance and Nonlinear Transformation

Yang He-fei Zheng Cheng-shi Li Xiao-dong
(*Institute of Acoustics*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)
(*Shanghai Advanced Research Institute*, *Chinese Academy of Sciences*, *Shanghai* 201210, *China*)

Abstract: In stereophonic Acoustic Echo Cancellation (AEC) systems, the strong correlation between the two stereophonic channels leads to nonuniqueness of adaptive solutions and further large filter misalignment. To solve this problem and preserve speech quality, the psychoacoustic spectral dominance effect is utilized to propose a novel hybrid decorrelation method for stereo AEC. According to spectral dominance, weak sinusoids are injected at the three lowest harmonics so as to reduce low-frequency coherence. Besides, the nonlinear transformation method is modified and applied to high-frequency decorrelation. Comparison test on several performances with traditional approaches is carried out. Results show that the proposed method can effectively improve filter misalignment together with convergence rate. Moreover, low speech distortion can be achieved simultaneously.

Key words: Speech signal processing; Stereophonic Acoustic Echo Cancellation (AEC); Decorrelation; Spectral dominance; Nonlinear transformation

1 引言

通信系统中由于存在声学回声问题而需要进行声学回声消除(Acoustic Echo Cancellation, AEC)，可利用自适应算法[1,2]实现该目的。立体声具有较强空间感，已广泛应用于语音通信系统。此时，远端房间内的两通道传声器信号间存在强相关性。这导致自适应滤波器的收敛值不唯一，从而使立体声声学回声消除(Stereo AEC, SAEC)系统的滤波器失调较大[3]。为了改善失调性能，需对两通道传声器信号进行去相关预处理。同时，为了保证通信质量，相应处理不应引入明显的语音失真。

目前已有大量的去相关方法可应用于SAEC。一类方法在信号中添加一部分其它信号。其中最典型的是非线性变换法，即在信号中添加一部分其本身的非线性变换[3]；虽然有多种非线性变换可供选择，但半波整流变换最有效[4]。亦有人提出在两个通道中添加相互独立的低强度随机噪声[5]，但为了不影响听觉感知，信噪比需足够高以至于去相关效果十分有限；此时可利用听觉掩蔽效应对添加噪声的频谱进行幅度调制[6]，以增强去相关效果。另一类方法通过直接改变原始信号来降低通道间相关性，如全通滤波法[7]、时变的相位调制法[8]和时间反转法[9]等。其中，时间反转法仅选择性地对能量足够弱的数据帧进行时域反转。为了在去相关与语音失真之间取得更好的折中，近年的SAEC研究开始关注心理声

2014-03-04收到，2014-11-04改回
国家自然科学基金(61201403)资助课题
*通信作者：李晓东 lxd@mail.ioa.ac.cn

学效应，并结合其它方法进行分频段处理。文献[10]基于听觉掩蔽效应在低频部分添加掩蔽噪声，并在高频段进行时变的全通滤波；文献[11]等结合基频遗失效应与非线性相位调制，其后期研究进一步将信号分为 3 个频段进行不同处理[12]。

本文基于心理声学效应提出一种混合 SAEC 方法。根据心理声学谱优势效应在 3 个低次谐频成分附近注入正弦信号，以降低低频相关性；在高频段则利用时变的非线性相位调制对非线性变换去相关法进行改进。为了简洁，下文中将所提出的方法简写为 SDNT (Spectral Dominance-Nonlinear Transformation)法。

2 SDNT 法

根据人耳听觉特性，人耳听觉系统在 1.5 kHz 以下对相位变化比较敏感[10]。因此，为了减小语音失真需尽量保持该频率范围内的信号相位不变，本文利用心理声学效应实现该目的。谱优势效应是一种心理声学效应，它是指：复音信号中，某些谐频成分对音高感知的影响更显著[13]。文献[14]指出，当复音信号的基频低于 350 Hz 时，其音高感知主要取决于四次及以上的谐频成分。因此可通过改变立体声某一通道信号中的 3 个低次谐频成分来达到低频去相关的目的，根据谱优势效应，如此操作不会引入明显的听觉损失。

综合上述分析，本文将 SDNT 法中的分频点取为 1 kHz。在 1 kHz 以下的低频段，基于谱优势效应改变立体声语音信号中某一通道(本文选左通道)的 3 个低次谐频成分。可利用滤波器实现改变谐频的目的，但实际滤波器的旁瓣影响将引入额外的失真。本文采用注入正弦信号的方法，分别在 3 个低次谐频成分附近注入能量较弱的正弦信号，且各正弦信号受相应谐频成分控制。

在 1 kHz 以上的高频段，由于双耳声级差对声源定位的影响增强，处理时应尽量保持信号幅值不变。本文基于非线性相位调制得到一种改进的非线性变换法。为了进一步降低相关性，调制相位随时间变化。考虑到人耳的相位敏感度随频率递减，将非线性相位调制成分在信号中的能量占比设定为随频率递增。

由此得到结合心理声学谱优势效应的混合 SAEC 方法，即 SDNT 法，其实现框图如图 1 所示，其中 $x_1(n)$与 $x_2(n)$分别表示左、右通道信号。低频段的谱优势处理仅应用于 $x_1(n)$；而高频处理则同时对两通道信号进行，但分别对 $x_1(n)$与 $x_2(n)$采用不同的调制相位。下面分别在 2.1 节和 2.2 节中对低、高频处理进行详尽阐述。

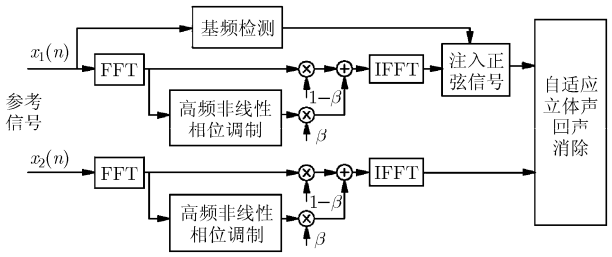


图 1 SDNT 法的实现框图

2.1 低频处理

2.1.1 基频检测 首先利用加权自相关法[15]检测基频。假设加权自相关函数为 $\psi(\tau)$，其最大值所对应的时延数 τ_0 即为基频周期的估计值。语音信号中，需预先限定基频取值范围。考虑到语音信号的基频最高约为 500 Hz[16]，但位于区间[60,300] Hz 内的概率高达 97%[17]；另外，谱优势效应仅在基频低于 350 Hz 时适用[14]。为了既适用于多数语音信号，又满足谱优势前提，本文将基频取值范围限定为[60,350] Hz。如此相当于

$$\tau_0 = \left\{\tau : \psi(\tau) = \max(\psi(\tau));\ f_s/350 \le \tau \le f_s/60\right\} \quad (1)$$

其中，f_s 为信号的采样频率。

式(1)得到的基频周期精度较低，仅为 1 个采样周期；为了提高检测精度，可进一步对以最大值为中心的相邻 3 个 $\psi(\tau)$ 值进行插值。设所用插值间隔为 Δ_τ，则所得基频的估计精度为

$$\frac{1}{\Delta_f} = 1\Big/\left(\frac{f_s}{\tau_0} - \frac{f_s}{\tau_0+\Delta_\tau}\right) = \frac{\tau_0(\tau_0+\Delta_\tau)}{f_s\Delta_\tau} \quad (2)$$

其值随 τ_0 的减小而降低，因此最低基频精度为

$$\begin{aligned}\left(1/\Delta_f\right)_{\min} &= \left(1/\Delta_f\right)\Big|_{\tau_0=f_s/350} \\ &= \left(f_s+350\Delta_\tau\right)\Big/\left(350^2\Delta_\tau\right)\end{aligned} \quad (3)$$

根据 τ_0 可得基频估计值为 $f_0 = f_s/\tau_0$，相应地，二、三次谐频分别为 $2f_0$ 和 $3f_0$。为了改变 3 个低次谐频成分，本文将 f_0 ,$2f_0$ 和 $3f_0$ 取为待注入正弦信号的频率。由于受基频检测算法的精度限制，f_0 存在一定的估计误差，从而使得被改变的不再是信号的 3 个低次谐频成分，这可能会引入较大的语音失真。考虑到人耳的最低可辨别特性[18]，当两个频率之差不大于阈值 d_f时，人耳不能对二者进行区分，其中 d_f取值为[18]

$$d_f = \begin{cases}1\ \text{Hz}, & f < 500\ \text{Hz} \\ 0.002f, & f \ge 500\ \text{Hz}\end{cases} \quad (4)$$

由此易知，f_0 ,$2f_0$ 和 $3f_0$ 的估计误差均应不高于 1 Hz。由于 d_f随单频信号持续时间的缩短而增大[18]，且语音信号中同一基频值的持续时间很短，本文将基频检测精度限定为不低于 1 Hz，结合式(3)得

$$\Delta_\tau \le f_s \big/ (350^2 - 350) \tag{5}$$

本文采样频率为 16 kHz，此时 $\Delta_\tau \le 0.13$，可将插值间隔设定为 0.1 个采样点。

2.1.2 基频和二、三次谐频处理　傅里叶变换得到左通道信号的频谱 $X_1(f)$。利用基频估计值检索 $X_1(f)$，分别得到基频与二、三次谐频信号的幅值和初相，并进行幅值加权得到

$$\left.\begin{aligned} A_i &= \left|X_1(i\cdot f_0)\right|\cdot\sigma_i \\ \varphi_i &= \text{angle}\left\{X_1(i\cdot f_0)\right\} \end{aligned}\right\} \tag{6}$$

其中，$i = 1, 2, 3$ 分别对应于基频和二、三次谐频，A,φ,σ 分别表示幅值、初相和幅值加权因子，函数 angle{•} 表示对复数取相位，σ_i 在[0,1]内取值。所需注入的 3 个正弦信号的频率分别为 $f_i = i\cdot f_0$，其幅值和初相采用式(6)所得结果。由此构造出 3 个正弦信号 $s_i(n)$，并通过时域加法注入到左通道信号中。$s_i(n)$ 负责改变第 i 次谐频，易知 σ_i 越大，第 i 次谐频的改变程度越大，语音失真越明显。

2.2 高频处理

非线性变换法的频域表示为

$$X_j'(f) = (1-\beta)\cdot X_j(f) + \beta\cdot F\left\{X_j(f)\right\} \tag{7}$$

其中，$X_j(f)$, j=1,2 为 $x_j(n)$的频谱；β 为非线性加权系数，在[0,1]内取值；函数 $F\{\cdot\}$ 表示非线性变换，传统的非线性变换去相关法中一般采用半波整流变换[3]。由于人耳听觉系统在高频对信号的幅值变化比较敏感，本文利用非线性相位调制得到一种改进的非线性变换法。

对于 1 kHz 以上的高频段，根据式(7)同时对两通道信号进行变换。函数 $F\{\cdot\}$ 采用非线性相位调制，且两通道的调制相位互为相反数，由此得到本文所采用的高频处理：

$$\left.\begin{aligned} X_1'(f) &= X_1(f)\cdot\left[(1-\beta)+\beta\cdot e^{i\Theta}\right] \\ X_2'(f) &= X_2(f)\cdot\left[(1-\beta)+\beta\cdot e^{-i\Theta}\right] \end{aligned}\right\} \tag{8}$$

其中，调制相位 Θ 随帧数 m 和频率 f 非线性变化。为了保证 IFFT 的实数性，$f_s/2$ 处的 Θ 值为 0，因此，

$$\Theta = \begin{cases} 0, & f \le 1\text{ kHz 或 } f = f_s/2 \\ \pi/2\cdot\sin(\vartheta), & 1\text{ kHz} < f < f_s/2 \end{cases} \tag{9}$$

其中，$\vartheta = 2\pi\cdot m\cdot f/f_s$。对于相邻两个数据帧，$\vartheta$ 值的相对差异为 $[(m+1)-m]/m = 1/m$，随 m 递减，即调制相位 Θ 的变化速度越来越慢；信号较长时，去相关效果将大为减弱。因此，将 ϑ 值优化为

$$\vartheta = 2\pi\cdot(m\%\Lambda)\cdot f/f_s \tag{10}$$

其中，Λ 表示变化周期，本文取为 200 帧，“%”表示将 m 对 Λ 求余数。

对于式(8)所示变换，变换前后信号能量几乎不变，此时通道间相干系数与频域互相关函数等价。下面利用互相关函数来评价高频去相关效果。变换后，两通道信号间的频域互相关函数 $C_{12}(f)$ 为

$$\begin{aligned} C_{12}(f) &= (\beta-1)^2\cdot\mathrm{E}\{X_1^*(f)\cdot X_2(f)\} \\ &\quad + [1-(\beta-1)^2]\cdot\mathrm{E}\{X_1^*(f)\cdot X_2(f)e^{-i2\Theta}\} \end{aligned} \tag{11}$$

其中，E{·} 表示数学期望，“*”表示对复数取共轭。式(11)的推导过程中利用了如下假设：

$$\begin{aligned} \mathrm{E}\{X_1^*(f)\cdot X_2(f)e^{-i\Theta}\} &\approx \mathrm{E}\{X_1^*(f)\cdot X_2(f)e^{-i2\Theta}\} \\ &\approx 0 \end{aligned} \tag{12}$$

观察式(11)的等号右边，$\mathrm{E}\{X_1^*(f)\cdot X_2(f)e^{-i2\Theta}\}$ 很小，接近于 0。由于 β 在[0,1]内取值，可知 β 越大，C_{12} 越小，去相关效果越强；但此时信号失真也大，折中的选择是 $\beta = 0.5$。构造一个立体声信号，两通道信号为完全相同的高斯白噪声。基于该信号统计得到 C_{12} 随 β 的变化关系，如图 2 所示。当 $\beta < 0.9$ 时，C_{12} 随 β 递减，与上述推断结果吻合；但当 $\beta > 0.9$ 时，C_{12} 略有增加，这可能是由式(12)的近似处理所导致。

在 1 ~ 2 kHz 范围内，人耳听觉系统对信号的相位和幅值均比较敏感[10]，因此将此频率范围内的 β 取为折中值 0.5。由于人耳听觉系统的相位敏感度随频率升高而降低，当频率高于 2 kHz 时，令 β 随 f 线性递增。结合上述统计结果，得到非线性加权系数 β 的值如式(13)所示，其值在 $f_s/2$ 处取最大值 0.9。得益于式(9)中调制相位的分段定义，β 在 1 kHz 以下的非零值并不会对低频信号产生影响。

$$\beta = \begin{cases} 0.5, & f \le 2\text{ kHz} \\ 0.5 + 0.4\cdot(f-2000)\big/(f_s/2-2000), & \\ \qquad 2\text{ kHz} < f \le f_s/2 & \end{cases} \tag{13}$$

3　σ_i 的选值

为了对比不同的 σ_i 取值方案，采用如下指标对其性能进行评价：

(1)通道间相干系数(Inter-Channel Correlation Coefficient, ICCC): ICCC 越小，相关性越弱。其定义为

$$\begin{aligned} \mathrm{ICCC}(f) &= \left|\mathrm{E}\left\{X_1^*(f)\cdot X_2(f)\right\}\right| \\ &\quad \Big/\sqrt{\mathrm{E}\left\{\left|X_1(f)\right|^2\right\}\cdot\mathrm{E}\left\{\left|X_2(f)\right|^2\right\}} \end{aligned} \tag{14}$$

(2)失调[19](MISalignment, MIS)：失调是评价自适应通道辨识能力的量化指标，其定义为

$$\mathrm{MIS}(n) = 20\cdot\lg\left(\sum_{j=1}^{2}\left\|\boldsymbol{h}_j - \hat{\boldsymbol{h}}_j(n)\right\|_2 \Big/ \sum_{j=1}^{2}\left\|\boldsymbol{h}_j\right\|_2\right) \tag{15}$$

其中，$\boldsymbol{h}_j$ 为物理脉冲响应，$\hat{\boldsymbol{h}}_j(n)$ 为第 n 个采样周期

内的自适应脉冲响应，$\|\bullet\|_2$表示2范数。

(3) 语音质量感知评估[20](Perceptual Evaluation of Speech Quality, PESQ): PESQ将主观音质客观量化，它在-0.5～4.5之间[20]取值，其值越大，表示失真越小。本文利用文献[21]中的PESQ算法计算其值。

3.1 σ_i 对低频去相关性能的影响

首先定义函数$\delta_m(f)$，当频率f在第m帧被检测为3个低次谐频之一时$\delta_m(f)$=1，否则$\delta_m(f)$=0；然后定义$S(f)=\sum_m\left|X_m(f)\right|^2$，以及$S_p(f)=\sum_m\delta_m(f)\cdot\left|X_m(f)\right|^2$。令$\mathrm{ICCC}_{\mathrm{low}}$表示不高于1 kHz的低频部分的ICCC值。考虑$X_1(f)=X_2(f)=X(f)$的情况，未进行低频去相关处理时$\mathrm{ICCC}_{\mathrm{low}}(f)\equiv 1$。为了便于观察$\sigma_i$的影响机理，令$\sigma_1=\sigma_2=\sigma_3=\sigma$，则处理后的低频ICCC值为

$$\mathrm{ICCC}'^2_{\mathrm{low}}(f)=\left\{1+\frac{S^2(f)/4-\left[S_p(f)-S(f)/2\right]^2}{\left[S_p(f)+S(f)/\sigma\right]^2}\right\}^{-1} \tag{16}$$

由此得到如下结论：

(1)σ越大，$\mathrm{ICCC}'_{\mathrm{low}}$越小。由式(16)易得此结论。统计处理后的语音测试信号的低频ICCC值，图3给出了125 Hz和250 Hz的ICCC-σ曲线。由图3可见，不论是125 Hz，还是250 Hz, ICCC都随σ递减，与此结论相符。

(2)频率f在3个低次谐频中的出现概率越大，$\mathrm{ICCC}'_{\mathrm{low}}$越小。将式(16)对$S_p(f)$求偏导数，并化简得到

$$\frac{\partial\,\mathrm{ICCC}'^2_{\mathrm{low}}(f)}{\partial S_p(f)}=-\frac{\sigma^2\left[S(f)+\sigma\cdot S_p(f)\right]\left[S(f)-(\sigma+2)\cdot S_p(f)\right]}{S(f)\left[S(f)+(\sigma^2+2\sigma)\cdot S_p(f)\right]^2} \tag{17}$$

由于语音信号的基频时刻变化，对于所有数据帧m, $\delta_m(f)=1$成立的次数很少，可得到$S_p(f)<S(f)/(2+\sigma)$。此时，式(17)的值为负，$\mathrm{ICCC}'_{\mathrm{low}}(f)$随$S_p(f)$递减。因此，频率$f$在3个低次谐频中的出现概率越大，相应地有$S_p(f)$越大，$\mathrm{ICCC}'_{\mathrm{low}}(f)$越小，即结论(2)成立。

以一段长20 min，采样率16 kHz的男声语音信号做为样本数据，利用2.1.1节所述方法对浊音段进行基频检测，统计得到各频率在基频与二、三次谐频中的出现概率如图4所示。当频率高于约350 Hz时，出现概率随频率升高而降低。因此，当$\sigma_1=\sigma_2=\sigma_3$时，根据结论(2)易知$\mathrm{ICCC}'_{\mathrm{low}}$将近似随频率升高而增大。

3.2 σ_i 的选取

为了确定σ_i的最佳取值，对不同取值方案进行性能对比。由于σ_i仅影响左通道信号的低频处理，此处讨论左通道PESQ和低频ICCC。但MIS性能将在全频带讨论。首先选择3种取值方案：SDNT, $\sigma_1=0.10$，$\sigma_2=0.20$，$\sigma_3=0.30$；SDNT-a，$\sigma_1=0.20$，$\sigma_2=0.35$，$\sigma_3=0.50$和SDNT-b，$\sigma_1=0.50$，$\sigma_2=0.50$，$\sigma_3=0.50$。对男、女声语音下的性能进行平均，得到上述3种取值方案的性能对比如图5所示。由图5(a)可见，$\sigma_1<\sigma_2<\sigma_3$时能获得较平坦的低频ICCC性能，且$\sigma_i$越大，去相关效果越好，这与3.1节的结论吻合。观察图5(b)，方案SDNT-a和SDNT-b下的失调性能并未比方案SDNT好，可能的原因是：当低频ICCC值在一定范围内变化时，整体ICCC-f曲线的形状对失调性能影响更大。由于本文旨在改善失调性能，综合图5可将σ_i之间的关系设定为$\sigma_3:\sigma_2:\sigma_1=3:2:1$。

令$\sigma_3:\sigma_2:\sigma_1=3:2:1$，不同$\sigma_1$值时的MIS曲线收敛值与PESQ如表1所示。结果显示，当σ_1从0增大至0.1时，失调得到明显改善且PESQ降低较慢；而当σ_1从0.1增大至0.2时则相反。$\sigma_1=0.1$时(即方案SDNT)，算法在失调性能与语音质量间取得了较好的折中；因此可认为方案SDNT是最佳取值方案之一。实际应用时，可根据系统要求调节σ_i的值。

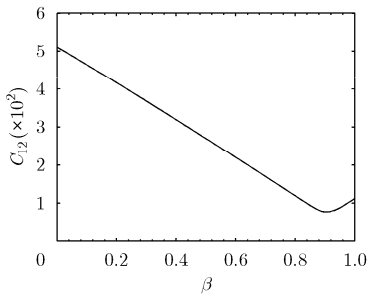


图2 互相关函数在高频段(>1 kHz)的平均值随非线性加权系数的变化曲线

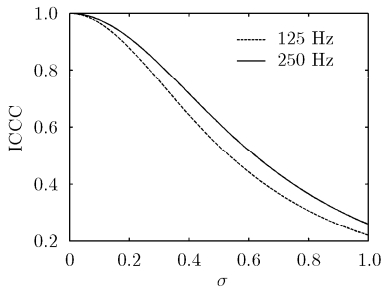


图3 ICCC随σ的变化曲线

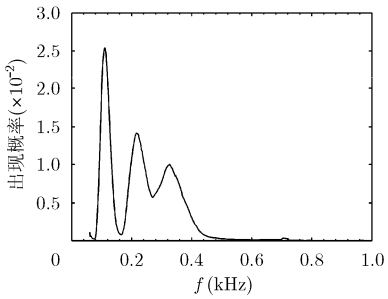
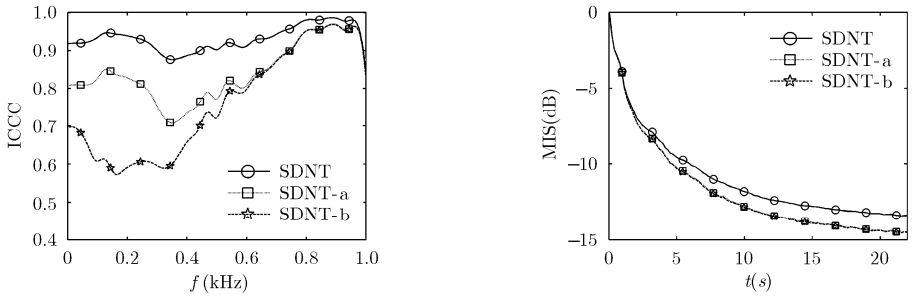


图4 不同频率值在3个低次谐频中的出现概率



(a)低频(<1 kHz)ICCC　　(b)全频带MIS

图 5 不同的幅值加权因子取值方案下的性能对比

表 1 不同 σ_1 值时的 MIS 曲线收敛值和左通道 PESQ

σ_1	0	0.03	0.06	0.10	0.20
MIS 收敛值(dB)	−9.3	−10.6	−12.2	−13.4	−14.4
PESQ	4.2	4.2	4.1	4.0	3.4

4 性能评价

考虑单向传输、远端房间中声源位置固定的 SAEC 场景。近端房间尺寸为 4×3×3 m^3，扬声器分别位于{1,2,1.2} m 和{3,2,1.2} m，传声器分别位于{1.8,1,1.2} m 和{2.2,1,1.2} m。采用信噪比为 30 dB 的高斯白噪声模拟近端房间内的背景噪声。测试用语音信号采用 TIMIT 数据库[22]中所有男声和女声信号。采样率取 16 kHz，房间混响时间设为 128 ms，利用虚源模型[23]得到扬声器至传声器的物理传递函数，长度为 1024 个采样点。自适应通道辨识时采用经典的的 NLMS 算法[2]，步长 0.4，自适应滤波器取 512 阶。

将 SDNT 法与以下方法进行性能比较：

(1)NLT(NonLinear Transformation)法：即采用半波整流变换的非线性变换法[3]。

(2)MFNP(Missing Fundamental-Nonlinear Phase modulation)法[11]：在低于 500 Hz 的低频段，利用心理声学上的基频遗失效应，在高频段则进行非线性相位调制。

(3)STR(Selective Time-Reversal)法[9]：对左通道信号中能量低于阈值的信号帧进行时间反转，右通道信号则保持不变。帧长取 256 点，能量阈值为 0.01。

NLT 法是一种非线性方法，SDNT 法的高频处理亦是，但二者采用不同的非线性变换：NLT 进行半波整流，而 SDNT 的高频处理利用非线性相位调制；另外，SDNT 在低、高频进行不同处理，而 NLT 则在全频带进行相同处理。MFNP 与 SDNT 均利用心理声学效应进行低频处理，但前者结合基频遗失效应，后者则基于谱优势效应；同时，二者采用不同的高频处理，前者直接进行相位调制，后者则利用非线性变换去相关法。STR 法与上述方法均不同，但由于其操作简便、技巧新颖，本文将其与 SDNT 法进行性能对比。

测试时，STR 法直接进行帧操作，NLT 法则逐点计算。其它方法均进行加权重叠求和，帧长 512 点，即 32 ms，重叠率 50%，所用窗函数 $w(n)$如式(18)[24]所示。

$$w(n)=\sin(0.5\cdot\pi\cdot\sin^2((n+0.5)\cdot\pi/L)) \quad (18)$$

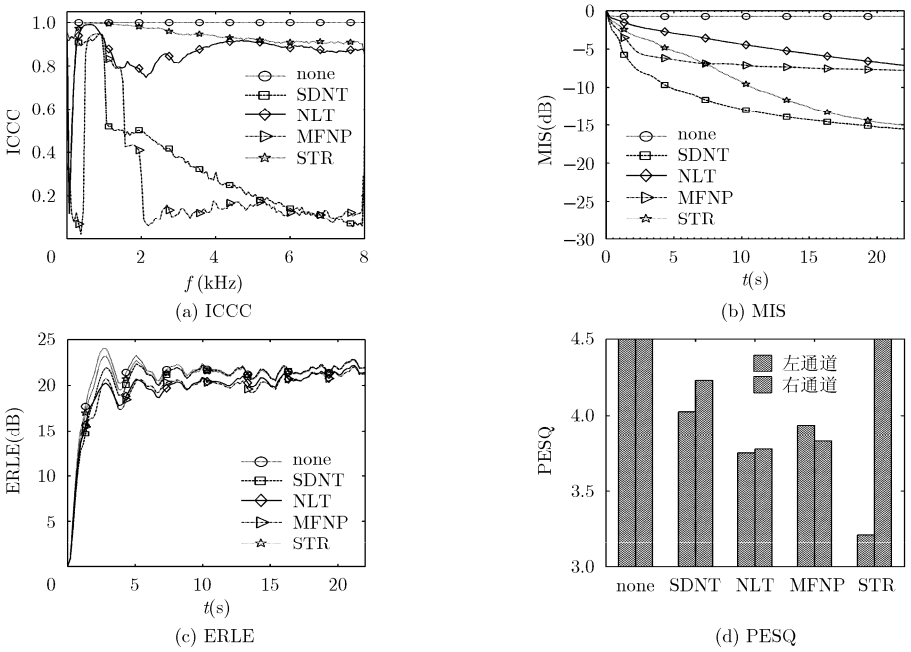
其中，L 等于帧长。根据第 3 节的结论，SDNT 法中的幅值加权因子取值为：$\sigma_1=0.10,\sigma_2=0.20$ 和 $\sigma_3=0.30$。

下面分别利用ICCC，失调，回声往返损耗增益 [9](Echo-Return-Loss Enhancement, ERLE) 和PESQ对各种SAEC方法的通道间相干性、失调、回声消除量和语音失真性能进行对比评价。对男、女声语音下的性能进行平均，得到的结果见图6，其中的“none”对应于未处理的原始信号。

4.1 通道间相干性

利用 ICCC 随频率的变化曲线检测各方法的去相关性能。计算 ICCC 时，对所有数据帧进行平均以估计数学期望，帧长 512 点，重叠率 80%。经前述各种方法处理后的立体声信号所对应的 ICCC 值如图 6(a)所示。

由于未考虑远端房间的染色效应，未处理时，两通道信号间的 ICCC 值接近 1。就去相关效果而言，STR 法最差，仅极低频和中高频相干系数略有减小；NLT 法次之，极低频与中频段 ICCC 值较低。MFNP 法在 500 Hz 以下利用基频遗失效应，从而获得强去相关效果；其在 500～2000 Hz 频段的 ICCC 值较高，高频段 ICCC 值再次降低。SDNT 法的高频去相关效果明显优于低频。由于进行分频段操作，SDNT 法与 MFNP 法的 ICCC 值分别在约 1 kHz 和 500 Hz 处发生突变，且 MFNP 法几乎在



(a) ICCC (b) MIS (c) ERLE (d) PESQ

图 6 4 种方法的性能对比

全频带具有最低的 ICCC 值。SDNT 法的去相关强度几乎与频率成正比；且由于谐频成分改变程度较低，其低频去相关效果差。

4.2 失调

仿真结果如图6(b)所示，4种方法的失调学习曲线与“none”曲线的对比验证了去相关预处理在SAEC中的必要性。SDNT法具有最快的收敛速度和最佳辨识精度，而NLT法性能最差。对照图6(a)易知，STR法虽然去相关效果最差，但其自适应性能优于NLT法，其收敛精度甚至高于MFNP法；MFNP法的去相关效果几乎最佳，但失调性能则不然。由此可见，不同频率处的通道间相干性对失调具有不同程度的影响。

4.3 回声消除量

好的去相关方法应该在改善失调性能的同时，不损伤系统的噪声消除性能。ERLE表征回声消除量的大小，它是评价回声消除算法性能的常用物理量，其定义为

$$\mathrm{ERLE}=10\lg\left(\mathrm{E}\left\{d^2(n)\right\}\Big/\mathrm{E}\left\{e^2(n)\right\}\right) \tag{19}$$

其中，$d(n)$为传声器信号；$e(n)$为经回声消除后的误差信号。ERLE 越大，回声消除越干净。仿真得到不同去相关方法下的 ERLE 曲线如图 6(c)所示。4 种去相关方法的回声消除性能差别不大，但 STR 法性能最佳，其 ERLE 曲线几乎与未处理时的重合。其次是 SDNT 法，其 ERLE 收敛略慢。NLT 法和 MFNP 法的 ERLE 性能相似，其收敛精度较低，且收敛速度较慢。

4.4 语音失真

图 6(d)给出了经过不同去相关方法处理后的语音信号的 PESQ 值。“none”方法对应于未处理，信号无失真，其 PESQ 值为理想值 4.5。4 种 SAEC 方法中，综合左右通道情况，SDNT 法的音质失真最小；且由于右通道的低频部分无处理，其所对应的 PESQ 值较左通道的高。STR 法仅对左通道信号处理，其右通道信号无失真，但左通道信号失真大。MFNP 法与 NLT 法同时对左右通道进行处理，在两通道中均存在较大失真，且前者失真较后者小；MFNP 法中，为了保持通道间同步，对右通道的低频部分进行时延补偿，但却破坏了低、高频间的信号同步，因此在图 6(d)中具有较小的右通道 PESQ 值。

5 结论

本文基于谱优势效应与改进的非线性变换法提出一种新型 SAEC 方法，并给出了一种最佳的幅值加权因子取值方案。在不同性能方面将本文方法与传统方法进行测试对比，结果表明其在去相关性、失调以及语音失真方面均具有较佳性能，能有效解决 SAEC 中的解失调问题。

参 考 文 献

[1] Malik S and Benesty J. Variationally diagonalized multichannel state-space frequency-domain adaptive filtering for acoustic echo cancellation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, 2013: 595–599.

[2] Nishiyama K. A unified view of adaptive algorithms for finite impulse response filters using the H∞ framework[J]. *Signal Processing*, 2014, 97(2014): 55–63.

[3] Benesty J, Morgan D R, and Sondhi M M. A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation[J]. *IEEE Transactions on Speech and Audio Processing*, 1998, 6(2): 156–165.

[4] Morgan D R, Hall J L, and Benesty J. Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation[J]. *IEEE Transactions on Speech and Audio Processing*, 2001, 9(6): 686–696.

[5] Romoli L, Cecchi S, and Piazza F. A combined approach for channel decorrelation in stereo acoustic echo cancellation exploiting time-varying frequency shifting[J]. *IEEE Signal Processing Letters*, 2013, 20(7): 717–720.

[6] Romoli L, Cecchi S, and Piazza F. Evaluation of a channel decorrelation approach for stereo acoustic echo cancellation [C]. 8th International Symposium on Image and Signal Processing and Analysis, Trieste, 2013: 783–788.

[7] Ali M. Stereophonic acoustic echo cancellation system using time varying all-pass filtering for signal decorrelation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seattle, 1998: 3689–3692.

[8] Herre J, Buchner H, and Kellermann W. Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Honolulu, 2007: 17–20.

[9] Nguyen D Q, Gan W S, and Khong A W H. Time-reversal approach to the stereophonic acoustic echo cancellation problem[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, 19(2): 385–395.

[10] Valin J M. Perceptually-motivated nonlinear channel decorrelation for stereo acoustic echo cancellation[C]. Hands-Free Speech Communication and Microphone Arrays, Trento, 2008: 188–191.

[11] Cecchi S, Romoli L, Peretti P, *et al.*. A combined psychoacoustic approach for stereo acoustic echo cancellation[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, 19(6): 1530–1539.

[12] Romoli L, Cecchi S, Peretti P, *et al.*. A mixed decorrelation approach for stereo acoustic echo cancellation based on the estimation of the fundamental frequency[J]. *IEEE Transactions on Audio, Speech and Language Processing*, 2012, 20(2): 690–698.

[13] Schwartz D A and Purves D. Pitch is determined by naturally occuring periodic sounds[J]. *Hearing Research*, 2004, 194(1): 31–46.

[14] Plomp R. Pitch of complex tones[J]. *The Journal of the Acoustical Society of America*, 1967, 41(6): 1526–1533.

[15] Shimamura T and Kobayashi H. Weighted autocorrelation for pitch extraction of noise speech[J]. *IEEE Transactions on Speech and Audio Processing*, 2001, 9(7): 727–730.

[16] 韩纪庆, 张磊, 郑铁然. 语音信号处理[M]. 北京: 清华大学出版社, 2013: 86–91.
Han Ji-qing, Zhang Lei, and Zheng Tie-ran. Speech Signal Processing[M]. Beijing: Tsinghua University Press, 2013: 86–91.

[17] Hu X, Wang S, Zheng C, *et al.*. A cepstrum-based preprocessing and postprocessing for speech enhancement in adverse environments[J]. *Applied Acoustics*, 2013, 74(12): 1458–1462.

[18] Zwicker E and Fastl H. Psychoacoustics: Facts and Models[M]. New York: Springer-Verlag, 1990: 185–187.

[19] Stanciu C, Benesty J, Paleologu C, *et al.*. A widely linear model for stereophonic acoustic echo cancellation[J]. *Signal Processing*, 2013, 93(2): 511–516.

[20] Hines A, Skoglund J, Kokaram A, *et al.*. Robustness of speech quality metrics to background noise and network degradations: Comparing ViSQOL, PESQ and POLQA[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, 2013: 3697–3701.

[21] Loizou P C. Speech Enhancement: Theory and Practice[M]. Boca Raton, FL: CRC Press, 2007: 514–524.

[22] Garofolo J S. DAPRA TIMIT acoustic-phonetic speech database [OL]. http://web.mit.edu/course/6/6.863/share/nltk_lite/timit/, 1990.10.

[23] Habets E A P. Room impulse response generator [OL]. http://home.tiscali.nl/ehabets/rir_generator.html, 2014.08.

[24] Montgomery C. Vorbis I specification[OL]. http://www.xiph.org/vorbis/doc/Vorbis_I_spec.html, 2013.08.

杨鹤飞： 女，1988 年生，博士生，研究方向为音频信号处理.

郑成诗： 男，1980 年生，博士，副研究员，主要研究方向为语音信号处理、音频信号处理.

李晓东： 男，1966 年生，博士，研究员，主要研究方向为语音信号处理、音频信号处理以及非线性声学.