基于近似 KLT 域的语音信号压缩感知

郭海燕　　杨　震
(南京邮电大学信号处理与传输研究院　南京　210003)

摘　要：压缩感知是近年来兴起的研究热点，该文基于语音信号在 KLT 域的稀疏特性，提出了基于模板匹配的近似 KLT，并在基于模板匹配近似 KLT 域上研究了语音信号的压缩感知性能。首先验证语音信号在基于模板匹配近似 KLT 域上的稀疏性，然后由语音信号与观测矩阵构造相应的观测，采取固定分配每帧观测个数和按帧能量自适应分配每帧观测个数两种方案，再以观测为已知条件利用 L1 优化算法重构语音信号在基于模板匹配近似 KLT 域的稀疏系数向量，进而重构原始语音信号。实验表明，语音信号在基于模板匹配的近似 KLT 域的压缩感知性能较好。

关键词：语音合成；压缩感知；稀疏性；L1 优化；Karhunen-Loeve 变换(KLT)
中图分类号：TN912.3　　**文献标识码**：A　　**文章编号**：1009-5896(2009)12-2948-05

Compressed Speech Signal Sensing Based on Approximate KLT

Guo Hai-yan　　Yang Zhen
(*Institute of Signal Processing and Transmission*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003, *China*)

Abstract: Compressed Sensing is a research focus rising in recent years. On the basis of the signal's sparse representation in the KLT domain, this paper proposes an approximate KLT method using template matching and studies on the corresponding compressed speech signal sensing. First, it verifies the sparsity of speech signal in the approximate KLT domain. Second, by speech signal and a measurement matrix, it arranges measurements of fixed or adaptive length according to frame energy. Third, according to the measurements, it finds the speech signal's sparsest coefficient vector through L1 optimization algorithm to recover the speech signal. Simulation results demonstrate that compressed speech signal sensing in the approximate KLT using template matching has good performance.

Key words: Speech synthesis; Compressed Sensing (CS); Sparsity; L1 optimization; Karhunen-Loeve Transform (KLT)

1　引言

压缩感知[1,2] (Compressed Sensing, CS)是近年来兴起的一种“边采样边压缩”技术，从某种意义上讲突破了奈奎斯特采样定律(Nyquist sampling theorem)的限制。奈奎斯特采样定律认为要使信号采样后能够不失真还原，采样频率必须大于或等于整个信号最高频率的两倍。CS 理论认为，如果信号在某一变换域上为 K-稀疏(K-sparse)的，我们就可以用此信号在某投影域上的观测向量(即投影，以下简称观测)来近似无损地重构原信号。可以看出，利用 CS 理论重构信号所需的观测个数与与信号的最高频率无关，与信号的稀疏性密切相关。另外 CS 理论的应用，并不需要分析信号在变换域的严格特性，只需验证信号在该变换域的稀疏性即可，这无疑是很有吸引力的。

自 Candes 等[3] 和 Donoho[1] 提出 CS 的相关理论以来，CS 在信息论与编码，信号恢复，有损压缩，机器学习，传感器网络等很多领域得到了广泛的研究，在图像信号处理也有广泛的应用。例如图像处理中，由于多数场合高频分量相对低频分量而言很小，图像信号时频系数可看作是稀疏的，利用 CS 技术来采集图像可以大大降低数码率[4,5]。目前针对语音信号的 CS 研究还不多，尚属于起步阶段。Gemmeke 和 Cranen 利用 CS 原理对噪声环境下的语音进行识别，实验结果证明识别系统的抗噪性能大大提高[6]，显示了将 CS 技术与语音处理技术结合的巨大前景。再者如果利用 CS 技术进行采样，数据量必然大大减少，在此基础上再研究后续的有损编码技术，这样就可能实现更低码率上的高质量

2008-12-15 收到，2009-05-18 改回
国家 863 计划重点项目(2006AA010102)国家自然科学基金(60971129)和江苏省普通高校研究生科研创新计划项目(CX0913-148Z)资助课题

语音编码。可以预测，将 CS 与语音信号处理结合，在研究语音信号 CS 性能的基础上再对观测用新的模型建模，那么语音压缩、语音识别、语音合成和语音增强等领域的现有理论和技术，都将会发生大的变革，因此对语音信号 CS 性能的基础研究具有重要的理论意义和实用价值。

本文针对语音信号在 KLT 域的稀疏特性，对近似 KLT 域的语音信号 CS 进行了研究。首先，提出了一种基于模板匹配的近似 KLT，然后，在这个近似 KLT 域对语音信号的稀疏性进行了实验上的验证，再对语音信号的 CS 性能进行了研究。另外，语音信号清音帧，浊音帧所包含的信息量不尽相同，没必要对每帧语音取相同个数的观测，本文根据每帧的能量自适应分配每帧的观测个数，并将自适应分配与固定分配观测个数时的 CS 性能进行了比较。

2　压缩感知

CS 理论认为如果信号在某已知变换域具有稀疏性，则通过原信号在某投影域的投影可以近似无损地重构原信号，要求投影域的基与已知变换域的基不相干[7]。表述如下。

已知离散信号 $\boldsymbol{x}=[x(1),x(2),\cdots,x(N)]^{\mathrm{T}}$ 的采样个数为 N，变换矩阵 $\boldsymbol{\Psi}=[\boldsymbol{\phi}_1,\boldsymbol{\phi}_2,\cdots,\boldsymbol{\phi}_N]$ 的列向量 $\{\boldsymbol{\phi}_i\}_{i=1}^{N}$ 正交，其中 $\boldsymbol{\phi}_i$ ($i=1,2,\cdots,N$) 为 $N\times1$ 的向量，则信号 $\boldsymbol{x}$ 可以表示为

$$\boldsymbol{x}=\boldsymbol{\Psi\Theta}=\sum_{i=1}^{N}\theta_i\boldsymbol{\phi}_i \tag{1}$$

其中 $\boldsymbol{\Theta}$ 为原信号在变换域的系数向量，当 $\boldsymbol{\Theta}$ 满足

$$||\boldsymbol{\Theta}||_0=K \tag{2}$$

时，信号 $\boldsymbol{x}$ 被称为 K-稀疏的，其中 $||\boldsymbol{\Theta}||_0$ 表示向量 $\boldsymbol{\Theta}$ 中非零元素的个数，此时可以用个数为 $M=cK$ (通常 c 取 3–4)[8] 的观测 $\boldsymbol{y}=[y(1),y(2),\cdots,y(M)]^{\mathrm{T}}$ 来近似无损地重构原信号 $\boldsymbol{x}$，通常 $M<N$。下面详细介绍观测的构造和原信号的重构。

已知 $\boldsymbol{\Phi}$ 为满足受限等距映射特性(Restricted Isometry Property， RIP)和不相关特性[6]的观测矩阵，由观测矩阵 $\boldsymbol{\Phi}$ 与原信号 $\boldsymbol{x}$ 可以得到个数为 M 的观测。

$$\boldsymbol{y}=\boldsymbol{\Phi x}=\boldsymbol{\Phi\Psi}^{\mathrm{T}}\boldsymbol{\Theta} \tag{3}$$

本文中取观测矩阵 $\boldsymbol{\Phi}$ 为满足高斯分布的随机矩阵[7]。

当 $M<N$ 时，式(3)无唯一解，本文用 L_0 优化算法重构原信号稀疏域的系数[1]。

$$(L_0)\ \boldsymbol{\Theta}=\min||\boldsymbol{\Theta}||_0,\ \text{s.t. } \boldsymbol{y}=\boldsymbol{\Phi x}=\boldsymbol{\Phi\Psi}^{\mathrm{T}}\boldsymbol{\Theta} \tag{4}$$

由于式(4)的求解需要列出所有满足限制条件的 $\boldsymbol{\Theta}$ 备选值，再从这些备选值中找出具有最少个非零元素的 $\boldsymbol{\Theta}$，复杂度很高且难以实现。Donoho 的研究表明，当 $\boldsymbol{\Theta}$ 满足一定条件时，L_1 优化算法与 L_0 优化算法同解[1]。所以我们通常转化为 L_1 优化问题来求解[1]。

$$(L_1)\ \boldsymbol{\Theta}=\min||\boldsymbol{\Theta}||_1,\ \text{s.t. } \boldsymbol{y}=\boldsymbol{\Phi x}=\boldsymbol{\Phi\Psi}^{\mathrm{T}}\boldsymbol{\Theta} \tag{5}$$

式(5)可以看作是式(4)的凸化，通过线性规划(Linear Programming, LP)算法可以方便地实现[9]。

3　基于近似 KLT 域的语音信号压缩感知

3.1　语音信号在 KLT 域的稀疏性

对语音信号 $\boldsymbol{x}=[x(1),x(2),\cdots,x(N)]^{\mathrm{T}}$ 进行 KLT，得到信号 $\boldsymbol{x}$ 在 KLT 域的系数 $\boldsymbol{\Theta}$：

$$\boldsymbol{\Theta}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{x} \tag{6}$$

其中 KLT 矩阵 $\boldsymbol{U}^{\mathrm{T}}$ 由 $\boldsymbol{x}$ 的自相关函数 $\boldsymbol{R}_x$ 进行特征值分解得到

$$\boldsymbol{R}_x=\boldsymbol{xx}^{\mathrm{T}}=\boldsymbol{U\Lambda}_x\boldsymbol{U}^{\mathrm{T}} \tag{7}$$

我们对采样率为 16 kHz，长度为 320 样点的清音帧信号和浊音帧信号进行 KLT 分解，发现语音信号在 KLT 域具有显著的稀疏性，如图 1 所示。

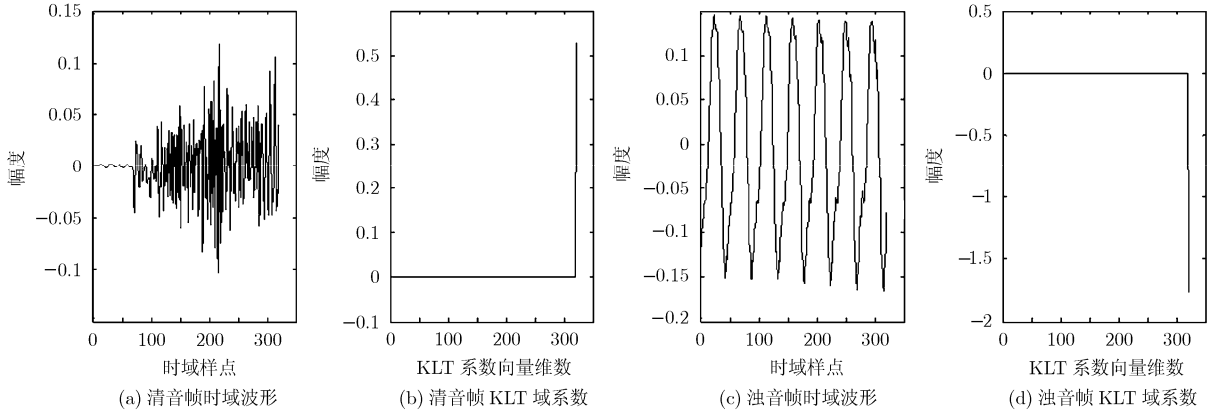


图 1 语音信号时域波形图及在 KLT 域的系数

尽管语音信号在 KLT 域具有惊人的稀疏性，但由于对每个信号 $\boldsymbol{x}$ 都要求对应的 KLT 矩阵，故基于 KLT 的语音信号 CS 方案实用性受到限制，下面我们寻求基于近似 KLT 的语音信号 CS 方案。

3.2 基于模板匹配近似 KLT 的语音信号压缩感知

为了避免对每个语音信号 $\boldsymbol{x}$ 分别求取相应的 KLT 分解矩阵，我们取若干组原说话人的语音进行训练，达到由训练语音来求取测试语音近似 KLT 分解矩阵的目的。由于 KLT 矩阵是由信号的自相关矩阵进行特征值分解得到的，所以对每段训练语音求取自相关矩阵，用与测试语音自相关矩阵最匹配的若干训练语音自相关矩阵的非齐次线性组合作为测试语音自相关矩阵的近似。为了减少存储量和计算量，将训练语音及其自相关矩阵对角元素构成的行向量作为模板而不是将训练语音及其自相关矩阵直接作为模板。下文称由训练语音组成的模板为模板 1，由自相关矩阵对角元素构成的行向量组成的模板为模板 2。

本文采用非齐次线性均方估计的方法[10]，用若干模板 2 元素的非齐次线性组合 $\boldsymbol{DR}_{x\text{train}}$ 来近似匹配测试语音自相关矩阵的对角元素构成的行向量 $\boldsymbol{DR}_{x\text{test}}$。

$$\boldsymbol{D}\widehat{\boldsymbol{R}}_{x\text{test}} = a_0 + a_1\boldsymbol{D}\widehat{\boldsymbol{R}}_{x\text{train}(i_1)} + a_2\boldsymbol{D}\widehat{\boldsymbol{R}}_{x\text{train}(i_2)}\cdots + a_{\text{num}}\boldsymbol{D}\widehat{\boldsymbol{R}}_{x\text{train}(i_{\text{num}})} \tag{8}$$

其中 $a_0, a_1, a_2, \cdots, a_{\text{num}}$ 为使均方误差

$$P = (\boldsymbol{DR}_{x\text{test}} - \boldsymbol{D}\widehat{\boldsymbol{R}}_{x\text{test}})(\boldsymbol{DR}_{x\text{test}} - \boldsymbol{D}\widehat{\boldsymbol{R}}_{x\text{test}})^{\text{T}} \tag{9}$$

最小的 num +1 个常数。式(8)，式(9)中 $\boldsymbol{DR}_{x\text{train}(i_1)}$，$\boldsymbol{DR}_{x\text{train}(i_2)}, \cdots, \boldsymbol{DR}_{x\text{train}(i_{\text{num}})}$ 为选取的与测试语音自相关矩阵的对角元素 l_∞-范数相差最小的 num 个模板 2 元素。由正交性原理[10]求取系数 $a_0, a_1, a_2, \cdots, a_{\text{num}}$（记 $\boldsymbol{A} = [a_0, a_1, \cdots, a_{\text{num}}]$）。

$$\boldsymbol{A} = [R_{01}, R_{02}, \cdots, R_{0\text{num}}] \begin{bmatrix} R_{i_0 i_0} & R_{i_0 i_1} & \cdots & R_{i_0 i_{\text{num}}} \\ R_{i_1 i_0} & R_{i_1 i_1} & \cdots & R_{i_1 i_{\text{num}}} \\ \vdots & \vdots & \vdots & \vdots \\ R_{i_{\text{num}} i_0} & R_{i_{\text{num}} i_1} & \cdots & R_{i_{\text{num}} i_{\text{num}}} \end{bmatrix}^{-1} \tag{10}$$

$$R_{0j} = \boldsymbol{DR}_{x\text{test}} \times \boldsymbol{DR}^{\text{T}}{}_{x\text{train}(i_j)} \tag{11}$$

$$R_{i_{j1} i_{j2}} = \boldsymbol{DR}_{x\text{train}(i_{j1})} \times \boldsymbol{DR}^{\text{T}}_{x\text{train}(i_{j2})} \tag{12}$$

其中 $\boldsymbol{DR}_{x\text{train}(i_0)}$ 为长度为 L 的行向量，各元素恒为 1。

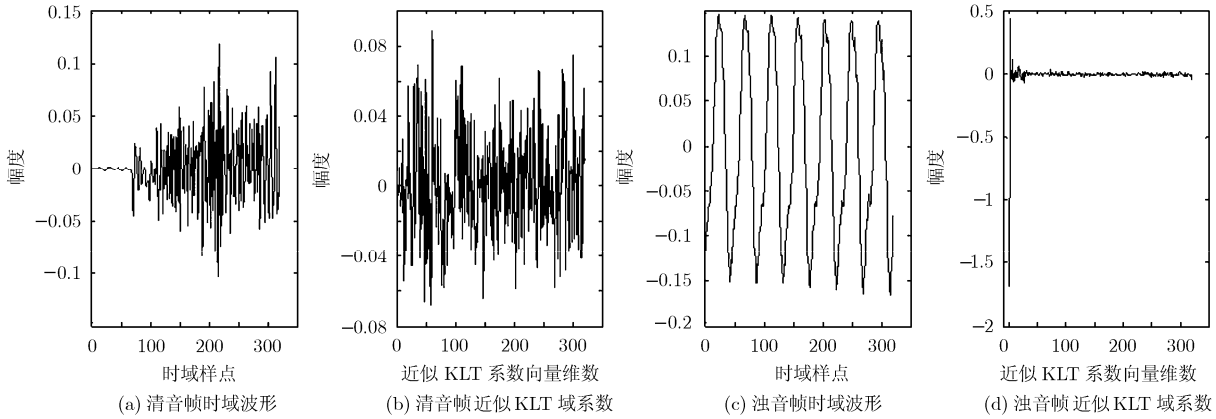
由模板 2 中 num 个元素所对应的 num 个模板 1 元素的自相关矩阵的非齐次线性组合得到测试语音的近似 KLT 矩阵 $\widehat{\boldsymbol{U}}^{\text{T}}$ 和近似 IKLT 矩阵 $\widehat{\boldsymbol{U}}$，

$$\widehat{\boldsymbol{R}}_{x\text{test}} = a_0 + a_1\boldsymbol{x}_{\text{train}(i_1)}\boldsymbol{x}^{\text{T}}_{\text{train}(i_1)} + a_1\boldsymbol{x}_{\text{train}(i_1)}\boldsymbol{x}^{\text{T}}_{\text{train}(i_1)} + \ldots + a_{\text{num}}\boldsymbol{x}_{\text{train}(i_{\text{num}})}\boldsymbol{x}^{\text{T}}_{\text{train}(i_{\text{num}})} \tag{13}$$

$$\widehat{\boldsymbol{R}}_{x\text{test}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{\Lambda}}_{x\text{test}}\widehat{\boldsymbol{U}}^{\text{T}} \tag{14}$$

本文用这种近似 KLT 矩阵 $\widehat{\boldsymbol{U}}^{\text{T}}$ 对语音信号进行分解，发现在浊音段仍然具有较强的稀疏性，而在清音段不具有稀疏性，这是由于清音帧样点序列变化快，用有限的训练码本的线性组合不能很好地匹配测试清音的细节信息，如图 2 所示。图 2 中每帧语音为 320 样点，采样率为 16 kHZ，num 值取为 32。由于语音信号中清音能量通常较小，语音的大部分信息都集中在浊音成分，故清音帧的非稀疏性不太影响近似 KLT 域的语音信号 CS 性能。

根据语音信号在基于模板匹配近似 KLT 域的稀疏性，本文在基于模板匹配的近似 KLT 域对语音信号的 CS 性能进行研究。在上述的基于模板匹配近似 KLT 域，分别对每帧语音信号采用 CS 方法重构。另外，为了用较少的观测更有效地重构语音信号，可以考虑自适应分配每帧观测个数，能量大的帧分配较多的观测，能量小的帧分配较少的观测。



(a) 清音帧时域波形 (b) 清音帧近似KLT 域系数 (c) 浊音帧时域波形 (d) 浊音帧近似KLT 域系数

图 2 语音信号时域波形图及在基于模板匹配的近似 KLT 域的系数

$$M_{i\mathrm{ad}} = \frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}{E(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})} M_s \tag{15}$$

其中 M_s 为固定分配的每帧观测个数，$M_{i\mathrm{ad}}$ 为自适应分配的第 i 帧观测个数。

Donoho 指出，当观测个数较少时，利用 CS 方法重构原信号会得到带噪的原始信号，并指出利用平移不变降噪的方法可以有效地去除重构信号中的噪声[2]。本文采用文献[11]提供的小波域平移不变降噪程序对 CS 重构信号进行降噪。首先对 CS 重构信号在 Haar 小波域进行平移不变小波变换(通过函数 FWT_TI 实现)，然后通过软阈值法(soft thresholding)对变换后的信号降噪(通过函数 SoftThresh 实现)，最后对降噪后的信号进行反平移不变小波变换得到降噪后的重构信号(通过函数 IWT_TI 实现)。具体程序见文献[11]。由文献[2]可以看出，采用小波域的平移不变降噪方法对 CS 重构信号进行降噪后，降噪后的信号幅度要比原始信号幅度低，故客观实验评价对象仍为降噪前的 CS 重构信号，主观实验评价对象为降噪后的 CS 重构信号。

4　实验结果及分析

本文定义压缩率 r 和重构信号 SegSNR 分别为

$$r = M/N \tag{16}$$

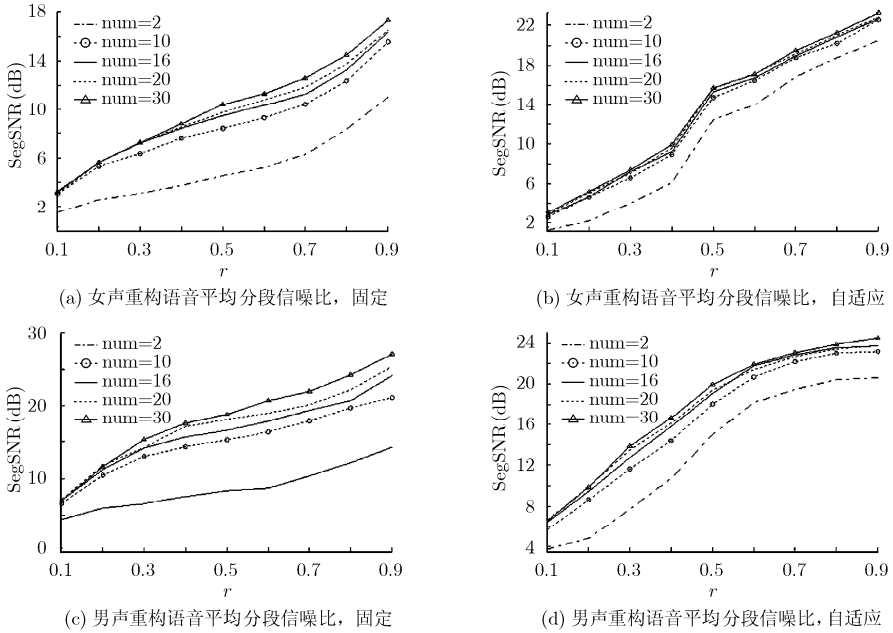
$$\mathrm{SegSNR}=\frac{1}{\mathrm{Nframe}}\sum_{i=1}^{\mathrm{Nframe}} 10\times\lg\left(\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i}{(\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i)^{\mathrm{T}}(\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i)}\right) \tag{17}$$

其中 Nframe 为原信号的总帧数。实验着重研究 r 与重构信号 SegSNR 的关系。实验环境为安静环境，实验对象为 4 位说话人的语音，男性两位，女性两位，采样率为 16 kHz。

本文对每个说话人各取 250 句短语结构的语音作为训练语音，50 句与训练语音不同的短语结构的语音作为测试语音。语音帧长 160 样点，帧间重叠 40 样点。由于 num 不同，对测试语音自相关矩阵对角元素向量的细节匹配性能会有差异，所以我们对不同 num 取值下 r 与重构信号 SegSNR 的关系进行了对比，如图 3 所示。图 3 中的“固定”表示固定分配每帧观测个数，“自适应”表示自适应分配每帧观测个数。

由图 3 可以看出，num 越大，相同 r 时重构信号 SegSNR 越大，这是因为 num 越大，测试语音自相关矩阵对角元素向量匹配的越准。但是 num 大到一定程度时效果的改善不再明显，在实验中取 num = 16，并对自适应分配和固定分配每帧观测个数情况下的 CS 性能进行比较。从图 3 可以看出，对于女声语音，自适应分配每帧观测个数下的性能要优于固定分配每帧观测个数下的性能，而男声语音正好相反，这是因为女声在基于模板匹配的近似 KLT 域的稀疏性比男声差且不稳定，自适应分配每帧观测个数利于能量较大帧的更好恢复从而改善整体性能，而男声语音在这个域上稀疏性好且稳定，固定分配每帧观测个数能兼顾到稀疏性不太好的帧从而有利于提高整体性能。

最后选用 P.862 标准对重构男女声语音做出主观评价，见表 1。从表 1 中可以看出，当 r =0.5 时，女声重构语音信号平均意见得分(Mean Opinion Score, MOS)分达到了 3.3，男声重构语音信号 MOS 分达到了 3.71。



(a) 女声重构语音平均分段信噪比，固定　(b) 女声重构语音平均分段信噪比，自适应

(c) 男声重构语音平均分段信噪比，固定　(d) 男声重构语音平均分段信噪比，自适应

图 3 不同 num 下 r 与重构信号 SegSNR 的关系

表 1 五级评分标准下重构语音的 MOS 分

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
固定，降噪，女声	2.48	2.88	3.15	3.28	3.21	3.44	3.45	3.58	3.90
自适应，降噪，女声	2.65	2.88	3.31	3.38	3.30	3.45	3.31	3.80	3.22
固定，降噪，男声	3.49	3.60	3.66	3.89	3.71	3.78	3.90	3.84	3.99
自适应，降噪，男声	3.05	3.21	3.28	3.37	3.43	3.38	3.49	3.48	3.63

5 结束语

本文提出了基于模板匹配的近似 KLT，在基于模板匹配的近似 KLT 域研究了语音信号的压缩感知性能，本文还针对语音信号各帧包含的信息不尽相同的特点，提出对各帧进行自适应分配观测个数进行压缩感知，并将自适应分配和固定分配每帧观测个数方案下的语音信号压缩感知性能进行了比较。实验证明语音信号在基于模板匹配的近似 KLT 域上是近似稀疏的，具有较好的压缩感知效果。但是由于语音信号在基于模板匹配的近似 KLT 域上并非绝对稀疏的，使压缩感知的优点未完全体现出来，今后将在用基于训练的方法寻找性能更接近于 KLT 的近似 KLT 等方面进行更深入的研究。

参 考 文 献

[1] Donoho D. Compressed sensing. *IEEE Transactions on Information Theory*, 2006, 52(4): 1289–1306.

[2] Tsaig Y and Donoho D. *Extensions of compressed sensing. Signal Processing*, 2006, 86(3): 533–548.

[3] Candès E, Romberg J, and Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 2006, 52(2): 489–509.

[4] Romberg J. Imaging via compressive sampling. *IEEE Signal Processing Magazine*, 2008, 25(2): 14–20.

[5] Duarte M, Davenport M, Takhar D, Laska J, Sun T, Kelly K, and Baraniuk R. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 2008, 25(2): 83–91.

[6] Gemmeke J F and Cranen B. Using sparse representations for missing data imputation in noise robust speech recognition. European Signal Processing Conf. (EUSIPCO), Lausanne, Switzerland, August 2008: 987–991.

[7] Baraniuk R G. Compressive sensing. *IEEE Signal Processing Magazine*, 2007, 24(4): 118–121.

[8] Baron D, Wakin M, Duarte M, Sarvotham S, and Baraniuk R. Distributed compressed sensing. Technical Report ECE-0612, Electrical and Computer Engineering Department, Rice University, December 2006.

[9] Chen S S, Donoho D L, and Saunders M A. Atomic decomposition by basis pursuit. *SIAM Review*, 2001, 43(1): 129–159.

[10] A·帕普里斯, S·U·佩莱著, 保铮, 冯大政, 水鹏朗译. 概率、随机变量与随机过程. 第 4 版, 西安: 西安交通大学出版社, 2004: 209–213.

[11] Coifman R R and Donoho D L. Translation-Invariant De-noising. New York: Springer-Verlag, 1995: 125–150.

郭海燕： 女，1983 年生，博士生，从事语音处理和盲源分离研究工作.

杨 震： 男，1961 年生，教授，博士生导师，从事无线通信与网络信号处理、语音处理与现代语音通信、信息安全技术研究工作.