第48卷第4期
2008年7月

大连理工大学学报
Journal of Dalian University of Technology

Vol. 48, No. 4
July 2008

电子与信息工程

文章编号：1000-8608(2008)04-596-06

一种基于粒子滤波的双模态语音提取方法

金乃高*，殷福亮

（大连理工大学 电子与信息工程学院，辽宁 大连 116024）

摘要：说话人的唇动信息有助于加强对语音的感知.根据说话人语音的双模态特性，将振动信息引入语音提取问题，提出了一种基于粒子滤波的贝叶斯融合架构的双模态语音提取方法.该方法融合说话人的语音和唇动信息，根据信息论中的最大互信息准则与盲源分离中的高阶统计量准则，将音视频互信息与语音峭度的乘积作为代价函数，利用粒子滤波估计混合矩阵，解决时变瞬时混合情况下的语音提取问题.仿真结果表明，该方法在低信噪比情况下仍然能够实现语音信号的有效提取.

关键词：语音提取；粒子滤波；高阶统计量；最大互信息

中图分类号：TN911.7　　**文献标志码**：A

0 引　言

基于麦克风阵列的语音信号提取是从多路混合语音中提取出一路感兴趣的源语音信号，其在复杂环境下的语音识别、高质量语音通信以及人机接口等领域具有广泛的应用前景.例如，在视频会议中经常出现多人同时说话的情形，这便需要系统从混合语音中提取出指定说话人的语音信号，经增强处理后再进行编码传输.现有的语音提取方法主要有波束形成方法[1]和盲信号提取方法[2]，分别根据声源的方向信息或语音源信号间的统计独立性进行语音提取.这两种语音提取方法都有一定的适用条件，研究如何提高实际环境中语音提取系统的性能是一项具有挑战性的工作.

在嘈杂的背景噪声或其他说话人干扰情况下，唇动等可视语音信息有助于增强听觉系统分离及提取感兴趣语音信号的能力，这是大脑对视听感知信息进行融合处理的结果.双模态语音处理方法[3]融合说话人的音频与视频信息，利用二者之间的相关性来提高系统的性能，已成功应用于复杂环境下的语音识别系统中[4].针对语音提取问题的双模态处理方法研究也取得了一些成果.Bub 等利用摄像机获取的视频信息确定说话人的位置，进而引导麦克风阵列波束形成的指向，以提取说话人语音[5].Sodoyer 等利用说话人发音过程中语音与唇动信息间的相关性，解决了盲源分离中存在的输出顺序不确定问题，同时也改善了语音提取系统的抗噪能力[6].Rajaram 等将卡尔曼滤波应用于双模态语音分离问题，在低信噪比下取得了较好的分离效果[7].

本文将音视频联合信号处理方法应用于说话人运动情况下的语音提取问题，在语音提取过程中融入说话人的唇动信息，采用粒子滤波实现语音信号的有序提取，以提高低信噪比下语音提取的质量.

1 粒子滤波

近年来，粒子滤波已经成为研究非线性、非高斯动态系统最优估计问题的有效方法[8].粒子滤波将贝叶斯理论与蒙特卡罗(Monte Carlo)方法相结合，使用非参数化的序贯蒙特卡罗方法实现递推贝叶斯滤波.贝叶斯滤波根据观测数据 $y_{1:k}$ 递推估计系统状态 x_k 的后验概率密度 $p(x_{1:k} \mid y_{1:k})$ 与滤波概率密度 $p(x_k \mid y_{1:k})$.

已知 $k-1$ 时刻的滤波概率密度为 $p(x_{k-1} \mid$

收稿日期：2006-10-25；　**修回日期**：2008-05-30.
基金项目：国家自然科学基金资助项目(60372082,60172073).
作者简介：金乃高*(1977-)，男，博士生；殷福亮(1962-)，男，教授，博士生导师.

$y_{1:k-1}$)，根据 Chapman-Kolmogorov 积分方程进行时间更新，则 k 时刻的预测概率密度

$$p(x_k \mid y_{1:k-1}) = \int [p(x_k \mid x_{k-1}) \cdot p(x_{k-1} \mid y_{1:k-1})] \mathrm{d}x_{k-1} \quad (1)$$

当获得最新测量值 y_k 后，通过贝叶斯公式进行量测更新，k 时刻滤波概率密度

$$p(x_k \mid y_{1:k}) = \frac{p(y_k \mid x_k) p(x_k \mid y_{1:k-1})}{\int p(y_k \mid x_k) p(x_k \mid y_{1:k-1}) \mathrm{d}x_k} \quad (2)$$

粒子滤波的核心思想是利用一系列随机样本的加权和来表示后验概率密度或滤波概率密度. 假设可从滤波概率密度函数 $p(x_k \mid y_{1:k})$ 中抽取 N 个独立同分布的随机样本 $x_k^{(i)}(i = 1, \cdots, N)$，任意函数 $g(x_k)$ 的最小均方误差估计可用下式来逼近，即 $\hat{x}_k = E[g(x_k) \mid y_{1:k}] =$

$$\int g(x_k) p(x_k \mid y_{1:k}) \mathrm{d}x_k = \frac{1}{N} \sum_{i=1}^{N} g(x_k^{(i)}) \quad (3)$$

粒子滤波算法采用重要性抽样方法，通过引入一个易于采样的重要性概率密度 $\pi(x_k \mid y_{1:k})$ 来加权逼近滤波概率密度 $p(x_k \mid y_{1:k})$. 令 $\{x_k^{(i)}, w_k^{(i)}, i = 1, \cdots, N\}$ 表示从重要性概率密度 $\pi(x_k \mid y_{1:k})$ 中抽样获取的支撑点集，其中 $w_k^{(i)}$ 为第 i 个粒子 $x_k^{(i)}$ 的权值，则滤波概率密度

$$p(x_k \mid y_{1:k}) = \sum_{i=1}^{N} w_k^{(i)} \delta(x_k - x_k^{(i)}) \quad (4)$$

其中 $\delta(\cdot)$ 是单位冲激函数，权值 $w_k^{(i)}$ 的计算公式为

$$w_k^{(i)} \propto p(x_k^{(i)} \mid y_{1:k}) / \pi(x_k^{(i)} \mid y_{1:k}) \quad (5)$$

粒子滤波采用序贯分析与重要性抽样相结合的序贯重要性抽样算法，从重要性概率密度函数 $\pi(x_k \mid x_{1:k-1}, y_{1:k})$ 中获取支撑点集，并随着测量值的依次到来序贯计算粒子的权值，即

$$w_k \propto w_{k-1} \frac{p(y_k \mid x_k) p(x_k \mid x_{k-1})}{\pi(x_k \mid x_{1:k-1}, y_{1:k})} \quad (6)$$

序贯重要性抽样算法通常存在退化问题，可采用重采样技术减小退化现象带来的不利影响. 对于重采样带来的粒子耗尽问题，可以采用粒子正则重采样方法或增加马尔可夫链蒙特卡罗移动步骤加以解决.

2　基于粒子滤波的双模态语音提取方法

2.1　双模态语音提取问题的描述

在视频会议场景中，通常存在背景噪声及多个说话人的交叉干扰，增加了语音提取系统的设计难度. 摄像机获取的人脸视频图像，可为语音提取提供有用信息. 双模态语音提取问题便是从多路混合语音中，提取出与指定说话人的唇动信息相关的语音信号. 本文研究时变瞬时混合情况下的双模态语音提取问题.

设 k 为帧数，$\boldsymbol{S}_k = \boldsymbol{s}_{1:n,k}$ 与 $\boldsymbol{X}_k = \boldsymbol{x}_{1:m,k}$ 分别为 n 路源信号向量与 m 路观测向量，$\boldsymbol{\omega}_k = \boldsymbol{\omega}_{1:m,k}$ 为观测噪声，$\boldsymbol{H}_k$ 为时变混合矩阵，则考虑观测噪声的时变瞬时混合过程可以描述为

$$\boldsymbol{X}_k = \boldsymbol{H}_k \boldsymbol{S}_k + \boldsymbol{\omega}_k \quad (7)$$

设 $\boldsymbol{v}_k$ 是摄像机获取的与第 1 路语音信号对应的唇动信息. 双模态语音提取问题就是从混合语音 $\boldsymbol{X}_k$ 中，提取出与唇动信息 $\boldsymbol{v}_k$ 相关的一路源语音信号 $\boldsymbol{s}_{i,k}$.

2.2　双模态语音提取方法

利用矩阵 $\boldsymbol{G}_k$ 对混合信号 $\boldsymbol{X}_k$ 进行白化预处理后，得到的信号 $\boldsymbol{B}_k$ 与源信号 $\boldsymbol{S}_k$ 之间是正交变换关系. 设 $\boldsymbol{D}_k$ 为正交矩阵，将其作用于预白化信号 $\boldsymbol{B}_k$，得到一组语音信号 $\boldsymbol{Y}_k$，即

$$\boldsymbol{Y}_k = \boldsymbol{D}_k \boldsymbol{B}_k = \boldsymbol{D}_k \boldsymbol{G}_k \boldsymbol{X}_k = \hat{\boldsymbol{S}}_k \quad (8)$$

从 $\boldsymbol{Y}_k$ 中任取一路语音信号，记为 $\boldsymbol{y}_k$，并将其作为与唇动信息 $\boldsymbol{v}_k$ 相关的源语音信号估计值.

可视语音作为语音的视觉表征，为语音提取提供了重要信息. 然而语音的视觉特征与音频特征之间关系复杂[9]，如何有效地融合说话人的语音与唇动信息，成为双模态语音提取方法研究的重点. 本文将信息论中的最大互信息准则(MMI)引入到双模态语音提取方法中，利用互信息来描述语音视觉特征与音频特征之间的关联程度. 在盲源分离问题中，语音信号分离的过程就是输出的各分量非高斯性增强的过程. 峭度(kurtosis)反映了一个随机变量偏离高斯随机变量的程度，可为语音提取提供重要信息. 本文综合盲源信号分离中的高阶统计量准则与信息论中的最大互信息准则，将语音峭度 $Kurt(\boldsymbol{y}_k)$ 与音视频互信息 $I(\boldsymbol{y}_k, \boldsymbol{v}_k)$ 的乘积作为目标函数 $F(\boldsymbol{D}_k)$，即

$$F(\boldsymbol{D}_k) = | Kurt(\boldsymbol{y}_k) I(\boldsymbol{y}_k, \boldsymbol{v}_k) | \quad (9)$$

综上所述，双模态语音提取问题可以转化为根据混合语音 $\boldsymbol{X}_k$ 与唇动信息 $\boldsymbol{v}_k$ 确定正交分离矩阵 $\boldsymbol{D}_k$ 的优化问题，即

$$\hat{\boldsymbol{D}}_k = \arg\max_{\boldsymbol{D}_k} F(\boldsymbol{D}_k) \tag{10}$$

当说话人运动时，语音提取方法应具有跟踪时变混合矩阵的能力. 本文采用粒子滤波跟踪时变混合矩阵，从而解决了时变瞬时混合下的双模态语音提取问题.

2.3 音视频互信息计算

音视频互信息描述了语音波形与唇动信息之间的相关性，可为语音提取提供重要信息. 为了计算音视频之间的互信息，需要对音频与视频信息进行特征提取. 因此，如何提取稳健的音视频特征成为解决双模态语音提取问题的重要环节.

Mel 尺度频率描述了人耳对语音频率感知的非线性特征，MFCC 系数(Mel frequency cepstral coefficients)[10]将人耳的听觉系统和语音的产生系统相结合，在一定程度上模拟了人耳对声音的处理特点. 本文采用 MFCC 系数作为计算音视频互信息所需的音频特征. 对于 MFCC 系数的提取，可以使用 Mel 尺度的三角滤波器组对短时傅里叶变换能量谱进行滤波，将滤波器组的输出能量取对数，然后做离散余弦变换，得到 MFCC 系数.

可视语音与说话人说话时唇、下颌及其面部肌肉的运动有关，其中以唇形的变化对发音的影响最为重要. 因此，可将唇动视为可视语音的主要信息源[11]. 本文选取嘴唇的宽度与高度作为计算音视频互信息所需的视频特征. 唇部形状可用参数化的可变模板进行建模，如图 1 所示. 唇部边缘用两条四次曲线 Y_u 与 Y_l 来描述，其轮廓的曲线方程为

$$\begin{aligned} Y_u &= h_1 \times \left(1 - \frac{x^2}{w^2}\right) + 4q_1 \times \left(\frac{x^4}{w^4} - \frac{x^2}{w^2}\right) \\ Y_l &= h_2 \times \left(1 - \frac{x^2}{w^2}\right) + 4q_2 \times \left(\frac{x^4}{w^4} - \frac{x^2}{w^2}\right) \end{aligned} \tag{11}$$

其中 h_1 与 h_2 分别为上下唇的高度，w 为唇宽. 参数 q_1 与 q_2 表示四次曲线偏离抛物线的程度. 嘴唇的宽度与高度的提取采用文献[12]提出的唇定位方法. 该方法首先检测双眼位置；然后用肤色模型预测唇的大致位置；最后利用可变模板算法实现精确唇定位，获取嘴唇宽度与高度信息.

设 A、V 分别为音频与视频特征，H 为特征矢量的熵，音视频互信息 $I(A,V)$ 定义为[13]

$$\begin{aligned} I(A,V) = H(A) + H(V) - H(A,V) = \\ -\sum_i p(a_i)\log p(a_i) - \\ \sum_j p(v_j)\log p(v_j) + \\ \sum_{i,j} p(a_i,v_j)\log p(a_i,v_j) \end{aligned} \tag{12}$$

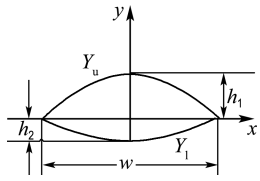


图 1 唇部模型

Fig. 1 The lip model

若音频特征矢量、视频特征矢量以及音视频联合特征矢量都服从局部高斯分布，设 $\boldsymbol{\Sigma}_A$、$\boldsymbol{\Sigma}_V$ 与 $\boldsymbol{\Sigma}_{AV}$ 分别为三者的方差矩阵，则音视频互信息 $I(A,V)$ 可以描述为

$$I(A,V) = \frac{1}{2}\log\frac{|\det(\boldsymbol{\Sigma}_A)||\det(\boldsymbol{\Sigma}_V)|}{|\det(\boldsymbol{\Sigma}_{AV})|} \tag{13}$$

2.4 双模态语音提取方法的粒子滤波实现

本文采用粒子滤波跟踪时变正交矩阵 $\boldsymbol{D}_k$，解决时变瞬时混合下的双模态语音提取问题. 假设正交矩阵 $\boldsymbol{D}_k$ 服从一阶马尔可夫过程，可用高斯随机游走模型加以描述. 将正交矩阵 $\boldsymbol{D}_k$ 向量化为 $\boldsymbol{d}_k$，即 $\boldsymbol{d}_k = \mathrm{vec}(\boldsymbol{D}_k)$，设 $\boldsymbol{\omega}_k$ 为高斯白噪声，则系统的状态方程为

$$\boldsymbol{d}_k = \boldsymbol{d}_{k-1} + \boldsymbol{\omega}_k \tag{14}$$

设 $\boldsymbol{C}_k = \boldsymbol{g}_{1:n,k}^{\mathrm{T}} \otimes \boldsymbol{I}_m$，$\otimes$ 表示 Kronecker 积，将式(8)重新写为

$$\boldsymbol{Y}_k = \boldsymbol{C}_k \boldsymbol{d}_k \tag{15}$$

设 $\boldsymbol{Z}_k = \{\boldsymbol{X}_k, \boldsymbol{v}_k\}$，在粒子滤波框架下解决双模态语音提取问题，便是根据观测语音信号 $\boldsymbol{X}_k$ 与可视语音信息 $\boldsymbol{v}_k$，计算分离矩阵 $\boldsymbol{d}_k$ 的滤波后验概率密度 $p(\boldsymbol{d}_k \mid \boldsymbol{Z}_k)$，得到分离矩阵 $\boldsymbol{d}_k$ 的估计值，从而实现语音信号的提取.

粒子滤波采用一组随机采样点及其对应权值来表示滤波概率密度 $p(\boldsymbol{d}_k \mid \boldsymbol{Z}_k)$. 令 $\{\boldsymbol{d}_k^{(i)}, w_k^{(i)}, i = 1,\cdots,N\}$ 表示从重要性概率密度函数中抽样获取的支撑点集，其中 $x_k^{(i)}$ 为 k 时刻的第 i 个状态，相应的权值为 $w_k^{(i)}$，则滤波后验概率密度可以表示为

$$p(\boldsymbol{d}_k \mid \boldsymbol{X}_k, \boldsymbol{v}_k) = \sum_{i=1}^{N} w_k^{(i)} \delta(\boldsymbol{d}_k - \boldsymbol{d}_k^{(i)}) \tag{16}$$

粒子滤波采用序贯重要性采样(SIS)算法从重要性概率密度函数中获取粒子集，并随着测量值的

依次到来序贯计算各粒子相应的权值 $w_k^{(i)}$，即

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(\boldsymbol{Z}_k \mid \boldsymbol{d}_k^{(i)}) p(\boldsymbol{d}_k^{(i)} \mid \boldsymbol{d}_{k-1}^{(i)})}{\pi(\boldsymbol{d}_{k-1}^{(i)} \mid \boldsymbol{d}_{1:k-1}^{(i)}, \boldsymbol{Z}_{1:k})} \quad (17)$$

采用系统转移函数 $p(\boldsymbol{d}_k \mid \boldsymbol{d}_{k-1})$ 作为重要性概率密度函数 $\pi(\cdot)$，根据系统的状态方程生成随机采样粒子 $\{\boldsymbol{d}_k^{(i)}\}_1^N$，将音视频互信息 $I(\boldsymbol{y}_k, \boldsymbol{v})$ 与语音峭度 $Kurt(\boldsymbol{y}_k)$ 的乘积作为观测似然函数 $p(\boldsymbol{Z}_k \mid \boldsymbol{d}_k)$，则粒子重要性权值 $w_k^{(i)}$ 的递推公式可以重新描述为

$$w_k^{(i)} \propto w_{k-1}^{(i)} \mid I(\boldsymbol{y}_k, \boldsymbol{v}_k) \cdot Kurt(\boldsymbol{y}_k) \mid \quad (18)$$

其中帧长为 L 的语音信号 $\boldsymbol{y}_k$ 的峭度计算公式为

$$Kurt(\boldsymbol{y}_k) = \sum_{j=1}^{L} y_k^4(j) - 3\Big[\sum_{j=1}^{L} y_k^2(j)\Big]^2 \quad (19)$$

通过粒子滤波算法计算出分离矩阵 $\boldsymbol{d}_k$ 的后验滤波概率密度 $p(\boldsymbol{d}_k \mid \boldsymbol{Z}_k)$ 后，便可得到分离矩阵 $\boldsymbol{d}_k$ 的最大后验(MAP)估计，即

$$\hat{\boldsymbol{d}}_k = \arg\max_{\boldsymbol{d}_k} p(\boldsymbol{d}_k \mid \boldsymbol{Z}_k) \quad (20)$$

计算出分离矩阵 $\hat{\boldsymbol{d}}_k$ 后，便可得到与可视语音信息 $\boldsymbol{v}_k$ 相关的语音信息 $\boldsymbol{y}_k$.

根据以上的推导，将基于粒子滤波的双模态语音提取方法的具体步骤归纳如下：

(1) 初始化

令 $k = 0$，粒子数为 N，从 $p(x_0)$ 中生成粒子集 $\{\boldsymbol{d}_0^{(i)}\}_{i=1}^N$，令所有粒子 $\boldsymbol{d}_0^{(i)}$ 的权值为 $1/N$；

(2)For $k = 1, 2, \cdots$

For $i = 1, 2, \cdots, N$:

① 粒子生成　从 $p(\boldsymbol{d}_k \mid \boldsymbol{d}_{k-1}^{(i)})$ 生成粒子 $\boldsymbol{d}_k^{(i)}$，求得相应的一路分离语音 $\boldsymbol{y}_k^{(i)}$；

② 计算音视频互相关系数 $I(\boldsymbol{y}_k^{(i)}, \boldsymbol{v}_k)$

a. 计算语音 $\boldsymbol{y}_k^{(i)}$ 的 MFCC 系数；

b. 提取嘴唇宽度与高度信息；

c. 计算音视频互信息 $I(\boldsymbol{y}_k^{(i)}, \boldsymbol{v}_k)$；

③ 计算语音信号的峭度 $Kurt(\boldsymbol{y}_k^{(i)})$；

④ 计算粒子的归一化重要性权值 $w_k^{(i)}$：

$\widetilde{w}_k^{(i)} \propto \widetilde{w}_{k-1}^{(i)} \mid Kurt(\boldsymbol{y}_k^{(i)}) \cdot I(\boldsymbol{y}_k^{(i)}, \boldsymbol{v}_k) \mid$；

$w_k^{(i)} = \widetilde{w}_k^{(i)} \Big(\sum_{i=1}^{N} \widetilde{w}_k^{(i)}\Big)^{-1}$；

End

⑤ 计算分离矩阵 $\hat{\boldsymbol{d}}_k$；

⑥ 输出语音 $\boldsymbol{y}_k$；

⑦ 粒子重采样　根据归一化权值 $\{w_k^{(i)}\}_{i=1}^N$ 更新粒子集 $\{\boldsymbol{d}_k^{(i)}\}_{i=1}^N$；

End

3　实验结果与分析

本文采用混合-分离系统全局矩阵的误差指数(error index, I_e)来客观评价语音提取方法的性能. 设全局矩阵为 $\boldsymbol{P} = \boldsymbol{DGH}$，误差指数的定义为

$$I_e = -10\log_{10}\Big\{\frac{1}{n}\Big[\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}\frac{p_{ij}^2}{\max_i p_{ij}^2} - 1\Big)\Big] + \frac{1}{n}\Big[\sum_{j=1}^{n}\Big(\sum_{i=1}^{n}\frac{p_{ij}^2}{\max_j p_{ij}^2} - 1\Big)\Big]\Big\} \quad (21)$$

仿真实验使用的原始语音及其对应的唇部信息取自 Carnegie-Mellon 大学提供的多媒体数据库[14]. 图像采集速度为 30 帧/s，本文提取嘴唇宽度与嘴唇高度作为视频特征. 音频采样率为 44.1 kHz，取 490 个采样点(11.1 ms)组成一帧计算 MFCC 系数与峭度. 由于音频与视频特征帧速不同，在计算音视频特征之间的互信息之前需要对视觉特征进行线性插值处理. 采用时变矩阵 $\boldsymbol{H}_k$ 来模拟说话人运动时混合矩阵的时变特性. 将原始语音信号加入高斯噪声，混合后两路信号的信噪比分别为 −0.2 dB 与 3.3 dB. 在粒子滤波算法中，粒子数取 1 000，采用残差重采样克服粒子滤波方法中的退化问题.

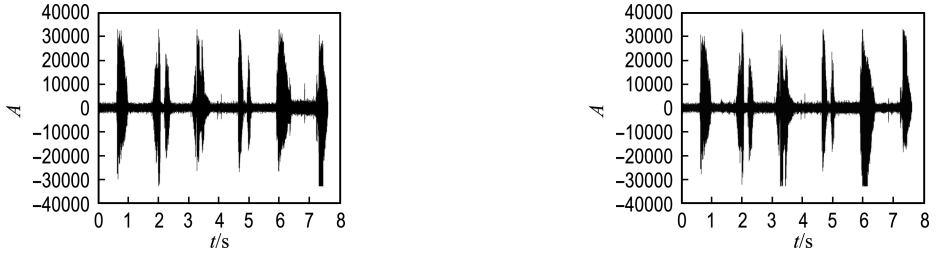
将本文方法与基于四阶累积量的联合对角化盲源分离方法[15](JADE)进行比较，来验证基于粒子滤波的双模态语音提取方法的有效性. 图 2 为计算机仿真实验的语音提取结果. 图 2(a)是 2 路原始语音信号，图 2(b)为英文发音对应的唇动信息. 图 2(c)为混合后的语音信号. 仿真实验分别应用 JADE 方法与本文方法进行语音提取，并采用谱减法对提取后的语音进行语音增强处理，实验结果分别如图 2(d)与图 2(e)所示. 从仿真结果可以看出，由于利用了说话人发音过程中音视频信息之间的相关性以及粒子滤波的跟踪能力，本文方法在时变混合情况下能够实现较好的语音提取效果. 图 3 是分离误差随时间的变化曲线. 与 JADE 盲源分离方法相比，本文的方法显示出较好的语音提取性能. 另外，实际听音效果也验证了本文方法的语音提取性能优于 JADE 盲源分离方法.

在不同信噪比情况下，本文方法、JADE 方法以及 FastICA 方法[16]的分离性能比较结果如表 1 所示. 从表中可以看出，在信噪比较低的情况下，本文方法由于利用了说话人的可视语音信息，仍然能够取得较好的语音提取效果.

(a) 两路原始语音信号

(b) 第一路语音信号对应的唇部信息

(c) 两路混合语音信号



(d) JADE 方法的提取结果　　(e) 本文方法的提取结果

图 2　语音信号提取结果

Fig. 2　The result of speech extraction

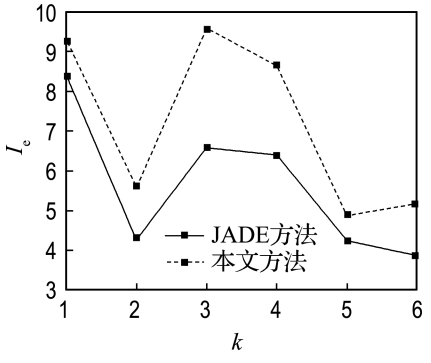


图 3　两种语音提取方法的误差指数比较结果

Fig. 3　The error index comparison of two methods

表 1　不同信噪比下 3 种方法的分离性能比较结果

Tab. 1　The comparison of three methods in different input SNR

SNR/dB	I_e		
	本文	JADE	FastICA
−2.24/−2.68	5.54	2.13	2.08
1.78/3.29	8.42	6.37	6.16
3.81/5.32	12.27	10.88	10.53
5.81/5.29	12.37	11.28	11.06
7.80/5.31	14.33	13.11	13.01

4 结　语

本文提出了一种基于粒子滤波的双模态语音提取方法，为视频会议等特定环境下的语音提取问题提供了新的解决方案. 该方法利用说话人语音与唇动信息之间的相关性，通过粒子滤波跟踪时变混合矩阵，实现语音信号的有序提取. 计算机仿真实验结果验证了本文方法的有效性. 利用双模态方法解决卷积混合情况下的语音提取问题，是作者下一步将要开展的工作.

参考文献：

[1] BRANDSTEIN M S, WARD D B. Cell-based beamforming for speech acquisition with microphone arrays [J]. **IEEE Transactions on Speech and Audio Processing**, 2000, **8**(6):738-742

[2] CRUCES-ALVAREZ S A, CICHOCKI A, AMARI S. From blind signal extraction to blind instantaneous signal separation:criteria, algorithms, and stability [J]. **IEEE Transactions on Neural Networks**, 2004, **15**(4):859-873

[3] CHEN T, RAO R. Audio-visual integration in multimodal communication [J]. **IEEE Processings**, 1998, **86**(5):837-852

[4] DUPONT S, LUETTIN J. Audio-visual speech modeling for continuous speech recognition [J]. **IEEE Transactions on Multimedia**, 2000, **2**(3):141-151

[5] BUB U, HUNKE M, WAIBLE A. Knowing who to listen to in speech recognition:visually guided beamforming [C] // **IEEE International Conference on Acoustics, Speech, and Signal Processing**. Detroit: IEEE, 1995:848-851

[6] SODOYER D, GIRIN L, JUTTEN C, *et al*. Developing an audio-visual speech source separation algorithm [J]. **Speech Communication**, 2004, **44**(1-4):113-125

[7] RAJARAM S, NEFIAN A V, HUANG T S. Bayesian separation of audio-visual speech sources [C] // **IEEE International Conference on Acoustics, Speech, and Signal Processing**. Quebec:IEEE, 2004: 657-660

[8] ARULAMPALAM M, MASKELL S, GORDON N, *et al*. A tutorial on particle filters for on-line nonlinear/non-Gaussian Bayesian tracking [J]. **IEEE Transactions on Signal Processing**, 2002, **50**(2): 174-188

[9] YEHIA H, RUBIN P, VATIKIOTIS-BATESON E. Quantitative association of vocal-tract and facial behavior [J]. **Speech Communication**, 1998, **26**: 23-43

[10] VERGIN R, OSHAUGHNESSY D, FARHAT A. Generalized Mel frequency cepstral coefficients for large vocabulary speaker independent continuous speech recognition [J]. **IEEE Transactions on Speech and Audio Processing**, 1999, **7**(5):525-532

[11] SUMBY W H, POLLAK I. Visual contributions to speech intelligibility in noise [J]. **Journal of the Acoustical Society of America**, 1954, **26**:212-215

[12] 姚鸿勋，高　文，李　静，等. 用于口型识别的实时唇定位方法[J]. 软件学报，2000, **11**(8):1126-1132

[13] COVER T, THOMAS J. **Elements of Information Theory** [M]. New York:Wiley, 1991

[14] CHEN T. Audiovisual speech processing [J]. **Transactions on Signal Processing**, 2001, **18**(1):9-21

[15] CARDOSO J F, SOULOUMIAC A. Blind beamforming for non-Gaussian signals [J]. **IEE Proceedings, Radar and Signal Processing**, 1993, **140**(6):362-370

[16] HYVARINEN A, OJA E. A fast fixed-point algorithm for independent component analysis [J]. **Neural Computation**, 1997, **9**(7):1483-1492

Bimodal speech extraction method based on particle filtering

JIN　Nai-gao*,　YIN　Fu-liang

(School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Lip movement information helps language comprehension when the auditory signal is degraded. A bimodal speech extraction method is presented based on the method of audio-visual signal processing. The particle filtering is used to construct a Bayesian fusion framework for bimodal speech extraction problem. By combining maximum mutual information criterion with higher-order statistics criterion of blind signal separation and estimating mixed matrices by particle filtering method, the proposed method can extract the interested instaneous time-varying speech signal by maximizing the product of kurtosis and audio-visual mutual information. Simulation results show that the proposed method improves the performance of the speech extraction system in the low SNR environment.

Key words: speech extraction; particle filtering; higher-order statistics; maximum mutual information