第 37 卷第 2 期　　声 学 学 报　　Vol. 37, No. 2
2012 年 3 月　　ACTA ACUSTICA　　Mar., 2012

基于谐波和能量特征的单声道浊语音分离方法*

赵立恒[1]　汪增福[1,2]
(1 中国科学技术大学自动化系 合肥 230027)
(2 中国科学院合肥智能机械研究所 合肥 230031)

2011 年 2 月 25 日收到
2011 年 4 月 27 日定稿

摘要 提出了一种基于谐波和能量特征的单声道浊语音分离方法。该方法将浊语音分离问题转化为声音在时频域的分类问题。首先，在已有谐波特征的基础上，引入能量特征。然后，对于谐波特征明显且能量大的时频单元，在分类器训练阶段复制它们的特征。实验结果表明该方法相比之前的方法有更好的信噪比增益。通过引入能量特征和特征复制，改善了浊语音的分离效果。

PACS 数：43.72

Monaural voiced speech separation based on harmonic and energy features

ZHAO Liheng[1]　WANG Zengfu[1,2]
(1 *Automation Department, University of Science and Technology of China* Hefei 230027)
(2 *Institute of Intelligent Machines, Chinese Academy of Sciences* Hefei 230031)

Received Feb. 25, 2011
Revised Apr. 27, 2011

Abstract A monaural voiced speech separation approach based on harmonic and energy features is proposed. The method casts voiced speech separation as sound classification problem in time-frequency domain. First, the energy feature is employed to assist harmonic features. Then, the harmonic and energy features of time-frequency units with obvious harmonicity and large energy are replicated in the process of classifier training. Experimental results show that the proposed method obtains better signal-to-noise ratio improvement compared with previous approaches. The voiced speech separation is improved by introducing energy feature and feature replication.

引言

背景噪声下的语音分离是一个具有挑战性的问题，这方面的研究在语音识别、多媒体检索和助听器设计等领域有着重要的意义。近年来，多通道信号处理方法被广泛用于解决语音分离问题[1−3]，并获得了巨大的成功。但是，许多实际应用场合中只有一个传感器信号可以利用，因此，如何解决单声道情况下的语音分离问题成为实际中的迫切需求，受到研究人员的重点关注。

尽管单声道语音分离充满挑战性，但人类的听觉系统还是展现出了优秀的单耳感知能力。因此，研究人员试图结合心理声学和心理生理学来分析人类听觉系统的感知机理，进而设计出模拟人耳听觉感知能力的计算机系统。1990 年，Bregman 首先提出了听觉场景分析 (Auditory Scene Analysis, ASA) 的概念[4]，为计算听觉场景分析 (Computational Auditory Scene Analysis, CASA) 提供了理论基础[5−8]，也为单声道语音分离问题提供了新的研究思路[9−11]。

语音信号通常由浊音和清音组成。其中，浊音是语音信号的主导成分。因此，浊语音的分离在语音分离问题中起关键作用。最近，在 CASA 的框架下，Hu 和 Wang 提出了一种单声道浊语音分离方法[12]，该方法基于浊音的谐波特性 (包括基音和多次谐波信息)，将浊语音分离问题转化为声音 (目标语音和噪声) 的分类问题。实验结果表明，Hu-Wang 方法优于之前的方法。

从声学特性来看，语音信号的大部分能量分布在浊音的谐波上，因此，浊音成分不仅含有丰富的

* 安徽省科技攻关计划 (语音专项，11010202192) 资助项目

谐波特性，而且含有能量特征。从听觉感知的角度来看，人耳对声音的感知主要依据三个要素，基音、响度和音色。其中，基音与谐波特性有关，响度与能量有关，而音色与谐波特性和能量都有关。因此，谐波特性和能量都是听觉感知中非常重要的特征。

基于上面的观察，我们提出了一种改进的单声道浊语音分离方法。与 Hu-Wang 方法相比，该方法有以下几个特点：(1) 在已有谐波特征的基础上，引入了能量特征；(2) 通过相对能量顺序来刻画能量特征；(3) 在分类器训练阶段，根据时频单元的能量特征和局部信噪比 (Signal-to-Noise Ratio, SNR) 来决定特征的训练方式。实验结果表明，该方法比 Hu-Wang 方法有更好的浊语音分离效果。

1 分离方法结构图

图 1 给出了本文提出的单声道浊语音分离方法的结构框图，主要由训练和分离两个部分组成。

在训练阶段，首先通过时频分解模块得到训练样本 (目标语音、噪声和混合信号) 的时频信息 (时频单元的滤波响应、响应包络和能量)。然后，根据时频信息提取时频单元的谐波和能量特征，并计算时频单元的局部 SNR。最后，通过时频单元的能量特征和局部 SNR 来决定特征的训练方式，进而训练出可以刻画特征分布的分类器。

在分离阶段，首先将混合信号通过时频分解和特征提取模块，得到每个时频单元的谐波和能量特征。然后，通过分类器将每个时频单元的特征映射到概率区间 [0,1] 中的一个概率值，该值反映了相应时频单元由目标语音主导的概率。紧接着，根据每个时频单元的概率值估计二值模 ('1' 表示目标语音主导，'0' 表示噪声主导)。最后，根据 Weintraub 方法从二值模和混合信号的滤波响应合成目标语音的时域波形 [13]，实现目标语音的分离。

2 时频分解和特征提取

时频分解和特征提取是训练和分离阶段所共有的模块。通过时频分解，输入的时域信号被转化为时频域的表示形式。再通过特征提取，得到输入信号在时频域的特征，为后继的分类器训练和语音分离提供输入。

2.1 时频分解

基于人耳的听觉感知机制，本文采用 128 个 gammatone 滤波器组成的滤波器组对输入声音信号进行带通滤波，滤波器的中心频率以等矩形带宽的方式分布在 80 Hz 到 5000 Hz 之间。然后，采用交叠分段的方法，以 20 ms 为帧长、10 ms 为帧移，对每一个频率通道的滤波响应做时域分帧处理，得到输入信号的时频域表示，即耳蜗图 (Cochleagram)[14]。通过半波整流和带通滤波，可进一步提取滤波响应的包络。

至此，我们得到了主要由耳蜗图表示的输入声音信号的有用时频信息：时频单元的滤波响应、响应包络和能量。

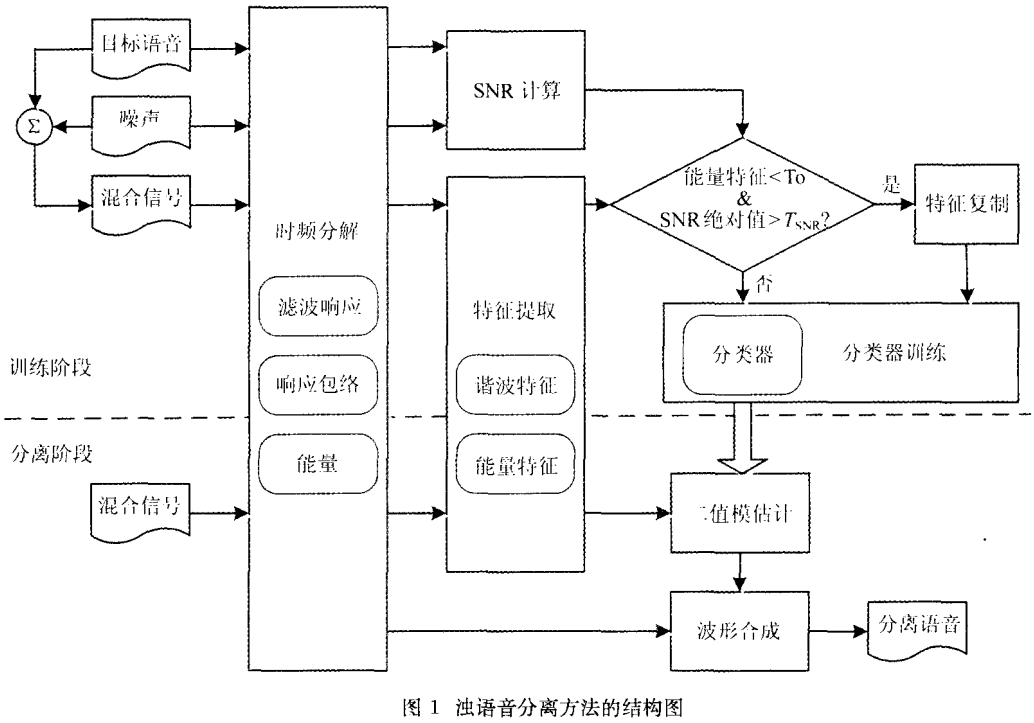

图 1　浊语音分离方法的结构图

2.2 谐波特征提取

语音信号中的浊音成分是由声带振动产生的，具有准周期性。因此，谐波特性 (包括基音和多次谐波) 是浊音的重要特征。

为了描述声音信号时频单元的谐波特性，和 Hu-Wang 方法一样，本文采用如下的 6 维谐波特征 [12]，

$$K_{cm}(\tau)=\begin{pmatrix} A(c,m,\tau),\overline{f}(c,m)\tau-\mathrm{int}(\overline{f}(c,m)\tau),\mathrm{int}(\overline{f}(c,m)\tau), \\ A_e(c,m,\tau),\overline{f}_e(c,m)\tau-\mathrm{int}(\overline{f}_e(c,m)\tau),\mathrm{int}(\overline{f}_e(c,m)\tau) \end{pmatrix}, \tag{1}$$

这里，c 和 m 分别表示时频单元所在的频率通道和时间帧。式中，$A(c,m,\tau)$ 表示滤波响应在时延 τ 的自相关函数值；$\overline{f}(c,m)$ 表示滤波响应的平均瞬时频率；int( ) 表示取整操作 (返回最近的整数)。用 $\tau_S(m)$ 表示基音周期在第 m 帧的估计值，如果滤波响应的周期接近 $\tau_S(m)$，那么前 3 维特征具有以下特性：

(1) $A(c,m,\tau)$ 在 $\tau=\tau_S(m)$ 附近有峰值；

(2) $\mathrm{int}(\overline{f}(c,m)\tau_S(m))$ 表示滤波器所响应的谐波序号；

(3) $\overline{f}(c,m)\tau_S(m)-\mathrm{int}(\overline{f}(c,m)\tau_S(m))$ 可以度量估计的基音周期和滤波响应周期的相似性。

在高频区域，gammatone 滤波器有较大的频率带宽，会同时响应一个周期信号的多个谐波成分，产生幅度调制现象 (Amplitude modulation)[15]。此时，相对于滤波响应，响应包络表现出更好的谐波特性。因此，6 维特征中的后 3 维用于描述响应包络的谐波特性。

2.3 能量特征提取

在语音信号中，浊音成分不仅含有丰富的谐波信息，而且占据了语音信号的大部分能量。因此，能量也是浊音的重要特征。

在本文中，时频单元的相对能量顺序被用作能量特征，它不仅可以度量时频单元的能量大小，也保证了不同声音信号的度量结果具有一致性。对于一个声音信号，可按照如下步骤提取时频单元的能量特征：

(1) 将所有时频单元的能量从大到小排序，得到排序后的时频单元能量序列 $\{E_S(c,m)\}$；

(2) 记每个时频单元的能量 $E(c,m)$ 在 $\{E_S(c,m)\}$ 中的序号为 $O(c,m)$；

(3) 计算每个时频单元的相对能量顺序 $O_r(c,m)$，并以此作为能量特征，

$$O_r(c,m)=O(c,m)/N, \tag{2}$$

其中，N 是声音信号中时频单元的总数。

据此，声音信号 (即使是不同的声音信号) 的能量特征都分布在 (0, 1] 区间中。而且，时频单元的能量越大，它所对应的能量特征值越小。

将能量特征加入 6 维谐波特征，可构成如下的 7 维特征向量：

$$K_{cm}(\tau)=\begin{pmatrix} A(c,m,\tau),\overline{f}(c,m)\tau-\mathrm{int}(\overline{f}(c,m)\tau),\mathrm{int}(\overline{f}(c,m)\tau), \\ A_e(c,m,\tau),\overline{f}_e(c,m)\tau-\mathrm{int}(\overline{f}_e(c,m)\tau),\mathrm{int}(\overline{f}_e(c,m)\tau),O_r(c,m) \end{pmatrix}, \tag{3}$$

在分类器的训练和测试过程中，用 $K_{cm}(\tau_S(m))$ 作为时频单元的 7 维特征，其中，$\tau_S(m)$ 表示目标语音的基音周期在第 m 帧的估计值。

3 分类器训练

在 Hu-Wang 方法中，时频单元的 6 维谐波特征被用来训练分类器。因此，在分离阶段，时频单元的谐波特征越强 (或弱)，分类器的输出概率越接近 1(或 0)，这些时频单元被错误分类的可能性也就越小。另一方面，在语音分离系统中，时频单元的能量越大，它的分离结果对系统性能的影响也越大。因此，降低能量大的时频单元被错误分类的可能性对提高分类器的性能有重要意义。

3.1 能量大的时频单元

从声学特性来看，语音信号的大部分能量分布在少数时频单元上。如 2.3 节所述，时频单元的相对能量顺序越小，它在声音信号中的能量越大。因此，在一个语音信号中，少数相对能量顺序小的时频单元占据了大部分能量。进一步，对于实验数据中的训练集和测试集 (数据集的具体描述参见 5.2 节)，分别统计混合声音信号在相对能量顺序区间的能量分布 (见表 1)，具体过程如下：

表 1　混合声音信号在相对能量顺序区间中的能量分布 (%)

样本集	相对能量顺序区间									
	(0,0.1]	(0,0.2]	(0,0.3]	(0,0.4]	(0,0.5]	(0,0.6]	(0,0.7]	(0,0.8]	(0,0.9]	(0,1]
训练集	77.64	89.51	94.48	96.99	98.38	99.19	99.66	99.90	99.99	100
测试集	83.39	93.08	96.72	98.36	99.17	99.59	99.81	99.92	99.98	100

(1) 根据 2.3 节中能量特征提取方法，计算混合信号所有时频单元的相对能量顺序 $\{O_r(c,m)\}$。其中，$O_r(c,m)$ 分布在 (0, 1] 区间；

(2) 记 $O_r(c,m)$ 落在区间 (0, T] 中的时频单元为 $U^T(c,m)$，且该时频单元的能量为 $E^T(c,m)$. 其中，$T \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$；

(3) 计算相对能量顺序位于区间 $(0,T]$ 中的所有时频单元的能量和 E^T，

$$E^T = \sum_U E^T(c,m), U \in \left\{U^T(c,m)\right\}, \tag{4}$$

由于 $O_r(c,m)$ 都位于 (0, 1] 区间，E^1 表示了混合信号中所有时频单元的能量和；

计算混合信号在相对能量顺序区间 (0, T] 中的能量分布 (百分比)P^T，

$$P^T = (E^T/E^1) \times 100\%. \tag{5}$$

(4) 计算数据集中所有混合信号在相对能量顺序区间 (0, T] 中的平均能量分布 P^T_{mean}，

$$P^T_{\text{mean}} = \sum_q P^T/Q, \quad q \in \{1, 2, \cdots, Q-1, Q\} \tag{6}$$

其中，Q 表示数据集中混合信号的总数。

从表 1 可以看出，平均来说，无论是训练集还是测试集中的混合声音信号，相对能量顺序位于 (0, 0.1] 区间的时频单元都占据了 70% 以上的能量。这表明，混合声音信号的大部分能量集中在相对能量顺序位于 (0, 0.1] 区间的时频单元上。为此，本文定义相对能量顺序小于 0.1 的时频单元为能量大的时频单元，即图 1 中的 T_o 取为 0.1 。

3.2 分类器的强化训练

一般而言，在训练过程中，如果某一种特征的数量越大，分类器对这种特征的刻画能力就越强，进而在测试过程中，分类器错误分类这种特征的可能性也就越小。因此，复制能量大的时频单元的特征，可以强化分类器对 '能量大' 的刻画能力，进而降低此类时频单元被错误分类的可能性。但是，某些能量大的时频单元的谐波特征却不明显 (既不强也不弱)，复制这些时频单元的 7 维特征，一定意义上弱化了分类器对谐波特征的刻画能力。

另一方面，在混合声音信号中，如果一个时频单元的能量大，而且局部 SNR 的绝对值足够大 (大于 10 dB，即图 1 中的 T_{SNR} 取为 10 dB)，那么该时频单元不是由目标语音的浊音成分绝对主导，就是由背景噪声绝对主导，此时，该时频单元的谐波特征明显 (强或弱) 。

基于以上考虑，本文提出了一种基于能量特征和局部 SNR 的分类器训练方法。该方法既能强化分类器对 '能量大' 的刻画能力，也能保持分类器对谐波特征的刻画能力。训练过程如下：

(1) 计算局部 SNR

对于每个特征 $K_{cm}(\tau_S(m))$，计算相应时频单元的局部 SNR，

$$\text{SNR}(c,m) = 10\log_{10}\{E_t(c,m)/E_n(c,m)\}, \tag{7}$$

$E_t(c,m)$ 和 $E_n(c,m)$ 分别表示该时频单元内目标语音和噪声的能量；

(2) 复制特征

对于每个特征 $K_{cm}(\tau_S(m))$，如果

$$\begin{cases} O_r(c,m) < T_o, & \text{(能量大)}, \\ \text{SNR}(c,m) \text{ 绝对值 } > T_{\text{SNR}}, & \text{(谐波特征明显)}, \end{cases} \tag{8}$$

则复制该特征 W 次 (例如，根据实验结果，W 取为 8)，否则保持特征数目不变；

基于新的特征集，对分类器进行训练 (见 5.1 节) 。

4 语音分离

本文假定，混合信号中的任意一个时频单元，或者由目标语音主导 (H_0)，或者由噪声主导 (H_1) 。因此，当且仅当不等式 (9) 成立，相应的时频单元才被认为由目标语音主导[12] 。

$$P(H_0|K_{cm}(\tau_S(m))) > P(H_1|K_{cm}(\tau_S(m))), \tag{9}$$

由于

$$P(H_0|K_{cm}(\tau_S(m))) + P(H_1|K_{cm}(\tau_S(m))) = 1, \tag{10}$$

不等式 (9) 可以转化为：

$$P(H_0|K_{cm}(\tau_S(m))) > 0.5. \tag{11}$$

根据分类器的输出概率 $P(H_0|K_{cm}(\tau_S(m)))$ 和不等式 (11), 我们可以估计混合信号的二值模 ('1' 表示目标语音主导, '0' 表示噪声主导)。然后, 从二值模和混合信号的滤波响应合成目标语音的时域波形[13], 实现目标语音的分离。具体而言, 首先将混合信号在每个频率通道的滤波响应做一次时域翻转, 并对翻转后的滤波响应再次 gammatone 滤波。紧接着, 对滤波器输出再做一次时域翻转 (两次时域翻转消除了滤波器输出在频率通道间的相位差)。然后, 以 20 ms 为帧长、10 ms 为帧移、升余弦函数 (raised cosine) 为窗函数, 对第二次翻转后的滤波响应做时域分帧处理。最后, 以二值模为权重将时频单元的滤波响应在频率轴加和, 进而得到目标语音的时域波形。

5 实验设计与结果评估

5.1 实验设计

在浊语音分离系统中, 基音估计和二值模估计是两个关键环节[12]。基音估计环节估计混合声音信号的基音周期 (如果目标语音和噪声都含有谐波成分, 就会有两个估计值, 一个对应于目标基音, 另一个对应于噪声基音)。紧接着, 在给定基音周期估计值的情况下, 提取混合声音信号中时频单元的谐波特征。然后, 通过分类器将每个时频单元的谐波特征映射到概率区间 [0,1] 中的概率值, 该概率值刻画了相应时频单元由目标语音或噪声主导的概率 (在一个时频单元中, 如果目标语音和噪声都含有谐波成分, 分类器就有两个输出概率, 一个表示目标语音主导的概率, 另一个表示噪声主导的概率)。最后, 根据每个时频单元的概率值估计二值模。具体而言, 当只有目标语音含有谐波成分时, 根据分类器的输出概率 $P(H_0|K_{cm}(\tau_S(m)))$ 和不等式 (11) 估计二值模; 当目标语音和噪声都含有谐波成分时, 根据不等式 (12) 估计二值模,

$$\begin{cases} P(H_0|K_{cm}(\tau_S(m))) > 0.5, \\ P(H_0|K_{cm}(\tau_S(m))) > P(H_1|K_{cm}(\tau'_S(m))), \end{cases} \tag{12}$$

其中, $\tau_S(m)$ 和 $\tau'_S(m)$ 分别表示目标基音和噪声基音在第 m 帧的估计值, $K_{cm}(\tau_S(m))$ 和 $K_{cm}(\tau'_S(m))$ 分别表示时频单元对应于目标基音和噪声基音的谐波特征, $P(H_0|K_{cm}(\tau_S(m)))$ 和 $P(H_1|K_{cm}(\tau'_S(m)))$ 表示分类器的输出概率。

本文希望通过改进特征提取和分类器训练来改善二值模估计, 进而提升分离系统的性能。为了避免基音估计对改进方法的影响, 用 Praat 提取纯净目标语音的基音[16], 并以此作为混合声音中目标基音的估计值。在给定目标基音的情况下, 直接根据不等式 (11) 估计二值模。

在每一个频率通道 (总共 128 个通道), 一个多层感知器 (Multiple Layer Perception, MLP) 被用作分类器[17]。每个 MLP 都由输入层、输出层和包含 5 个隐节点的单一隐层组成。在训练阶段, 输入特征是训练样本的 7 维特征, 期望输出是理想二值模 (Ideal Binary Mask, IBM)。IBM 是基于听觉掩蔽效应的二值矩阵, 其中 '1' 表示相应时频单元由目标语音主导 (局部 SNR 大于 0), '0' 表示相应时频单元由噪声主导 (局部 SNR 小于 0)[18]。在测试阶段, 先通过训练好的 MLP 把测试样本的 7 维特征映射到概率区间 [0,1] 中的概率值, 然后根据输出概率和不等式 (11) 估计二值模, 最后合成目标语音。

5.2 实验数据

实验数据分为训练集和测试集, 并且所有声音信号都重采样到 16 kHz。

在训练集中, 一共有 200 个混合信号, 由 10 个目标语音和 20 个噪声在 0 dB 下混合得到。其中, 20 个语音信号 (10 个目标语音和 10 个语音噪声) 来自 TIMIT 数据集的训练集[19], 10 个非语音噪声分别来自 Guoning 噪声集的 10 类噪声[12]。

在测试集中, 一共有 300 个混合信号, 由 10 个目标语音和 10 个噪声在 3 种信噪比 (0 dB, −5 dB, −10 dB) 下混合得到。其中, 10 个目标语音来自 TIMIT 数据集的测试集, 10 个噪声分别是: N0- 警报声, N1- 电话铃声, N2- 风声, N3- 咳嗽声, N4- 鸡尾酒会噪声 (Cocktail party noise), N5~N9-TIMIT 测试集中的男性语音或女性语音。

特别地, 在训练集和测试集中, 不仅目标语音没有重复 (在说话人和语音内容方面), 噪声也没有重复。

5.3 分离结果评估与比较

理想二值模被广泛用作 CASA 系统的设计目标[8,10,12,18]。因此, 基于 IBM 的 SNR 通常被用来评估 CASA 系统的分离结果, 计算方法如下:

$$\mathrm{SNR}_{\mathrm{mix}} = 10\log_{10}\frac{\sum_k [S_{\mathrm{ibm}}(k)]^2}{\sum_k [S_{\mathrm{ibm}}(k) - S_{\mathrm{mix}}(k)]^2}, \tag{13}$$

$$\mathrm{SNR}_{\mathrm{est}} = 10\log_{10}\frac{\sum_k [S_{\mathrm{ibm}}(k)]^2}{\sum_k [S_{\mathrm{ibm}}(k) - S_{\mathrm{est}}(k)]^2}, \tag{14}$$

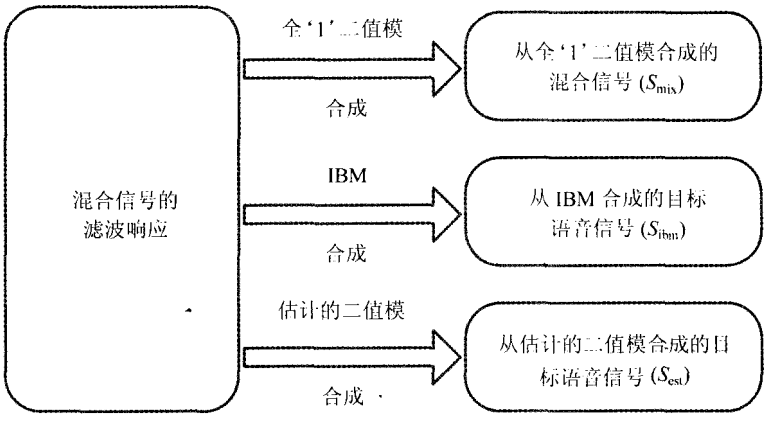

图 2　从二值模和混合信号的滤波响应合成声音信号

表 2　不同混合信噪比和特征复制次数情况下的 SNR 结果 (dB)

混合信噪比	Mixture	Hu-Wang	Proposed				
			0	4	8	12	16
0	−0.0351	7.2034	7.5866	7.6976	**7.8114**	7.7165	7.7381
−5	−5.5044	2.9114	3.8640	4.2990	**4.5321**	4.4343	4.4873
−10	−11.0280	−1.9643	−0.6281	0.1120	**0.3951**	0.2992	0.3848

其中，$S_{\mathrm{mix}}(\cdot)$, $S_{\mathrm{ibm}}(\cdot)$ 和 $S_{\mathrm{est}}(\cdot)$ 表示从相应二值模和混合信号的滤波响应合成的声音信号，图 2 给出了它们的合成方法。

基于上面的评估方法，表 2 给出了不同混合信噪比和特征复制次数情况下的 SNR 结果。其中，Mixture 表示原始混合信号的结果，Hu-Wang 表示 Hu-Wang 方法的结果，Proposed 表示本文方法的结果。表 2 中每个数据是 100 个混合信号的平均结果。

从表 2 可以看出，在所有混合信噪比和特征复制次数情况下，混合信号经过本文的分离算法后，SNR 都有明显提升。特别地，对不同特征复制次数 W 情况下的 SNR 结果进行观察，我们有以下分析：

(1) 当 W 取为 0 时，本文方法相比于 Hu-Wang 方法有 SNR 提升。该提升效果得益于能量特征的引入；

(2) 当 W 取为 4 和 8 时，SNR 有进一步的提升。该提升效果得益于分类器训练过程中特征复制方法的引入；

(3) 当 W 取为 12 和 16 时，SNR 没有继续提升，甚至稍有下降。我们认为，随着特征复制次数 W 的继续增加，分类器的训练可能出现 "过拟合" 现象，从而影响分类器的整体区分性。

当混合信噪比为 0 dB 且特征复制次数为 8 时，表 3 给出了不同噪声环境下的 SNR 结果。其中，每个数据是相应噪声环境下 10 个混合信号的平均结果，最后一行还给出了所有噪声环境下 100 个混合信号的平均结果。

从表 3 可以看出，在所有噪声环境下，混合信号经过本文的分离算法后，SNR 都有明显改善，平均提升约 7.8 dB 。特别地，在窄带噪声 (例如 N0, N1, N3, N5~N9) 环境下，本文方法比 Hu-Wang 方法在 SNR 提升方面有较大优势，但在宽带噪声 (例如 N2 和 N4) 环境下，该优势不明显。一般而言，相比于宽带噪声环境，窄带噪声环境下混合信号中能量大的时频单元，其局部 SNR 的绝对值较大，相应的谐波特征也较明显。另一方面，本文方法通过强化训练，降

表 3　不同噪声环境下的 SNR 结果 (dB)

噪声	Mixture	Hu-Wang	Proposed
N0	−0.2389	6.8630	8.1368
N1	−0.0908	10.7604	11.3453
N2	−0.2454	6.6792	6.9719
N3	−0.0956	7.8783	8.6143
N4	0.4533	7.4517	7.4331
N5	−0.1582	7.0199	7.5301
N6	−0.1057	6.4536	7.0226
N7	0.3030	5.8982	6.7812
N8	−0.0053	6.9468	7.5501
N9	−0.1679	6.0824	6.7287
平均	−0.0351	7.2034	7.8114

低了能量大且谐波特征明显的时频单元被错误分类的可能性 (见 3.2 节)。因此，相比于宽带噪声环境，本文方法在窄带噪声环境下有更明显的改善效果。

6 结论

基于声音信号的声学特性和分类器的基本特点，本文提出了一种改进的单声道浊语音分离方法。特征提取方面，通过谐波和能量特征共同描述声音信号的声学特性。分类器训练方面，对于混合信号中谐波特征明显且能量大的时频单元，通过复制它们的特征来强化训练，从而降低它们被错误分类的可能性。实验结果表明，相比之前的方法，该方法有更好的浊语音分离效果。

另一方面，一个完整的浊语音分离系统包含基音估计和二值模估计两个环节。在本文中，通过改进特征提取和分类器训练来改善二值模估计，进而改善浊语音分离效果。为了避免基音估计对改进方法的影响，用 Praat 提取纯净目标语音的基音，并以此作为混合信号中目标基音的估计值。因此，本文方法改善了给定目标基音情况下的浊语音分离系统的性能，在实用方面有局限性。加入基音估计算法将增加本文方法的实用性。

致谢

衷心感谢内蒙古大学的张学良老师给予的指导和帮助。

参 考 文 献

1 Sawada H *et al.* Blind extraction of dominant target sources using ICA and time-frequency masking. *IEEE Trans. Audio, Speech, Lang. Process.*, 2006; **14**(6): 2165—2173

2 Saruwatari H *et al.* Blind source separation combining independent component analysis and beamforming. *EURASIP J. Appl. Signal Process.*, 2003; **2003**(11): 1135—1146

3 Benesty J *et al.* On microphone-array beamforming from a MIMO acoustic signal processing perspective. *IEEE Trans. Audio, Speech, Lang. Process.*, 2007; **15**(3): 1053—1065

4 Bregman A S. Auditory scene analysis: The perceptual organization of sound. MA: The MIT Press, 1994

5 Wang D L, Brown G J. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Netw.*, 1999; **10**(3): 684—697

6 Brown G J, Cooke M. Computational auditory scene analysis. *Comput. Speech Lang.*, 1994(8): 297—336

7 Ellis D P W. Prediction-driven computational auditory scene analysis. Ph.D. thesis, MIT, 1996

8 Kim G *et al.* An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *J. Acoust. Soc. Am.*, 2009; **126**: 1486—1494

9 Hu G N, Wang D L. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Netw.*, 2004; **15**(5): 1135—1150

10 张学良等. 改进谐波组织规则的单通道浊语音分离系统. 声学学报, 2011; **36**(1): 88—96

11 Li P *et al.* Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. *IEEE Trans. Audio, Speech, Lang. Process.*, 2006; **14**(6): 2014—2023

12 Hu G N, Wang D L. A tandem algorithm for pitch estimation and voiced speech segregation. *IEEE Trans. Audio, Speech, Lang. Process.*, 2010; **18**(8): 2067—2079

13 Weintraub M. A theory and computational model of auditory monaural sound separation. Ph.D. thesis, Stanford University, 1985

14 Wang D L, Brown G J. Computational auditory scene analysis: principles, algorithms, and applications. John Wiley & Sons, Inc., 2006: 15—19

15 Helmholtz H. On the sensation of tone. Second English ed., New York: Dover Publishers, 1863

16 Boersma P, Weenink D. Praat: Doing phonetics by computer. 2009

17 Duda R O, Hart P E, Stock D G. Pattern classification. John Wiley & Sons, Inc., 2001: 282—349

18 Wang D L. On ideal binary mask as the computational goal of auditory scene analysis. *Speech Separation by Humans and Machines*, Kluwer Academic Pub, 2005: 181—197

19 Garofolo J *et al.* DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, 1993