第 33 卷　第 18 期　　计　算　机　工　程　　2007 年 9 月
Vol.33　No.18　　Computer Engineering　　September 2007
·人工智能及识别技术·　文章编号：1000—3428(2007)18—0211—03　文献标识码：A　中图分类号：TP391

基于计算听觉场景分析的语音混叠信号分离

王　珊，许　刚
(华北电力大学(北京)信息工程系，北京　102206)

摘　要：基于计算听觉场景原理，提出了一种混叠语音信号分离算法模型，对两个说话者的混叠声音进行分离。该模型对低频区和高频区的分离分别采用了不同方法，避免了因采用同样方法处理低频高频区而导致对高频段语音不能很好分离的结果。实验结果表明，该模型具有很好的应用效果。

关键词：计算听觉场景分析；语音分离；基音估计；幅度调制

Speech Separation Based on Computational Auditory Scene Analysis

WANG Shan, XU Gang
(Dept. of Information Engineering, North China Electrical Power University, Beijing 102206)

【Abstract】This paper proposes a model of speech separation based on the principle of computational auditory scene analysis, to separate a mixture of two speeches. Different from other models, the model uses different methods to separate resolved and unresolved harmonics, so as to avoid the problem of inability to separate unresolved harmonics successfully. Experimental results show excellent performance of the model.

【Key words】computational auditory scene analysis(CASA); speech separation; pitch evaluation; amplitude modulation

盲信号处理及听觉场景分析是混叠语音分离的 2 种主要方法。在盲信号分离中，由于计算时需要对信号的统计特性和混叠方式进行假设，而在实际声学环境中，这些条件又常常难以满足，因此该方法具有局限性[1]。鉴于此，利用计算听觉场景分析(CASA)建立混叠语音分离的模型[2～4]是近年来诸多学者主要的研究目标，他们建立的模型都基于人类感知和组织声音机制中潜在的准则，使CASA在根本上区别于其他语音分离方法。

同以往的模型一样，本文采用听觉滤波器组对信号进行时频分析，并且提取其周期特性，作为下一步分组的基础。但以往模型分离高频部分的语音不是特别理想，因为这些系统采用了同样的方法来分离低频区和高频区的语音。心理声学的证据表明[5]，人类听觉系统对低频和高频区谐波是采用不同的感知机制来处理的，且低频区和高频区谐波对听者语音分离的作用也不同。因此，本文采用了计算听觉场景分析来分离混叠的语音信号，效果良好。

1　模型概述

混叠语音分离模型的构建是按照计算听觉场景分析准则的 3 个步骤来完成的，即分解、组合和合成，如图 1 所示，系统输入为两个说话者的混叠声音信号。

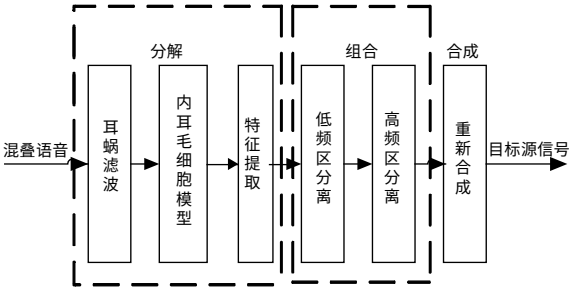


图 1　混叠语音分离模型示意图

(1)分解阶段，即语音听觉处理和特征提取。将输入信号通过一组耳蜗滤波器来模拟人耳的外围听觉处理，选择gammatone滤波器来模拟心理学记录的听觉神经纤维的冲激响应。本文使用中级听觉表示法[6]，通过计算混叠语音信号的相关图谱，对经外围听觉处理后的混合声音提取周期特性。

(2)组合阶段，即混叠的语音信号分解为两个独立的声音流。语音分离是本模型的核心。由于上一阶段产生的相关图谱挖掘出了耳蜗滤波器输出的周期特性，对基音估计是十分有效，因此选取基音信息作为线索对语音进行分类。但在高频区域，耳蜗滤波器较宽，在同一滤波器内许多谐波相互作用产生幅度调制(AM)，改变了自相关结构。如果对高频段也使用基音作为归类线索，则很难组合高频成分，导致高频区分离结果不理想。因此，可对高频区采用AM作为线索来进行语音分离[7]。最后对两个频段的分流进行合并，得到最终分流结果。

(3)重新合成阶段，即对听觉表示进行转化，然后重建分离后信号的时域波形。这样，就能用听音实验或者包含原始和重建信号的比较信息来评测分离后的信号。

在实验部分，采用英文和中文两部分共 60 个数据，分别用客观、主观评测法对模型的分离结果进行检测，并比较了分离前后的声音信号时域波形。为了进一步说明本模型的优越性，以Wang-Brown模型[3]为例，对分离前后的SegSNR和恢复语音能量百分比分别进行了比较，并用主观DRT测试[8]进一步说明分离后语音的可懂度。

2　混叠语音分离的实现

2.1　外围听觉处理和特征提取

通过一组通带重叠的带通滤波器来模拟外围听觉的频率

作者简介：王　珊(1982－)，女，硕士，主研方向：语音信号处理与分析；许　刚，教授

收稿日期：2006-09-30　**E-mail：**pixyss@gmail.com

选择特性已经得到了广泛的认同。本文使用 128 频段的 gammatone 滤波器组，中心频率在 80Hz～5 000Hz 之间呈准对数分布，滤波器参数的选择通常要与心理物理学数据相匹配，其将阶数选定为 4，与心理物理学与神经生理学的听觉频率选择性估计达到了很好的一致。

在得到混叠语音信号的二维时频表示后，以模拟听觉神经发放的方式提取以下的周期信息：

(1)相关图谱。该图谱是基音信息从低频和高频谐波区域相结合得到的一种基音模型。相关图谱在听觉外围和语音分离之间提供了一个有效的中级听觉，该听觉表示为

$$A_H(c,m,\tau)=\frac{1}{N_c}\sum_{n=0}^{N_c-1}h(c,mT-n)h(c,mT-n-\tau) \tag{1}$$

其中，延迟 $\tau\in[0,12.5\text{ms}]$；$h(c,n)$ 表示在时间阶 n 频段 c 的毛细胞输出；$T=160$ 对应 10ms。

(2)和相关图谱。对于周期声音而言，相关图谱中所有有效滤波的自相关会在相应于周期的延迟上达到峰值。对相关图谱按频段相加来强化该基音相关的结构，得到和相关图谱。

$$s(m,\tau)=\sum_c A_H(c,m,\tau) \tag{2}$$

定义主基音周期 $\tau_D(m)$ 对应有效基音范围[2ms,12.5ms]之内 $s(m,\tau)$ 的最大值的延迟。

(3)包络相关图谱。计算包络相关图谱可得到下一阶段将要用到的 AM 信息，即

$$A_E(c,m,\tau)=\frac{1}{N_c}\sum_{n=0}^{N_c-1}h_E(c,mT-n)h_E(c,mT-n-\tau) \tag{3}$$

其中，$h_E(c,n)$ 为时间 n 频段 c 的毛细胞输出包络。

(4)交叉频段互相关。由于相邻滤波频段的互相关可以表示该滤波频段是否属于同一声源，因此对同一时频单元，计算各频段间的互相关来对频段之间的相似分量进行量化。频段 c 和 $c+1$ 之间的交叉频段互相关的定义如下：

$$C_H(c,m)=\sum_{\tau=0}^{L-1}\hat{A}_H(c,m,\tau)\hat{A}_H(c+1,m,\tau) \tag{4}$$

其中，$\hat{A}_H(c,m,\tau)$ 表示 $A_H(c,m,\tau)$ 归一化得到的零均值、单位方差的自相关函数。

(5)交叉频段包络互相关。它用来表示 AM 之间的相似度，对判断在高频范围内两个频段是否对应同一周期源是一个很好的线索。

$$C_E(c,m)=\sum_{\tau=0}^{L-1}\hat{A}_E(c,m,\tau)\hat{A}_E(c+1,m,\tau) \tag{5}$$

其中，$\hat{A}_E(c,m,\tau)$ 表示归一化的 $A_E(c,m,\tau)$。

2.2 低频和高频区语音分离

由于相邻频段的通带有很大的重叠，低频谐波通常存在于相邻频段，这样会带来很高的交叉频段互相关。另外，由于浊音部分通常持续时间超过 50ms，因此依据瞬时连续性和交叉频段互相关，将时频单元进行合并形成段。首先，只考虑满足 $A_H(c,m,0)>\theta_H^2$ 和 $C_H(c,m)>\theta_c$ 的时频单元，其中，$\theta_H=50$ 接近于Meddis模型[9]中听觉神经发放率；$\theta_c=0.985$。高频区时频单元的交叉频段互相关非常小，因此，这样形成的段仅限于低频成分的语音。

如果一个时频单元 u_{cm} 满足式(6)，就称 u_{cm} 和 $\tau_D(m)$ 相符。

$$\frac{A_H(c,m,\tau_D(m))}{A_H(c,m,\tau_p(c,m))}>\theta_p \tag{6}$$

其中，$\theta_p=0.95$；$\tau_p(c,m)$ 为有效基音范围之内相应于 $A_H(c,m,\tau)$ 最大值的延迟。选取最长的段作为种子流，在一个给定的时间帧，一个段如果和最长的段同时和主基音相符或者不相符，那么就称该段和最长段相符。如果一个段内超过一半的重叠帧和最长段都相符，那么该段在最长段持续时间内的时频单元被组合成为种子流，否则被组合成竞争流；如果它在超过一半的时间帧内和主基音相符，那么它就有可能包含该主基音对应的声音源，因此，将包含最长段的流记为前景流 S_F^0，竞争流记为背景流 S_B^0。前景流对应的是目标声音源，背景流对应背景声音源。这样的分组成功地组合了大部分低频区的段。

主基音和目标源信号的真正基音在很多帧内还存在偏离，为了得到更准确的基音曲线，有必要将从前景流中得到的基音曲线进行修正，仅对 S_F^0 中的时频单元，计算和相关图谱 $s_F(m,\tau)$，$\tau_s(m)$ 就是有效基音范围内 $s_F(m,\tau)$ 的最大值。由于 S_F^0 也包含一些背景声音的成分，因此 $\tau_s(m)$ 不是十分准确，其准确度需要进一步检验。一个准确的基音周期应该和各个组成单元的周期一致，如果一个估计的基音周期是可信的，前景流相应帧中至少有一半单元和它相符，即应满足式(7)。

$$\frac{A_H(c,m,\tau_s(m))}{A_H(c,m,\tau_p(c,m))}>\theta_p \tag{7}$$

以上是低频区的语音分离，该分离将混叠语音信号分解为单独的两个声音流。对于高频区，由于存在幅度调制现象，上面的周期准则并不适合。由于对应于高频谐波的滤波器响应包络以声源 $F0$ 的速率波动，因此本文以 AM 为线索来分离高频区语音。

2.3 重新合成

将每一个滤波器的输出再次通过 gammatone 滤波器来去除交叉频段相位误差，将 gammatone 滤波器的输出按 20ms 的长度进行分块，每两个块间有 10ms 的重叠。若目标声音信号强于背景声音信号，权值为 1；反之，权值为 0，加权的输出求和即可得到重新合成的语音信号。

3 实验结果与分析

本次实验将语音分为英文和中文两部分，采样频率为 16kHz，大小为 16bit。为了便于与其他分离模型相比较，英文部分选择Cooke[10]采集的已经用来测试过其他一些CASA系统的 100 个混合声音中的 30 个，这些混合声音是通过将 10 个浊音发声的波形与 3 个背景发声波形分别相叠加而形成的，其中的 10 个浊音发声由两名男性说话者S1、S2 分别朗读 5 个句子，背景说话者为一名男性说话者N1 和两名女性说话者N2、N3。中文部分也采用 30 个混合声音，由两名女性说话者S3、S4 分别朗读 5 个词，背景说话者为两名男性说话者N4、N5 和一名女性说话者N6。图 2 为S1 朗读的“ I'll willingly marry Marilyn ”与N1 混合声音波形形式的分离结果，其中，图 2(a)为干净的前景说话者声音，图 2(b)为分离后的声音，可知，分离出的声音波形和干净的声音波形非常相近。最后分别采用客观和主观评估方法对本模型语音分离性能进行评估。

客观测度是指原始语音和分离后语音之间的距离，使用分段信噪比SegSNR进行分离语音的评估。图 3 显示了测试的结果。同一个背景说话人条件下，对每一个前景说话人计算得出的SegSNR比较相近，因此，图中显示的是在同一背景说话者条件下每一前景说话者SegSNR的平均值。在每一个背景说话者条件下，分离之后的SegSNR都提高了。为进

一步体现本模型的性能，可与Wang-Brown模型[3]进行比较，结果如图 3 所示，可知，对每一种背景说话者，本模型分离后的SegSNR较之Wang-Brown模型平均提高 5.3dB。

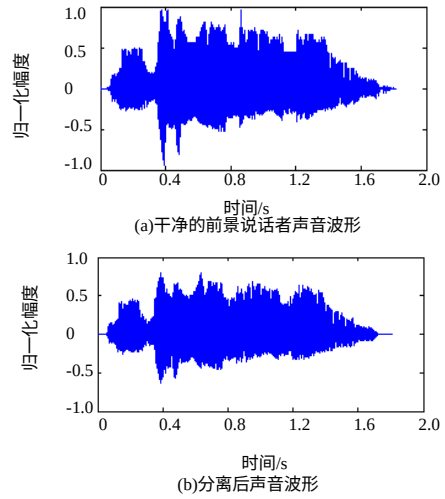


图 2　混合声音波形分离结果

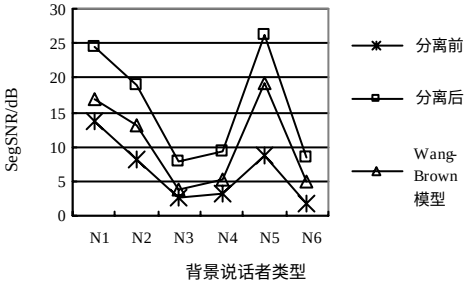


图 3　分离前后 SegSNR 以及与 Wang-Brown 模型的比较

本文同样计算了从每一混合声音中恢复的语音能量百分比，并且与 Wang-Brown 模型进行了比较，如图 4 所示。该百分比也是在同一背景说话者条件下对每一前景说话者求得的平均值，可知，较之 Wang-Brown 模型，本模型在分离之后恢复的语音能量百分比对每一背景说话者情况下都有一定的提高，平均值达到 80.5%。

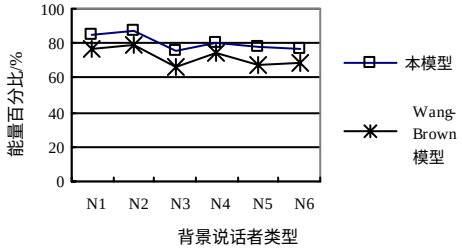


图 4　恢复的语音能量百分比以及与 Wang-Brown 模型比较

除了客观的评估方法，本文也采用了主观评估方法来进一步说明分离语音信号的可懂度高低。该方法是DRT(diagnostic rhyme test)测试法：选择 DRT 单词表中的 veal 和 bean 两单词分别作为目标语音，由一名男性朗读，并分别以一名女性朗读的 sense 为背景相混合的声音进行分离，将原始语音与分离后语音分别进行 DRT 单词对 veal-feal, bean-pean 的 DRT 测试，单词对由同一人朗读；选用测试者 20 人，10 名男性 10 名女性，测试在 79dB SPL 较为舒适的声压级环境下进行；在原始语音和分离后语音两种情况下，由测试者分别在听到的单词对中对两个单词进行判断，根据测试者判断的对错由式(8)打出分数。

$$DRT = \frac{N_{正确} - N}{\mathrm{N}} \times 100\% \qquad (8)$$

最终可懂度可按所求得的分数来衡量，表 1 就是对 20 名测试者在两种情况下进行的两组 DRT 测试的分数，可见对分离后语音进行 DRT 测试的分数与原始语音相接近，经本模型分离后的语音可懂度还是较高的。

表 1　两种情况下两组 DRT 测试结果

DGR 测试内容	得分
原始 veal-feal	90
分离后 veal-feal	85
原始 bean-pean	95
分离后 bean-pean	85

4　结束语

本模型和过去一些 CASA 模型类似，即利用瞬时连续性和谐波线索来分离语音，不同的是采用不同的线索来分离高频段和低频段的语音，并且从一个分离语音流中估计目标基音，这也可以看作是一个反复进行基音估计的过程，目的是确保得到更加准确的基音信息。通过和 Wang-Brown 模型的对比，可以看出无论是在分离后的信噪比还是所能恢复的语音能量百分比都比 Wang-Brown 模型有一定的提高，同时从客观的评测和主观的听音实验，本模型都能达到很好的分离效果。但本文所能实现的两个说话者声音分离仅仅局限于浊音，对于混叠清音的分离是今后将进一步研究的问题。

参考文献

1 Van der Kouwe J W, Wang D L, Brown G L. A Comparison of Auditory and Blind Separation Techniques for Speech Segregation[J]. IEEE Trans. on Speech Audio Processing, 2001, 9(3): 189-195.

2 Roman N, Wang D L. Binaural Sound Segregation for Multisource Reverberant Environment[C]//Proc. of Int'l Conference on Acoustics, Speech, and Signal Processing. 2004: 373-376.

3 Wang D L, Brown G L. Separation of Speech from Interfering Sounds Based on Oscillatory Correlation[J]. IEEE Trans. on Neural Networks, 1999, 10(3): 684-697.

4 赵鹤鸣, 朱美虹, 俞一彪. 一种适于计算声场景分析的混叠语音基音检测方法[J]. 电子学报, 2003, 31(1): 4-5.

5 Carlyon R P, Shackleton T M. Comparing the Fundamental Frequencies of Resolved and Unresolved Harmonics: Evidence for Two Pitch Mechanisms?[J]. Journal of the Acoustic Society of America, 1994, 95(6): 3541-3554.

6 Ellis D P W, Rosenthal D. Mid-level Representations for Computational Auditory Scene Analysis: The Weft Element[C]//Proc. of Int'l Joint Conference on Artificial Intelligence. Mahwah, NJ: Lawrence Erlbaum,1998.

7 Hu G, Wang D L. Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation[J]. IEEE Trans. on Neural Networks, 2004, 15(5): 1135-1150.

8 Voiers W D. Evaluating Processed Speech Using the Diagnostic Rhyme Test[J]. Speech Technology, 1983, 1(4): 30-39.

9 Meddis R. Simulation of Auditory-neural Transduction: Further Studies[J] Journal of the Acoustic Society of America, 1988, 83(3): 1056-1063.

10 Cooke M P. Modeling Auditory Processing and Organization[D]. CS Dept., Univ. of Sheffield, 1991.