基于多基音跟踪的单声道混合语音分离*

李 鹏[a], 关 勇[b], 刘文举[b], 徐 波[a,b]
(中国科学院 自动化研究所 a.数字媒体内容技术研究中心; b.模式识别国家重点实验室, 北京 100080)

摘 要: 针对许多计算听觉场景分析系统无法很好地解决多说话人混合语音信号分离的问题,提出了一种基于多基音跟踪的单声道混合语音分离系统。该系统充分利用了多基音跟踪研究的最新成果,通过将多基音跟踪得到的目标语音和干扰语音的基音轨迹信息结合到分离系统中,有效地改善了分离系统在包括多说话人混合在内的多种干扰情况下的分离效果,为多说话人语音分离问题的解决提供了新的思路。

关键词: 计算听觉场景分析; 多基音跟踪; 语音分离
中图分类号: TN912.3 **文献标志码**: A **文章编号**: 1001-3695(2008)06-1660-03

Monaural speech separation based on multi-pitch tracking

LI Peng[a], GUAN Yong[b], LIU Wen-ju[b], XU Bo[a,b]
(*a. Digital Media Content Technology Research Center, b. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: This paper proposed a multi-pitch tracking based monaural speech separation system to solve the problem that many computational auditory scene analysis(CASA) systems could not process the mixture signal of multi-speaker. The proposed system made use of the latest achievement in the field of multi-pitch tracking. Through combing the pitch contours of target speech and interference acquired by multi-pitch tracking algorithm into the CASA system, improved obviously the performance of separation system for various mixture conditions, including two-speaker mixture condition, which offered a good solution to separate the mixture speech of multi-speaker.

Key words: computational auditory scene analysis(CASA); multi-pitch tracking; speech separation

0 引言

在语音信号处理中,一个重要的问题就是如何从混合语音信号中分离出人们感兴趣的语音。这方面的研究大体上集中在盲源分离(BSS)及计算听觉场景分析(CASA)两方面。其在语音识别、多媒体检索以及语音增强等领域都有着重要的意义[1,2]。

计算听觉场景分析对人类听觉的处理过程进行建模,从而使计算机具备从混合声音中分离出各物理声源并作出合理解释的能力。它的出现大大激发了人们对人类听觉系统研究的兴趣,使长期以来一直困扰研究人员的技术难题(如语音识别系统在复杂现实环境中的应用)有了突破的可能。

近年来,基于计算听觉场景分析的混合语音分离研究取得了快速的发展,相继研制出了许多具有不同特色的分离系统。虽然这些系统在许多噪声情况下具有很好的分离性能,但是对于多个说话人语音混合的情况,系统的性能并不令人满意[2~4]。针对这一情况,本文提出了一种结合了多基音跟踪算法的单声道混合语音分离系统,提高了系统从多说话人混合语音中分离出目标语音的能力,并给出了详细的评估结果。

1 多基音跟踪算法介绍

在多说话人语音混合的情况下,混合语音中存在多个基音,因此如果能够准确地提取出每个说话人的基音,并利用提取出的基音对各说话人的语音进行组织的话,将有助于提高分离系统的性能。为此本文采用 Wu 等人[5]提出的多基音跟踪算法与 CASA 系统进行结合来提高 CASA 系统的分离性能。选择该算法主要基于以下两方面的原因:

a)算法具有很好的跟踪性能,能够从具有多个基音的混合语音中比较准确地估计出其中的基音个数以及相应的基音轨迹。这一特点对于所提取基音的准确性对系统性能具有重要影响的 CASA 系统而言,有非常重要的意义。

b)由于算法在预处理阶段采用了与许多 CASA 系统类似的处理方式[2~4],使用该算法进行多基音跟踪时,可以充分利用 CASA 系统预处理的结果,减少因算法引入带来的计算和资源消耗。

多基音跟踪算法由四个阶段组成,如图1所示。算法的第一个阶段是前端处理阶段,该阶段首先使用一组听觉感知模型在各个通道内对信号进行滤波,提取出高频通道内滤波后的信号包络,然后计算归一化的相关图[3,4]。

收稿日期: 2007-05-09; **修回日期**: 2007-08-11 **基金项目**: 国家"973"计划资助项目(2004CB318105)

作者简介:李鹏(1978-),男,河南洛阳人,博士研究生,主要研究方向为语音信号处理、计算听觉场景分析、语音识别(pengli@hitic.ia.ac.cn);关勇(1980-),男,博士研究生,主要研究方向为计算听觉场景分析、说话人识别;刘文举(1960-),男,副研究员,博导,博士,主要研究方向为语音识别与合成、说话人识别、语音增强、计算听觉场景分析;徐波(1966-),男,研究员,博导,博士,主要研究方向为多媒体内容管理、语音信号处理、语音识别与合成、统计机器翻译.

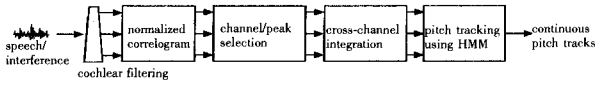

图 1　多基音跟踪算法结构图

算法的第二阶段由通道选择和峰值选择构成。对于带噪语音,部分通道受噪声干扰比较明显。通过选择那些受影响较小的通道,可以明显改善系统的鲁棒性。通常算法中的通道选择是在中心频率高于 1 270 Hz 的中、高频通道内进行的。Wu 等人将通道选择的思想扩展到了低频通道,提出了一种对所有通道进行选择的改进方法。此外,考虑到归一化相关图中的峰值揭示了信号的周期性,Wu 等人还在算法中引入了峰值选择来去除那些所给出的周期信息并不能反映信号真实周期的峰值。

算法的第三个阶段是通道周期信息结合。传统的在一个时间帧内对所有通道的自相关(或归一化自相关)进行累加的周期信息结合方法虽然实现简单,但是包含在通道中的周期性信息并没有被充分利用。通过对真实基音周期与峰值选择阶段所选定的峰值时间延迟之间的统计关系进行研究,Wu 等人用公式描述了通道支持某一基音假设的概率,并采用一个统计结合的方法来产生给定假设基音条件下观测信号在某一时间帧内的条件概率。

多基音跟踪算法的最后一个阶段是使用隐马尔可夫模型(HMM)形成连续的基音轨迹。Weintraub[6] 使用 HMM 来决定信号中究竟出现了 0 个、1 个还是 2 个基音。Gu 等人[7] 使用 HMM 来组织自下而上的基音确定算法所提出的基音候选,并形成连续的基音轨迹。Tokuda 等人[8] 基于多空间概率分布,利用 HMM 对基音模式进行建模。在这些研究中,基音均被视为观测量,因此 HMM 的转移概率和观察概率都必须进行训练。而在 Wu 等人提出的多基音跟踪算法中,基音被明确地建模为隐藏状态,因此只需从自然语音中提取基音的统计特性来确定转移概率,然后利用 Viterbi 算法就可以获得最优的基音轨迹。有关多基音跟踪算法的详细介绍参见文献[5]。

2　多基音跟踪与 CASA 系统的结合

2.1　系统描述

多基音跟踪算法与 CASA 系统的结合如图 2 所示。与 Hu-Wang[2] 系统类似,基于多基音跟踪的单声道语音分离系统也由分解和特征提取、初始分离、基音跟踪和时频单元标记、最终分离以及再合成五个阶段组成。混合信号在经过前端预处理后进入多基音跟踪模块。经过多基音跟踪模块的处理,得到目标语音和干扰的基音轨迹。这些基音轨迹接下来被结合到初始分离阶段中,并被用来指导初始分离的进行。

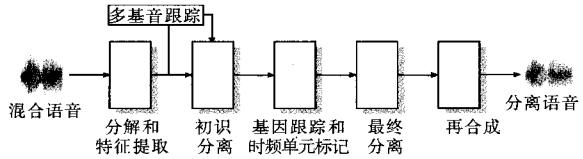

图 2　基于多基音跟踪的单声道混合语音分离系统框图

Hu-Wang 系统在初始分离阶段首先从滤波后的信号中估计出一个大致的全局基音轮廓;然后利用这一估计出的粗糙的基音轮廓对初始切分所形成的片段进行分组;之后,系统再进一步从分组得到的前景流中估计更可靠的基音轨迹,并利用这一估计出来的相对更可靠的基音轨迹对语音片段重新进行分组,从而为后续的处理提供可靠的前景流和背景流[2]。

与 Hu-Wang 系统不同,本文提出的分离系统已经通过多基音跟踪算法获得了有关目标语音和干扰的相对可靠的基音轨迹。因此在初始分离阶段,系统可以直接利用多基音跟踪的结果,对初始切分得到的语音片段进行分组,而无须再像 Hu-Wang 系统那样通过复杂的处理来不断地组织和调整前景流与背景流。

需要强调的是,由于多基音跟踪算法可以同时给出语音中的多个基音轨迹,在使用多基音跟踪的结果时,需要首先从算法估计出的多个基音轨迹中确认出目标语音的基音轨迹。考虑到本研究所要分离的目标语音都由浊音组成,因此可以很容易地在跟踪得到的多个基音轨迹中,选择其中连续的基音轨迹作为目标语音的基音轨迹。确定了目标语音的基音轨迹后,就可以利用它来进行前景流和背景流的划分了。

在使用多基音跟踪的结果对前景流和背景流进行组织时,除了目标语音的基音轨迹可以被充分利用外,干扰的基音轨迹也可以被利用。事实上,干扰基音轨迹的引入为前景流和背景流的划分提供了更多的线索,有助于提高两者的分组准确性。为此,本文对前景流与背景流的组织方法进行了相应的改进。改进后的方法不仅考虑了目标语音的基音与时—频单元的符合程度,还考虑了干扰语音的基音与时—频单元的符合程度。

2.2　结合方式

具体而言,对于前景流和背景流的分组是通过比较时—频单元的响应与目标语音以及干扰的基音周期来完成的。假设多基音跟踪算法估计的目标语音的第 m 帧的基音周期为 $\tau_s(m)$,干扰噪声在第 m 帧的基音周期为 $\tau_N(m)$,混合语音在通道 c 内相应的自相关函数为 $A_H(c,m,\tau)$,那么,时—频单元 u_{cm} 的响应周期与目标语音和干扰的基音周期的比较可以按照如下步骤进行:

a)如果时—频单元响应的周期与对应的目标语音基音周期相当,则该时—频单元与目标语音的基音相符合。也就是说,如果在似真基音范围内,$A_H(c,m,\tau_s(m))$ 与 $A_H(c,m,\tau)$ 的最大值满足

$$A_H(c,m,\tau_s(m))/A_H(c,m,\tau_p(c,m)) > \theta_p \tag{1}$$

则 u_{cm} 与 $\tau_s(m)$ 相符合;否则,继续 b)中的比较。

b)如果时—频单元响应的周期与干扰的基音周期相当,则时—频单元与干扰基音相符合。也即,如果在似真基音范围内,$A_H(c,m,\tau_N(m))$ 与 $A_H(c,m,\tau)$ 的最大值满足

$$A_H(c,m,\tau_N(m))/A_H(c,m,\tau_p(c,m)) > \theta_p \tag{2}$$

则 u_{cm} 与 $\tau_N(m)$ 相符合;否则,继续 c)中的比较。

c)如果时—频单元响应周期满足

$$A_H(c,m,\tau_s(m))/A_H(c,m,\tau_p(c,m)) > \theta'_p \tag{3}$$

且

$$A_H(c,m,\tau_s(m))/A_H(c,m,\tau_p(c,m)) > (A_H(c,m,\tau_N(m))/A_H(c,m,\tau_p(c,m)) + \theta_B) \tag{4}$$

则时—频单元与目标语音的基音相符合,即单元 u_{cm} 与 $\tau_s(m)$

相符合;否则,单元 u_{cm} 不符合 $\tau_s(m)$。

上述比较中,$\theta_p=0.95$,$\theta'_p=0.75$,$\theta_B=0.1$;$\tau_p(c,m)$ 为在似真基音周期[2 ms,12.5 ms]内,使 $A_H(c,m,\tau)$ 取最大值所对应的延时。

完成上述比较后,可以根据比较的结果按照如下方法对语音片段进行分组:对于初始切分形成的任意一个语音片段,如果其中某一帧内超过一半的时—频单元与该帧的目标语音的基音相符合,则称该片段在这一帧上与目标语音的基音相符合。由于本研究中的目标语音全部是浊音,在切分形成的语音片段中,可以选择最长的片段作为种子流(seed stream)。在某一帧内,如果某个片段与最长的片段同时符合或同时不符合目标基音的话,则称这一片段与最长的片段在该帧内相符合。如果某个片段与最长的片段在两者交叠的帧内有一半以上的帧相符合,那么该片段在最长片段的持续时间内的所有时—频单元被分组到种子流中;否则,该片段被分组到竞争流中。最长的片段也被用来确定哪一个流对应目标语音。如果它有超过一半的帧与目标语音的基音相符合,那么它将非常可能包含了主要的目标语音。在这种情况下,将包含有最长片段的流视为前景流,记做 S_F^0;而将竞争流视为背景流,记为 S_B^0。否则,将上述两个流的名称互换。

上述处理完成后,系统将按照与 Hu-Wang 模型相同的处理方法对混合语音进行进一步的处理,最终形成分离出的目标语音[3]。

3 评估与比较

本文使用了英国谢菲尔德大学 Cooke 搜集的 100 句混合语音数据集[9]对系统进行了评估。所使用的数据集由 10 句浊音句子与 10 种不同干扰噪声组成,它被广泛用于 CASA 系统的性能评估[2~4]。其中,10 种干扰噪声分别是:a) N0,1 kHz 纯音;b) N1,白噪声;c) N2,突发噪声;d) N3,鸡尾酒会噪声;e) N4,摇滚乐;f) N5,警报声;g) N6,电话颤音;h) N7,女说话人语音;i) N8,男说话人语音;j) N9,女说话人语音。这里使用信噪比(SNR)作为标准量化评估所提出的分离系统的性能。为了检测分离前后语音的信噪比,使用混合前的目标语音作为纯净语音计算分离前语音的信噪比。为了补偿合成过程中幅度和失真的影响,目标语音进行全 1 掩蔽[2,3]后的合成语音被用来作为纯净语音计算分离后语音的信噪比。

此外,为了明确本文所提出的基于多基音跟踪的单声道混合语音分离系统相比于其他分离系统的性能,笔者还将系统的分离结果与在 Hu-Wang 系统中使用真实基音的 true pitch 系统以及使用理想二值掩蔽的 ideal mask 系统进行了比较[2,3]。

表 1 给出了不同干扰情况下原始混合语音的信噪比以及所提出的系统、true pitch 系统和理想二值掩蔽(ideal mask)系统所得分离语音的信噪比。其中,最后一行给出了各种噪声条件下的平均信噪比。从表中可以看出,本文提出的基于多基音跟踪的分离系统分离语音的信噪比相比原始混合语音在所有干扰条件下均得到了明显的改善,平均信噪比提高约为 10.65 dB。特别地,对于两说话人的情况(N7、N8 和 N9),系统分离后的语音信噪比也得到了明显的提高。另外,本文提出的系统与 true pitch 系统分离结果的平均信噪比 11.508 dB 相比仅相差 0.444 dB,这表明系统在性能上已经非常接近使用基音作为分离线索的分离方法的上限。但是相比以二值掩蔽思想为基础的分离方法的上限——ideal mask 方法的平均信噪比 14.571 dB,系统在性能上还有一定的上升空间。

表 1 SNR 结果

SNR	mixture	multi-pitch	true pitch	ideal mask
N0	-7.418	15.973	16.140	20.043
N1	-8.272	5.288	5.825	7.213
N2	5.616	14.655	14.740	18.459
N3	0.800	5.756	6.366	8.123
N4	0.676	8.536	9.338	11.562
N5	-9.999	14.298	14.307	17.37
N6	-1.619	15.373	15.576	19.896
N7	3.844	10.017	10.808	14.158
N8	9.527	13.897	14.522	17.679
N9	2.745	6.843	7.461	11.21
ave	-0.410	11.064	11.508	14.571

上述评估和比较的结果证明:使用多基音跟踪算法在解决多混合语音的分离,特别是说话人混合语音分离问题上是行之有效的。

4 结束语

本文提出了一种基于多基音跟踪的单声道混合语音分离系统。该系统采用多基音跟踪算法对混合语音中出现的多个基音进行估计,并将提取出来的多个基音的轨迹一同作为分离线索结合到计算听觉场景分析系统中以指导分离。对系统的评估结果表明,该系统能够很好地处理不同干扰条件下的语音分离问题。特别是对于多个说话人语音混合的情况,该系统能够明显提高分离后语音的信噪比,因而为多说话人语音分离研究提供了很好的解决思路。

参考文献:

[1] DENBIGH P N, ZHAO J. Pitch extraction and separation of overlapping speech[J]. Speech Communication, 1992, 11(2-3): 119-125.

[2] LI Peng, GUAN Yong, XU Bo, *et al*. Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech[J]. IEEE Trans on Audio, Speech, and Language Processing, 2006, 14(6): 2014-2023.

[3] HU Guo-ling, WANG De-liang. Monaural speech segregation based on pitch tracking and amplitude modulation[J]. IEEE Trans on Neural Networks, 2004, 15(5): 1135-1150.

[4] WANG De-liang, BROWN G J. Separation of speech from interfering sounds based on oscillatory correlation[J]. IEEE Trans on Neural Networks, 1999, 10(3): 684-697.

[5] WU Ming-yang, WANG De-liang, BROWN G J. A multi-pitch tracking algorithm for noisy speech[J]. IEEE Trans on Speech and Audio Processing, 2003, 11(3): 229-241.

[6] WEINTRAUB M. A computational model for separating two simultaneous talkers[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Tokyo: [s. n.], 1986: 81-84.

[7] GU Y, BOKHOVEN W M G van. Co-channel speech separation using frequency bin nonlinear adaptive filter[C]//Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington DC: IEEE Computer Society 1991: 949-952.

[8] TOKUDA K, MASUKO T, MIYAZAKI N. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling[C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington DC: IEEE Computer Society 1999: 229-232.

[9] COOKE M. Modeling auditory processing and organization[D]. Sheffield: University of Sheffield, 1991.