重叠语音的帧同步分离研究

戴礼荣, 宋 彦, 王仁华
(中国科学技术大学电子工程与信息科学系, 合肥 230027)

摘 要: 重叠语音在本文中定义为由两个及两个以上原始语音叠加所形成的混合语音. 重叠语音分离, 由于不能利用方位信息, 困难更大. 但人的听觉系统却能够有效感知和分离这种重叠语音. 本文利用语音的谐波特性对重叠语音分离进行研究. 提出基于语音信号谐波分析的帧同步分离方法, 并给出实验结果.

关键词: 语音分离; 基音频率; 谐波分析
中图分类号: TN391.42 文献标识码: A 文章编号: 0372-2112 (2002) 10-1552-03

A Study on the Frame Synchronized Segregation of Overlapped Speech

DAI Li-Rong, SONG Yan, WANG Ren-Hua
(Dept. Of Electronic Engineering & Information Science, University of Science & Technology of China, HeFei, Anhui 230027, China)

Abstract: In this article, the overlapped speech is defined as the hybrid of two or more original speeches. Because there is not any direction information, it is more difficult to segregate the overlapped speech. But the human auditory system can successfully perceive and segregate this kind of overlapped speech. This paper addresses the problem of the overlapped speech segregation based on the speech harmonicity. A frame synchronized segregation method based on harmonic analysis is proposed and some corresponding experimental results are given.

Key words: speech segregation; fundamental frequency; harmonic analysis

1 引言

在多个背景语音的嘈杂环境下, 人的听觉系统具有分离及提取感兴趣的语音信号的能力. 尽管在这种分离中人们利用了方位信息, 但人的听觉系统信号分离功能不完全依赖方位信息. 事实上, 人的听觉系统对于在同一信道(如广播, 电视等) 中传输的两个重叠语音仍具有分离的能力. 本文研究的重叠语音分离不同于根据信号盲分离理论[1] 的语音分离, 在信号盲分离理论中, 主要利用了有关信号源的方位信息.

当一个语音信号是由两个及两个以上的语音信号叠加所形成的混合语音信号时, 在本文中称为重叠语音. 对于语音识别及其它语音通信应用, 重叠语音的实时分离具有重要意义. 重叠语音分离的难点一方面是重叠信号占据相同的频带, 另一方面不能利用信号源的方位信息. 唯一可利用的信息可能是同一信号源产生的语音信号具有某种共同的性质. 如, Bregman[2] 指出, 语音信号的谐波特性是人的听觉系统对重叠语音信号进行分离的一个重要依据. 据此, 有些学者, 对这个问题进行了研究. 如, M. Unoki[3] 等基于语音信号的调频-调幅(AM-FM) 模型通过将谐波特性等作为模型求解的约束条件进行了这方面的研究. 重叠语音信号分离是一个非常困难的问题, 目前这方面的研究尚没有很好的成果[4]. 本文将基于语音信号谐波分析法对重叠语音分离进行研究. 所提出的分离算法是按帧实现的, 即帧同步分离, 非常适合按帧处理的语音识别前端. 文中给出的实验结果说明了分离方法的有效性.

2 基于谐波分析的重叠语音分离原理

设语音信号 $s(n)$, 是由语音信号 $s_1(n)$ 和语音信号 $s_2(n)$ 迭加而成, 即

$$s(n) = s_1(n) + s_2(n) \tag{1}$$

若 $s_1(n)$ 和 $s_2(n)$ 均具有谐波结构, 则对应的分离信号 $\tilde{s}_1(n)$, $\tilde{s}_2(n)$ 可表示为若干谐波的叠加, 采用复数表示形式为:

$$\tilde{s}_1(n) = \sum_{q=1}^{Q} A^1(q\omega_0^1)\, e^{j(q\omega_0^1 n + \Phi_q^1)} \tag{2}$$

$$\tilde{s}_2(n) = \sum_{p=1}^{P} A^2(p\omega_0^2)\, e^{j(p\omega_0^2 n + \Phi_p^2)} \tag{3}$$

其中 ω_0^1, ω_0^2 是基音频率, $A^1(q\omega_0^1)|_{q=1}^{Q}$, $A^2(p\omega_0^2)|_{p=1}^{P}$ 是谐波幅度, $\Phi_q^1|_{q=1}^{Q}$, $\Phi_p^2|_{p=1}^{P}$ 是谐波相位, Q, P 分别是谐波数, 等于信号带宽除基音频率. 由于 $s(n)$ 是 $s_1(n)$ 和 $s_2(n)$ 的迭加, $s(n)$ 则可表示为若干正弦信号的迭加(事实上, 由 K-L 变换分析表明[4], 语音信号可有效地表示为正弦信号的迭加, 只要正弦信号的频率间隔小于一定值, 如 100Hz). 采用复数表示形式为:

$$s(n) = \sum_{l=1}^{L} A_l e^{j(\omega_l n + \Phi_l)} \tag{4}$$

正弦信号参数 (A_l, ω_l, Φ_l) 可通过检测 $s(n)$ 的幅度谱峰点确定 ($l = 1, 2, \cdots, L$; L 为正弦信号数, 对于带宽为 300Hz–3400Hz 的语音信号, 实验表明, 一般取 $L_{max} = 60$ 足够). 按式 (2), (3) 从 $s(n)$ 分离 $s_1(n)$ 和 $s_2(n)$ 的均方误差 MSE 定义为:

$$MSE = 1/(N+1) \sum_{n=-N/2}^{n=N/2} |s(n) - \tilde{s}_1(n) - \tilde{s}_2(n)|^2 \tag{5}$$

其中, N 为语音帧的帧长, 在已知 $s(n)$ 是两个具有谐波特性的 $s_1(n)$ 和 $s_2(n)$ 的叠加的条件下, 从 $s(n)$ 中分离 $s_1(n)$ 和 $s_2(n)$ 问题可看作是选择参数 $[\omega_0^1, \omega_0^2, A^1(q\omega_0^1)|_{q=1}^{Q}, A^2(p\omega_0^2)|_{p=1}^{P}, \Phi_q^1|_{q=1}^{Q}, \Phi_p^2|_{p=1}^{P}]$, 使式(5) 的 MSE 最小化.

收稿日期: 2001-02-13; 修回日期: 2001-12-25

3 分离算法

定义长度为 N 的序列 $s(n)$ 的付里叶变换为:

$$S(\omega)=1/(N+1)\sum_{n=-N/2}^{n=N/2}s(n)\,e^{j\omega n} \tag{6}$$

结合式(2), (3), (4) 可得

$$S(\omega)=\sum_{l=1}^{L}A_l e^{j\Phi_l}D(\omega-\omega_l) \tag{7}$$

$$\tilde{s}_1(\omega)=\sum_{q=1}^{Q}A^1(q\omega_0^1)\,e^{j\Phi_q^1}D(\omega-q\omega_0^1) \tag{8}$$

$$\tilde{s}_2(\omega)=\sum_{p=1}^{P}A^2(p\omega_0^2)\,e^{j\Phi_p^2}D(\omega-p\omega_0^2) \tag{9}$$

其中, $D(\omega-\omega_l)=W_R(\omega-\omega_l)$, $W_R(\omega)=\sin[(N+1)\omega/2]/[(N+1)\sin\omega/2]$; 如果语音信号在进行付里叶变换前, 采用 Hamming 加窗, 则

$$D(\omega-\omega_l)\approx\begin{cases}0.54W_R(\omega-\omega_l)+0.23[W_R(\omega-\omega_l-2\pi/N)\\ \quad +W_R(\omega-\omega_l+2\pi/N)],\ |(\omega-\omega_l)|<4\pi/N\\ 0,\ \text{other}\end{cases} \tag{10}$$

将以上结果代入式(5) 并作适当近似后得:

$$\begin{aligned}MSE\approx&\sum_{l=1}^{L}|A_l|^2+\sum_{q=1}^{Q}|A^1(q\omega_0^1)|^2+\sum_{p=1}^{P}|A^2(p\omega_0^2)|^2\\&-2\mathrm{Re}\sum_{q=1}^{Q}A^1(q\omega_0^1)\,e^{j\Phi_q^1}\sum_{l=1}^{L}A_l^*\,e^{-j\Phi_l}D(q\omega_0^1-\omega_l)\\&-2\mathrm{Re}\sum_{p=1}^{P}A^2(p\omega_0^2)\,e^{j\Phi_p^2}\sum_{l=1}^{L}A_l^*\,e^{-j\Phi_l}D(p\omega_0^2-\omega_l)\\&-2\mathrm{Re}\sum_{p=1}^{P}A^2(p\omega_0^2)\,e^{j\Phi_q^2}\sum_{q=1}^{Q}A_l^*\,(q\omega_0^1)\,e^{-j\Phi_q^1}D(p\omega_0^2-q\omega_0^1)\end{aligned} \tag{11}$$

由于 ω_l 是 $S(\omega)$ 的幅度谱峰值频率点, 当采用 Hamming 窗时, 可认为谱峰间隔满足: $\omega_{l+1}-\omega_l>4\pi/(N+1)$, $l=1,\ldots,(L-1)$. 结合式(10), 则式(11) 右边的第四项中和第五项中的第二个 L 项求和, 对 MSE 求和分别只有一项有贡献, 即满足下式的项:

$$\begin{aligned}&|q\omega_0^1-\omega_{l_q}|<4\pi/(N+1),\ l_q\in[1,\ldots,L]\text{及}|p\omega_0^2-\omega_{l_p}|\\&\quad<4\pi/(N+1),\ l_p\in[1,\ldots,L]\end{aligned} \tag{12}$$

同时, 对于式(11) 右边的第六项中的第二个求和式, 由于一般地满足: $\omega_0^1>4\pi/(N+1)$ 及 $\omega_0^2>4\pi/(N+1)$. 所以对于固定 p 值, 第六项中的第二个求和式中至多只有一项对求和有贡献, 记为 q_p. 首先, 考虑选择相位参数使式(11) 最优化:

(1) 如果, 对于给定的 p, 存在 q 满足:

$$|p\omega_0^2-q\omega_0^1|<4\pi/(N+1) \tag{13}$$

记为 q_p, 其中, $p\in[1,\ldots,P]$, $q\in[1,\ldots,Q]$, 则取: $\Phi_p^2=\Phi_{l_p}=\Phi_{q_p}^2\approx\Phi_q^1=\Phi_{l_q}$. 显然, 满足这一条件的 p 和 q 是一一对应的, 记为: $p_i=p_1, p_2,\ldots,p_I$, $q_i=q_1, q_2,\ldots,q_I$. 满足此条件的 $s_1(n)$ 和 $s_2(n)$ 的正弦分量称为重叠分量.

(2) 如果, 对于给定的 p, 不存在 q 满足(13) 式, 则式(11) 的最后一项等于零, 此时的 p 和 q 记为: $p'_j=p'_1, p'_2,\ldots,p'_{PJ}$, $q'_j=q'_1, q'_2,\ldots,q'_{QJ}$; 此时, 取: $\Phi_q^1=\Phi_{l_q}$, $\Phi_p^2=\Phi_{l_p}$.

利用以上结果, 在给定 ω_0^1, ω_0^2 时, 当取: $A^1(q'_j\omega_0^1)=A_{l_{q'_j}}D(q'_j\omega_0^1-\omega_{l,q'_j})$, $j=1,\ldots,QJ$; $A^2(p'_j\omega_0^2)=A_{l_{p'_j}}D(p'_j\omega_0^2-\omega_{l,p'_j})$, $j=1,\ldots,PJ$; 及当 $q_i\omega_0^1\neq p_i\omega_0^2$ 时, 取

$$A^1(q_i\omega_0^1)=\frac{A_{l,q_i}D(q_i\omega_0^1-\omega_{l,q_i})+A_{l,p_i}D(p_i\omega_0^2-\omega_{l,p_i})\,D(p_i\omega_0^2-q_i\omega_0^1)}{[1-D^2(q_i\omega_0^1-p_i\omega_0^2)]},\quad i=1,2,\ldots,I$$

$$A^2(q_i\omega_0^2)=\frac{A_{l,p_i}D(p_i\omega_0^2-\omega_{l,p_i})+A_{l,q_i}D(q_i\omega_0^1-\omega_{l,q_i})\,D(q_i\omega_0^1-p_i\omega_0^2)}{[1-D^2(p_i\omega_0^2-q_i\omega_0^1)]},\quad i=1,2,\ldots,I$$

时, 式(11) 取极值:

$$\begin{aligned}MSE=&\sum_{l=1}^{L}|A_l|^2-\sum_{j=1}^{PJ}[A_{l,p'_j}D(p'_j\omega_0^2-\omega_{l,p'_j})]^2\\&-\sum_{j=1}^{QJ}[A_{l,q'_j}D(q'_j\omega_0^1-\omega_{l,q'_j})]^2\\&-\sum_{i=1}^{I}[A_{l,q_i}D(q_i\omega_0^1-\omega_{l,q_i})]^2+[A_{l,p_i}D(p_i\omega_0^2-\omega_{l,p_i})]^2\\&+2[A_{l,q_i}D(q_i\omega_0^1-\omega_{l,q_i})A_{l,p_i}D(p_i\omega_0^2-\omega_{l,p_i})\\&D(p_i\omega_0^2-q_i\omega_0^1)]/[1-D^2(p_i\omega_0^2-q_i\omega_0^1)]\end{aligned} \tag{14}$$

从以上结果可见, 如果某帧两原始语音的某两个重叠分量完全重叠(即两谐频率相等 $q_i\omega_0^1=p_i\omega_0^2$), 则单由谐波性不能对该两个重叠分量实现分离. 为此, 需根据语音谱的连续性消除完全重叠分量分离的歧义. 即: 首先对非完全重叠分量进行分离, 然后对完全重叠分量进行分离; 在对完全重叠分量进行分离时, 如果某完全重叠分量的两个相邻谐波分量不是完全重叠分量, 则运用平滑内插的方法求其幅值, 重复该过程, 直到没有该类型的完全重叠分量; 如果仍有完全重叠分量, 则由其最相邻的若干谐波分量通过外推的方法求其幅值.

4 实验

在实现以上描述的重叠语音分离算法中, 基音频率搜索范围是 60Hz–450Hz. 为减少运算量, 采用两级搜索机制. 信号为 8KHz, 16bit 采样. 帧长 35ms, 帧移 10ms. 由于语音的基音频率是语音感知的一个重要参数, 所以实验中, 采用两种测度衡量分离效果. 一是由此算法在分离语音的同时得到的分离语音的基音频率和原始语音的基音频率的相对平均误差:

$$RFFE=\frac{1}{NF}\sum_{i=1}^{NF}|\tilde{f}_{0,j}-f_{0,i}|/f_{0,j}\times 100\%.$$

二是谱失真测度[3](Spectrum Distortion, dB):

$$(1/NF)\sum_{i=1}^{NF}\sqrt{\frac{1}{W}\sum_{\omega}^{W}\left(20\log\tilde{F}_i(\omega)/F_i(\omega)\right)^2}.$$

其中, NF 是语音的帧数, W 是语音带宽. $\tilde{f}_{0,i}, f_{0,i}$ 分别是第 i 帧的分离后的和原始的语音基音频率; $\tilde{F}_i(\omega)$, $F_i(\omega)$ 分别是第 i 帧根据分离后的参数由式(2) 和式(3) 合成后的和原始的语音幅度谱.

实验 **1**　具有谐波特性的重叠信号的分离. 信号 $s_1(n)$ (实验中记为 $s1$) 和 $s_2(n)$ (实验中记为 $s2$) 是程序产生的均具有三个谐波分量的合成信号. 信号 $s_1(n)$ 的基频为 280Hz, 信号 $s_2(n)$ 的基频为 200Hz.

实验 **2**　实际重叠语音的分离. 语音 $s_1(n)$ (实验中记为

$v1$)和 $s_2(n)$(实验中记为 $v2$)分别是由一男声和女声“阿 a”的发音.

对 $s_1(n)$和 $s_2(n)$的相对平均功率 PR(dB)(Power Ratio)为- 15dB 到 20dB 的不同情况下进行了实验. PR 定义为:

$$PR=10\log\left[\left[(1/N1)\sum_{n=1}^{N1}s_1^2(n)\right]/\left[(1/N2)\sum_{n=1}^{N2}s_2^2(n)\right]\right] \quad (15)$$

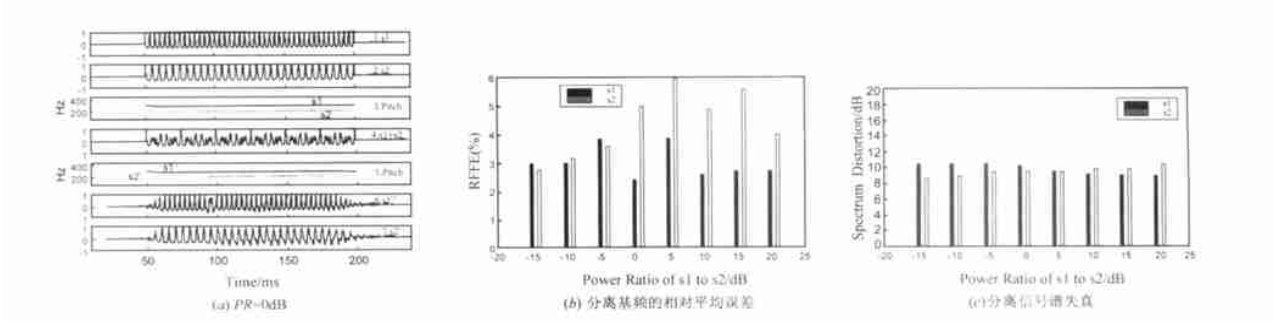

图1 具有谐波特性的重叠信号的分离

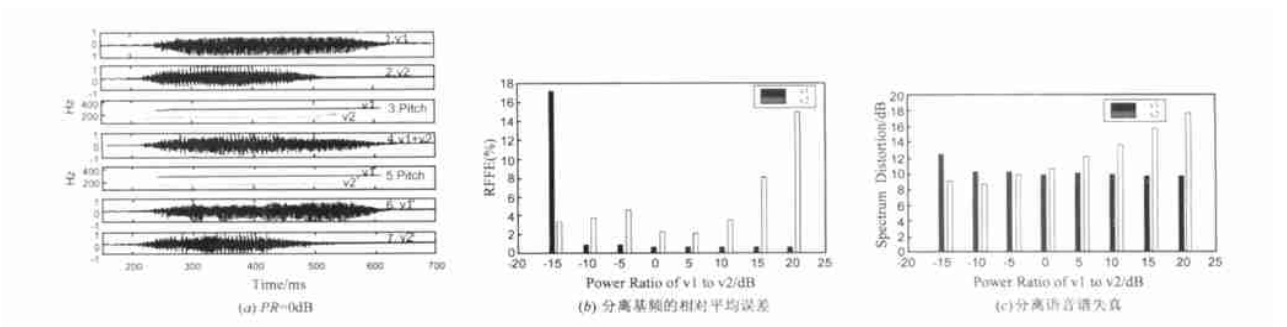

图2 实际重叠语音的分离

实验结果见图 1 和图 2. 图 2(a)和图 1(a)为在 $PR=0$dB 时:两个原始语音(信号)波形 $v1$ 和 $v2$($s1$ 和 $s2$);对语音(信号)直接进行基音检测所得的基音频率(基频)曲线;重叠语音(信号)波形 $v1+v2$($s1+s2$);分离的两个语音(信号)的基音频率曲线;根据分离后的参数按式(2)和(3)合成的分离后语音(信号)波形 $v1'$ 和 $v2'$($s1'$ 和 $s2'$);图 1(b)、(c)和图 2(b)、(c)分别给出了在不同相对平均功率 PR(dB)的条件下,实验 1 和实验 2 中分离基音频率相对平均误差 RFFE 和分离语音谱失真.

图 1(b)、(c)和图 2(b)、(c)结果表明,当 $S_1(n)$和 $S_2(n)$的相对平均功率在 10dB 内时,分离基频的误差和谱失真同相对平均功率 PR 关系不大,这种无关性是由于分离算法基于信号的谐波特性的缘故,此时,较弱信号的主要谐波分量仍然对 MSE 有较大的贡献;但由于语音信号的谐波特性比合成信号弱,因此,语音信号的分离基频误差和谱失真的这种无关性所对应的相对平均功率 PR 的范围要小. 这也反应了强语音信号对弱语音信号在超过一定强度后的掩蔽效应. 在 $PR=0$dB 时,两个分离语音信号的谱失真约为 10dB. 这种较大的失真和采用精细的谱失真测度(见式(15))和对语音采用谐波分析(见式(2)和(3))有一定的关系. 事实上,通过测听,原始语音和分离后的语音非常接近.

5 结论

本文提出了基于语音谐波分析法的重叠语音分离方法,该分离方法是帧同步的. 实验结果表明了本文中提出方法的有效性. 如何利用同一源产生的语音共同性质实现重叠语音的分离,应该来说,还有很多需要研究的问题:如,如何利用同一语音相邻帧频谱的连续性,以及同一语音各个频率分量在起、止时间上的相关性等进行重叠语音分离. 另外,重叠语音清音段的分离,也是极具挑战性的需要研究的问题.

参考文献:

[1] J F Cardoso. Blind signal separation: statistical principles [J]. IEEE Proceedings, 1998, 86(10): 2009- 2069.

[2] Bregman, A S. Auditory scene analysis: The perceptual organization [M]. Cambridge, MA: MIT Press, 1990.

[3] Masashi Unoki, Masato Akaki. A method of signal extraction from noisy signal based on auditory scene analysis [J]. Speech Communication, 1999, 27(3- 4): 261- 279.

[4] Nakatani T, Okuno G. Harmonic sound stream segregation using localization and its application to speech stream segregation [J]. Speech Communication, 1999, 27(3- 4): 209- 222.

作者简介:

戴礼荣 男,1962 年生于安徽,副教授,1983 年在西安电子科技大学电子工程系获得学士学位,1985 年在合肥工业大学获得通信电子系统专业硕士学位,1996 年获得中国科技大学通信与信息系统专业博士,主要从事语音信号处理、语音编码与通信、人机语声对话的研究及 DSP 技术应用.