ISSN 1000-0054 清华大学学报 (自然科学版) 2001 年 第 41 卷 第 7 期 1/34
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2001, Vol. 41, No. 7 1-4

一种基于听觉掩蔽效应的语音增强方法

张金杰, 曹志刚, 马正新

(清华大学 电子工程系, 微波与数字通信国家重点实验室, 北京 100084)

摘 要: 为提高增强语音的听觉效果, 研究了一种基于听觉掩蔽效应的语音增强方法。推出了一个功率谱域的基于听觉掩蔽效应的不等式准则, 并用这个准则动态地选择一个作为语音短时谱幅度估计器的非线性函数的参数值, 通过这个参数自适应变化的非线性函数对语音谱幅度进行估计实现语音增强。在此基础上, 设计实现了一个单声道语音增强算法。对增强语音的客观测试和非正式听音测试表明: 相对于传统的减谱法和对数短时谱幅度最小均方误差估计增强法, 基于听觉掩蔽效应的语音增强方法能更好地抑制背景噪声。

关键词: 语音增强; 噪声抑制; 听觉模型; 掩蔽效应

中图分类号: TN 912.3

文章编号: 1000-0054(2001)07-0001-04 **文献标识码**: A

Speech enhancement method based on auditory masking

ZHANG Jinjie, CAO Zhigang, MA Zhengxin

(State Key Lab on Microwave & Digital Communications, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: This paper presents a speech enhancement method based on the masking properties of the human auditory system and a non-linear filter. The speech phase is ignored to identify enhanced criterion in the power spectral domain based on masking. The criterion is then applied to determine the parameter values for a non-linear filter employed to estimate the short-time spectral amplitude of speech. The method is then used to develop a single-channel speech enhancement algorithm. Objective measurements combined with informal subjective listening tests show that the proposed algorithm can suppress audible noise more effectively than the popular power spectral subtraction method and the minimum mean-square error log-spectral amplitude estimation method.

Key words: speech enhancement; noise suppression; auditory model; masking

语音增强应用于在语音通信领域中, 比如噪杂环境下的噪声抑制、语音压缩、语音识别等。

本文研究只有一路带噪语音信号可以利用的单声道语音增强方法。单声道语音增强方法中目前常用的是一类基于短时谱幅度(STSA)估计的语音增强方法, 该类方法认为语音信号的 STSA 对语音的感知起主导作用, 从而在语音增强中需要精确估计, 而相位对语音的感知并不重要, 没有必要精确估计。文[1]通过实验为此提供了一定的依据, 文[2]则证明在一定条件下语音相位的最小均方误差(MMSE)估计值就是带噪语音相位本身, 因此, 基于 STSA 估计的语音增强方法一般都直接采用带噪语音的相位作为增强语音的相位。基于 STSA 估计的语音增强法包括减谱法及其各种变形[3]、MMSE 估计法[2,4,5]等。减谱法通过从带噪语音的 STSA 中直接减去噪声的平均谱幅度来得到增强语音的 STSA, 实现起来简单, 但是剩余噪声大, 并且产生不舒服的"音乐噪声"。后来, Ephraim[2,4,5]等人提出了 STSA 的MMSE 估计法, 部分解决了"音乐噪声"问题, 但在带噪语音信噪比(SNR)较低时其剩余噪声还是很大, 尤其是当信噪比小于 5 dB 时。

近年来, 人们针对听觉外周提出了一些计算模型, 并在语音编码、音频压缩和音质的客观度量等方面获得了应用, 同时, 基于人类听觉特性的语音增强研究也取得了一定的进展[6]。目前, 在语音增强中用得比较成功的是听觉掩蔽效应, 它指出语音信号能够掩蔽与其同时进入听觉系统的一部分能量较小的噪声信号, 而使得这部分噪声不为人所感知到。因此从掩蔽效应的角度看, 语音增强应该通过改变带噪语音的 STSA 使得所有噪声成分都能被语音信号掩盖掉。据此, 本文提出了一种基于掩蔽效应和 STSA 非线性估计的语音增强方法。

收稿日期: 2000-03-17

基金项目: 国家自然科学基金资助项目 (60072011); 清华大学创新基金项目

作者简介: 张金杰(1973-), 男(汉), 浙江, 博士研究生。

1 算法描述

带有背景噪声的语音 $y(n)$ 一般可以写为

$$y(n)=x(n)+d(n),$$

其中: $x(n)$ 是纯净语音, $d(n)$ 是平稳加性噪声。$x(n)$ 和 $d(n)$ 两者互不相关。

由于增强是按帧进行的, 把上述模型写成帧的形式为

$$y(n,i)=x(n,i)+d(n,i),$$
$$i=1,2,\cdots;\quad n=0,1,\cdots,N-1,$$

其中: i 为帧号, N 为帧长。另外设 $\hat{x}(n,i)$ 为该帧增强语音, 并令 $d(n,i)$, $x(k,i)$, $y(k,i)$ 和 $\hat{x}(n,i)$ 的离散频谱和功率谱分别为 $D(k,i)$, $X(k,i)$, $Y(k,i)$, $\hat{X}(k,i)$ 和 $D_p(k,i)$, $X_p(k,i)$, $Y_p(k,i)$, $\hat{X}_p(k,i)$, 其中 $k=0,1,\cdots,K-1$, 是离散频率下标, K 为快速 Fourier 变换(FFT)长度。

根据听觉掩蔽效应, 语音增强应该使得增强语音的噪声分量都小于纯净语音的掩蔽阈。设 $\hat{d}(n,i)$ 为增强语音 $\hat{x}(n,i)$ 中的噪声, 其频谱和功率谱分别为 $\hat{D}(k,i)$ 和 $\hat{D}_p(k,i)$, $T(k,i)$ 为纯净语音的掩蔽阈, 那么基于掩蔽效应的语音增强准则应为 $\hat{D}_p(k,i)\leq T(k,i)$。另外, 若 $\hat{d}(n,i)$ 一般可定义为

$$\hat{d}(n,i)=\hat{x}(n,i)-x(n,i),$$

则

$$\hat{D}_p(k,i)=|\hat{D}(k,i)|^2=|\hat{X}(k,i)-X(k,i)|^2. \tag{1}$$

基于 STSA 估计的语音增强方法, 一般把带噪语音的相位作为增强语音的相位, 所以有

$$\arg\hat{X}(k,i)=\arg Y(k,i).$$

另外, 假定纯净语音相位用带噪语音相位代替将不会对它带来听觉上的不同, 于是听觉上剩余噪声频谱可用

$$\hat{D}(k,i)\approx\{|\hat{X}(k,i)|-|X(k,i)|\}\exp[\arg Y(k,i)]$$

近似。这样, 由式(1)可以得到约去了相位的近似剩余噪声功率谱表达式

$$\hat{D}_p(k,i)=\left[\sqrt{\hat{X}_p(k,i)}-\sqrt{X_p(k,i)}\right]^2, \tag{2}$$

同时, 基于掩蔽效应的语音增强准则可以写为

$$\left[\sqrt{\hat{X}_p(k,i)}-\sqrt{X_p(k,i)}\right]^2\leq T(k,i). \tag{3}$$

鉴于上述增强准则只是一个不等式, 为方便求解, 选择下面的单参数非线性函数作为估计器, 其形式为

$$\hat{X}_p(k,i)=\frac{Y_p(k,i)}{Y_p(k,i)+a(k,i)}Y_p(k,i), \tag{4}$$

其中 $a(k,i)>-Y_p(k,i)$, 是一随时间和频率改变的参数。把式(4)代入式(3)得 $a(k,i)$ 的取值范围为

$$\begin{cases}a_1(k,i)\leq a(k,i)\leq a_h(k,i), & X_p(k,i)>T(k,i),\\ a_1(k,i)\leq a(k,i), & X_p(k,i)\leq T(k,i),\end{cases} \tag{5}$$

其中 $a_1(k,i)$ 和 $a_h(k,i)$ 的取值分别为

$$\begin{cases}a_1(k,i)=\dfrac{Y_p^2(k,i)}{\left(\sqrt{X_p(k,i)}+\sqrt{T(k,i)}\right)^2}-Y_p(k,i),\\ a_h(k,i)=\dfrac{Y_p^2(k,i)}{\left(\sqrt{X_p(k,i)}-\sqrt{T(k,i)}\right)^2}-Y_p(k,i).\end{cases} \tag{6}$$

需要指出的是, 上述结果只有理论指导意义, 并不能实际使用, 因为它包含了纯净语音的短时功率谱值, 而这个值在实际应用时是无法得到的。不过在实际应用中可以得到纯净语音短时功率谱的一个估计值, 但是由于估计误差, 上面的结果并不能直接使用, 需要作进一步改进, 而增强效果也会差于理论值。下面给出这样一个实际算法。

纯净语音的短时功率谱估计可以通过功率谱域的减谱法结合移动平均来实现, 也就是

$$X_p(k,i)=\eta X_p(k,i-1)+(1-\eta)\max Y_p(k,i)-\beta\times D_p(k,i),0, \tag{7}$$

其中: $0<\eta<1$, $\beta\geq 1$, $D_p(k,i)$ 在无声段进行更新。

$$D_p(k,i)=\epsilon D_p(k,i-1)+(1-\epsilon)Y_p(k,i), \tag{8}$$

其中 $0<\epsilon<1$, 是遗忘因子, 用于控制更新速率。另外, 有声/无声检测采用了文[7]的算法, 而语音掩蔽阈的计算则与文[6]类似。

由于语音和噪声的功率谱都是通过估计得到的, 直接应用式(6)增强效果并不好。研究发现, 如果在式(6)中显式地包含噪声功率谱, 则增强效果要好很多, 具体地说就是用 $X_p(k,i)+D_p(k,i)$ 去近似 $Y_p(k,i)$, 这样做的直接好处是 $a(k,i)$ 代回式(4)时避免约去分式中的 $Y_p(k,i)$ 项。替换之后有

$$\begin{cases}a_1(k,i)=\dfrac{X_p(k,i)+D_p(k,i)^2}{\left(\sqrt{X_p(k,i)}+\sqrt{T(k,i)}\right)^2}-\\ \qquad X_p(k,i)+D_p(k,i),\\ a_h(k,i)=\dfrac{X_p(k,i)+D_p(k,i)^2}{\left(\sqrt{X_p(k,i)}-\sqrt{T(k,i)}\right)^2}-\\ \qquad X_p(k,i)+D_p(k,i).\end{cases} \tag{9}$$

另外, 为统一起见, 在 $X_p(k,i)\leq T(k,i)$ 时, 给 $a(k,i)$ 定义一个有限的上界。由式(9)易知 $a_h(k,i)>a_1(k,i)$, 所以这个上界不妨也取为 $a_h(k,i)$。在具

体实现时, $a(k,i)$ 可以简单地取为

$$a(k,i) = \lambda[\alpha|a_1(k,i)| + (1-\alpha)|a_h(k,i)|], \quad (10)$$

其中: $0\leq\alpha\leq1$, $\lambda>0$。λ 用于补偿 $a_1(k,i)$ 和 $a_h(k,i)$ 的估计误差, 它的取值通过实验确定。

综上所述, 基于听觉掩蔽效应的 $a(k,i)$ 值可由式(7～10)估计得到, 然后应用式(4)估计出语音的短时功率谱, 从而得到语音的短时谱为 $\hat{X}(k,i) = \sqrt{\hat{X}_p(k,i)}\exp[\arg Y(k,i)]$, 再经反 FFT 并与前一帧作重叠加就可以得到增强语音。

2 结果与讨论

实验中采用的语音材料选自中文语音平均意见得分(MOS)测试库, 共 6 位发音人(3 男 3 女), 每人一句。噪声材料选自 NOISEX-92 数据库, 噪声类型包括白噪声、类似语音的噪杂声噪声、喷气战斗机机舱噪声和工厂噪声等。语音和噪声信号经 8 kHz 采样、16 bit 量化为数字信号, 并在计算机中按一定比例混合生成不同信噪比(10 dB, 5 dB, 0 dB 和 - 5 dB)的带噪语音。带噪语音通过长度为 256 点(32 ms)的汉明(Hamming)窗形成长度为 256 点的语音帧, 相邻两帧之间重叠 128 点, 然后对每一帧带噪语音逐帧进行增强处理。

增强性能评测采用客观测试结合非正式听音测试来进行。采用的客观测试指标包括分段信噪比(SEGSNR)和噪声掩蔽阈比(NMR)[6]。文[6]指出 NMR 与主观评测有较高相关度, 一般其值越小语音音质越好。需要说明的是, 为方便评测比较, 客观测试指标给出的是语音增强前后的差值, 具体地说, 就是分段信噪比的提高量和噪声掩蔽阈比的减少量。

经过分析和实验比较, 算法中的一些参数值取为 $\lambda=5$, $\alpha=0.4$, $\beta=6$, $\eta=0.999$, $\epsilon=0.995$。表 1 给出了 4 种噪声对于不同带噪声信噪比 γ_{SNR} 的客观测试指标结果, 作为对照, 同时也给出了功率谱相减法(PSS)和对数短时谱幅度最小均方误差估计法(LOG-MMSE)[4]这两种目前常用的语音增强方法的增强结果。由表 1 可见, 对于所有噪声类型和不同的信噪比, 增强语音的 SEGSNR 都有不同程度的提高, 而 NMR 也有很大程度地降低, 这说明本方法确实能有效地抑制背景噪声。进一步, 比较表 1 中各种增强方法的增强结果可以看出, 对于 SEGSNR 指标, 在带噪语音信噪比较低时, 本方法提高得较多, 而在其信噪比较高时则

表 1 各种噪声类型下增强语音的客观测试指标

噪声类型	γ_{SNR}/dB	Δ_{SEG}/dB			Δ_{NMR}/dB		
		PSS	LOG-MMSE	MASK	PSS	LOG-MMSE	MASK
白噪声	- 5	3.77	11.7	14.1	3.94	15.8	28.5
	0	3.47	9.57	10.4	3.89	15.0	25.3
	5	3.04	7.24	6.88	3.84	14.0	22.1
	10	2.45	5.01	3.50	3.74	12.7	18.4
噪杂声噪声	- 5	1.61	5.17	7.24	2.09	6.61	9.46
	0	1.57	4.62	5.97	2.20	6.44	8.79
	5	1.55	4.08	4.51	2.34	6.18	7.91
	10	1.31	3.05	2.28	2.61	5.85	6.69
F16 机舱噪声	- 5	2.94	9.56	12.2	3.42	13.2	21.1
	0	2.74	8.08	9.29	3.48	12.6	19.1
	5	2.33	6.10	6.08	3.49	11.5	16.2
	10	1.55	3.68	2.54	3.38	9.98	12.7
工厂噪声	- 5	5.50	12.4	13.9	3.77	12.3	17.7
	0	5.24	10.4	10.4	3.91	11.3	14.8
	5	4.66	7.83	6.97	3.97	9.80	11.3
	10	3.41	4.75	3.22	3.58	7.43	7.42

γ_{SNR} 表示合噪语音信噪比; Δ_{SEG} 表示分段信噪比; Δ_{NMR} 表示噪声掩蔽阈比; PSS 表示功率谱相减法; LOG-MMSE 表示对数短时谱幅度最小均方误差估计法; MASK 表示本文提出的基于听觉掩蔽效应的方法。

比LOG-MMSE 提高得少; 但是对于NMR 指标, 本方法对于所有噪声类型和所有信噪比都要比另两种算法好。非正式听音测试表明, 本方法增强语音的剩余噪声在听觉上都要比另两种小很多。另外需要指出的是, 尽管本方法在带噪语音信噪比为 10 dB 时其分段信噪比提高得比LOG-MMSE 少, 非正式听音测试却表明其增强语音的音质比它好, 主要表现为前者对于白噪声、飞机噪声和工厂噪声几乎听不到剩余噪声, 对于类似语音的噪杂声噪声其剩余噪声也相对小很多(该类噪声较不平稳, 其增强语音的剩余噪声也相对大一些), 这说明 SEGSNR 与主观评测并不完全一致。但是, 10 dB 时本方法的NMR 指标比另两种方法都要好, 这与非正式听音测试的结果基本一致, 部分证实了文[6]中提到的NMR 与音质的主观评测有较高程度相关性的结论。总之, 客观测试和非正式听音测试均表明, 本文基于听觉掩蔽效应的语音增强方法能在听觉上更好地抑制语音中的背景噪声。

3 结 论

推出了一个功率谱域的基于听觉掩蔽效应的不等式准则, 然后应用这个准则去动态选择作为语音短时谱幅度估计器的非线性函数的参数值, 使得它能自适应地对带噪语音实施滤波以实现语音增强功能。在此基础上, 设计实现了一个单声道语音增强算法。对增强语音的客观测试和非正式听音测试表明, 相对于传统的减谱法和对数短时谱幅度最小均方误差估计增强法, 本文方法能更好地抑制背景噪声。

参考文献 (References)

[1] Wang D L, Lim J S. The unimportance of phase in speech enhancement [J]. IEEE Trans Acoust Speech Signal Processing, 1982, 30(4): 679-681.

[2] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator [J]. IEEE Trans Acoust Speech Signal Processing, 1984, 32(6): 1109-1121.

[3] Lim J S, Oppenheim A V. Enhancement and bandwidth compression of noisy speech [J]. Proceedings of the IEEE, 1979, 67(12): 1586-1604.

[4] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator [J]. IEEE Trans Acoust Speech Signal Processing, 1985, 33(2): 443-445.

[5] 曹志刚, 郑文涛. 基于短时谱最小均方误差估计的语音增强和剩余噪声衰减 [J]. 电子学报, 1993, 21(4): 7-12.
CAO Zhigang, ZHENG Wentao. Speech enhancement based on minimum mean-square error short-time spectral estimation and residual noise reduction [J]. Acta Electronica Sinica, 1993, 21(4): 7-12. (in Chinese)

[6] Tsoukalas D E, Mourjopoulos J N, Kokkinakis G. Speech enhancement based on audible noise suppression [J]. IEEE Trans Speech and Audio Processing, 1997, 5(6): 497-514.

[7] Lynch J F, Josenhans J G, Crochiere R E. Speech/silence segmentation for real-time coding via rule based adaptive endpoint detection [A]. Odell P L, Hunt L R. Proceedings of ICASSP'87 [C]. Piscataway, NJ: IEEE Acoustic, Speech and Signal Processing Society, 1987. 1348-1351.

书 讯

《数字视频处理》

数字视频是用数字手段提供全运动视频图像的高新技术, 近十余年来推动了多媒体, 虚拟现实, 视频通信, VCD 等产业的飞速发展; 在即将来临的信息社会中, 还将给计算机、通信、影像等产业以巨大的推动。为帮助读者在未来破浪前进, 这本及时问世的书首次全面讲述了数字视频处理的原理以及面向各种应用的主要算法。全书分为 6 个部分: 数字视频表示, 包括视频图像模型和空域——时域采样; 二维运动估计; 三维运动估计; 视频滤波; 静图像压缩; 视频压缩。该书是在为研究生和高年级学生讲课基础上写成的, 取材全面系统, 表述精练, 插图丰富, 并有详尽的文献索引。对于所用的数学原理, 作者进行了仔细处理和精心安排, 特别便于自学。

该书由清华大学出版社出版。