计算机科学 1997Vol. 24No. 6

语音识别中的噪声抑制方法

Noise Suppression Methods in Speech Recognition

王承发 韩纪庆 徐近霈 高 文

(哈尔滨工业大学计算机科学与工程系 哈尔滨 150001)

摘 要 In this paper, we review the noise suppression methods in the speech recognition, and divide the methods into four major areas: removing noise estimation, environment normalization technique, updating model technique, and building noise model. We can not review all possible techniques of noise suppression in speech recognition, because there are far too many methods in use today. Instead, we only review those techniques most commonly used.

关键词 Noise suppression, Speech recognition Signal-noise-ratio

1. 引言

当前有很多语音识别系统和产品,但绝大部分是工作在安静环境下的,一旦在噪声环境下使用,语音信号中混入严重的背景噪声,信噪比就大为下降,影响了参数的稳定性,而且通常采用的语音特征在噪声条件下并不稳定,其性能明显下降;另一方面,当背景噪声进入人耳后,将使话者产生情绪上的变化和心理上的紧张,因此其发音将明显改变,即产生 Lombard 效应和 Loud 效应[1]。因此,对噪声的稳定性已成为语音识别系统走向实际应用的主要障碍之一。这也吸引了语音技术领域的众多技术人员进行这方面的研究工作。

通常我们假设有两种环境噪声使得语音识别系统的性能下降:加性噪声和卷积噪声。前者是由机械噪声或由于话者之外其他人的语音而产生的背景噪声,后者是由麦克风或传输通道等滤波过程引起的噪声。

为设计稳定的语音识别系统,人们提出了很多方法,这些方法可概括为四个方面:第一,假设噪声频谱是叠加在语音频谱之上的,因此可从混噪语音频谱中去除噪声的频谱估计来抑制噪声,谱减技术是典型的这种方法。第二,使用环境规正技术来抑制卷积噪声,如:倒谱规正技术、倒谱平均减技术、以及 RASTA-PLP 技术。第三,不修正语音信号,而只是调整识别器的模型使之适合噪声,如信号偏移去除的方法以及修正模板的方法。第四,建立噪声模型,并在识别的过程中处理噪声。本文将综述语音识别中的噪声抑制方法。

2. 谱减技术

谱减技术是基于噪声频谱叠加在语音频谱之上,因而可以先估计噪声频谱,而后从混噪语音的频谱中减去噪声的频谱估计。目前有两种类型的谱减:线性谱减和非线性谱减。线性谱减可用如下的公式来表示[2]:

$$|Y(w)|=|X(w)|-|N(w)| \tag{1}$$

这里 $|X(w)|$ 是混噪语音幅值谱的短时估计,$|N(w)|$ 是利用语音间隙估计出的噪声谱,$|Y(w)|$ 是去噪后语音频谱。为保证谱减后不出现负的频谱值,通常采用一个频谱下限 ε,因而公式(1)变为:

$$|Y(w)|=\begin{cases}|X(w)|-|N(w)|; \\ \quad \text{if } X(w)|-|N(w)>\varepsilon \\ \varepsilon \quad \text{其它}\end{cases} \tag{2}$$

线性谱减适用于去除平稳的噪声,但当语音间隙中的噪声变化很快时,这种方法存在许多问题,为此,人们提出了非线性谱减的方法[3]。在非线性谱减中,首先确定一个改进的噪声模型 $(N(w),\alpha(w))$,其中 $\alpha(w)$ 是一个过估计系数,它可利用语音间隙与噪声幅值谱一起被估计出来。非线性谱减的实现过程如下:

$$|Y(w)|=|X(w)|-\mathrm{opt}(\alpha(w))|N(w)| \tag{3}$$

这里 $\mathrm{opt}(\alpha(w))$ 是一个在特定频段上按照 SNR 给出的加权系数 $(1\leqslant \mathrm{opt}(\alpha(w))\leqslant 3)$,它在高 SNR 区域采用最小值,而在低 SNR 区域采用最大值。类似

于线性谱减，如果谱减后的结果小于一个频谱下限值，则用这个下限值代替计算值。

3. 环境规正技术

环境规正技术常用于处理因麦克风、传输通道，以及其他训练和测试环境不同而引起的卷积噪声。通常有两种环境规正技术：一是基于倒谱修正的，如倒谱规正、倒谱平均减；另一个是 RASTA—PLP 技术。

3.1 倒谱规正

这种方法假定训练和测试环境的不同可通过倒谱向量上的一个加性校正来反映。A. Acero 和 R. Stern 提出了两种倒谱规正的方法：信噪比相关的倒谱规正（SDCN）和码本相关倒谱规正（CDCN）。SNCN 基于可以使用同步记录的训练和测试环境数据语料，它通过计算在各个 SNR 下训练和测试环境平均倒谱的差来获得补偿向量。当一个新语音输入系统，逐帧计算其瞬时 SNR，并依此 SNR 对各帧使用不同的补偿向量。尽管这种方法能获得满意的性能，但它要求使用同步记录的训练和测试环境的训练数据库，因而不能适应未知的测试环境。CDCN 方法可分为两个步骤：第一步通过使观测到的噪声倒谱向量的似然度达到最大来估计环境的参数值，第二步是在给定噪声语音的条件下利用最小均方估计来获取纯净语音的倒谱。CDCN 具有不需要测试环境的先验知识的优点，但它的计算要比 SDCN 复杂。

A. Acero 和 R. Stern 提出了固定码本相关的倒谱规正（FCDCN），集中了 SDCN 和 CDCN 两者的许多优点：类似于 SDCN，补偿系数等于训练和测试环境倒谱的差，同时类似于 CDCN，对不同的 VQ 码本补偿系数不同。FCDCN 比较有效，但它也要求使用同步记录的训练和测试环境的训练数据库，因而也不能适应未知的测试环境。

文[4]中提出了多重固定码本相关的倒谱规正（MFCDCN）。这种方法采用 FCDCN 的训练过程并行地预先计算出与各个测试环境相关的补偿向量。当一个未知环境的发音输入后，系统首先确定当前的测试环境与训练数据中的哪个测试环境最相似，而后用被选中的测试环境的补偿向量去规正输入的发音。尽管这种方法很有效，但它需要大量的训练。

3.2 倒谱平均减（CMS）

倒谱平均减基于输入语音特征流的总体平均为零这一假设，采用修正倒谱系数的方法使得由于通道畸变而引起的训练和测试数据的不匹配达到最小。这种方法先求出语音数据的倒谱系数平均值，而后每一语音帧的倒谱系数都与之减去。然而在许多实际情况下，用于训练和测试的语音数据是有限的、因此，语音的倒谱平均为零均值的这一假设并不成立，对一个不正确的通道倒谱估计，使用 CMS 将导致从每一语音帧中削弱有用的倒谱信息，并使得系统识别率下降。

D. Naik[5]发现窄带极点对总体长时倒谱平均的贡献是使频谱内容产生了较大的偏移，在这一倒谱域的 CMS 将削弱有用的频谱信息；而另一方面，宽带极点对长时平均的贡献是仅仅补偿由于通道产生的频谱的陡变。D. Naik 提出了极点滤波的倒谱系数（PFCC）以减少倒谱平均上的语音内容，因而改进了通道的估计。PFCC 是通过扩展窄带极点的带宽而同时保持其频率不变而后求出来的基于 LP 的倒谱系数。因此，基于 PFCC 的 CMS 方法是从每个语音帧的 LP 倒谱系数中减去修正后的长时倒谱平均，而不是减去通常的长时倒谱平均。CMS 能有效地补偿未知线性滤波的影响，但它不能直接补偿加性噪声和线性滤波的混合影响。

3.3 RASTA—PLP

PLP 是由 H. Hermansky 提出的基于听觉特性的语音分析方法，它先将许多听觉特性，如：临界带频谱分辨率、等响度曲线、强度响度功率规则等加以工程模拟，而后使用 AR 模型来近似计算听觉频谱。PLP 方法在短时频谱值受到通道频率响应的影响时并不是稳定的。RASTA 是一种能使 PLP 有效抑制频谱畸变的方法[6]，RASTA—PLP 先计算临界带对数频谱的差分，而后再计算其重积分。经过上述处理后，这种方法有效地减少了对数频谱域上的加性部分，尤其是各个频率通道上常数的或变化缓慢的部分；然而经过 RASTA 后，那些在频谱域上的与信号无关的加性噪声部分，经过对数运算后变得与信号有关了，因此不能用 RASTA 有效地去除。为了解决这一问题，RASTA 中的对数变换被修正如下：

$$Y=\ln(1+JX) \qquad (4)$$

其中 X 是临界带频谱，J 是依赖于信号的正常数，式(4)在 $J<1$ 时是线性的，在 $J>1$ 时是对数的。

试验表明最优的 J 值与每个特定的噪声级别有关，而噪声级别在训练和测试时通常是不同的，因此存在一个如何选择 J 值的问题。J 的选择方法可分为三类：

(1)多个识别器：训练多个识别器，每个采用不同的 J 值，识别时，通过估计噪声来确定一个 J 值，

并使用与之相关的识别器。这种方法对含噪声语音的试验结果好于 PLP 或 RASTA，但对纯语音数据其识别性能有所下降。

(2)一个识别器适合于多个 J 值：识别器在训练时具有足够的自由度，训练数据能够在一定范围内变化以适应不同的 J 值，而 J 值是通过测得的平均噪声能量来计算的。这种方法无论是对纯语音还是对混噪语音，其性能都优于多个识别器的方法，但它要求训练数据的多重处理。

(3)频谱映射：使用线性映射将与噪声有关的 J 值所对应的频谱映射到与纯语音有关的 J 值所对应的频谱上。识别时通过局部噪声估计来确定一个较优的 J 值，而后对临界带的输出使用先前算得的映射系数进行映射。这种方法性能较好，但它要求具有所有可能的噪声环境的先验知识。

尽管 RASTA—PLP 有效地抑制了卷积噪声，但它需要复杂的计算，不利于实时处理。文[2]采用了直接利用临界带频谱计算倒谱系数，而不是计算全极点模型以获得 LPC 系数，而后再求倒谱系数，这种改进有利于实现实时处理。

4. 修正模型技术

这种技术通过修正识别器的模型参数来达到适应噪声的目的，信号偏移去除和修正参考模板是其中两种典型的方法。

4.1 信号偏移去除

文[7]中，M. Rahim 和 B. Juang 提出了一种修正偏移去除的方法，并将其运用到一个离散密度的 HMM 系统中。该方法假设 VQ 码本中心保持常数而将偏移作为一个能用最大似然方法估计的未知量，因此可用一个迭代过程来修正信号，而后使用通常的 LBG 算法计算一个局部优化的码本向量集。这种方法具有只通过修正 VQ 码本就能消除噪声而不必重新训练 HMM 基本系统的优点。

4.2 修正参考模板

文[8]中 S. Dvorak 和 T. Hormann 提出了一种修正参考模板的方法。该方法在识别过程中先从词模式中找出与实际发音最相似的参考模板，而后通过计算实际的测试模板与被识别出的参考模板的加权和来得到新的参考模板，接着用新计算出的参考模板代替原有的参考模板。这种方法能连续地吸收有关信息，如：由于紧张而引起的语音变化、Lombard 语音、甚至参考模板中噪声的不同变化，因此它能在有变化的情况下改变识别性能。

文[9]中，K. Takagi 提出了一种基于频谱均等的快速环境自适应算法。该算法按照时间一致来跟踪测试发音与最近的参考模板之间的差，而后将这些差应用到所有的参考模板，使之自适应测试环境，接着用修正后的参考模板对输入发音再进行一次识别。该方法具有不需要迭代过程，并且允许使用测试发音进行修正的特点。

5. 建立噪声模型

该方法认为一个信号是由两个独立的部分叠加而成，这两个独立部分可分别用通常的 HMMs 来表达，而由这两部分混合而产生的信号可以用一个它们混合输出的函数来表示。因此这种方法是先建立噪声的 HMMs，而后在识别时从信号中区分出噪声和语音。

Y. Ephraim 等提出了处理加性的、统计独立的噪声语音的一种建立噪声模型的方法，该方法将纯净语音表示为带有混合高斯 AR 输出处理的一个 HMM 过程，把噪声作为一个平稳的统计独立的高斯 AR 向量。模型参数的估计分别用纯净语音和噪声的训练序列来进行。模型参数的估计和增强的过程都使用了最大似然估计。

文[10]中提出了一种用 HMMs 进行信号分解的方法。该方法先建立语音和噪声两者的 HMMs，而后同步地识别语音和噪声。对两部分的混合情况，同步 HMM 混合影响的观测概率如下：

$$\text{Observation Probability} = P(\text{Observation} \mid M1 \otimes M2) \quad (5)$$

其中 M1 和 M2 是并行的 HMMs，$\otimes$ 是任何的混合运算符。识别时通过扩展通常的 Viterbi 解码算法来实现在两个模型混合的状态空间上搜索。

小结 本文综述了一些常用的噪声抑制方法，即去除噪声估计、环境规正技术、修正模型技术、以及建立噪声模型。所有的方法都能在一定程度上抑制噪声，提高识别系统的性能，但这并不是说语音识别中的噪声抑制问题已经解决。目前的语音识别系统性能远远低于人类本身语音处理的能力。随着语音识别系统从实验室走向实际应用，噪声抑制问题正逐步得到重视。我们相信还有很多工作要做，我们将面对严峻的挑战。

参考文献

[1]B. Stanton et al., Acoustic Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions, ICASSP, 1988

计算机科学 1997 Vol.24 №.6

非语音声音在人机交互技术中的应用

The Applications of Non-speech Audio in Human computer Interaction Technology

方志刚 吴晓波 徐新民 金文光

(浙江大学CAD&CG国家重点实验室 杭州 310027) (杭州大学电子工程系 杭州 310028)

摘 要 Presentationg information in non-speech audio could take advantages of the superiority of human auditory sensory modality. This paper discesuses the user listening model, the applications of non-speech in human-computer interaction technology, and some audio display technology.

关键词 Non-speech audio, Human-computer interaction, Visualization, Scientific computing

1 引言

人机交互技术已经历了字符用户界面(CUI),图形用户界面(GUI)和多媒体用户界面阶段,目前正朝着三维,多通道用户界面以及虚拟现实(VR)技术的方向发展。从人的特点考虑,人机交互技术发展的几个阶段可分别归结为:(1)符号阶段,用户面对的只有单一文本符号,虽然有视觉参与,但视觉是非本质的,本质的东西只有符号和概念;(2)视觉阶段,借助计算机图形学技术使人机交互大量利用色彩,形状等视觉信息;(3)听觉阶段,早期计算机系统只有单调的蜂鸣声,虽然多媒体技术将声频及视频形式带进计算机,但仍缺少听觉交互手段,即人处于被动收听状态,声音缺少位置、方向等变化。上述每个阶段的早期阶段所用的交互手段在后续的阶段中仍然存在。人们在日常交往和信息传递过程中大量运用视觉信息、听觉信息、语言和概念符号,人学会运用视觉听觉可能不分先后,但语言及概念符号却一定是后天获得的。显然人机交互的历史基本上与人认识世界的过程相反,这可归结为以下原因:(1)计算机作为计算工具具有高度抽象性;(2)技术限制;(3)更重要的是声音通讯的抽象性和短暂性。当前,在人机交互中结合进视觉的和听觉的以及更多的通道是必然趋势,特别是将听觉通道作为补充的或替换的信息通道已显示出重要性和优越性[1]

研究表明,听觉通道有许多优越性,如:(1)听觉信号检测快于视觉信号检测的速度;(2)人对声音随时间的变化极其敏感;(3)声音信号所具有的全向特性可作为引导视觉对目标进行细调分析的粗调机制,即所谓"听觉是视觉的眼睛";(4)听觉信息与视觉信息同时提供可使人获得更强烈的存在感和真实感;等等。

[2]Z. Yang et al., Study on the Method and System Implemeniation of An Isolated Word Recognition System Used in Noise Environment, Chinese Journal of Acoustics(English Edition), 15(2) 1996

[3]J. A. Nolazco Flores, S. J. Yong, Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation, ICASSP, 1994

[4]F. Liu et al., Environment Normalization for Robust Speech Recognition Using Direct Cepstral Comparison, Same to [3]

[5]D. Naik, Pole-Filtered Cepstral Mean Subtraction, ICASSP, 1995

[6]H. Hermansky et al., RASTA-PLP Speech Analysis Technique, ICASSP, 1992

[7]M. Rahim, B. Juang, Signal Bias Removal for Robust Telephone Based Speech Recognition in Adverse Environments, Same to [3]

[8]S. Dvorak, T. Hormatsu, High Performance Speech Recognition in Noise by Continuously Updated Reference Templates, Eurospeech, 1991

[9]K. Takagi et al., Rapid Environment Adaptation for Robust Speech Recognition, ICASSP, 1995

[10]A. P. Varga, R. K. Moore, Hidden Markov Model Decomposition of Speech and Noise, ICASSP, 1990