第六届全国人机语音通讯学术会议，35-38页，2001年11月20-22日，深圳 **(大会优秀青年论文二等奖)**

基于子带信息的鲁棒语音特征提取框架

张欣研，王帆，郑方，徐明星，吴文虎

清华大学智能技术与系统国家重点实验室 语音技术中心
清华大学计算机系，北京，100084
{zxy,wangf,fzheng,xumx,wuwh}@sp.cs.tsinghua.edu.cn

摘要

本文提出一种鲁棒语音特征提取框架。通过使用一种基于子带能量分布的噪声估计方法，无需静音段，就可以估计出带噪语音的子带噪声，同时提出结合谱减和谱加权方法对特征进行处理，最终生成具有较高鲁棒性的特征。

实验证明，在语音识别系统中，这种特征可以有效提高语音识别的鲁棒性，在噪声较强（信噪比0dB到15dB）的情况下，识别率可以提高20％以上；并且，在干净语音的情况下又能保证识别率没有大的下降；同时，这种特征上的处理方法对各种噪声的适应能力都很强，无需对噪声进行预先分类即可得到很好的抗噪效果。

1. 引言

语音识别系统往往由于其训练使用不含噪的语音而导致识别的准确率在含噪的真实环境中会有大幅度的下降，其原因主要在于训练集和测试集的差别。因此，语音识别系统的鲁棒性是其走向实用必须解决好的一个重要问题。近年来，很多技术力图提高语音识别的鲁棒性。

子带语音识别（SubBand Speech Recognition）[1]是一种基于子带特征（Sub-band Mel-frequency cepstral coefficients）的方法。在子带语音识别中，语音信号在频域被切成几个不重复的频段，称为子带，对这些子带分别做特征提取，然后用一些合并规则将几种特征合成一个特征进行识别，或者分别对几种特征进行识别，然后将结果使用加权等方式综合。

子带语音识别基于下面的事实：在普通的特征提取中，一个子带中的噪音将影响整个特征，而在子带特征提取的情况下只会影响子带对应的特征。因此，这种方法在窄带噪音的情况下能有效的提高识别的准确率。

然而，子带语音识别在现代语音识别系统中并不常用，这是由下面的原因造成的。

- 对干净语音的识别准确率较低

  这几乎是所有鲁棒语音识别方法的通病。鲁棒语音识别会在模型空间将语音识别单元的边界混淆，以此来达到鲁棒的效果。不幸的是在干净语音的情况下处于边界处的语音识别单元往往会识别错误。

  对于子带语音识别来说，它还有其方法的独特问题。那就是，单独子带的识别会丢失子带之间的信息。尽管可以使用多子带特征合并成一个特征识别的方法，但这种合并仍然比基于DCT变换的特征合成方法差。这样就使干净语音的识别准确率下降了。

- 子带无法自由调节

  在子带语音识别中，子带的个数和分割方法对子带识别的效果有很大的影响。然而，这些要素无法轻易的调节，改变子带的方法必须将整个模型重新训练，因此是很难方便的实现的。同时，这也限制了各种自适应方法的使用。

本文提出一种方法，将子带信息结合到特征提取中。通过使用谱减和谱加权，吸取子带识别方法的优势，避免了子带方法的缺点。实验证明新方法和标准方法相比具有很好的鲁棒性。

本文的篇章是这样组织的，第二节详细叙述了提出的鲁棒语音特征提取框架，第三节给出了这种方法的实验及结果，第四节是结论和未来工作。

2. 基于子带信息的鲁棒特征提取框架

各种鲁棒特征方法的根本思想是抑制语音中噪声的影响，子带语音识别方法就可以有效的抑制含噪频段的影响。从这个想法出发，我们提出一种基于MFCC的鲁棒语音特征提取方法。图1就是这个方法的框图，阴影部分是和标准MFCC计算不同的部分。

采样率为8或16 kHz的语音，经过预加重和哈明窗之后，被切割成一些互相重叠的帧。对每帧进行傅立叶(FFT)变换之后，我们就可以开始对语音进行鲁棒特征处理了。首先我们估计频带语音的噪声强度，根据噪声强度，对语音进行谱减和谱加权，从而达到抑制噪声的目的，最后按照标准方法继续计算MFCC特征。

2.1. 噪声估计

对于噪声处理来说，噪声强度估计是一项必须而且重要的环节。我们使用了一种基于子带频谱能量分布的方法来估计噪声强度，这种方法最早在[6]中有所描述。图2显示了干净语音和带噪语音的频谱图和子带频谱能量分布图，可以明显看出分布图的峰的位置正处

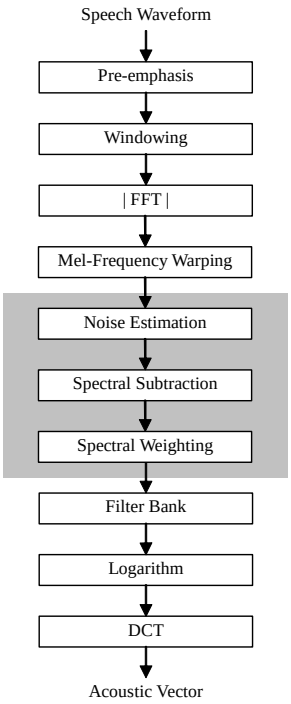

图 1: 基于MFCC的鲁棒语音特征提取

于噪声能量上。因此，我们可以使用下面的算法估计子带噪声强度。

子带噪声强度估计算法

1. 计算频谱能量分布

   定义S_ω为子带ω的短时频谱，S_ω的分布函数可以定义为：

$$D(S_\omega) = pdf(S_\omega)$$

2. 子带噪声能量$\hat{N}_\omega$处于$D(S_\omega)$的峰值位置：

$$\hat{N}_\omega = \arg\max_{S_\omega}(D(S_\omega)) \quad (1)$$

3. 信噪比（Signal to Noise Radio，SNR）$SNR_\omega(t)$的平滑计算方法

$$SNR_\omega(t) = \tau SNR_\omega(t-1) + (1-\tau)\frac{|S_\omega|}{|\hat{N}_\omega|} \quad (2)$$

2.2. 谱减

被噪声严重污染的子带会对识别的准确率产生很大的影响。谱减（Spectral subtraction, SS）方法可以降低被噪声污染的子带中的噪声强度。然而，在清除噪声的过程中，谱减方法很容易清除过多而造成子带中新的噪音，例如“音乐噪音”。在我们的方法中，我们使用一种平滑的谱减方法(3)来减少这种额外噪声的产生。

$$SS(S_\omega) = \begin{cases} (S_\omega^\gamma - \alpha\hat{N}_\omega^\gamma)^{\frac{1}{\gamma}} & S_\omega^\gamma - \alpha\hat{N}_\omega^\gamma \geq 0 \\ 0 & S_\omega^\gamma - \alpha\hat{N}_\omega^\gamma < 0 \end{cases} \quad (3)$$

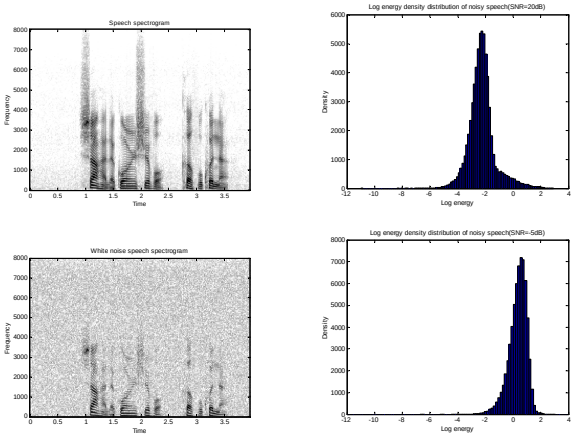

图 2: 干净语音和带噪语音（白噪音，−5dB）的频谱图和子带能量分布图

2.3. 谱加权

谱减方法的主要问题在于假定了噪声的稳定性。当这个假定不成立时，噪声估计是很难准确的。在这种情况下，谱减方法的效果很差。

另外，由于噪声叠加的不可恢复性，谱减会破坏原始语音的频谱。特别是噪声较大时，谱减往往在抹去噪声的同时，也将语音抹去了很多，因此原始语音的频谱很难恢复。从这个角度讲，经过去噪处理以后，噪声能量越的子带往往会被破坏的越严重。

含噪语音的识别实验[4]表明如果忽略含噪的部分特征，识别准确率会有明显的提高。1953年Flecher[3]的实验也表明人类的听觉系统通常能够在含噪语音中的干净子带部分得到足够多的信息，而忽略噪声污染的部分。以上的原因给我们以启示，我们提出一种加权的方式，作为对谱减方法的补充，实现这种忽略噪声部分的模式。

最基本的思想是很简单的，我们通过降低含噪子带的能量来达到忽略噪声影响的目的。子带加权根据噪声能量大小来降低子带能量，噪声能量越大，我们认为子带被破坏越大，从而能量降低就更多。在极端的情况下，如果子带被严重污染，我们可以降低能量到很小，甚至降低到能量为0，从而使未受或少受污染的子带能够更显著的在识别中发挥作用。这个方法和谱减方法的不同点在于，谱减的方法只是对子带自身降低能量，而我们的方法借鉴了子带语音识别的思想，根据含噪大小突出了子带之间的区别。

显而易见，如图1所示，MFCC的计算中，在对频谱进行求对数之前都是线性操作。因此我们在频域实施加权方法和在时域是等同的。这样我们就称这种方法为谱加权（Spectral Weighting, SW）。谱加权方法可以用以下的式子表示：

$$SW(S_\omega) = S_\omega \cdot W(SNR_\omega) \quad (4)$$

这里，$W(SNR_\omega)$是加权函数，其中SNR_ω代表子带ω中的信噪比，可以由(1)式计算而得。

从加权函数的意义来看，$W(SNR_\omega)$应该

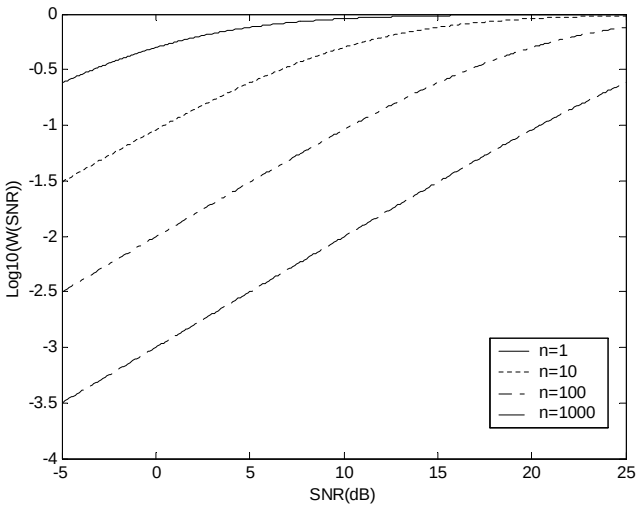

图 3: $W(SNR)$在各种n值下的对数坐标曲线

是SNR_ω的递增函数。而且$W(SNR_\omega)$应满足下面的式子：

$$\lim_{SNR\to 0} W(SNR_\omega) = 0$$
$$\lim_{SNR\to \infty} W(SNR_\omega) = 1$$

因此，我们很容易得到一组加权函数：

$$W(SNR_\omega) = \frac{SNR_\omega}{n + SNR_\omega} \quad (5)$$

式子中的n可以调节加权力度的强弱，n越大，加权的强度就越大。图3显示了$W(SNR)$在各种n值下的对数坐标曲线。

2.4. 对训练语音的处理

在语音识别中，任何一种噪音处理方法都会在子带中残留有噪音的能量，从谱图中我们可以观察到去噪后的语音和干净语音明显的不同。这主要是由于噪声估计的不准确而造成的。因此，一种比较有效的方法就是在训练集实施这些噪声处理方法。尽管我们的训练集基本上是没有噪声的，但去噪方法一样会对语音进行处理。这样，训练语音的频谱就会和测试语音更相似，因此在识别中会有更好的效果。

3. 实验

3.1. 数据和实验准备

我们进行的实验是孤立数字识别，使用安静环境下录制的16 kHz数据。训练数据包括由24个不同的人阅读的1520个孤立数字语音，测试数据包括由4个人读出的252个孤立数字语音。为了对噪声环境下的数据进行测试，我们在测试集的每个语音中手工加入了信噪比为20，15，10，5，0和-5dB的噪声，为了方便对比，我们定义干净语音的信噪比为25dB。噪声数据是从NOISEX 92标准噪声数据库中得到的。我们选择了四种典型噪声，进行我们的噪声实验。包括稳定的噪声：白噪声（White）和飞机噪声（F16 corkpit），以及不稳定的噪声：说话噪音（Babble）和工厂噪音（Factory1）。

实验中，语音被切割成32ms的小段，相邻段之间有16ms的重叠，然后在频域分割成26个频段，最后求出13个特征参数以及其差分和二阶差分。特征中我们还使用了频谱归一化（Cepstral Mean Normalization）的方法。我们使用隐马尔可夫模型（HMM）进行识别，识别中共使用了10个孤立数字模型，分别是数字零到九。每个模型包含6个状态，所有的状态都使用8个高斯混合。

3.2. 实验结果和分析

首先，我们使用谱减和谱加权方法对干净语音进行了测试，结果如表1所示，其中，B表示标准MFCC特征，SS表示对特征进行了谱减操作，SW表示在谱减的基础上进行了谱加权。

从实验的结果可以看出，由于噪声估计的误差，谱减方法使干净语音的识别准确率下降了4%，而谱加权的方法在这里显示出很好的适应性，在噪音估计不准确的情况下，并没有使识别准确率下降，反而在谱减的基础上上升了0.8%。

根据谱加权的原理，由于此方法只考虑误差的相对性，而不考虑误差的绝对性，因此噪声估计的不准确性对其影响不是很大。

接下来，对各种噪声情况我们使用谱减和谱加权方法进行测试。稳定噪声和不稳定噪声的结果分别列在表2和表3中。可以看出，无论在何种情况下，谱减方法都能有效的提高识别的准确率，而谱加权的方法在谱减的基础上可以更进一步提高识别率。

在实验结果中，我们可以观察出如下几个现象：

- 在极低信噪比下噪声处理的效果较差

  从结果中可以看出，在信噪比为-5或0dB时，经过谱减后，白噪声和工厂噪声的识别准确率几乎没有提高，说话噪声和飞机噪声也只是由于其基准很低而有限的提高。可以看出，这种情况主要是由于噪声处理效果较差而造成的。首先，在这种情况下，噪声估计很难准确，其次由于此时噪声的能量已经比语音能量大，噪声处理很难回复原始语音频谱，因此很难得到好的效果。

- 谱加权方法对非谱均衡的噪声效果较好

  由于谱加权针对噪声谱中的不均衡性进行加权，因此，从结果中可以看出，对于均衡谱的噪声，例如白噪声，并没有什么效果。

  但实际上，除了白噪声以外，其余噪声均是不均衡的噪声。因此在所列出的各种其他噪声结果中，性能都有明显的提高。

- 谱加权方法对不稳定的噪声效果较好

  谱减对不稳定的噪声具有一定的局限性，性能会相比稳定噪声差。而谱加权方法虽然也有这

表 1: 干净语音的识别准确率

SNR	Baseline	SS	SW
Clean	98.81	94.05	94.84

表 2: 含稳定噪声（White，F16）语音的识别准确率

SNR	White			F16 Corpkit		
	B	SS	SW	B	SS	SW
20	92.9	92.46	93.25	89.3	93.56	92.86
15	86.5	91.67	93.25	68.3	83.33	90.08
10	71.8	83.73	85.71	39.7	65.48	80.95
5	45.2	70.24	75.00	23.4	40.48	57.14
0	28.2	53.57	57.14	19.1	30.95	30.16
-5	17.0	39.29	40.08	14.7	26.98	27.78

表 3: 含不稳定噪声（Babble，Factory1）语音的识别准确率

SNR	Babble			Factory1		
	B	SS	SW	B	SS	SW
20	94.4	89.29	93.25	95.24	93.25	93.25
15	83.3	86.51	89.68	83.73	88.1	90.48
10	58.7	70.24	77.78	63.1	75.79	84.13
5	36.5	45.24	60.71	38.1	58.33	66.27
0	21.4	37.7	41.27	32.54	38.1	48.81
-5	18.3	25	27.38	26.59	25	31.35

方面的影响，但相对要少一些。从实验结果来看，无论是稳定噪声还是不稳定噪声，在谱减的基础上，谱加权方法都有比较稳定的性能增长。

4. 结论及未来工作

本文提出的特征提取框架，区别于标准的MFCC，和其他鲁棒语音识别方法比较，有如下的优点：

- 有效性
  方法能够有效的提高鲁棒语音识别的准确率。如图4所示，在噪声较强（信噪比15dB及以下）的情况下，识别率可以提高20％以上。
- 适应性
  方法在干净语音的情况下又能保证识别率没有大的下降；同时，对各种噪声，包括稳定噪声和不稳定噪声均能有较好的效果。
- 可行性
  方法可以作为标准语音系统的一个预处理部分，直接嵌入系统，无需对代码进行大量改动。同时，方法无需对噪声进行预先分类即可得到很好的抗噪效果。

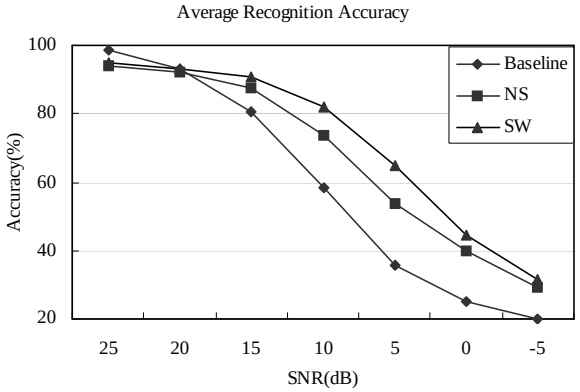

图 4: 识别结果的平均曲线

- 可扩展性
  方法可以和其他各种鲁棒方法，例如模型和其他信号方法结合使用。同时，方法由于具有模块性，可以方便的修改其中某个模块，以达到特殊要求和更好的效果。

未来工作主要是在方法中添加一定的自适应性，能够适应噪声大小和类型的变化。同时结合其他鲁棒方法，做更深一步的测试。最后希望能够探索极低信噪比噪声下的噪声处理方法。

5. 参考文献

[1] H. Hermansky, S. Tibrewala, and M. Pavel, “Towards ASR on partially corrupted speech” Proceedings International Conference on Spoken Language Processing, Philadelphia, October 1996.

[2] Okawa, S., E. Bocchieri and A. Potamianos, “Multi-band speech recognition in noisy environments,” in Proceedings of ICASSP ’98, pages 641–644, Seattle, Washington, USA. IEEE.

[3] H. Fletcher, “Speech and Hearing in Communication,” Krieger, New York, 1953

[4] M. Cooke, A. Morris, and P. Green, “Recognising occluded speech,” Proceedings of the ESCA Workshop on the auditory basis of speech perception, pages 297-300, 1996.

[5] L. Deng, A. Acero, M. Plumpe and X. Huang, “Large-Vocabulary Speech Recognition under Adverse Acoustic Environments,” in Proc. Int. Conf. on Spoken Language Processing. Beijing, China, Oct, 2000

[6] H. G. Hirsch, and C. Ehrlicher, “Noise Estimation Technique for Robust Speech Recognition,” Proc. ICASSP ’95, pp. 153–156, May 1995.