基于谱熵的语音检测 [1]

吴启晖 王金龙

(解放军理工大学通信工程学院移动通信教研室 南京 210016)

摘 要 根据离散余弦变换 (DCT) 特性和最大离散熵定理，该文提出利用短时语音余弦变换谱的谱熵进行语音信号检测 (Voice activity detection)．在强噪声环境下，传统的能量，过零率，相关等检测方法将会失效，而谱熵法则具有稳健的抗噪特性．计算机模拟显示这是一种比较好的抗噪语音检测方案

关键词 语音检测，DCT 变换，谱熵，抗噪

中图号 TN912.3

1 引 言

语音信号检测 (VAD) 目的是能够正确区分语音与各种背景噪声 [1]．在语音信号处理、通信等领域，它有着十分重要的意义．在典型的电话、多媒体通信中，无语音段压缩后，可以利用话音信道进行数据传输 [2]；在移动通信系统中，采用不连续突发方式如 GSM，语音信号检测可以减少邻道干扰、增加系统容量、延长充电电池工作时间；在 UMTS(Universal Mobile Telecommunication Systems)，可以减少平均比特速率，提高系统容量；在军用软件无线电中，实现数字静噪；战术电台接入网中，实现战术短波电台接入。实际场合中，常常各种背景噪声和语音信号混在一起，影响着语音检测的性能。因此，语音检测的抗噪性能也越来越引起人们的重视。在移动通信，野战通信中尤为突出。寻求一种抗噪性能好的语音检测方法是非常重要的。有良好抗噪特性的语音信号检测算法在变速率语音编码中，可以降低平均码率。码率的降低将意味着增加移动蜂房通信系统的容量，在 CDMA 移动通信系统中更为明显。在单工电台入网中，它则直接影响到通信的可靠性。单工电台入网军标征求意见稿中，提出要做到 0dB 的可靠语音检测。

语音信号检测算法主要依据语音与噪声的不同特性进行语音 / 噪声判决．传统的方法有能量 [3]、过零率、零能比、时域、频域的基音检测 [4,5] 等方法，这些算法都是建立在相对比较理想的条件下，要求背景噪声保持平稳，信噪比较高。实际工作中，这些条件很难得到满足，经常会遇到背景噪声干扰。采用神经元网络 [6] 和最大似然法可以进行语音检测。前者因需要训练过程，而不能在很多场合得到应用。后者在因信道原因基音缺损时，将不能很好地进行检测，且抗正弦能力弱。

本文对含噪语音的短时余弦谱进行谱分析，用谱熵的方法检测谱的平坦程度，从而达到检测语音的目的。同时，检测可靠性与含噪语音信号的大小无关，只与信噪比有关。模拟显示在强噪声环境下，用谱熵法检测语音效果良好。

2 谱熵检测原理

谱熵检测法原理方框图如图 1 所示。

[1] 1999-10-19 收到，2000-04-07 定稿

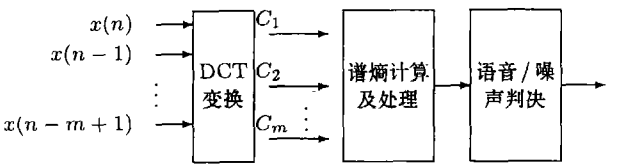

图 1 谱熵语音检测原理

本节从以下几个方面说明谱熵在语音检测中的基本原理.

2.1 DCT 变换

离散余弦变换 (DCT) 在图像处理, 数据压缩, 模式识别等领域有着广泛的应用. 其性能仅次于卡南-洛伊夫变换 (KLT 变换), 明显优于傅里叶变换, 哈尔变换、 Hadamard 变换、斜变换. 正交变换是将时域信号反映到系数空间 (频域) 上. 时域内具有强相关的信号反映到频域上, 常常是某些特定区域内的能量被集中在一起. 我们需要一种正交变换能够最大限度地把时域的语音信号的相关特性反映到频域的集中特性上去, 以便用谱熵法进行检测. 人们经过研究, 发现 KLT 变换是去时域相关特性实现频域能量集中的最佳正交变换. 但 KLT 运算量极大, 且没有快速算法, 而 DCT 变换是次最佳的, 又有快速算法. 这正是人们所需要的.

DCT 变换并不仅仅是简单地将傅里叶变换的复数运算转化到易于计算的实数域中进行运算方法的产物. DCT 是以 Chebyshev 多项式 $T_k(z_n)$, $k,n=0,1,\cdots,N-1$, 在 $T_N(z_n)$ 的各零点上的离散值为基向量定义的. 前 N 个 Chebyshev 多项式是

$$\left.\begin{aligned} T_0(z) &= 1/\sqrt{N} \\ T_k(z) &= \sqrt{2/N}\cos[k\cos^{-1}(z)], k=1,2,\cdots,N-1 \end{aligned}\right\} \tag{1}$$

第 $(N+1)$ 个 Chebyshev 多项式 $T_N(z)=\sqrt{2/N}\cos[N\cos^{-1}(z)]$. 为了求出 $T_N(z)$ 的各个零点, 令 $T_N(z)=0$, 可得 $z_0=\cos[(2n+1)\pi/(2N)]$, $n=0,1,\cdots,N-1$. 计有 N 个零点, 代入 (1) 式, 就得到

$$\left.\begin{aligned} T_0(n) &= 1/\sqrt{N} \\ T_k(n) &= \sqrt{2/N}\cos[(2n+1)k\pi/(2N)] \end{aligned}\right\} \tag{2}$$

其中 $n=0,1,\cdots,N-1$, $k=0,1,\cdots,N-1$.

根据 (2) 式可以定义 DCT:

$$C_x(0)=\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}x(n)$$

$$C_x(k)=\sqrt{\frac{2}{N}}\sum_{n=0}^{N-1}x(n)\cos\frac{(2n+1)k\pi}{2N},\quad k=1,2,\cdots,N-1$$

研究表明 [7], DCT 有同 KLT 较为相近的基向量, 其去数据相关性的能力仅次于 KLT, 且快速算法 [8,9] 运算量为 $O(N\log 2N)$ 与 FFT 相近.

图 2 和图 3 显示噪声、浊音的 DCT 和 FFT 变换谱. 其中纵坐标 $\log C$ 表示归一化的变换系数取以 2 为底的对数. $C_i=|c_i|/\sum_{i=1}^{N}|c_i|$, c_i 为变换系数.

图 2 噪声的 FFT 谱谱熵为 6.8218, DCT 谱谱熵为 6.6112. 图 3 浊音的 FFT 谱谱熵为 5.0424, DCT 谱谱熵为 4.1897. 由此可见, 与常用的 FFT 相比, 用 DCT 谱谱熵来区分噪声与浊音则具有更大的隔离度. 因而, 我们采用 DCT 变换.

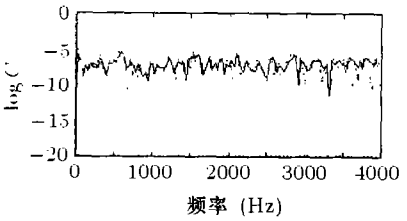

图 2　噪声谱
— 实线为 FFT 变换谱，-- 虚线为 DCT 变换谱

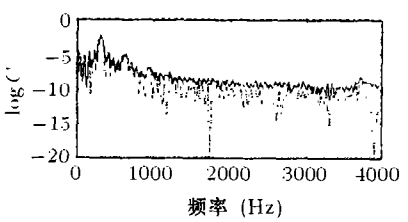

图 3　浊音谱
— 实线为 FFT 变换谱，-- 虚线为 DCT 变换谱

2.2 谱熵

熵这个字来源于统计热力学，是紊乱程度的测度。信息论借用它来表示信源的平均不确定性。信源的平均不确定性的变化可以用熵函数来表示。设离散信源 X，其概率空间为

$$\begin{bmatrix} X \\ P(x) \end{bmatrix} = \begin{bmatrix} a_1, a_2, \cdots, a_q \\ p_1, p_2, \cdots, p_q \end{bmatrix}$$

则熵函数为：$H(P) = H(p_1, p_2, \cdots, p_q) = -\sum_{i=1}^{q} p_i \log p_i$，其中 $P = (p_1, p_2, \cdots, p_q)$ 是 q 维矢量，并满足 $\sum_{i=1}^{q} p_i = 1$ 和 $p_i \geq 0$，故常称 P 为概率矢量。熵函数有个很重要的特性——极值性 [10] 即 $H(p_1, p_2, \cdots, p_q) \leq H(1/q, 1/q, \cdots, 1/q) = \log q$。也就是等概分布时，熵达到极大值。这表明等概分布时信源的平均不确定性为最大，这一结论被称为最大离散熵定理。

现在令归一化 DCT 系数 C_k 为频率点 K 的出现概率，则谱熵为 $H(C) = -\sum_{k=1}^{N} C_k \log C_k$，其中 $C_k = |c_i| / \sum_{i=1}^{N} |c_i| = |c_i|/A$，$c_i, i = 1, 2, \cdots, N$ 是 N 个语音样值的 DCT 变换系数。当归一化 DCT 系数均相等（即谱最平坦）时，有最大谱熵。谱熵函数可以很方便地描述谱的平坦特性。由图 2 和图 3 可以看出噪声谱较为平坦，其谱熵较大。语音能量集中在低频段，其谱熵较小。谱熵大的为噪声，小的为语音，这就是谱熵检测语音的原理。同时，我们还可以看到，由于 DCT 系数被归一化了，所以谱熵检测法可靠性不会受信号大小的影响，只与信噪比有关。

3 谱熵检测原理的应用与计算机模拟

由于应用场合不同，对语音检测的性能要求也不同，因而基于谱熵检测原理需采用不同的方法。

3.1 变速率语音编码

在变速率语音编码中，不仅需要检测出浊音段 (voiced segment) 和无声段 (silent segment)，还需要检测出清音段 (unvoiced segment)。对于不同的语音段采用不同的编码速率，从而在达到一定语音质量的前提下，使平均码率降低。这就需要计算全频带，高频带，低频带的谱熵来区分各语音段 (表 1)。

表 1　各语音段各频带谱熵比较

	全频带 (谱熵)	高频带 (谱熵)	低频带 (谱熵)
无声段	高	高	高
浊音段	低	高 / 低	低
清音段	低	低	高

3.2 战术单工电台接入网

本文着重讨论谱熵检测原理在单工电台入网中的应用，并进行计算机模拟。单工电台入网中要求 SNR=0dB 时，语音检测可靠性为 95% 以上，但允许有 150ms 左右的延迟。换句话说，

SNR=0dB，150ms 内，检测到有 / 无语音的正确概率为 95%．针对单工电台入网的特点，我们采用以下一些处理方法来减小运算量提高检测性能．

(1) 由于浊音与噪声有着明显的区别，通过检测有无浊音来达到检测语音的目的，所以只需计算全频带的谱熵．与变速率语音编码相比，运算量减小了近 1/2．

(2) 对 $-x\log_2 x$ 求极值可知，$x=1/2$ 时 $(\log_2 x=-1)$ 有极大值．$x<1/2$ 时 $(\log_2 x<-1)$，$-x\log_2 x$ 成单调上升．从图 2 和图 3 可知，噪声谱较为平坦，浊音能量集中在 1000Hz 以下的低频段．在低频段内，图 3 的 $\log C$ 基本上大于图 2 的 $\log C$．所以在低频段内，根据单调性可知图 3 浊音谱熵 (3.3978) 大于图 2 噪声谱熵 (1.5839)．根据谱熵检测原理可知，利用全频带的谱熵进行检测则低频段的谱熵起了相反的作用．又由于短波通信中，话带带宽常常是 300Hz~3400Hz(有时为 300Hz~2700Hz)．因而我们采用 1000Hz~2700Hz 谱熵的 2 倍作为全频带的谱熵．这样减小了约 1/2 的计算量，且提高了检测的性能．从图 4 中曲线 2 和曲线 3 可以看出，对于噪声段，两者相近．对于浊音段，曲线 3 的谱熵远小于曲线 2 的谱熵．因而采用低频段的谱熵来检测语音，则具有更佳的性能．

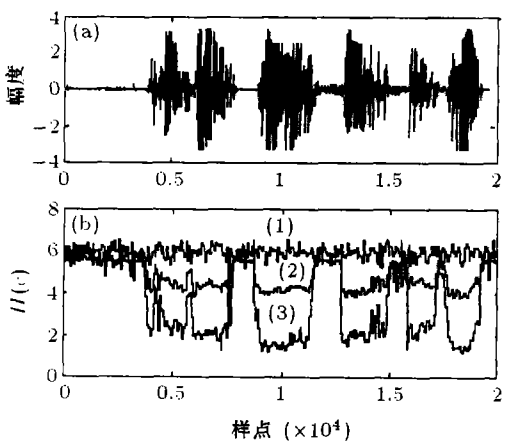

图 4 (a) 为高信噪比下的话音 8kHz 采样，20000 个样点话音内容为 "长江，我是黄河"
(b) (1) 为噪声谱熵，(2) 为对应于 (a) 图的全频带谱熵，(3) 为对应于 (a) 图的 1000Hz~2700Hz 频带谱熵的 2 倍

(3) 由于单工电台入网中对浊音的起止并不需要精确的检测出来，同时噪声的个别帧的谱熵较低如图 5 所示．

因而可以采用光滑技术求得短时平均谱熵，来减小噪声个别帧的影响．经过光滑处理以后的模拟结果如图 6 所示．实验结果表明，针对单工电台入网，利用谱熵检测法对 SNR=0dB 的语音进行检测，正确概率为 99%．

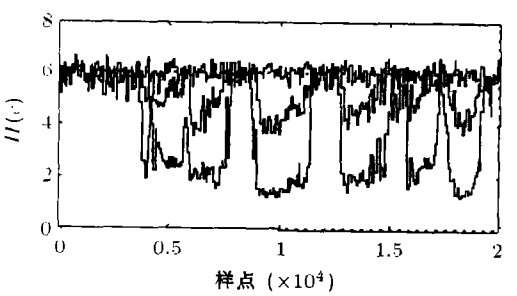

图 5 (1) 噪声谱熵，(2)SNR=0dB 的语音谱熵，(3) 高信噪比的语音谱熵

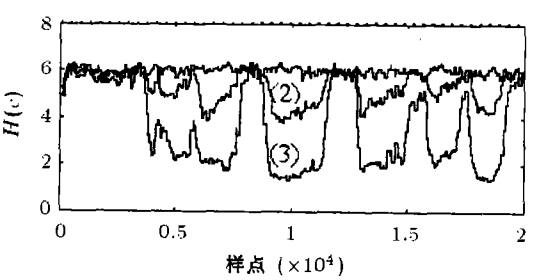

图 6 (1) 噪声谱熵，(2)SNR=0dB 的语音谱熵，(3) 高信噪比的语音谱熵

4 结束语

本文针对强噪声下的语音信号检测问题，提出采用谱熵方法进行检测，并阐述了谱熵检测的原理。结合战术单工电台接入网研制课题及特殊性，对谱熵法进行进一步改进，计算机上模拟显示，效果良好。低信噪比下的语音检测技术在一些领域是很重要的，谱熵检测法则是一种较好的选择。

参 考 文 献

[1] F. Beritelli, S. Casale, A. Cavallaro, A robust voice activity detector for wireless communications using soft computing, IEEE J. on SAC, 1998, SAC-16(9), 1818–1829.

[2] R. V. Cox, P. Kroon, Low bit-rate speech coders for multimedia communication, IEEE Commun. Mag., 1996, 34(1), 34–41.

[3] Qualcomm, Inc., Digital Cellular System CDMA–Analog Dual-Mode Mobile Station–Base Station Compatibility Standard, March 5, 1992.

[4] L. R. Rabiner, On the use of autocorrelation analysis for pitch detection, IEEE Trans. on Acoust., Speech, Signal Processing, 1977, ASSP-25(1), 24–33.

[5] S. Seneff, Real-time harmonic pitch detector, IEEE Trans. on Acoust, Speech, Signal Processing, 1978, ASSP-26(2), 358–365.

[6] A. Bendiksen, K. Steiglitz, Neural networks for voiced/unvoiced speech classification, ICASSP'90., Bonn, Germany, 1990, 521–524.

[7] 高文，多媒体数据压缩技术，北京：电子工业出版社，1992, 48.

[8] Byeong Gi Lee, A new algorithm to compute the discrete cosine transform, IEEE Trans. on Acoust., Speech, Signal Processing, 1984, ASSP-32(6), 1243–1245.

[9] N. Ahmed, T. Natarajan, K. R. Rao, Discrete cosine transform, IEEE Trans. on Computer, 1974, C-23(1), 90–93.

[10] 傅祖芸，信息论基础，北京：电子工业出版社，1989 年，23.

VOICE DETECTION BASED ON SPECTRAL ENTROPY

Wu Qihui Wang Jinlong

(*Teaching Section of Radio Communication, ICE, Nanjing* 210016, China)

Abstract An approach is introduced to voice detection that differs from normal approach via energy, correlation and zerocrossing criteria. By measuring the gross shape of the short-term speech spectrum using spectral entropy to detect voice segment, it is shown that the spectral entropy can be used effectively even in heavy background noise. The simulation results show that the approach via spectral entropy has good performance for anti-noice.

Key words Voice detection, DCT transformation, Spectral entropy, Robustness

吴启晖： 男，1970 年生，博士，曾获军队科技进步一等奖一项，主要从事软件无线电、短波通信、数字信号处理、移动通信等方面的研究工作.

王金龙： 男，1963 年生，博士，教授，博士生导师，曾获军队科技进步一等奖两项，二等奖一项，三等奖三项，主要研究方向为短波通信、数字通信、数字信号处理、移动通信等.