一种自适应分组丢失率的多描述正弦语音编码器

郎 玥 赵胜辉 匡镜明
(北京理工大学信息科学技术学院 北京 100081)

摘 要：多描述（MD）语音编码器可以在不可靠的信道如 internet 上稳定地传输语音信号。然而，当前的 MD 语音编码器一般对不同的分组丢失率采用固定的多描述结构，不能很好地适应实际网络环境中分组丢失率的实时变化。该文提出一种自适应多描述正弦编码器(AMDSC)，可根据分组丢失率的大小在两个描述间动态地分配冗余，从而使最终的重建失真最小。仿真结果表明，AMDSC 的重建语音质量相对于其他固定结构的 MD 编码器有明显改善。

关键词：语音编码；多描述；正弦编码
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2008)06-1354-05

An Improved Multiple Descriptions Sinusoidal Speech Coder Adaptive to Packet-loss Rate

Lang Yue Zhao Sheng-hui Kuang Jing-ming
(*School of Information Science and Technology*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: Multiple Descriptions (MD) speech coder can provide robust speech communication over unreliable channels such as the internet. However, existing MD speech coders use fixed scheme for different packet-loss environments. A novel Adaptive Multiple Descriptions Sinusoidal Coder (AMDSC) is proposed, which can vary in coding multiple descriptions according to the network packet-loss rate with optimally adding redundancy between two descriptions in order to make the final distortion minimum. Simulation results show that the proposed AMDSC outperforms existing MD speech coders by taking network loss characteristics into account.

Key words: Speech coding; Multiple descriptions; Sinusoidal coding

1 引言

在过去的几年里，VoIP 获得了广泛的重视，并取得了巨大的成功。但是在不可靠的分组网络上，由于分组丢失的存在，传输的语音质量还不尽如人意。传统的处理分组丢失的方法是重传。但是当分组丢失率较高时，重传会导致更加拥塞的环境，并且不能满足实时性的要求。与重传不同，多描述编码(MDC)可以显著提高传输的稳定性，而又不引入明显的时延[1]，是一种有效的解决分组丢失的方法。

MDC 的思想已经成功地应用到语音编码中。近年来，研究人员提出了许多多描述语音编码器[2−5]。这些编码器通过在描述间增加冗余来提高语音的传输稳定性，因为如果一个描述丢失了，解码器仍然能够从其它正确接收的描述中获得部分信息，从而部分地恢复原始信号。但是现有的 MD 语音编码器以一种固定的模式来增加冗余，而不考虑分组丢失率的变化。本文的想法是描述间的冗余应该自适应于分组丢失率，以两个描述的情况为例：在网络的分组丢失率比较高时，每次只收到一个描述的概率相对较大，因此希望两个描述之间的冗余较大，这样单个描述得到的重建质量相对高一些；而当分组丢失率比较低时，两个描述同时被收到的概率相对较大，所以应使两个描述之间的冗余减小，这样由两个描述得到的重建质量更高。然而，在给定比特率的条件下，不可能使单个描述与两个描述的恢复质量同时达到最优，所以需要动态调整两个描述之间的冗余量，从而达到单个描述和两个描述的恢复质量之间的最佳折衷。

作者提出了一个分组丢失自适应的多描述正弦编码器（AMDSC），该编码器可以通过动态地调整冗余来实现上述的质量折衷。它基于多描述正弦编码器(MDSC)[6]，通过对 MDSC 进行简单的率失真分析得到最优的冗余量，从而在分组丢失变化的情况下，使重建失真最小。

2 多描述编码(MDC)

一个基本的 MDC 的结构如图 1 所示。语音信号被编码形成两个或者多个描述并独立传输。每个描述可以单独解码，部分地恢复原始信号（也就是图 1 中的输出 1 和输出 2）。如果多个描述同时正确接收，它们可以联合起来获得一个更好的质量(输出 0)。

MDC 的基本思想是在两个描述之间引入相关性，也就是冗余。从而当一个描述丢失时，解码器可以从正确接收的描述里估计丢失的分量。文献[2，3] 中提出的编码器是波形多描述编码器。它将奇数和偶数样点分解到不同的描述。这

2006-11-09 收到，2007-03-20 改回

类编码器简单而有效，但是它们利用样点之间内在的固定冗余来恢复丢失的信息，因此不能自适应于分组丢失率的变化。而文献[4，5]中提出的编码器是参数多描述编码器，这类编码器不能灵活的产生两个描述，因此调整冗余的大小也不容易。

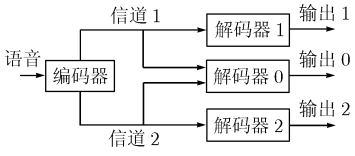

图 1 多描述编码框架

3 多描述正弦编码(MDSC)

文献[6]提出了 MDSC。该编码器主要基于正弦模型。对于一个普通的正弦模型，每一帧语音信号可以表示成一组正弦信号的和。因此，对于第 k 帧，有

$$\tilde{s}^k[n]=\sum_{i=1}^{N}A_i^k\cos(\omega_i^k n+\phi_i^k) \tag{1}$$

其中 N 为提取的正弦总数。参数集合包括幅度 $A=\{A_i\}$,频率 $\omega=\{\omega_i\}$ 和相位 $\phi=\{\phi_i\}$ 。它们通过迭代的贪婪算法提取出来。每次迭代时，具有最大内积的正弦信号被提取出来，然后得到误差信号。每次提取的正弦参数使得误差信号的能量最小。因此经过 m 次迭代之后的误差信号由

$$r_m[n]=r_{m-1}[n]-A_m\cos(\omega_m n+\phi_m) \tag{2}$$

给出，而误差信号的能量为

$$E_m=\sum_n\left|\omega_a[n]\{r_{m-1}[n]-A_m\cos(\omega_m n+\phi_m)\}\right|^2 \tag{3}$$

其中 $\omega_a[n]$ 是分析窗， n 在分析帧的范围内。该算法可参见文献[7]。

MDSC 编码器的结构如图 2 所示。正弦参数从每个分析语音帧中提取出来，然后进行标量量化。量化之后，正弦参数按照能量从大到小的顺序排列。序号为奇数和偶数的两组正弦参数被分解到两个不同的描述。前 L 个正弦的参数作为冗余存在于两个描述中，用以提高编码器的稳定性。等效矩形带宽(ERB)噪声模型[8]用来表示噪声状的误差信号。模型输出一个 20 维的噪声能量矢量，该矢量通过分裂矢量量化后，索引分解到两个描述中。最后两个描述中参数被分别打包并独立传输。

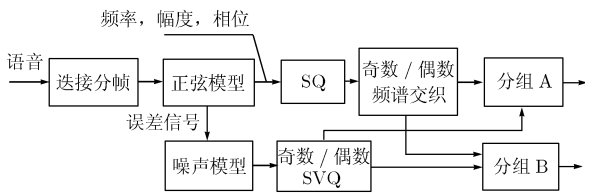

图 2　MDSC 编码器结构

4 分组自适应多描述正弦语音编码器(AMDSC)

本文对 MDSC 进行了改进。根据分组丢失率的大小，最优的选取两个描述中的冗余参数，从而使原始信号与重建信号之间的误差最小。

假定各个描述的丢失相互独立。在总比特率固定的限制下，需要确定两个描述间最优的冗余度，从而达到最好的重建质量，为此我们对 MDSC 进行了率失真分析。在 MDSC 的率失真分析中，用来量化每根正弦参数的比特数是固定的，速率的控制和速率的预测非常直接。失真的预测也很简单，因为每增加一根正弦，失真就会减少相应正弦幅度的平方。因此 MDSC 最优化的过程相比其他 MD 语音编码器要简单。

总的失真由下式给出：

$$D=p(1-p)D_1+p(1-p)D_2+(1-p)(1-p)D_0+ppD^* \tag{4}$$

其中 p 是每个描述分组的丢失概率， $D_i, i=0,1,2$ 依次是原始语音与中心解码器以及两个边解码器输出的重建语音之间的失真。 D^* 是当信息全部丢失时的失真。

速率基于如下的限制：

$$R_1+R_2\le R \tag{5}$$

其中 $R_i, i=1,2$ 是描述 1 和描述 2 的传输速率。 R 为总的可用传输速率。因为对噪声参数进行编码的比特数是固定的，所以以上速率不包括噪声参数的编码比特数。

式(4)中的最后一项可以忽略，因为它不包含任何自由变量。这是一个在非线性约束条件下的非线性优化问题，但是可以通过简化限制条件而使问题得到简化。因为用来量化每根正弦参数的比特数是一定的，从而编码每一帧语音信号的比特数正比于正弦的根数。又因为较高的速率会对应较低的失真，任何算法为了最小化失真，应该使用所有可用的带宽。基于这样的分析，速率的限制可以重新写成

$$N+L=T/K \tag{6}$$

其中 N 是每一帧语音信号提取的正弦的数目， L 是在两个描述之间重复的正弦数目，也就是冗余的数目， T 是每帧中量化正弦参数所需的总比特数， B 是每个正弦参数需要的量化比特数。所以式(4)中的 $D_i, i=0,1,2$ 可以写成：

$$D_0=\frac{\sum_{m=1}^{M}\left(S[m]-\sum_{i=1}^{N}\frac{\tilde{A}_i}{2}(e^{j\tilde{\phi}_i}W(m+\tilde{\omega}_iM)+e^{-j\tilde{\phi}_i}W(m-\tilde{\omega}_iM))\right)^2}{\sum_{m=1}^{M}S[m]^2} \tag{7}$$

$$D_1=\left[\sum_{m=1}^{M}\left(S[m]-\sum_{i=1}^{L}\frac{\tilde{A}_i}{2}(e^{j\tilde{\phi}_i}W(m+\tilde{\omega}_iM)+e^{-j\tilde{\phi}_i}W(m-\tilde{\omega}_iM))\right.\right.$$
$$\left.\left.-\sum_{\substack{k=L+2n+1\\n=0,1,2\cdots}}^{N}\frac{\tilde{A}_k}{2}(e^{j\tilde{\phi}_k}W(m+\tilde{\omega}_kM)+e^{-j\tilde{\phi}_k}W(m-\tilde{\omega}_kM))\right)^2\right]$$
$$\Big/\sum_{m=1}^{M}S[m]^2 \tag{8}$$

$$D_2=\left[\sum_{m=1}^{M}\left(S[m]-\sum_{i=1}^{L}\frac{\tilde{A}_i}{2}(e^{j\tilde{\phi}_i}W(m+\tilde{\omega}_iM)+e^{-j\tilde{\phi}_i}W(m-\tilde{\omega}_iM))\right.\right.$$
$$\left.\left.-\sum_{\substack{k=L+2n\\ n=1,2\cdots}}^{N}\frac{\tilde{A}_k}{2}(e^{j\tilde{\phi}_k}W(m+\tilde{\omega}_kM)+e^{-j\tilde{\phi}_k}W(m-\tilde{\omega}_kM))\right)^2\right]$$
$$\Big/\sum_{m=1}^{M}S[m]^2 \tag{9}$$

其中 M 为对语音进行 FFT 变换的样点数。W 为叠接分析窗(汉明窗)的频谱。$\tilde{A}_k$，$\tilde{\phi}_k$ 和 $\tilde{\omega}_k$ 分别为解码之后的幅度，相位和频率。以上失真为频域信噪比，比较适合正弦模型。

当提取的正弦数目 N 增大时，每个描述中重复的正弦数目 L 因受式(6)的限制而减小。当 N 增大时，根据式(7)-式(9)，D_0 降低，而 $D_i, i=1,2$ 增加。因为作为冗余的正弦是非常重要的，它们的能量很高，对于合成语音质量的贡献大。这些正弦中丢失任意一个都会显著降低语音的质量。当 L 增加时，D_0 增加，因为此时用于合成语音的正弦数量降低。$D_i, i=1,2$ 的变化不是单调的。首先，随着 L 的增加失真会减小，当 L 继续增长时，由于式(6)的限制，N 将会减小。此时没有足够的正弦参数来合成语音，因此描述 1 和描述 2 的失真会增加。从上面的分析可以看出，存在最优的冗余量。但是从式(4)中直接求取最优的冗余 L 非常困难，因此，我们通过仿真来估计最优的 L 。

5 仿真

5.1 仿真语料

本文采用 8 个语句来估计最优的冗余，它们选自 NTT-AT 宽带中文语音数据库，包括各种说话人类型和语句，以便覆盖所有的语音特征。语音采用 16kHz 采样 16bit 量化精度。

5.2 仿真速率

为了能够跟其他编码器在相同的速率下进行比较，本文将 AMDSC 的速率限制在 12.5kps。分析帧长为 32ms，两帧之间有 50%的叠加。所以每帧参数需要 200 个比特来编码。并且每根正弦参数需要 15 个比特来量化[6]。在去掉噪声参数的量化比特数(20)之后，式(6)中的 $N+L$，也就是每帧信号需要编码的正弦数目为 12。

5.3 仿真平台

本文中所使用的网络仿真环境采用 NS2 仿真软件[9]搭建，以获得比较真实的分组丢失状况。图 3 显示了网络的拓扑结构，这是一个常见的哑铃型拓扑。网络流量由以太网一些节点 N_i 产生，然后传输到另外的网络节点 S_i 中，两个网络的边缘节点(图 3 中的灰色节点)由一条 $E1$ 线路相连接。这条链路是网络的瓶颈，分组丢失主要发生在这里。

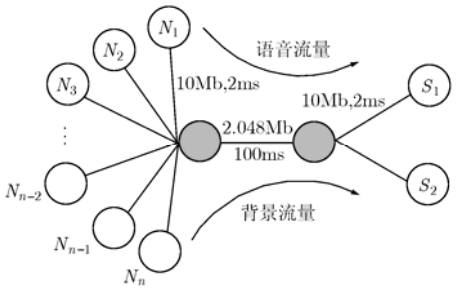

图 3 网络拓扑结构

仿真平台上有两种流量类型，由于正常会话中，只有大约 40%的时间真正包含语音，其余时间全是背景噪声[10]。所以在建模过程中，选择指数分布 ON/OFF 流量模型来进行模拟，设定开启和关闭的平均周期分别为 1.004s 和 1.587s，这相当于 38.53%的时间为通话时间， 符合 ITU-T 的标准[11]。分组的大小包括 40bit 的 IP 包头，以及一个描述所需要的比特数 100bit，速率设定为 12.5kps。另外一种是背景流量，由 CBR 模型来模拟，该流量占用大约 90%的带宽。不同的分组丢失率可以通过改变语音业务的用户数来实现。

5.4 仿真结果及分析

以分组丢失率 0.05 为步长下的冗余与失真的关系如图 4 所示。从图中可以看出，当没有分组丢失时，不需要增加任何的冗余，但是当有分组丢失时，增加冗余可以提高 MDSC 的性能，不同的丢失率，对应不同的最优冗余。为了进一步确定最优冗余与分组丢失率的关系，以分组丢失率 0.01 为步长，得到两个描述间最优的冗余数 L 与分组丢失率的关系如图 5 所示。

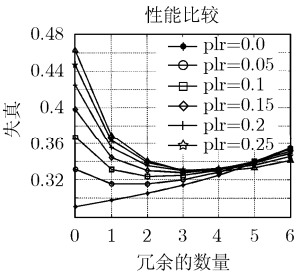

图 4 不同丢失率下的冗余与失真的关系

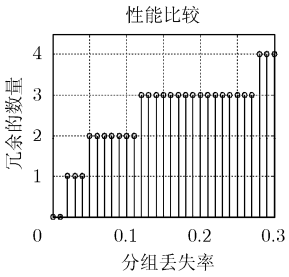

图 5 不同丢失率下最优的 L

从图 5 可以得到不同的丢失率范围下对应的最优的冗余数，从图中可以看出，最优冗余数随分组丢失率的变化并不剧烈，一种冗余分配方案可以在一定的分组丢失率范围内适用，所以冗余分配方案的调整边界不用十分严格。以分组丢失率 0.01 为步长，对应的冗余分配方案如表 1 所示：

表 1 不同丢失率范围对应的最优冗余数

分组丢失率 p 的范围	[0 0.02)	[0.02 0.05)	[0.05 0.12)	[0.12 0.28)	[0.28 0.3]
最优 L	0	1	2	3	4

根据表 1，我们重新设计了 MDSC，使得它可以自适应于信道的丢失率。AMDSC 的结构如图 6 所示。

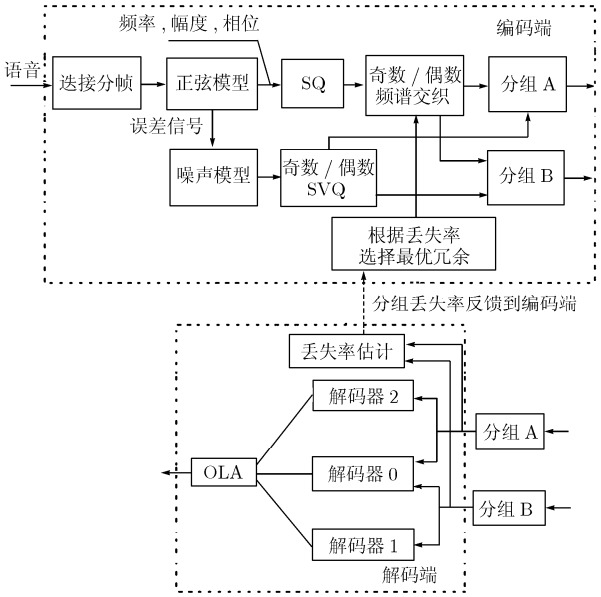

图 6　AMDSC 编解码器结构

在解码端增加了一个分组丢失估计模块。该模块每隔一个固定的时间间隔，将信道的状况反馈到编码器端。编码端增加丢失率到最优冗余的映射模块。该模块根据表 1 将丢失率映射到最优的冗余 L 。编码器根据这个信息进行参数调整。最优冗余的分配信息不需要传输到解码端，因为当两个描述都正确接收时，两个描述间参数相同的部分就是冗余。而当只有一个描述接收到时，该描述中所有的正弦参数都将用于合成语音信号。

为了对 AMDSC 的性能进行评价，将它与其它编码器 (13.5kbps 的 MD AMR-WB[4]，12.65kbps 的 AMR-WB 和 12.5kbps 的 MDSC (冗余固定为 1)[6])进行了比较。解码后的语音质量由宽带 PESQ[12]来评测。8 个 NTT-AT 的语句用于测试。测试语句与估计最优冗余的语句不同。仿真结果如图 7 所示。从图 7 可以看出：

(1)重建语音信号的 PESQ 值随着分组丢失率的增加而降低。AMDSC 的性能优于其他编码器。

(2)当没有分组丢失时，AMR-WB 和 MD-AMR 获得比其他两种编码器更好的重建质量。因为该类编码器所基于的 CELP 结构比正弦模型更适合语音信号。

(3)当没有分组丢失时，AMDSC 的性能优于 MDSC。这是因为两个描述之间没有引入冗余，编码效率获得了提高。

(4)随着分组丢失率的增加，AMDSC 相对于 MDSC 的性能增益接近常数。由此可见，AMDSC 的性能提升非常稳定。

(5)当分组丢失率在 10%–15%之间时，AMDSC 相对于 AMR-WB 获得了最大的性能增益。该分组丢失率是互联网上进行长距离通信的典型分组丢失率，这意味着 AMDSC 具有非常好的应用前景。

(6)AMDSC 和 MDSC 对分组丢失率不敏感，所以当分组丢失率不断增加的时候，编码器的性能衰减比较缓慢，而 AMR-WB 与 MD AMR-WB 在有分组丢失的情况下，性能衰减非常剧烈。这是因为 AMDSC 与 MDSC 基于正弦模型，信号的重建不受解码器状态的影响，具有较高的稳定性。

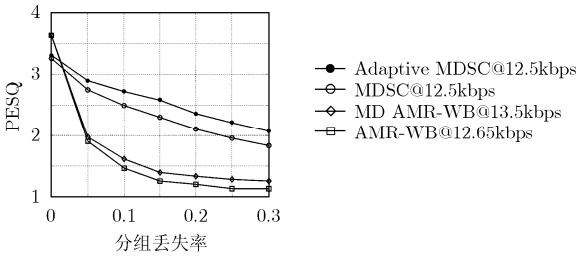

图 7　不同的编码器在近似速率下的性能比较

6　结束语

本文对 MDSC 进行了率失真分析，提出了 AMDSC 编码器，该编码器针对不同信道分组丢失率进行了优化，根据信道的分组丢失率动态地调整两个描述之间的冗余。与其它编码器在相同分组丢失率条件下的性能比较表明 AMDSC 具有最好的抵抗分组丢失的性能。

为了简单起见，本文所采用的最优化的准则为最小化频域信噪比，而进一步，还需要考虑人耳的感知特性。最优化的工作是在无记忆信道下进行的，而在有记忆信道下的性能优化还需要进一步完善。

致谢　感谢北京理工大学与瑞典爱立信公司之间的合作项目对本文工作的资助。

参 考 文 献

[1] Goyal V K. Multiple description coding: compression meets the network. *IEEE Signal Processing Magazine*, 2001, 18(5): 74–93.

[2] Jayant N S and Christensen S W. Effects of packet losses in waveform coded speech and improvements due to an odd-even sample-interpolation procedure. *IEEE Trans. on Commun.*, 1981, 29(1): 101–109.

[3] Ingle A and Vaishampayan V A. DPCM system design for diversity systems with applications to packetized speech. *IEEE Trans. on Speech and Audio Processing*, 1995, 3(1): 48–58.

[4] Dong H, Gersho A, Gibson J D, and Cuperman V. A multiple description speech coder based on AMR-WB for mobile ad hoc networks. IEEE International Conference on ICASSP04 Processing, Montreal Canada, May 2004, Vol.1, I: 277–280.

[5] Wah B W and Lin Dong. LSP-based multiple-description coding for real-time low bit-rate voice over IP. *IEEE Trans. on Multimedia*, 2005, 7(1): 167–178.

[6] Lang Y, Xie X, and Kuang J M. A novel sinusoidal speech codec using multiple descriptions. IEEE International Conference on ICSP06 Processing, Guilin China, Dec 2006, Vol.1: I–662–666.

[7] George SE B and Smith M J T. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Trans. on Speech Audio Processing*, 1997, 5(1): 389–406.

[8] Goodwin M. Residual modeling in music analysis-synthesis. IEEE International Conference on ICASSP96 Processing, Atlanta, USA, 1996, 2(7): 1005–1008.

[9] The Network Simulator-ns-2, http://www.isi.edu/nsnam/ns.

[10] Srinivasan K and Gersho A, Voice activity detection for cellular networks. IEEE Speech Coding Workshop Proc, Quebec , October 1993: 85–86.

[11] ITU-T Recommendation P.59 Artificial conversational speech, 1993.

[12] Recommendation P. 862.3, Application Guide for Objective Quality Measurement Based on Recommendation P. 862. ITU T, 2005.

郎 玥： 男，1980 年生，博士生，从事语音、音频编码的研究.

赵胜辉： 男，1970 年生，副教授，从事移动通信及语音、音频信号处理的研究和教学工作.

匡镜明： 男，1943 年生，教授、博士生导师，从事数字通信及数字信号处理的研究和教学工作.