基于信号处理的电话语音模拟

左国玉[1,2]，刘文举[1]，阮晓钢[2]
（1.中国科学院 自动化所模式识别国家重点实验室，北京 100080；
2.北京工业大学 电子信息与控制工程学院，北京 100022）

摘 要：针对电话语料比较缺乏的问题，提出了一种完全由软件模拟实现由纯净语音向电话质量语音转换的算法．该算法采用滤波器设计技术来模拟电话线路连接中各种模拟传输设备频率响应特性，并对电话通道环境中各种噪音行为等电话语音现象进行了模拟研究．频谱失真分析和识别实验结果表明，通过模型参数的合理设置和调整，该算法能有效地实现纯净语音向电话质量语音的近似模拟，使用由纯净数据生成的模拟语音可以获得与真实语音同等的识别性能．

关键词：电话语音模拟；信号处理；滤波器；频谱失真；语音识别
中图分类号：TN912.34 **文献标识码**：A **文章编号**：0254－0037(2003)02－0182－06

随着信息技术的发展，电话语音识别在自然语言处理很多新兴领域中得到了广泛的应用．在真实电话语料比较缺乏而纯净语料相对充沛的情况下，建立一个由纯净语音生成电话质量语音的系统可以解决电话语料稀少的问题．有不少研究者在电话传输通道模拟方面作了很多工作[1-3]，但是这些方法只能模拟一种或少量几种电话语音通道条件，或者需要大量的硬件设备支持．本文根据国际电信联盟(ITU-T)的有关标准[4]提出了一种由纯净语音向电话质量语音转换的纯软件模拟方法．该方法通过设计各种滤波器使其尽可能准确地反映电话通道中各模拟设备的信号传输特性，同时模拟各种噪音行为以及其他电话语音现象．在对模拟方法的有效性分析时，选取单帧语音对数频谱距离和整句话全部语音帧似然失真COSH距离[5]两个层面来客观地考察真实语音和转换语音的差异．在做语音识别实验分析时，语音识别器由纯净语音经电话语音板转录得到的语音训练而成，在一个真实测试集和几个由本文方法产生的电话语音集上进行了识别性能的对比测试．

1 模拟方法

1.1 方法描述

图1为经过公共电话交换(PSTN)网络的语音流在64 kb/s架构的端对端电话通道中的传输流程模型示意图．模型中考虑了大多数传输损害因素(如衰减损失、回音、时延以及由语音编码方案引起的量化失真)．除此之外，系统也模拟了脉冲噪音和串音效应的影响．特别地，电话连接中各模拟传输设备根据其典型的物理特性建立了相应的滤波器模型，以反映语音信号所产生的频谱变化．由于本文所用普通话纯净863语料库的数据采样率为16 kHz，因此在语音转换之前首先须将从该数据库中取得的待转换纯净语音降采样到8 kHz带宽．数字语音信号在编码解码时对输入信号幅度水平比较敏感，因而会影响输出语音的可懂度．为减少仅仅由于语音水平不一致造成的语音质量编解码效应影响，必须将每句话的有声段语音信号水平根据预设值进行归一化．

收稿日期：2003-01-17．
基金项目：国家自然科学基金资助项目(60172055)；中国科学院自动化所领域前沿基金资助项目(1M02J05)．
作者简介：左国玉(1971－)，男，博士生．

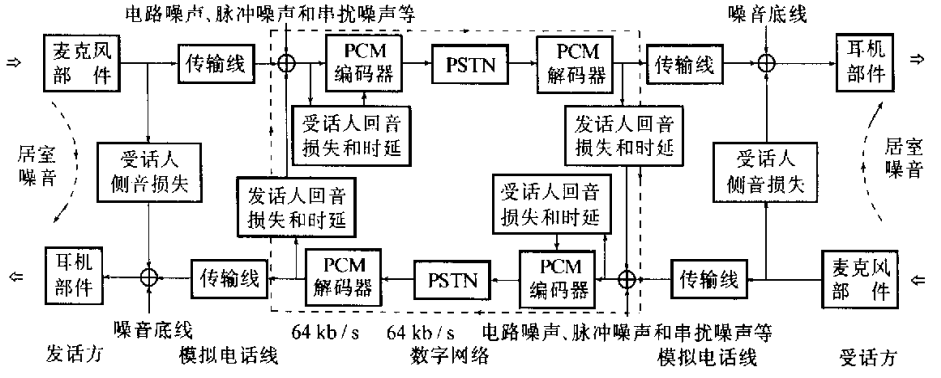

图1　端对端电话通道中的语音传输流程示意图

1.2　模拟信道中的滤波器设计

普通电话装置中的碳粉麦克风部件是造成电话连接中语音信号非线性失真的一个主要来源．ITU-T建议采用改进型中间参考系统(IRS)的发送特征来描述送受话器(handset)中各部件频率特性．这种滤波行为模仿了一个典型电话机麦克风部件的频率响应．同样，受话方送受话器的耳机部件的频响行为采用类似的IRS滤波器来描述．在假定居室噪音水平较低的情况下，模拟方法采用最小p范数最优无限脉冲响应(IIR)滤波器方法模拟典型送受话器的特性，减少了数据转换时的计算量负担，而频响特性与文献[6]中给出的IRS发送和接受部件的名义敏感度值相一致．

线路特性在这里是指电话传输线的电气性质．根据电路传输理论，信号线将非线性失真引入所加载信号，使语音信号的频谱发生变化[7]．给定负载阻抗Z_L，线路的传递特性可以定义为$T(w)=Z_L/[Z_L\cosh(\gamma d)+Z_0\sinh(\gamma d)]$．其中：$d$表示传输线的长度；$\gamma$和$Z_0$分别由单位长度传输线的串联电感$L$、并联电容$C$和并联电导$G$表示．加载信号高频部分的衰减随着线路长度增加而显著增加．

在64 kb/s脉冲编码调整制(PCM)结构中，PCM编码器实现了3种功能：信号的反混叠滤波；采样；将采样值转换为A律或μ律二进制码字．在PCM解码器中，接收到的码字被解码，转换为模拟信号，再进行低通滤波．模拟系统中用一个五阶低通和一个三阶高通的椭圆滤波器的级联组成编码器中的反混叠滤波器．椭圆滤波器能比巴特沃斯和契比雪夫滤波器提供更陡峭的高低频衰减(rolloff)特性，而且通带和止带内的纹波比较均匀，在满足同样性能指标下，滤波器阶数最低．式(1)、(2)表示在一次模拟行为中采用的低通滤波器$H_L(z)$和高通滤波器$H_h(z)$表达式：

$$H_L(z)=\frac{0.4931+2.3699z^{-1}+4.6484z^{-2}+4.6484z^{-3}+2.3699z^{-4}+0.4931z^{-5}}{1+3.4636z^{-1}+5.0773z^{-2}+3.8205z^{-3}+1.4535z^{-4}+0.2079z^{-5}} \tag{1}$$

$$H_h(z)=\frac{0.7968-2.3839z^{-1}+2.3839z^{-2}-0.7968z^{-3}}{1-2.5480z^{-1}+2.1875z^{-2}-0.6259z^{-3}} \tag{2}$$

可以验证，由上述两滤波器组成的反混叠滤波器的幅度和群时延频响特性满足ITU-T Rec. G712的要求．PCM解码器中采用了与$H_L(z)$完全相同的低通滤波器．在一定范围内调节这两个滤波器以及线路滤波器和IRS滤波器的系数，可以获得模拟整个电话通道特性的期望频率曲线形状．

1.3　回音和侧音行为的计算

由图1可见，除了前向传输通道外，模拟系统考虑了侧音损失和回音损失也对语音信号流造成损害[8]．侧音损失发生在同一部电话装置的由麦克风到耳机部件的电子路径中．呼叫中心的近端和远端接口处耦合电路阻抗的不匹配导致发话人回音和受话人回音现象的发生．衰减为A_m的受话人侧音比(listener sidetone rating)为$R_L=R_m+D$，表示送受话器对周围噪音的抑制程度．其中，$R_m=S_t+R_t+A_m-1$表示侧音屏蔽比(sidetone mask rating)，反映由使用者经侧音路径到耳朵的响度损失．S_t和R_t表示送受话器发

送部件和接受部件的特性；A_m为侧音平衡返回损失的加权平均．D为缺省值，对于线性麦克风，一般可取1.5～4 dB．具有单向时延T和衰减因子I_c的发话人回音响度比(talk echo loudness rating)是衡量线路中发话人回音损失程度的参数，由$R_T = R_s + R_r + I_c$计算得到．其中发送方响度比R_s和接收方响度比R_r在模拟系统中表现为传输线特性和IRS滤波．

1.4 噪声模拟和编码方案

语音质量在语音传输过程中受到各种不同类型噪声的侵害，主要包括电路噪声、噪声底线(noise floor)、脉冲噪音、居室环境噪音、串扰噪音以及编码效应等．在模拟系统中，这些噪音源被描述为A加权的平均功率水平．分布于电话连接中的电路噪声在模拟系统中被建模成300～3 400 Hz的窄带白噪声，用宽带白噪声模拟接收方的噪声底线．模拟系统中，居室环境噪音对语音质量的影响效应由在频域加入受话方语音信号来近似表示．邻近电话线路的耦合作用会造成串音现象，由于长距离的远端串音影响较弱，系统中只考虑近端串音效应，其大小按频率的1.5次幂规律增加．电磁场干扰和电话连接中的设备切换会产生语音信号中的脉冲噪音干扰，其表现形式为一个随机波形，幅度大大高于背景噪音．系统脉冲幅度、脉冲持续时间和两次脉冲时间间隔作为3个随机量模拟噪音的产生，时间间隔取为1～5 min．

语音的编码方案造成的信号失真也降低了语音质量．模拟系统实现电话连接中传输语音的各种编码方案包括PCM(G.711)和ADPCM(G.721，G.726，G.729)以及它们之间可能的级联等．为了确认多种不同方法的性能影响，模拟方法中也考察了ETSI(欧洲电信标准)的GSM6.10编码效应．在图1的PSTN位置处取代各种编码方案，模型中研究了调制噪音参考单元(MNRU)产生的信号相关噪音的性能．这种信号调制噪音造成的信号失真对语音质量的影响会产生与上述典型语音编码方案相类似的效果．调制语音信号可表示为：$x(t) = s(t)[1 + mn(t)]$；$Q = -20\log_{10} m$．其中：$s(t)$表示初始语音信号；$n(t)$表示均值为0方差为1的白噪声；Q为调制深度．

图2给出了一次由纯净语音转换为模拟语音的仿真行为事例．待转换样本为一女声片段"避免东欧"．可以看到，纯净语音转换为模拟语音时发生了较大的波形失真，而且出现模拟电话信道背景噪音．语谱图除表达转换语音的带宽变化信息外，也显示了有背景噪音加入到信号中．

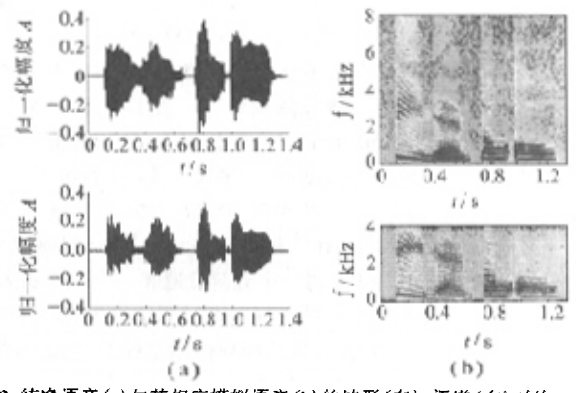

图2 纯净语音(a)与其相应模拟语音(b)的波形(左)、语谱(右)对比

2 模拟方法有效性的频谱失真分析

用频谱失真度客观度量语音模式距离具有数学的可处理性和计算的有效性．在进行识别实验前，先分别从单个语音帧对数频谱距离和一句话所有语音帧的COSH距离两个方面来考察本模拟方法的有效性．

2.1 频谱失真测度表示形式

考虑由全极点模型$\sigma / A(z)$表示已加汉明窗的语音帧X_n，$A(z) = \sum_{i=0}^{p} a_i z^{-i}$，其中$\boldsymbol{a}^T = (a_0, a_1, \cdots, a_p)$为线性预测系数，$a_0 = 1$．于是$p$阶全极点模型的频谱就表示为

$$S(w) = \sigma_p^2 / |A(e^{jw})|^2 \tag{3}$$

其中：$\sigma_p^2 = \boldsymbol{a}^T \boldsymbol{R}_p \boldsymbol{a}$为该语音帧的残差能量；$\boldsymbol{R}_p$是$(p+1)\times(p+1)$维自相关矩阵；由$X_n$的自相关序列$r = (r(0), r(1), \cdots, r(p))$求得．于是可计算出频谱$S(w)$和$S'(w)$关于频率$w$的对数距离：

$$V(w)=\log S(w)-\log S'(w)=\log(\sigma^2/|A|^2)-\log(\sigma'^2/|A'|^2) \tag{4}$$

由于 Itakula-Saito 距离的非对称性，在计算一句话所有语音帧的频谱失真时采用 COSH 距离来衡量：

$$d_{\mathrm{COSH}}(S,S')=\left[d_{\mathrm{IS}}\left(\frac{\sigma^2}{|A|^2},\frac{\sigma'^2}{|A'|^2}\right)+d_{\mathrm{IS}}\left(\frac{\sigma'^2}{|A'|^2},\frac{\sigma^2}{|A|^2}\right)\right]/2 \tag{5}$$

其中 d_{IS} 是 Itakura-Saito 距离.

2.2　频谱失真测量

对于一段纯净语音(初始语音)片断，通过大量试听以调整不同参数获取相对应的转换语音段，使其中有些话语从听觉上接近电话质量的语音．不失一般性，从这些已产生的语音片断中选取与真实电话语音(目标语音)频谱失真最小的一条话语的参数调整过程来说明本模拟方法的有效性．

考虑图 2 示例语音，选取在调整过程中到达最小频谱失真语音段所生成的一些相关语音片断．也不失一般性，再从这些片断中选取一个对应位置的发音帧(一帧的大小取为 30 ms 的 240 样本点，帧移为 7.5 ms 的 60 样本点)．根据公式 (3)、(4)，图 3 显示了各帧 LPC 频谱及其与真实语音对应帧的对数频谱距离．

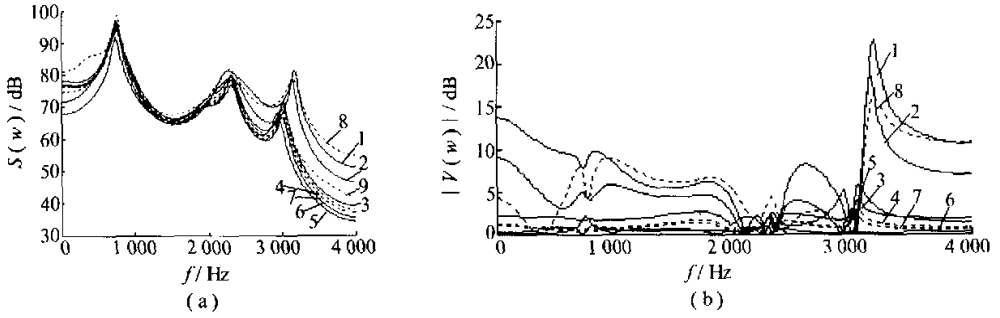

图3　各语音帧的LPC频谱$S(w)$(a)和真实语音帧与其他各帧语音的对数频谱距离$|V(w)|$(b)

考察在现有信号只加 IRS 滤波作用时生成的转换语音 1(除文中提及的待调整参数外，其他未提及的参数均设为缺省值)，可以看到低频段发生了较大的能量衰减，影响了第 1 共振峰的形状，而第 3 共振峰能量相对增强，这种现象与文献 [6]描述的送受话器频响特性一致．图 3 的线性预测系数(LPC)频谱分析显示，语音 1 的频谱失真与初始语音的失真几乎相差无几．表 1 中数据(单帧COSH 距离项)表明，在一定范围内调整 IRS 滤波器系数，信号频谱特性可以获得一定改善(语音 2)．在对 IRS 滤波后的信号做模拟编解码器内反混叠滤波和低通滤波后，再调整不同的电话传输线特性参数(主要是长度参数的模拟)模拟获得各种转换语音 5、6、7、3 和 4，其中语音 7、3 和 4 传输线特性设置相同，但分别加载了不同的编码方案以及噪音，这些频谱曲线表明了电话线的传输线特性：高频信号部分的衰减随着线路长度增加而显著增加．频谱幅度 $S(w)$和频谱失真$|V(w)|$显示，这 5 条语音频谱线的第 3 共振峰发生显著衰减并向第 2 共振峰位置靠拢，第 2 共振峰也产生衰减并略有右移、共振峰宽度也有所降低，但与目标语音的频谱更加接近，反映了实际电话通道的带限特性和衰减特性．由表 1 中语音 3、4 和 7 的单帧距离数据和图 3 的频谱可以看到，编码方案、热噪声以及较低的环境噪音，对语音信号频谱形状影响不太大．目标语音 9 与语音 4 频谱的 COSH 距离是 0.040 6，而与初始语音 8 的距离达到了最大值 11.885 1．

表1　目标语音与初始语音及各转换语音的似然失真值

转换语音	1	2	3	4	5	6	7	8(初始语音)	9(目标语音)
单帧COSH距离	10.114 5	2.864 0	0.044 4	0.040 6	0.100 6	0.020 7	0.025 1	11.885 1	
平均COSH距离	20.159 1	7.144 2	1.138 6	0.082 8	0.442 6	0.573 1	0.202 3	55.935 8	

表 1 中，平均 COSH 距离项描述了图 2 示例初始语音 8 及其上述各转换语音的全部语音帧与目标语音 9 的似然失真的平均值．从表 1 可以看到目标转换语音 9 与其他各语音平均频谱失真值的大小，与单帧距离各

值大小基本相对应，所以前面数据样本选取的不失一般性假设能够成立．表1平均频谱距离显示初始语音8的d_{COSH}值为55.935 8，明显大于其他转换语音的失真度．随着各种模型参数的设置和调整，转换语音的频谱和失真可以渐近地趋向于目标语音．而且语音4与目标语音的平均COSH值达到最小值0.082 8．

以上频谱失真分析均表明，可以通过合理地设置和调整不同的模型参数，使转换语音频谱接近于真实语音频谱实现电话质量语音的模拟．下面的识别实验分析进一步讨论了该算法的可行性．

3 训练和测试数据集的选取

电话语音识别器的训练语音库，是从863男声语音库中取出80人43 096条语句共180 626 s，通过局域电话网经电话语音板转录而成(这里做了一个假设，认为通过电话语音板获得的语音更接近于真实的电话语音条件)．实验中共采用6个数据集进行识别性能的测试．从男声863库余下的数据中取出4人共240条语句共1 032 s作为基本集，通过相同处理获得的语音作为测试用的真实电话语音(+)．

根据前面的频谱分析可知，不同编码方案和各种合理水平噪音的设置对语音信号频谱形状影响不太显著，主要原因在于LPC频谱不能像快速傅里叶变换(FFT)频谱那样较多地反映语音频谱的更多细节．实验选取语音4所对应的各滤波器模型参数作为生成测试数据集用的模型滤波器参数，其他参数设置如表2所示，这样获得另外5个转换语音测试集(A、B、C、D和E)，分别描述了模拟通道的一些电话通道的特性(见注释部分说明)．另外选中这5个测试集是因为在正式实验前所作的一次10人非正式的主观评测中，这5种生成语音在听觉上均被认为与真实电话通道语音的音质比较接近．

表2 通过设置不同模型参数集获得由纯净语音转换的各语音测试集

测试集	电路噪声/dB	A加权的环境噪声/dB	编解码方案	注 释
A	35	45	G711	低噪声，脉冲调制编码
B	45	40	G711	噪声，脉冲调制编码
C	35	40	G726(32 kb/s)	低噪声，数字电路倍增设备
D	35	45	GSM610	噪声，移动设备
E	20	35	MNRU，Q=20 dB	信号相关噪音

4 识别实验与结果分析

为了进一步探讨由纯净数据转换为模拟数据的可能性，训练语音识别器时，语音信号声学特征用12维美尔频标倒谱系数(MFCC)、一维能量特征和一维归一化基频特征以及对应的一阶和二阶差分来表示，形成一个42维特征向量．声学模型采用基于决策树有调类三音子模型共10 801个类，2 591个输出分布(16个高斯混合数)．语言模型采用退化的Trigram，大约从10亿字节的语料库训练得到，共80 Mb(Trigram有45 Mb，Bigram有35 Mb)．表3列出了对各测试集的识别性能评估结果．

表3 不同测试集上的识别性能

测试集	识别性能/%				
	正确	替代	删除	插入	总错误率
+	74.3	22.8	2.9	0.0	25.7
A	74.1	22.5	3.5	0.3	26.3
B	68.0	29.7	4.3	1.9	35.9
C	76.8	19.8	3.3	0.1	23.2
D	73.7	23.0	3.3	0.3	26.5
E	74.3	22.6	3.1	0.1	25.8

对于B测试集，识别正确率只有68.0%，而相同的识别器对其他测试集的正确率均在74%左右，与真实语音的识别性能差别不太明显．因此结果表明，通过设置不同的模型参数，可以产生与真实电话语音识别率相近的不同模拟数据集，也就是通过模型参数的调整，系统能够产生近似于真实语音的电话噪声语音．比较B与A、C和D集，数据显示，除电话信道的频宽限制以外，电话传输信道中的电路热噪音和环境噪音在一定程度上影响了语音质量，而这一点在频谱分析中不能有效地反映出来．在系统中合适模拟各种噪音会使模拟语音更加接近真实情况．A和D相比较显示出识别其性能受到语音编码方案的影响，但是实验结果

显示性能差别并不显著，可能因为这里所设置的背景噪音水平的影响效应比编码效应更大一些.

5 结束语

本文讨论了一种由纯净语音产生实现模拟电话质量语音的算法，通过控制各种模型参数近似模拟各种不同环境下的真实电话语音. 为使模拟语音更加真实地逼近于实际语音，还需要进一步研究各传输设备特性、噪声影响和信道的时变特性. 在模拟的实现方法上，可以考虑将声音变换技术中的各种方法应用到电话语音模拟技术中来. 如何不依赖于主观试听方式而能尽快找到合适的模型的模拟参数也成为后续工作所要解决的问题.

参考文献：

[1] MORENO P J, STERM R M. Source of degradation of speech recognition in the telephone network[A]. Proc of ICASSP'94[C]. Piscataway: IEEE Service Center, 1994. 109-112.

[2] TARCISIO C, DANIELE F, ROBERTO G, et al. Use of simulated data for robust telephone speech recognition[A]. Proc of EUROSPEECH'99[C]. Budapest: ESCA, 1999. 2825-2828.

[3] MÖLLER S, BOURLARD H. Real-time telephone transmission simulation for speech recognizer and dialogue dystem evaluation and improvement[A]. Proc of ICSLP'2000[C]. Beijing: China Military Friendship Publish, 2000. 750-753.

[4] ITU-T Recommendation G.107, The E-Model, a Computational Model for Use in Transmission Planning[S].

[5] GRAY R M, BUZO A, GRAY A H Jr, et al. Distortion measures for speech processing[J]. IEEE Trans Acoustics, Speech, Signal Proc, 1980, 28(4): 367-376.

[6] ITU-T Recommendation P.48, Specifications for an Intermediate Reference System[S].

[7] STARR T, CIO J M, SILVERMAN P. Understanding Digital Subscriber Line Technology[M]. Upper Saddle River: Prentice-Hall, 1999.

[8] ITU-T Recommendation G.111, Loudness Ratings (LRs) in an International Connection[S].

Telephone Speech Simulation Using Signal Processing Method

ZUO Guo-yu[1,2], LIU Wen-ju[1], RUAN Xiao-gang[2]

(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China;
2. College of Electronics Information and Control Engineering, Beijing University of Technology, Beijing 100022, China)

Abstract: Owing to the lack of telephone speech data, this paper proposes a software simulation implementation of converting clean speech sounds into telephone-quality ones. Filter design technology is adopted to simulate the frequency response characteristics of various analogue transmission equipments in telephone circuit connection and to make simulation study on such telephone speech phenomena as different noise behaviours in telephone channels. The spectral distortion analysis and recognition experiment results show that through the reasonable setting and regulating, the algorithm can effectively realize the approximate simulation from clean speech to telephone-quality ones so that the simulated speech sounds generated from clean data can achieve as good recognition performance as real speech.

Key words: telephone speech simulation; signal processing; filter; spectral distortion; speech recognition