语音识别中带宽失配的补偿研究

何勇军[1),2)] 韩纪庆[1)]

[1)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

[2)](哈尔滨理工大学计算机科学与技术学院 哈尔滨 150080)

摘 要 目前的语音识别系统在训练环境与测试环境匹配的情况下具有很高的识别率,而当环境失配时,其性能将急剧下降.作者研究发现,带宽失配,即训练语料和测试语料带宽不一致,也是引起环境失配的主要原因之一.当测试语音带宽比训练语音带宽窄时,丢失的频段不可逆,且其影响在倒谱域或对数频谱域上是时变的,因而无法用目前的信道补偿方法补偿.文章在分析丢失频段对梅尔频率倒谱系数影响的基础上,提出了用频谱折叠方法对窄带测试语音进行补偿.在此基础上给出了语音带宽检测算法和带宽补偿统一框架.在 AN4 和 TIMIT/NTIMIT 数据库上的实验表明,该框架能有效增强语音识别系统在带宽失配情况下的鲁棒性.

关键词 带宽失配;畸变补偿;梅尔倒谱;鲁棒性;语音识别

中图法分类号 TP319 **DOI 号**: 10.3724/SP.J.1016.2011.01629

Research on Bandwidth Mismatch Compensation in Speech Recognition

HE Yong-Jun[1),2)] HAN Ji-Qing[1)]

[1)](*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

[2)](*School of Computer Science and Technology, Harbin University of Science and Technology, Harbin* 150080)

Abstract Speech recognition systems obtaining high recognition rates in clean environments perform badly in mismatch environments without compensation. Based on the research, we found that bandwidth mismatch, namely the bandwidth difference between the training and test conditions, is one of the main factors leading to environment mismatch. When the bandwidth of the test speech is narrower than that of the training speech, the distortion is non-invertible and time-varying in the logarithm spectrum and cepstrum domains. So it could not be compensated with current channel compensation methods. After analyzing the Mel-frequency cepstrum coefficient distortion caused by the lost frequency band, we propose a compensation method based on spectral fold. Furthermore, we provide an algorithm for speech bandwidth detection and a unified compensation framework. Experiments on the AN4 and TIMIT/TIMIT databases show that the proposed framework improved the robustness of speech recognition underbandwidth mismatch conditions.

Keywords bandwidth mismatch; distortion compensation; MFCC; robustness; speech recognition

1 引 言

让机器能像人一样感知和理解语音一直是人类的梦想,语音识别为这一梦想带来了希望.经过几十年发展,语音识别技术取得了巨大成就,从最初的孤立词识别到如今的大词表连续语音识别(Large Vocabulary Continue Speech Recognition, LVCSR),

收稿日期:2009-06-03;最终修改稿收到日期:2011-07-25.本课题得到国家"八六三"高技术研究发展计划项目基金(2006AA010103)与国家"九七三"重点基础研究发展规划项目基金(2007CB311100)资助.**何勇军**,男,1980 年生,博士研究生,讲师,主要研究方向为鲁棒语音识别、说话人识别和声学事件检测. E-mail: heyongjun@hit.edu.cn. **韩纪庆**,男,1964 年生,教授,博士生导师,主要研究领域为语音处理、人工智能. E-mail: jqhan@hit.edu.cn.

语音识别技术已经迈出了实验室并逐步走向应用.在理想环境下,目前的小词表以及中等词表语音识别系统的识别率能达到99%以上,LVCSR系统识别率也能超过95%[1],但在环境失配情况下,识别率将急剧下降.数十年来,研究者们尝试用各种方法来增强语音识别系统的鲁棒性,虽取得了一定进展,但目前的语音识别系统仍然难以适应复杂的应用环境.环境失配是指识别系统的训练环境和测试环境存在差异.造成环境失配的主要因素在于噪声的存在,这种噪声可能是加性的,可能是卷积性的,也可能是两者的混合[2].一般认为,外部环境噪声呈加性,信道影响呈卷积性.诸如训练环境没有噪声而测试环境存在噪声,或者训练环境和测试环境存在不同的噪声,都会引起环境失配.环境失配必然导致语音特征参数的分布存在偏差进而影响系统性能.

为了增强语音识别系统的环境鲁棒性,研究者们提出了大量的方法.这些方法大致可分为两类,即特征增强和模型补偿.特征增强试图从畸变语音中提取鲁棒特征.这类方法或先对信号去噪然后提取特征,例如谱减[3]、维纳滤波、卡尔曼滤波[4]、子空间法[5]等,或者直接补偿特征,例如倒谱均值方差规正(Cepstral Mean Normalization,CMN)[6-7]、特征弯折[8]、短时高斯化[9]、相关谱滤波(RelAtive SpecTrAl,RASTA)[10]、非线性滤波[11]等.模型补偿则试图修改声学模型来适应环境,典型的有并行混合模型[12]、泰勒级数展开(Vector Taylor Series,VTS)[2,13]以及各种自适应方法如最大似然线性回归(Maximum Likelihood Linear Regression,MLLR)[14]、最大后验概率(Maximum A-Posteriori,MAP)[15]等.

测试语音和训练语音的带宽不同也会造成环境失配,比如训练语料采用16 kHz采样率具有0～8 kHz带宽,测试语料为10 kHz具有0～5 kHz带宽.语音的带宽失配广泛存在,除了采样率的影响,不同编码格式,不同编码速率以及不同麦克风或传输网络,都会造成带宽差异.一般来说,采样率越低,带宽越窄(奈奎斯特定理),编码速率越低,带宽也越窄(放弃了对某些频段的编码).带宽失配补偿的重要性体现在:

(1)网络语音识别.互联网的发展使网络语音识别及相关应用应运而生,网络上的语音数据带宽复杂;

(2)分布式语音识别.普适计算的兴起,分布式语音识别系统成为发展趋势.语音识别服务器要处理来自各种语音采集前端的数据,这些数据带宽多变;

(3)语音检索.语音检索已经成为当前研究热点,处理的数据跨度数十年甚至上百年,有各种编码、采样率,必然存在带宽失配.

因此,研究带宽失配补偿有着重要意义.虽然,在有额外标注数据的情况下,传统的模型自适应方法能有效补偿带宽失配.但上述应用环境复杂,且带宽随时变化,为每种可能出现的带宽都采集数据然后标注训练是不切实际的.我们需要的是能自动检测带宽,快速补偿不同带宽数据的方法.目前这方面的研究较少,文献[16]曾在丢失数据理论的基础上,提出用少量宽带语音和大量窄带语音训练桌面语音识别系统,却未给出窄带测试语音的补偿方法.

带宽失配问题可以由窄带信道的传输引起,似乎可以采用目前的信道畸变补偿方法进行补偿.但这类方法大多假定信道在计算特征的频段上可逆,且信道影响在倒谱域的偏差为常量[6-7,17].基于上述假定的方法在丢失的频段是不成立的:一方面,在丢失频段上信道不可逆[18];另一方面,本文的分析表明,丢失频段在梅尔频率倒谱参数上引起的偏差是时变的.因此,传统的信道补偿方法事实上无法消除带宽差异对语音识别系统的影响.本文在分析丢失频段对梅尔倒谱特征参数影响的基础上,提出了采用频谱折叠方法补偿丢失频段的统一框架并给出了快速带宽检测算法.最后用实验验证了本文方法的有效性.

2 带宽失配对语音特征分布的影响

在语音识别中,测试语音的带宽比训练语音的带宽宽或窄都会引起带宽失配.前者并不影响识别率,因为计算特征的梅尔滤波器组所覆盖的频段范围是固定的,范围以外的带宽已被忽略;对于后者,则需要重建丢失频段的信息.基于人耳听觉特性的梅尔频率倒谱系数(Mel-Frequency Cepstrum Coefficients, MFCC)及其差分在语音识别中处于统治地位.因此,本节将定量分析带宽失配对梅尔频率倒谱特征的影响.不失一般性,假定有宽带语音 $s(n)$(这里的宽带语音是指带宽较宽的语音,非一般意义的宽带,窄带语音也类似),其采样率为 f_s(Hz),带宽为 $f_s/2$(Hz),分帧后各帧数据为 $s_1,s_2,\cdots,s_M$.计算梅尔频率倒谱特征,首先做短时傅立叶变换并取幅值(有的文献取功率谱,但这不影响分析结果):

$$S_m[k]=|DFT(s_m)| \tag{1}$$

其中：$m=1,\cdots,M$ 为帧序号；$DFT(\cdot)$为短时傅里叶变换；$S_m[k]$为第 m 帧的第 k 个频谱分量的幅值，$k=1,2,\cdots,K$ 为离散傅立叶变换序号，且 K 对应奈奎斯特频率. 由于其带宽/Hz 为 $f_s/2$，$S_m[k]$满足：

$$S_m[k]\begin{cases}>0, & k=1,2,\cdots,K\\ =0, & \text{否则}\end{cases} \tag{2}$$

该语音在某一环境影响下，其带宽变窄，相当于存在滤波器其传递函数 $H[k]$满足

$$H[k]=\begin{cases}1, & k=1,2,\cdots,K'\\ 0, & \text{否则}\end{cases} \tag{3}$$

其中 $K'<K$，需要注意的是，本文只分析带宽失配问题，不讨论频谱幅值的改变，但这种影响在信道不匹配情况下是存在的. 滤波器 $H[k]$的作用仅仅在于将宽带语音变成窄带语音. 在频域，宽带变窄带表现为信号与该滤波器的乘积运算，窄带语音的幅度谱为

$$S'_m[k]=S_m[k]H[k] \tag{4}$$

显然，有如式(6)成立

$$S'_m[k]\begin{cases}>0, & k=1,2,\cdots,K'\\ =0, & \text{否则}\end{cases} \tag{5}$$

在梅尔倒谱的计算中，一组梅尔滤波器被用来模拟人的听觉特性. 梅尔滤波器分布在频谱分量 $1\sim K$ 范围内，则宽带语音经过梅尔滤波器组后，其输出表示为

$$Y_m[i]=\sum_k W_i[k]S_m[k] \tag{6}$$

其中 $Y_m[i]$为第 m 帧宽带语音在第 i 个梅尔三角滤波器上的输出，$i=1,2,\cdots,I$，且 I 为最后一个梅尔滤波器序号，对于窄带语音有

$$Y'_m[i]=\sum_k W_i[k]S'_m[k]=\sum_k W_i[k]S_m[k]H[k] \tag{7}$$

其中 $Y'_m[i]$为第 m 帧窄带语音在第 i 个梅尔滤波器上的输出. 结合式(3)，则当 $i=1,2,\cdots,I'$时，其中 I' 为覆盖窄带语音截止频率的梅尔滤波器序号，下式成立

$$Y'_m[i]=\begin{cases}Y_m[i], & i\le I'\\ 0, & \text{否则}\end{cases} \tag{8}$$

如图 1 所示，三角形为梅尔滤波器，虚线为某帧宽带语音的幅度谱，在频谱分量 $0\sim K$ 上有不为零的能量，丢失频段以后，频谱分量 $0\sim K'$(对应梅尔滤波器输出为 $Y_m[1]\sim Y_m[I']$)的能量保持不变，频谱分量 $K'+1\sim K$(对应梅尔滤波器输出为 $Y_m[I'+1]\sim Y_m[I]$)上的能量丢失. 可见频宽较窄的语音在末尾几个梅尔滤波器所覆盖的范围，失去了宽带语音的全部信息，仅剩下信道噪声①级别的能量，即当 $i=I'+1,I'+2,\cdots,I$ 时 $Y_m[i]\neq 0$，而 $Y'_m[i]\approx 0$，在对数运算时将溢出. 语音识别系统通常采用一个很小的阈值代替，也就是令 $Y'_m[i]=\varepsilon>0$，而 ε 接近于信道噪声并远小于宽带语音在该处的值. 宽带语音的梅尔倒谱系数为

$$\boldsymbol{c}=\boldsymbol{C}\ln\boldsymbol{Y}=[\boldsymbol{C}_{I\times I'}\quad \boldsymbol{C}_{I\times(I-I')}]\begin{bmatrix}\ln\boldsymbol{Y}_{I'\times 1}\\ \ln\boldsymbol{Y}_{(I-I')\times 1}\end{bmatrix}=\boldsymbol{C}_{I\times I'}\ln\boldsymbol{Y}_{I'\times 1}+\boldsymbol{C}_{I\times(I-I')}\ln\boldsymbol{Y}_{(I-I')\times 1} \tag{9}$$

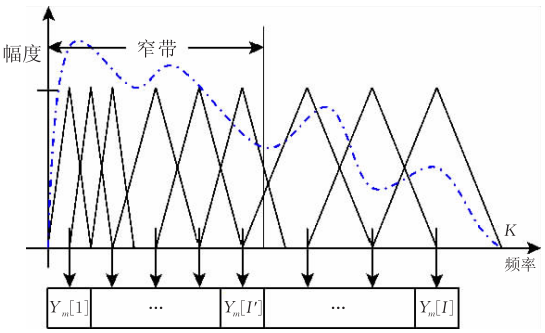

图 1　梅尔滤波器组与语音带宽

其中：$\ln(\cdot)$表示对输入向量的每个元素取对数，$\boldsymbol{C}$为离散余弦变换矩阵，$[\boldsymbol{C}_{I\times I'}\quad \boldsymbol{C}_{I\times(I-I')}]$是其分块矩阵的形式，各下标为行乘列形式. 类似的，窄带语音的梅尔倒谱系数为

$$\boldsymbol{c}'=\boldsymbol{C}\ln\boldsymbol{Y}'=[\boldsymbol{C}_{I\times I'}\quad \boldsymbol{C}_{I\times(I-I')}]\begin{bmatrix}\ln\boldsymbol{Y}'_{I'\times 1}\\ \ln\boldsymbol{Y}'_{(I-I')\times 1}\end{bmatrix}=\boldsymbol{C}_{I\times I'}\ln\boldsymbol{Y}'_{I'\times 1}+\boldsymbol{C}_{I\times(I-I')}\ln\boldsymbol{Y}'_{(I-I')\times 1}=\boldsymbol{C}_{I\times I'}\boldsymbol{Y}'_{I'\times 1}+\boldsymbol{C}_{I\times(I-I')}\ln\boldsymbol{\varepsilon} \tag{10}$$

其中 $\boldsymbol{Y}$ 和 $\boldsymbol{Y}'$分别为宽带语音帧及其对应的窄带语音帧的梅尔滤波器组的输出，皆为 I 维列向量，在这里省略了帧序号 m，$\boldsymbol{\varepsilon}=[\varepsilon\quad \varepsilon\quad \cdots\quad \varepsilon]^{\mathrm{T}}$为$(I-I')$维列向量. $\boldsymbol{c}$ 和 $\boldsymbol{c}'$分别为两种语音对应的倒谱系数向量. 根据式(7)～(9)，二者差值为

$$\begin{aligned}\boldsymbol{d}&=\boldsymbol{c}-\boldsymbol{c}'=\\ &(\boldsymbol{C}_{I\times I'}\ln\boldsymbol{Y}_{I'\times 1}+\boldsymbol{C}_{I\times(I-I')}\ln\boldsymbol{Y}_{(I-I')\times 1})-\\ &(\boldsymbol{C}_{I\times I'}\ln\boldsymbol{Y}'_{I'\times 1}+\boldsymbol{C}_{I\times(I-I')}\ln\boldsymbol{Y}'_{(I-I')\times 1})=\\ &\boldsymbol{C}_{I\times(I-I')}(\ln\boldsymbol{Y}_{(I-I')\times 1}-\ln\boldsymbol{\varepsilon})=\\ &\boldsymbol{C}_{I\times(I-I')}\left(\ln\frac{\boldsymbol{Y}_{(I-I')\times 1}}{\boldsymbol{\varepsilon}}\right)\end{aligned} \tag{11}$$

其中 $\boldsymbol{d}$ 为 I 维误差向量，代表每一维倒谱系数产生的偏差. 从式(11)可以看出：(1) 宽带语音变成窄带语音在倒谱参数上的影响是扩散性的，即某些频段的缺失将使倒谱域上的每个参数产生偏差；(2) 丢失的频段越宽，能量越大，产生的畸变就越严重. (3) 偏差值并非常量，而是与该帧丢失的频段有关，

① 信道噪声为信道内的加性噪声或引起带宽丢失的处理所形成的残差，其能量远低于对应频段上的宽带语音的能量.

因为 $\boldsymbol{d}$ 随着 $Y_{(I-I')\times 1}$ 变化. 由于各帧丢失部分的能量各异,在倒谱参数上产生的偏差也就不同,因此这个偏差是随着帧不同而不同的,是时变的. 目前的信道补偿方法,在倒谱域补偿特征或在模型域调整均值,其内在的假定都是信道对语音的影响在倒谱域是一个加性的常量,这对可逆时不变信道有效,而对丢失频段则是无效的.

3 丢失频段的重建

将带宽较窄的语音信号准确恢复为带宽较宽的原始信号是不可能的,因为被衰减到零(严格说是信道噪声)的频段是不可逆的. 所幸的是,语音特征参数是根据人的听觉特性计算的. 从式(6)可以看出,梅尔滤波器的输出是其覆盖频段下能量的加权,如果能将丢失的高频部分的能量恢复到丢失前的水平,就能减小偏差,进而提高系统识别率.

频谱扩展是一种将窄带信号扩展为宽带信号的方法,它利用窄带信号的频谱分量,根据窄带与宽带频谱能量之间的相关性来获得高频段能量. 这类方法虽然不以准确恢复语音信号为目的,但却能使重建后的信号在高频段的能量与丢失前相近. 文献[19]报道这类方法能显著改善窄带语音的主观听觉和客观能量误差. 本文将频谱扩展方法用于补偿丢失的高频段频谱,以提升系统在不同带宽情况下的识别率.

频谱扩展方法可分为基于模型的和非基于模型两类. 基于发声模型的方法常采用线性预测(Linear Prediction Coding, LPC)模型对声道建模,并用LPC分析然后合成的方式扩展频谱. 该类方法用窄带语音的残差和LPC系数计算丢失的高频语音的残差和LPC系数,然后合成高频语音. 典型的有基于高斯混合模型[20]、隐马尔可夫模型[21]、人工神经网络[22]和基于码本映射[23]等方法. 非基于模型的方法有频谱折叠、频谱搬移、非线性函数法[19]等,这类方法通过某种变换,将窄带语音低频部分的非零频谱"搬移"到高频以达到频谱扩展的目的. 本文采用频谱折叠补偿丢失频段主要出于以下考虑:

(1)实时性要求. 基于发生模型的方法需要训练模型,然后结合模型补偿,计算复杂度高,而频谱折叠只需要插0计算和滤波计算,运算量低;

(2)工作量要求. 基于模型的方法需要宽带语音和对应的语音数据库训练模型,这需要大量人力物力,在大多数应用中无法满足,频谱折叠方法无需训练数据;

(3)应用可行性. 基于模型的方法主要针对带宽不随时间变化时的频谱扩展,当带宽发生改变,还必须重新采集数据集训练模型,这在现实应用中是无法做到的,而频谱折叠却无此限制.

频谱折叠首先在时域对语音隔点插零,获得与低频频谱成镜像的高频频谱. 假定 $s_n[n], n=0,1,\cdots,N-1$ 为窄带信号,$s_f[n], n=0,1,\cdots,2N-1$ 为折叠后的信号,有

$$s_f[n]=\begin{cases}2s_n\left[\dfrac{n}{2}\right], & n=0,2,4,\cdots,2(N-1)\\ 0, & \text{否则}\end{cases} \tag{12}$$

从式(12)可以看出,频谱折叠对信号在时域插入0值后,信号的采样率变成了原信号的一倍. 为了统一频点,我们首先将窄带信号进行过采样,即

$$s'_n[n]=\{s_n[1],[s_n[1]+s_n[2]]/2,s_n[2],\cdots,[s_n[N-2]+s_n[N-1]]/2,s_n[N-1]\} \tag{13}$$

对窄带信号进行离散傅立叶变换有

$$\begin{aligned}|S_n[k]|^2&=\left|\sum_{m=0}^{2N-1}s'_n[m]e^{-\mathrm{j}\frac{2\pi}{2N}mk}\right|^2=\\ &\left|\sum_{m=0}^{N-1}s_n[m]e^{-\mathrm{j}\frac{2\pi}{N}mk}+\sum_{m=0}^{N-1}\frac{s_n[m]+s_n[m+1]}{2}e^{-\mathrm{j}\frac{2\pi}{2N}(2m+1)k}\right|^2=\\ &\left|\sum_{m=0}^{N-1}2s_n[m]e^{-\mathrm{j}\frac{2\pi}{N}mk}\right|^2\end{aligned} \tag{14}$$

由傅立叶变换的性质有,$|S_n[k]|^2$ 以 $k=N$ 为对称轴. 由于过采样不增加频宽,因此 $|S_n[k]|^2$ 在 $k=1,2,\cdots,N/2$ 时幅度值不为零,在其它频点幅度为零. 对频谱折叠后的信号进行离散傅立叶变换有

$$\begin{aligned}|S_f[k]|^2&=\left|\sum_{m=0}^{2N-1}s_f[m]e^{-\mathrm{j}\frac{2\pi}{2N}mk}\right|^2=\\ &\left|\sum_{m=0}^{N-1}2s_n[m]e^{-\mathrm{j}\frac{2\pi}{N}mk}\right|^2\end{aligned} \tag{15}$$

当 $k=1,2,\cdots,N/2$ 时,

$$|S_f[k]|^2=\left|\sum_{m=0}^{N-1}2s_n[m]e^{-\mathrm{j}\frac{2\pi}{N}mk}\right|^2=|S_n[k]|^2 \tag{16}$$

而当频点大于 $N/2$ 时

$$\begin{aligned}|S_f[N-k]|^2&=\left|\sum_{m=0}^{N-1}2s_n[m]e^{-\mathrm{j}\frac{2\pi}{N}m(N-k)}\right|^2=\\ &\left|\sum_{m=0}^{N-1}2s_n[m]e^{-\mathrm{j}\frac{2\pi}{N}mk}\right|^2=\\ &|S_n[k]|^2\end{aligned} \tag{17}$$

可见频谱折叠后，高频部分存在与低频部分成镜像的频谱（如图 2、3 所示）. 因为语音的能量大部分集中在低频段，高频段的幅度大致随频率递减，故通过隔点插零处理后的语音，高频段频谱幅度并不接近真实语音. 要使折叠后的语音接近真实语音，必须对高频部分能量进行修正. 一个合理的整形函数能使折叠后的信号在高频部分非常接近真实语音. 频谱整形可以用下面式子表示

$$S_w[k] = H[k]S_f[k] \tag{18}$$

其中，$k=0,2,\cdots,2N-1$，$S_f[k]$为频谱折叠后语音的频谱，$S_w[k]$为滤波整型后语音的频谱，$H[k]$为整形滤波器. 整形滤波器的目的在于削弱高频部分的镜像频谱，使其接近真实语音的高频频谱. 我们采用有限冲击响应低通滤波器达到这一目的. 滤波器在设计时设定其上限截止频率比频谱对称点大 100 Hz左右. 为了使滤波对高频的衰减稍慢，其阶数设置为 5 比较适中. 如图 4 所示，频谱整形以后，高频部分有了可用的频谱.

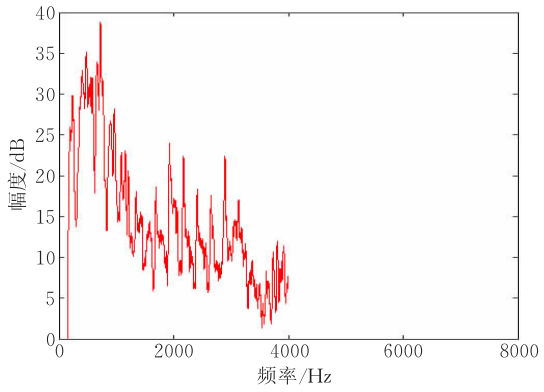

图 2　窄带语音频谱，带宽为 0.1～4.0 kHz

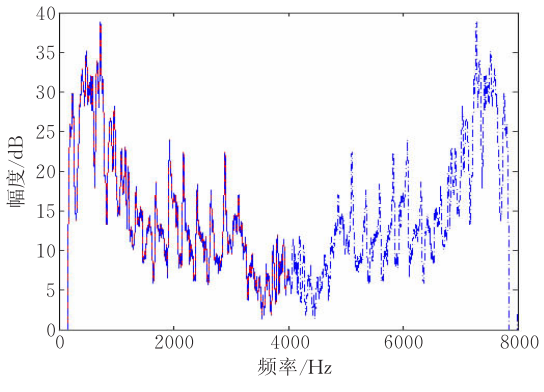

图 3　频谱折叠后效果（在高频部分出现了镜像频谱）

需要注意的是，频谱折叠后，高频部分和低频部分是以奈奎斯特频率（采样频率的 1/2）为对称点的. 如果语音最大不为零的频率值低于奈奎斯特频率，频谱折叠就会在奈奎斯特频率附近产生一个空隙，在空隙中信号能量仍然为零，无法被重建. 因此，

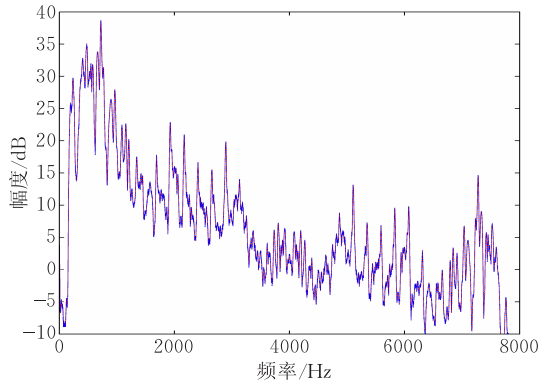

图 4　对折叠后频谱整形滤波，获得的宽带语音频谱

在频谱折叠之前，首先需要对语音进行重采样，重采样的频率为语音实际带宽上限截止频率的两倍.

4　语音带宽检测

引起语音带宽变窄的因素众多，各种因素对语音带宽的影响差别较大，因此只用语音的采样率，根据奈奎斯特定理来确定语音的带宽是不可靠的. 而用频谱折叠重建高频段需要准确知道语音的带宽，否则就会在奈奎斯特频率附近形成频谱空隙. 如果语音的带宽/Hz 为 $f_l \sim f_h$，带宽检测的任务在于确定 f_l和 f_h的值. 考虑到低频段损失很小，系统采用 $f_l=0.25$ kHz 的设置，因此，带宽检测的主要任务在于确定 f_h.

首先通过语音活动检测获取一段活动语音 $\{X_1,X_2,\cdots,X_{T_1}\}$和另一段静音$\{Y_1,Y_2,\cdots,Y_{T_2}\}$，各帧都已经过 $2K$ 点傅立叶变换，取第 1 到第 K 个值（频率采样定理）并取模. 带宽检测如算法 1 所示.

算法 1.

for $i=1$ to K

$\varepsilon[i] \leftarrow \frac{1}{T_2}\sum_{t=1}^{T_2}\boldsymbol{Y}_t[i]$;

$\varepsilon_{\max}[i] \leftarrow \max_t(\boldsymbol{Y}_t[i])$;

$\boldsymbol{x}[i] \leftarrow \frac{1}{T_1}\sum_{t=1}^{T_1}\boldsymbol{X}_t[i]$;

endfor

$i \leftarrow K$;

while $\boldsymbol{x}[i] \leq \varepsilon_{\max}[i]$

$i \leftarrow i-1$;

endwhile

$f_h=(f_s/2K)\times i$;

其中 f_s为未知带宽语音的采样率. 算法通过比较$\varepsilon_{\max}[i]$和 $\boldsymbol{x}[i]$的值来确定 f_h所在的位置. 对于窄带语音，语音段均值 $\boldsymbol{x}$ 的高频分量（衰减为信道噪声）近似等于向量 $\boldsymbol{\varepsilon}$ 内对应元素的值，这些值又必然小

于 ε_{max} 对应的值，而 $\boldsymbol{x}$ 低频部分的值（未被衰减）又大于 ε_{max}. 算法最后一行将截止频率所在的傅立叶变换序号换算成频率. 该算法自动统计门限，运算复杂度低.

5 带宽失配补偿框架

补偿带宽失配就是要减小测试语音和训练语音的带宽不一致所造成的影响，提高识别率. 因此，在测试时，只需要对那些带宽比训练语音带宽窄的语音进行频谱折叠补偿. 在系统的构建上，必须为训练语音确定一个合理的带宽. 一般而言，桌面语音识别系统采用 0.2～8 kHz 的带宽，电话语音识别系统采用 0.3～3.5 kHz 带宽. 如果采用桌面系统的带宽训练语音，在面对电话语音时需要重构 3.5～8 kHz 的频段，而频谱分量相关性随频率距离增加而减弱[24]，在用低频频谱重建高频谱时，如高低频谱频率距离太大将导致重构误差增大. 如果采用电话系统的频谱训练语料，将所有频谱的高频部分都去掉，虽然无需频谱重构，但其性能是次优的[16]. 我们需要确定一个训练带宽，既确保宽带语音被准确识别，又使得窄带语音不至于重构太宽的频带. 本文通过在 AN4 数据库上的实验来确定训练模型使用的带宽. AN4 中语音采样率为 16 kHz，带宽为 0.13～8 kHz. 使用低通滤波器将训练集和测试集处理成为带宽为 0.25～7.5 kHz、7 kHz、6.5 kHz、6 kHz、5.5 kHz 5 个版本的数据，分别用来训练模型，然后用相应的测试集测试. AN4 数据库的介绍以及语音识别系统参数设置见 6.1 节. 如图 5 所示，实验结果表明随着带宽变窄，带宽匹配情况下识别率逐渐降低，但在 6.5 kHz 以上时，下降并不明显，因为6.5～8 kHz 内能量低，畸变较小（式(10)）. 训练语音带宽确定为 0.25～6.5 kHz 是一个折中而合理的选择. 因此，本文系统训练时用采样率 16 kHz 的语音，并使梅尔滤波器组分布的频率范围为 0.25～6.5 kHz.

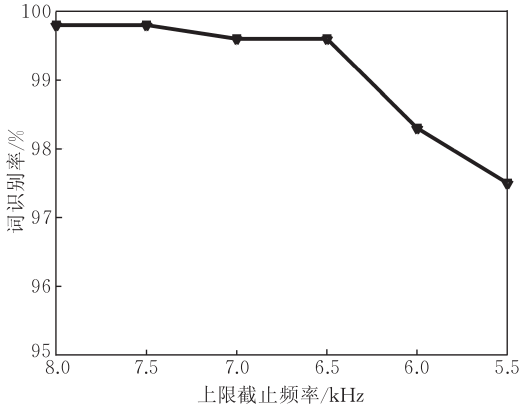

图 5 不同带宽对识别率的影响

系统的补偿框架如图 6 所示，测试语音首先经过带宽检测模块，采用算法 1 检测带宽，如果测试语音带宽高于训练语音带宽，则直接重采样，使其采样率与训练语音的采样频率一致并送入特征提取模块. 对于带宽上限截止频率低于 6.5 kHz 的语音，先重采样，使其采样频率为其上限截止频率的二倍以避免频谱折叠时引起频段空隙，然后用前述的隔点插零以及滤波整形重构丢失频段，紧接着重采样使其采样率为 16 kHz，进而使用高通滤波器仅保留丢失频段部分，最后将原始语音重采样后与重构语音进行叠加，形成补偿后的语音用于特征提取.

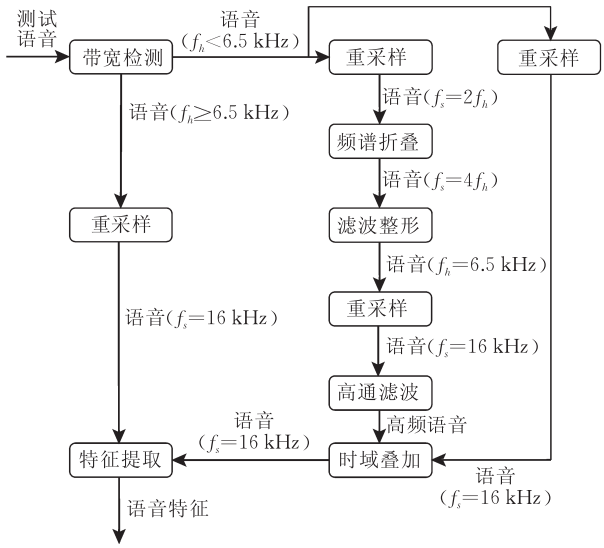

图 6 带宽失配补偿流程，其中 f_h 为语音带宽上限截止频率，f_s 为语音采样频率

6 实验与分析

6.1 在 AN4 上的实验

本节采用 AN4 数据库对本文方法进行测试. AN4 是卡耐基梅隆大学的一个语音数据库，采样率为 16 kHz，内容为数字串和字母串. 训练集含有 74 个发音人（53 男，21 女）的 948 句语音，测试集含有 74 个发音人的 130 句语音. 实验采用 CMU 的 Sphinx-4，并用 AN4 中所有语音文本训练 3 元统计语言模型. 在声学模型的训练方面，采用 3 音子绑定声学模型. 将 AN4 训练语料通过带通滤波器处理，使其带宽在 0.25～6.5 kHz 范围内，用来训练声学模型. 将 AN4 语料分别通过带通滤波器处理，使其

形成 0.25～6.0 kHz、5.5 kHz、5.0 kHz、4.5 kHz、4.0 kHz、3.5 kHz、3.0 kHz 等 7 个窄带语料库，用于自适应或测试. 前端处理的预加重系数为 0.97，梅尔滤波器组的滤波器个数为 40，短时傅立叶变换点数为 512，帧长为 30 ms，帧移 10 ms. 采用 13 维梅尔频率倒谱系数及其一阶，二阶差分参数构成 39 维特征参数. 训练和所有测试中的梅尔倒谱参数都经过 CMN 与 RASTA 处理. 各用于用于对比实验的方法标记如下：

Baseline：用宽带声学模型识别各窄带测试语音，训练数据和测试数据经 CMN 和 RASTA 处理；

MLLR：用宽带模型和窄带数据进行 MLLR 自适应，然后识别对应的窄带语音；

MAP：用宽带模型和窄带数据进行 MAP 自适应，然后识别对应的窄带语音；

SF：用本文方法对窄带语音重建后用宽带声学模型识别.

对于 MLLR 和 MAP 采用监督学习的方法，而为了测试本文方法的带宽测试能力，带宽信息对 SF 是不可知的. 也就是识别时，SF 以句子为单位，检测语音带宽作为频谱重构的输入信息.

从各训练集选择性别均衡的 40 人的语音（约 480 句）作为对应带宽的自适应数据. 从图 7 可以看出，使用 0.25～6.5 kHz 训练的模型在测试语音带宽匹配的情况下识别率最高达到 99.6%，随着测试语料带宽变窄，Baseline 的识别率明显下降，到了 3 kHz时降到了 69.2%. 经过 CMN 与 RASTA 处理后，识别率仍然随着带宽降低而快速降低，这种结果符合式(11)的分析. MLLR 和 MAP 在有较多标注数据时能有效补偿带宽失配，随着带宽变窄，对识别率的提高尤其显著. 自动检测带宽时，SF 在上限截止频率为 6 kHz 和 5.5 kHz 时与模型自适应方法相当，随着带宽继续降低时，其性能优于模型自适应方法. 另外，在带宽给定情况下，SF 的性能要略优于自动检测带宽时的识别率. 这说明经一步提高带宽检测准确率有助于提高本文方法的性能.

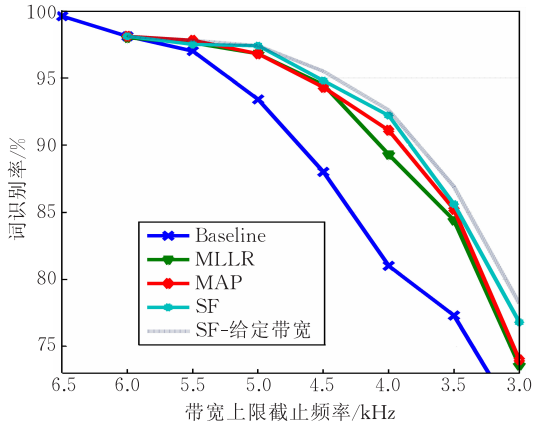

图 7　AN4 上的实验结果

6.2　在 TIMIT/NTIMIT 上的实验

为了进一步验证本文算法的有效性，我们在 TIMIT/NTIMIT 数据库上做了实验. TIMIT 数据库中包含来自美国的 8 个方言区的 630 个发音人（其中男 438，女 192）的语音数据，每个人作 10 次发音，每次发音内容为一个句子，语音采样率为 16 kHz. 库中的 6300 个发音被分成训练集和测试集，其中训练集 4620 句，测试集 1680 句. NTIMIT 是 TIMIT 在窄带电话环境的版本，由 TIMIT 中的数据通过美国本土电话网络后重新录制而成. 与 TIMIT 相比，NTIMIT 数据库虽然采样频率也为 16 kHz，但其真实带宽为电话信道的带宽，即 0.3～3.2 kHz. 除此之外，NTIMIT 的频谱在幅值上也受到电话信道的影响. 本实验训练带宽和特征参数的选择与 AN4 数据库上的实验相同，CMN 与 RASTA 也用于前端处理. 系统采用 TIMIT 训练集训练 3 音子绑定声学模型，用 TIMIT 上所有文本训练 3 元统计语言模型，分别用来识别 NTIMIT 测试集以及用本文方法补偿后的 NTIMIT 测试集. 选择 NTIMIT 测试集的 40 人（男 20，女 20，共 400 句）的语音作为自适应数据为 MLLR 和 MAP 做自适应. 实验结果如表 1 所示，可以看出，Baseline 使用 CMN 与 RASTA 消除信道影响后在 NTIMIT 测试集上识别率只有 56.2%. 这进一步证实了这两种方法补偿带宽失配的局限性. MLLR 和 MAP 的识别率分别达到了 59.4%和 60.6%，相比之下，本文的方法进行补偿后识别率比 Baseline 提高了 6.5 个百分点.

表 1　TIMIT/NTIMIT 上的实验结果

方法	词识别率/%
Baseline	56.2
MLLR	59.4
MAP	60.6
SF	62.7

6.3　实验小节

当丢失带宽较多时，带宽失配对识别率影响比较严重，必须加以补偿. 模型自适应是目前常用且有效的方法，在有足够带标注数据情况下，补偿效果较为明显. 可以预料，随着自适应语料的增加，这类方法的识别率会继续上升甚至超过本文方法. 然而在带宽失配情况下，这类方法存在很大局限性. 复杂应

用环境，诸如网络，语音检索等应用，带宽不可预测而且随时发生变化. 自适应算法需要重新采集数据进行标注然后用于训练. 这使得该类方法难以使用. 本文方法工作在语音识别的前端，随时检测带宽，无需任何标注数据直接补偿，且算法复杂度低，适合实时应用.

7 结 论

训练语音与测试语音带宽失配在应用中广泛存在. 本文研究表明，带宽失配对梅尔倒谱特征分布的影响是扩散性的，且损失的频段越宽，其能量越大，产生的畸变也越大. 这种畸变是时变的，因此无法用加减信道偏移量或在模型域移动高斯均值来实现补偿. 模型自适应方法在带宽不变时效果明显，在带宽变时难以使用. 本文分析了带宽失配对梅尔特征参数的影响并提出了基于频谱折叠的补偿框架，其补偿效果明显. 进一步的实验表明：本文提出的框架能有效提高语音识别系统在带宽失配情况下的鲁棒性.

参 考 文 献

[1] Walker W, Lamere P. Sphinx-4: A flexible open source framework for speech recognition. Sun Microsystems, 2004: 1-15

[2] Li J Y, Deng L, Yu D. A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions. Computer Speech and Language, 2009, 23: 389-405

[3] Ortega-Garcia J, Gonzalez-Rodriguez L. Overview of speaker enhancement techniques for automatic speaker recognition// Proceedings of the International Conference on Spoken Language Processing. Philadelpia, 1996: 929-932

[4] Benesty J, Chen J D, Huang Y T. Microphone Array Signal Processing. Berlin Heidelberg: Springer-Verlag, 2008

[5] You C H, Rahardja S, Member S. Audible noise reduction in eigendomain for speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(6): 1753-1765

[6] Furui S. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing, 1981, 29(2): 254-272

[7] Barras C, Gauvain J L. Feature and score normalization for speaker verification of cellular data//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Hongkong, 2003, 2: 49-52

[8] Pelecanos J, Sridharan S. Feature warping for robust speaker verification//Proceedings of the ISCA Workshop Speaker Recognition. 2001: 213-218

[9] Xiang B, Chaudhari U V, Navrátil J, Ramaswamy G N, Gopinath R A. Short-time Gaussianization for robust speaker verification//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, Florida, 2002, 1: 681-684

[10] Hermansky H, Morgan N, Bayya A, Kohn P. RASTA-PLP speech analysis technique//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Minneapolis, Minnesota, USA, 1992: 1121-1124

[11] Thomas F Q, Douglas A Reynolds, Gerald C. Estimation of handset nonlinearity with application to speaker recognition. IEEE Transactions on Speech and Audio Processing, 2000, 8(5): 567-584

[12] Gales M, Young S. HMM recognition in noise using parallel model combination//Proceedings of the European Conference on Speech Communication and Technology. Berlin, Germany, 1993: 342-346

[13] Moreno P J, Raj B, Stern R M. A vector Taylor series approach for environment independent speech recognition// Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. New York, 1996: 733-736

[14] Cui X, Alwan A. Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR. IEEE Transactions on Speech Audio Processing, 2005, 13(6): 1161-1172

[15] Huo Q, Chan Q, Lee C H. Bayesian adaptive Learning of the parameters of Hidden Markov Model for speech recognition. IEEE Transactions on Speech and Audio Processing, 1995, 3(5): 334-345

[16] Michael L S, Alex A. Training wideband acoustic models using mixed-bandwidth training data for speech recognition. IEEE Transactions on Speech and Audio Processing, 2007, 15(1): 235-245

[17] Milner B P, Vaseghi S V. Bayesian channel equalization and robust features for speech recognition. IEEE Vision, Image and Signal Processing, 1996, 143(4): 223-231

[18] Saeed V. Advanced Digital Signal Processing and Noise Reduction. Chichester, England: John Wiley & Sons Inc., 2005: 376-377

[19] Eberhard H, Gerhard S. Speech and Audio Processing in Adverse Environments. Berlin Heidelberg: Springer-Verlag, 2008: 135-183

[20] Qian Y, Kabal P. Dual-mode wideband speech recovery from narrowband speech//Proceedings of the 8th European Conference on Speech Communication and Technology. Washington, DC, 2003: 1433-1436

[21] Jax P, Vary P. On artificial bandwidth extension of telephone speech. Signal Processing, 2003, 83(8): 1707-1719

[22] Juho K, Laura L, Paavo A. Neural network-based artificial bandwidth expansion of speech. IEEE Transactions on Speech and Audio Processing, 2007, 15(3): 873-881

[23] Carl H, Heute U. Bandwidth enhancement of narrow-band speech signals//Proceedings of the EUSIPCO'94. Edinburgh, Scotland, UK, 1994: 1178-1181

[24] Cassia V, Bruno S, Luiz P. Frequency extension of telephone narrowband speech signal using neural networks//Proceedings of the IMACS' 06. Beijing, 2006: 1576-1579

HE Yong-Jun, born in 1980, Ph.D. candidate, lecturer. His research interests include robust speech recognition, speaker recognition and acoustic event detection.

HAN Ji-Qing, born in 1964, Ph. D., professor, Ph. D. supervisor. His research interests are in speech processing and artificial intelligence.

Background

This research is partly supported by the National Basic Research Program (973 Program) of China under grant No. 2007CB311100, the National High Technology Research and Development Program (863 Program) of China under grant No. 2006AA010103.

With the development of speech processing, speech recognition is widely used in many fields such as human-computer interface, robot speech understanding, speech control system, machine hearing, internet information security, etc. Recently, speech recognition in Web environment and other complex application environments has become a hotspot. However, the environment robustness of a speech recognition system is still a challenge. As one of the major factors degrading the performance of a speech recognizer, bandwidth mismatch exists in many applications, especially, in narrowband transition and Internet. Current channel distortion compensation methods work well in invertible and time-unvarying channel while lose their effectiveness in bandwidth mismatch condition. This paper addresses bandwidth mismatch and proposes a new compensation method based on frequency fold.