基于回声状态网络的音频频带扩展方法

刘 鑫,鲍长春
(北京工业大学电子信息与控制工程学院,北京 100124)

摘 要: 宽带音频通信系统对传输信号有效带宽的限制会降低重建音频的主观质量和自然程度. 本文提出了一种基于回声状态网络的宽带向超宽带音频盲目式频带扩展方法. 该方法借助回声状态网络来模拟音频信号高低频频谱参数间的映射关系,并依据网络模型中的时延递归结构连续更新系统状态来近似描述音频特征的时域演变过程,有效地估计了高频成分的频谱包络. 同时,结合频谱复制方法得到的高频频谱细节,该方法实现了宽带向超宽带音频的有效扩展. 测试结果表明,本文所提方法提升了宽带音频的听觉质量;对于多数测试数据,该方法在静态和动态失真方面获得了优于高斯混合模型扩展方法的扩展性能.

关键词: 音频编码; 音频频带扩展; 回声状态网络; 频谱复制
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2016)11-2758-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.11.027

Audio Bandwidth Extension Method Based on Echo State Network

LIU Xin, BAO Chang-chun
(*School of Electronic Information and Control Engineering, Beijing University of Technology. Beijing* 100124, *China*)

Abstract: The bandwidth limitation in wideband audio communication systems degrades the subjective quality and naturalness of the reproduced signals. In this paper, a wideband to super-wideband audio bandwidth extension method was proposed by using echo state network. The echo state network is adopted to model the mapping function between the low-and high-frequency spectral coefficients of audio signals, and the temporal evolution of audio features is represented by continuously state updating on the basis of the recursive structure in the network, for effectively estimating the high-frequency spectral envelope. By combining the high-frequency fine spectrum extended by spectral translation, the proposed method can effectively extend the bandwidth of wideband audio to super-wideband. Evaluation results show that the proposed method upgrades the auditory quality of wideband audio, and gains better extension performance than the Gaussian mixture model-based bandwidth extension method in terms of both static and dynamic distortions for most test data.

Key words: audio coding; audio bandwidth extension; echo state network; spectral translation

1 引言

受到通信网络传输速率的限制,感知音频编码方法通常限制音频有效带宽,优先编码其低频成分,以提升编码效率[1]. 然而,人们并不满足于现有的宽带音频通信质量,并期望获得更加明亮而富有表现力的音频服务. 为此,如何使宽带音频系统获得或接近超宽带音频的主观听感成为了音频通信领域亟待解决的问题.

作为有效的音频增强方法,频带扩展在不改变信源编码和网络传输的前提下,在解码器重建信号中人为地增添高频成分,以实现信号带宽的扩展[2]. 近十几年来,相关学者从频谱包络和频谱细节两个方面提出了众多频带扩展解决方案. 非正式听力测试结果表明,高频频谱包络估计的准确性对重建音频听觉质量的提升十分重要[3]. 因此,可借助统计学习方法拟合高低频频谱间的映射关系. 1994 年,Y M Cheng 等学者提出利用统计恢复函数来预测高频频谱,初步改善了重建音频的质量[4]. 同年,H Carl 借助低频特征和高频频谱包络的联合码本模拟两者的一对一映射,提出了基于码本映射的频谱包络估计方法[5]. 该方法降低了扩展后音频频谱失真. 在其基础上,有学者相继提出了内插、软判决和分裂码本映射等方法,以降低单一码本造成的

收稿日期:2015-06-15;修回日期:2015-11-15;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61072089, No.61471014)

频谱失真[6~8]. 2000 年,K Park 和 H S Kim 提出了基于高斯混合模型(Gaussian mixture model,GMM)的频谱包络估计方法[9],该类方法利用 GMM 来近似高低频特征联合概率密度,并在均方误差最小准则下实现了高频频谱包络的估计. 该方法基于软聚类的连续统计模型,抑制了码本映射等离散映射方法重建音频频谱的非自然间断. 此外,有学者利用前向神经网络来估计高频频谱包络[10,11]. B Iser 等学者则将前向神经网络方法和码本映射方法进行了对比,结果表明两者扩展后音频的听觉质量没有显著差异,而前向神经网络方法计算复杂度明显降低[12].

上述方法均着重去发掘当前音频帧内部高低频的相关性,更侧重于频谱静态特性的展现. 而 P Jax 等利用隐马尔科夫模型来模拟音频频谱包络时域动态演变[13,14],将帧间相关性引入到频谱包络估计中[15,16]. 但是,该方法仅利用离散的状态来分段模拟实际音频频谱的时间演变,其重建音频仍然存在动态失真. 为此,有必要在频谱包络估计中引入连续动态模型. 本文提出了一种基于回声状态神经网络(echo state network,ESN)的频谱包络估计方法,借助递归结构的非线性特性连续更新系统状态,进而描述音频特征的动态演变,并依据高维空间的线性映射来拟合高低频特征参数间的非线性关系. 结合基于频带复制的频谱细节扩展方法,实现了宽带音频向超宽带音频的有效扩展.

2　基于回声状态网络的音频频带扩展方法

本文所提方法的基本原理如图 1 所示. 输入信号为 16kHz 采样 7kHz 带宽的宽带音频信号. 该信号首先经过上二采样和低通滤波,获得 32kHz 采样 7kHz 带宽的音频信号,并按照 32ms 帧长、16ms 帧移分帧,加汉明窗. 然后,加窗后信号 $s_{wb}(i)$,$i=0,\cdots,1023$ 经过离散傅里叶变换(Discrete Fourier transform,DFT)转换到频域,并在梅尔频率尺度上利用三角滤波器组将 64 ~ 7000Hz 频率范围内的音频频谱 $A(k)$ 均匀地划分为 20 个通道,

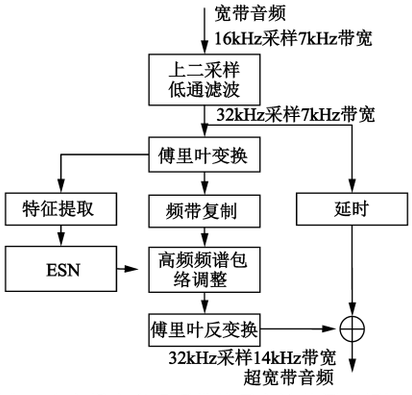

图1　本文提出音频频带扩展方法的原理框图

进而提取 20 维梅尔频率倒谱参数(Mel Frequency Cepstral Coefficient,MFCC)$F_{MFCC}(i)$,$i=0,\cdots,19$. 接下来,对提取得到的 F_{MFCC} 进行区间归一化处理,并输入到预先训练好的 ESN 中实现高频频谱包络的估计. 此处,高频频谱包络采用 7 ~ 14kHz 范围内 4 个不交叠非均匀子带的均方根值 $F_{RMS}(i)$,$i=0,\cdots,3$ 表示,如下式所示.

$$F_{RMS}(i)=\sqrt{\frac{1}{h(i)-l(i)+1}\sum_{k=l(i)}^{h(i)}A^2(k)},\ i=0,1,\cdots,3 \tag{1}$$

式中,$A(k)$ 为音频频谱幅度值,$h(i)$ 和 $l(i)$ 分别为第 i 个子带上下限频率对应的频点序号. 各子带的中心频率分别位于 8470Hz、9338Hz、11653Hz 以及 13657Hz.

高频频谱细节则采用频谱复制方法,将低频频谱直接复制到高频频谱,并根据估计得到的 F_{RMS} 来调整扩展后高频频谱包络. 最终,利用离散傅里叶逆变换和叠接相加技术将重建高频转换到时域中,并结合适当延迟后的宽带音频信号,重建出超宽带音频.

2.1　基于 ESN 的频谱包络估计

令 $\boldsymbol{F}_X(m)$ 表示第 m 帧宽带音频的 MFCC,其维数为 $d_X=20$,$\boldsymbol{F}_Y(m)$ 表示第 m 帧高频子带均方根值,其维数为 $d_Y=4$. 通过 $\boldsymbol{F}_X$ 估计 $\boldsymbol{F}_Y$ 的过程可用某个映射函数 $F(\cdot)$ 表示,

$$\boldsymbol{F}_Y=F(\boldsymbol{F}_X) \tag{2}$$

为了逼近高低频参数间的真实映射,本文引入了 ESN[17~20],其网络结构如图 2 所示. 首先利用隐含层中预生成的大规模递归结构将 $\boldsymbol{F}_X$ 转换到高维空间中,进而借助高维隐含状态 $\boldsymbol{S}_{hidden}$ 的连续更新来描述 $\boldsymbol{F}_X$ 的动态演变. 在此基础上,ESN 分别从 $\boldsymbol{F}_X$ 与 $\boldsymbol{S}_{hidden}$ 中获取音频低频成分的静态和动态特性,进一步借助高维空间中的线性映射逼近 $\boldsymbol{F}_X$ 与 $\boldsymbol{F}_Y$ 间的非线性映射.

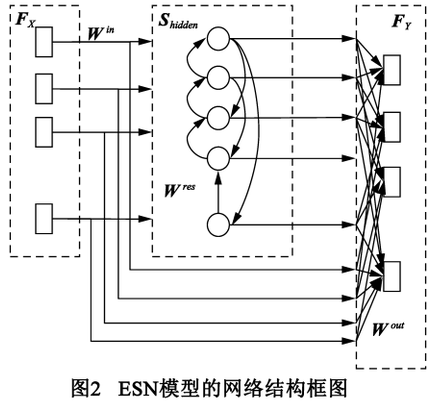

图2　ESN模型的网络结构框图

2.1.1　ESN 的数据模型

ESN 可分为隐含状态更新和高维空间映射两个部分.

隐含状态更新中,ESN 采用 leaky-integrated 函数作

为隐含层的非线性激活单元. 该函数结合非线性变换和时间递归结构,实现对隐含状态 $\boldsymbol{S}_{hidden}$ 的动态更新,具体过程[19]如下式所示,

$$\tilde{\boldsymbol{S}}_{hidden}(m)=\tanh\left(\boldsymbol{W}^{in}\begin{bmatrix}1\\ \boldsymbol{F}_X(m-1)\end{bmatrix}+\boldsymbol{W}^{res}\boldsymbol{S}_{hidden}(m-1)\right) \tag{3}$$

$$\boldsymbol{S}_{hidden}(m)=(1-\alpha)\boldsymbol{S}_{hidden}(m-1)+\alpha\tilde{\boldsymbol{S}}_{hidden}(m) \tag{4}$$

式中,$\boldsymbol{S}_{hidden}(m)$ 和 $\tilde{\boldsymbol{S}}_{hidden}(m)$ 分别为第 m 帧 ESN 网络隐含状态及其更新值,其维数 d_S 又称为隐含层规模. 实际应用中随着 d_S 增大,ESN 能保留更丰富的宽带音频特征,从而提升高频频谱包络估计的准确性.

输入权值矩阵 $\boldsymbol{W}^{in}$ 表征了 $\boldsymbol{F}_X$ 与 $\boldsymbol{S}_{hidden}$ 间的关联性,其维数为 $d_S\times(d_X+1)$. 通常,$\boldsymbol{W}^{in}$ 中元素的取值范围限制在$[-a_{in},a_{in}]$之间. 若 a_{in} 趋近于 0,leaky integrated 函数呈现近似线性特性. 随着 a_{in} 增大,$\boldsymbol{F}_X$ 在驱动 $\boldsymbol{S}_{hidden}$ 的更新过程中则呈现出更多的非线性特性,进而提升 ESN 对高低频频谱真实映射的拟合能力.

递归权值矩阵 $\boldsymbol{W}^{res}$ 则表征了前后帧 $\boldsymbol{S}_{hidden}$ 的关联性,其维数为 $d_S\times d_S$,实际应用中可通过人为调节 $\boldsymbol{W}^{res}$ 谱半径 a_{res} 的大小来控制 ESM 递归结构的稳定性. 此外,a_{in} 和 a_{res} 共同决定了 $\boldsymbol{S}_{hidden}$ 更新过程中 $\boldsymbol{F}_X$ 和 $\boldsymbol{S}_{hidden}(m-1)$ 的相对重要性. 若 a_{res} 较大,$\boldsymbol{S}_{hidden}(m-1)$ 在状态更新过程的作用中更为显著,网络结构会保留 $\boldsymbol{F}_X$ 中更多的长时相关性,有助于改善重建高频频谱包络的时间连续性.

隐含层内部神经元间通常采用稀疏递归链接的方式. $\boldsymbol{W}_{res}$ 中元素的稀疏程度 $f_{sparsity}$ 表示了 $\boldsymbol{S}_{hidden}$ 内相互之间存在连接的神经元占所有神经元总数目的百分比. 理论上讲,ESN 采用稀疏的 $\boldsymbol{W}_{res}$ 能够改善网络泛化能力,提升高频频谱包络估计的准确性,并进一步减低 ESN 对模型参数存储的需求.

参数 α 为泄漏速率,它表征了 $\boldsymbol{S}_{hidden}$ 的动态更新速率. 当 $\alpha=1$ 时,leaky-integrated 非线性函数退化为 tanh 函数;随着 α 逐渐减小,ESN 中 $\boldsymbol{S}_{hidden}$ 的更新会明显减慢,增加了递归网络的短时记忆长度.

经过 leaky-integrated 非线性单元的逐帧更新,$\boldsymbol{S}_{hidden}$ 获取了 $\boldsymbol{F}_X$ 的动态特性. 在此基础上,ESN 网络将 $\boldsymbol{F}_X$ 和 $\boldsymbol{S}_{hidden}$ 相结合构成高维特征矢量,进一步借助高维空间中的线性映射来逼近 $\boldsymbol{F}_X$ 与 $\boldsymbol{F}_Y$ 间的非线性映射[17],如下式所示,

$$\hat{\boldsymbol{F}}_Y(m)=\boldsymbol{W}^{out}\begin{bmatrix}1\\ \boldsymbol{F}_X(m)\\ \boldsymbol{S}_{hidden}(m)\end{bmatrix} \tag{5}$$

式中,$\boldsymbol{W}^{out}$ 表示输出权值矩阵,其维数为 $d_Y\times(1+d_X+d_S)$. 最终,$\hat{\boldsymbol{F}}_Y(m)$ 可作为当前帧高频频谱包络的估计值.

2.1.2 ESN 的训练方法

根据上述模型,可采用适当的训练方法来求取 ESN 中的模型参数($\boldsymbol{W}^{in}$、$\boldsymbol{W}^{res}$ 和 $\boldsymbol{W}^{out}$). 传统基于梯度下降的训练方法并不能保证 ESN 的稳定性,并且计算复杂,收敛慢. 鉴于此,有学者针对 ESN 网络结构提出了一种启发式参数训练方法[17].

该方法首先在初始化阶段随机生成 $\boldsymbol{W}^{in}$ 和 $\boldsymbol{W}^{res}$. 由于 $\boldsymbol{W}^{res}$ 的矩阵谱半径 a_{res} 直接影响了 ESN 的稳定性,因此需要根据实际应用条件对其人为调整[17]. 令 $\boldsymbol{W}$ 为一个随机生成的稀疏矩阵,$\lambda_{\max}$ 为 $\boldsymbol{W}$ 的谱半径,则 $\boldsymbol{W}^{res}$ 可以表示为,

$$\boldsymbol{W}^{res}=\frac{a^{res}\boldsymbol{W}}{|\lambda_{\max}|} \tag{6}$$

相关研究结论[21]表明,递归神经网络的训练中输出权值矩阵会根据梯度变化而迅速改变,隐层内部连接的权值则呈现出高度耦合,其变化较为缓慢. ESN 隐含层中递归节点数目庞大,网络复杂,因此在参数训练中 $\boldsymbol{W}^{in}$ 和 $\boldsymbol{W}^{res}$ 呈现出显著的强耦合性,不随梯度剧烈改变[17]. 鉴于此,启发式训练方法可令 $\boldsymbol{W}^{in}$ 和 $\boldsymbol{W}^{res}$ 在其后参数训练中保持固定不变,而通过修正 $\boldsymbol{W}^{out}$ 的方式调整 ESN 模型对 $\boldsymbol{F}_X$ 和 $\boldsymbol{F}_Y$ 间非线性映射的拟合能力,简化递归神经网络的训练过程.

假定 $\boldsymbol{F}_X(m)$ 和 $\boldsymbol{F}_Y(m)$ 分别是训练数据集中的宽带音频特征和高频频谱包络参数,$m=0,1,\cdots,N_{train}-1$ 为音频帧序号,N_{train} 为数据总帧数. 参数训练方法可以依据式(3)、式(4)来驱动 ESN 实现对 $\boldsymbol{S}_{hidden}(m)$ 的逐帧更新. 在实际应用中,通常会随机设置 $\boldsymbol{S}_{hidden}(-1)$,这样必然会影响网络的稳定性. 为此本文设定了网络稳定时间 $T_0=200$ms,并假设当 ESN 超过该时刻后达到渐进稳定. 从 T_0 开始,逐帧收集 $\boldsymbol{F}_X(m)$、$\boldsymbol{S}_{hidden}(m)$ 以及 $\boldsymbol{F}_Y(m)$,并分别构成状态收集矩阵 $\boldsymbol{B}$ 和期望输出矩阵 $\boldsymbol{Q}$. 其中,$\boldsymbol{B}$ 的维数为 $(1+d_X+d_S)\times(N_{train}\text{-}T_0)$,其每列元素为 $[1,\boldsymbol{F}_X(m)^{\mathrm{T}},\boldsymbol{S}_{hidden}(m)^{\mathrm{T}}]^{\mathrm{T}}$,包含了每一帧的宽带音频特征的静态和动态特性;而 $\boldsymbol{Q}$ 的维数为 $(d_Y)\times(N_{train}\text{-}T_0)$,其每列元素为该帧音频的 $\boldsymbol{F}_Y(m)$.

依据 $\boldsymbol{B}$ 和 $\boldsymbol{Q}$,可以采用均方误差最小准则来估计 ESN 网络的 $\boldsymbol{W}^{out}$,使得 $\hat{\boldsymbol{F}}_Y$ 逼近真实的高频频谱包络. 然而,实验观测发现,若 $\boldsymbol{W}^{out}$ 中元素的数值较大,$\boldsymbol{W}^{out}$ 会放大 $\boldsymbol{F}_X(m)$ 的微小差异,使得所训练网络对样本数据过度拟合. 为此,本文在误差函数中引入了正则项,利用岭回归来求解 $\boldsymbol{W}^{out}$[17],如下式所示,

$$\boldsymbol{W}^{out}=\arg\min_{\boldsymbol{W}^{out}}\left(\frac{1}{N_{train}-T_0}\sum_{m=0}^{N_{train}-T_0-1}\|\boldsymbol{F}_Y-\hat{\boldsymbol{F}}_Y\|^2+\beta\|\boldsymbol{W}^{out}\|^2\right) \tag{7}$$

式中,$\beta\|\boldsymbol{W}^{out}\|^2$ 为正则项,可在兼顾最小化预测均方误差的前提下对 $\boldsymbol{W}^{out}$ 中数值较大的元素进行惩罚;正则因

子β用于控制上式两项间的相对重要性.

对上式求解,可以得到最终的$\boldsymbol{W}^{out}$,

$$\boldsymbol{W}^{out}=\boldsymbol{Q}\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}+\beta\boldsymbol{I})^{-1} \tag{8}$$

式中,$\boldsymbol{I}$为单位矩阵.

根据所获得的$\boldsymbol{W}^{res}$、$\boldsymbol{W}^{in}$和$\boldsymbol{W}^{out}$,可构建出一个完整的 ESN. 在实际扩展中,利用每一帧提取的$\boldsymbol{F}_{MFCC}$连续更新$\boldsymbol{S}_{hidden}$,进而借助高维线性映射有效估计高频频谱包络.

2.2　高频成分的重建

本文采用频谱复制来扩展高频频谱细节,即将 0 ~ 7kHz 范围内的频谱细节直接复制到 7 ~ 14kHz 的高频中. 而低频频谱细节可采用归一化幅度谱参数$A_{norm}(k)$,$k=0,\cdots,223$,来表示,

$$A_{norm}(k)=\frac{A(k)}{\boldsymbol{F}_{RMS\_WB}(i)},\ i=\lfloor k/N_{subband}\rfloor \tag{9}$$

式中,$A(k)$为音频幅度谱;$\boldsymbol{F}_{RMS_WB}(i)$为低频子带均方根值,其计算方式与式(1)相近,可初步描述音频低频频谱包络. 此处,为了保证$A_{norm}(k)$的频谱平坦度,低频频谱子带采用均匀划分方式. 0 ~ 7*kHz* 的频率范围分为为 14 个子带,每个子带包含$N_{subband}=16$个频点. 那么,经过频带复制,扩展后高频频谱细节可表示为,

$$A_{norm}(k)=A_{norm}(k-224),\ k=224,\cdots,447 \tag{10}$$

由于高频频谱包络采用 ERB 尺度分带,因此高频频谱细节采用相同方式划分为四个互不交叠的子带,并结合 ESN 估计到的$\hat{F}_{RMS}(i)$,$i=0,\cdots,3$,通过高频频谱包络调整,生成扩展后的高频频谱$A_{swb}(k)$,$k=224,\cdots,447$,如下式所示,

$$A_{swb}(k)=A_{norm}(k)\hat{F}_{RMS}(i),\ i=Subband(k) \tag{11}$$

式中,$Subband(k)$表示第k个频点所在高频子带的序号.

高频频谱相位$\theta(k)$,$k=224,\cdots,447$,同样采用频谱复制方法获得,如下式所示,

$$\theta(k)=\theta(k-224),\ k=224,\cdots,447 \tag{12}$$

最终,根据 IDFT,高频频谱转换到时域. 而上采样后的宽带音频经过适当的延时后,与人为生成的高频信号相结合,重建出超宽带音频.

3　回声状态网络模型参数对扩展性能的影响

本文针对 2.1.1 节中涉及到的网络参数($\boldsymbol{W}^{in}$缩放因子a_{in}、$\boldsymbol{W}^{res}$谱半径a_{res}、$\boldsymbol{W}^{res}$稀疏度$f_{sparsity}$、leaky-integrated 函数泄漏率α、岭回归正则因子β、储备池规模d_S等)对 ESN 方法性能的影响进行了初步测评. ESN 训练数据源自于 4 小时时长现场音乐会转录的无损音频,其中包括对话、音乐、人声演唱、实况背景音效等类型. 声音采样率为 32kHz,有效带宽为 14kHz,采用 16 比特 PCM 进行存储. 该超宽带数据库经过低通滤波、下采样和时间延迟进一步获得平行宽带数据库. 分别从平行宽带和超宽带音频数据中提取 20 维 MFCC 和 4 维高频子带均方根参数作为 ESN 的输入特征矢量$\boldsymbol{F}_X$和期望输出矢量$\boldsymbol{F}_Y$. 所获得的 50% 样本数据用于模型训练,而另 50% 数据用于性能测试.

此外,本文选择了 7 ~ 14kHz 频率范围内频带扩展方法处理后音频信号与原始超宽带音频信号的对数谱失真(log spectral distortion,LSD)作为客观测度对 ESN 的预测准确度进行评价. LSD 可以直接利用 DFT 功率谱计算得到[22],如下式所示,

$$d_{LSD}(i)=\sqrt{\frac{1}{N_{high}-N_{low}+1}\sum_{n=N_{low}}^{N_{high}}\left[10\log_{10}\frac{P_i(n)}{\hat{P}_i(n)}\right]^2} \tag{13}$$

式中,$d_{LSD}(i)$为第i帧的 LSD 值;P_i和$\hat{P}_i$分别为原始超宽带音频和扩展后音频的 DFT 功率谱;N_{low}和N_{high}为 7 ~ 14kHz 高频频带频率上下限对应的序号. 在计算 LSD 值之前,所有数据需要重采样到 32kHz,并和原始音频进行精确地时域对齐. 接下来,逐帧计算 7 ~ 14kHz 频率范围高频 LSD 值. 最终,将整段数据的平均 LSD 值作为其客观质量测度. 下面将根据参数重要性逐一评价不同储备池参数对预测均方误差的影响.

(1)$\boldsymbol{W}^{in}$缩放因子a_{in}

缩放因子a_{in}决定了 leaky-integrated 激励函数的非线性特性. 本文利用实验测试的手段来经验性地确定a_{in}. 分别设定$a_{res}=1$、$f_{sparsity}=1$、$\alpha=1$、$\beta=1$、$d_S=4\times d_X=80$,并在 LSD 测度下针对不同的a_{in}值进行测试,如表 1 所示. 当$a_{in}=1/8$时,ESN 获得最小的 LSD. 而当非线性函数趋近于线性或二值函数时,LSD 值均会增加. 由此可见,$\boldsymbol{F}_X$与$\boldsymbol{F}_Y$之间确实存在一定的非线性关系.

表 1　不同的a_{in}下 ESN 模型的 LSD 值

a_{in}	d_{LSD}(dB)
1/32	6.3477
1/16	6.3455
1/8	6.2991
1/4	6.3170
1/2	6.3423
1	6.3512
2	6.3465
4	6.3499
8	6.3491

(2)$\boldsymbol{W}^{res}$谱半径a_{res}

a_{res}是$\boldsymbol{W}^{res}$的谱半径,它决定了 ESN 的稳定性. 本文借助 LSD 测度经验性地确定$\boldsymbol{W}^{res}$谱半径a_{res},如表 2 所示. 本文将其他参数分别设定如下,$a_{in}=1/8$、$f_{sparsity}=1$、$\alpha=1$、$\beta=1$、$d_S=4\times d_X=80$. 当$a_{res}=0.6$时,模型的 LSD 值最小;而当$a_{res}>1$时,ESN 的 LSD 值逐渐增大,部分

帧估计的高频频谱包络和原始音频具有较大的差异;而当 a_{res} 较小时,储备池中内部神经元的递归作用减弱,也会导致模型的 LSD 有所增加. 由此可见,在保证 ESN 网络稳定的条件下,适当引入递归特性有助于提升 ESN 对宽带音频特征时间动态特性的描述能力.

表 2 不同的 a_{res} 下 ESN 模型的 LSD 值

a_{res}	d_{LSD}(dB)
0.2	6.3399
0.4	6.3201
0.6	6.2752
0.8	6.2937
1	6.2931
1.2	6.3152
1.4	6.3187
1.6	6.3513
1.8	6.3645

(3) $\boldsymbol{W}^{res}$ 稀疏度 $f_{sparsity}$

令其他参数分别设定为,$a_{in}=1/8$、$a_{res}=0.6$、$\alpha=1$、$\beta=1$、$d_S=4\times d_X=80$,本文进一步针对稀疏度 $f_{sparsity}$ 进行评价,如表 3 所示. 当隐藏状态神经元之间采用全递归连接的方式,网络模型获得最小的 LSD 值;而在 $f_{sparsity}$ 较低的情况下,LSD 会有所增加;而当 $f_{sparsity}$ 低于 10% 左右时,LSD 值将降低到 6.28dB 附近. 由此可见,增加 $\boldsymbol{W}^{res}$ 的稀疏程度不能改善 ESN 重建音频的客观质量. 然而,采用较小的 $f_{sparsity}$(如 0.025)可以在不过多加重 LSD 的前提下提升 ESN 的训练效率,并降低模型的存储需求.

表 3 不同的 $f_{sparsity}$ 下 ESN 模型的 LSD

$f_{sparsity}$	d_{LSD}(dB)
1	6.2752
0.8	6.3499
0.6	6.3528
0.4	6.3185
0.2	6.4271
0.1	6.3105
0.05	6.2850
0.025	6.2824
0.0125	6.2907

(4) leaky-integrated 函数泄漏率 α

泄漏率 α 表征了 $\boldsymbol{S}_{hidden}(m)$ 的动态更新速度. 本文分别设定 $a_{in}=1/8$、$a_{res}=0.6$、$f_{sparsity}=1$、$\beta=1$、$d_S=4\times d_X=80$,并测试了不同 α 对模型性能的影响,如表 4 所示. 结果表明,α 对 LSD 值的影响不大,即 $\boldsymbol{S}_{hidden}$ 更新过程中涉及的 $\boldsymbol{F}_X$ 长时记忆性对 ESN 的性能没有明显的改进作用.

(5) 岭回归正则因子 β

$\boldsymbol{W}^{out}$ 可采用岭回归计算,以防止过度拟合. 设置 $a_{in}=1/8$、$a_{res}=0.6$、$f_{sparsity}=1$、$\alpha=1$、$d_S=4\times d_X=80$,则 β 与 LSD 间的关系如表 5 所示. 基于岭回归方法训练模型的 LSD 值明显低于线性回归方法($\beta=0$);当 $\beta=3.5$ 时,ESN 获得最优的性能.

表 4 不同的 α 下 ESN 模型的 LSD 值

α	d_{LSD}(dB)
1	6.2752
0.9	6.2762
0.8	6.2767
0.7	6.2767
0.6	6.2784
0.5	6.2763
0.4	6.2766
0.3	6.2784
0.2	6.2813
0.1	6.2847

表 5 不同的 β 下 ESN 模型的 LSD 值

β	d_{LSD}(dB)
0	7.4324
0.5	6.2824
1	6.2752
1.5	6.2723
2	6.2733
2.5	6.2724
3	6.2702
3.5	6.2692
4	6.2701
4.5	6.2709
5	6.2708

(6) 隐含层规模 d_S

一般来说,如果采用适当的正则化方法来抑制过度拟合,那么 d_S 越大可获得更好的性能. ESN 的参数训练方法计算简单,因此 d_S 通常在数百左右. 然而考虑到实际存储需求,仍需适当控制其规模. 令 $a_{in}=1/8$、$a_{res}=0.6$、$f_{sparsity}=1$、$\alpha=1$、$\beta=3.5$,本文针对网络隐含层规模进行了探讨,如表 6 所示. LSD 测试结果表明,随着 d_S 逐渐增大,ESN 方法所重建高频频谱失真逐渐降低,而其最小值出现在 $d_S=24\times d_X=480$ 处.

表 6 不同的 d_S 下 ESN 模型的 LSD

d_S	d_{LSD}(dB)
20	6.2703
40	6.2785
80	6.2692
160	6.2647
320	6.2619
480	6.2430
640	6.2494
800	6.2477
960	6.2470
1120	6.2493

综合上述评测结果,本文最终确定网络参数为 a_{in}

$=1/8$、$a_{res}=0.6$、$f_{sparsity}=1$、$\alpha=1$、$\beta=3.5$、$d_S=24\times d_X=480$.

4　性能评价与测试结果

本文首先根据扩展后超宽带音频与原始超宽带音频高频子带均方根值之间的均方误差来初步评价高频频谱包络估计方法的准确性. 在此基础上,进一步从对数谱失真(log spectral distortion,LSD)、双曲余弦测度(COSH)和差分对数谱失真(differential log spectral distortion,DLSD)三个方面对所提方法和基于 GMM 的频带扩展参考方法重建音频的客观质量进行对比. 此外,本文依据主观偏爱测试和计算复杂度对频带扩展方法进行评价.

4.1　参考算法与音频数据

除了频谱包络估计模块,GMM 参考方法和图 1 所示的扩展原理基本一致. 在 GMM 方法中,每帧提取的 MFCC 输入到基于 GMM 的最小均方误差估计器. 其中 GMM 包含 128 个高斯分量,并采用对角方差矩阵. 而高频频谱细节同样采用频谱复制方法.

参考方法与本文方法所需训练数据均源自 4 小时现场音乐会转录的无损音频,其中包括对话、不同类型的音乐、人声演唱以及现场背景等. 相关测试结果表明,进一步增加训练数据的长度对频带扩展方法主客观性能的提升并不明显. 对该音频数据进行重采样和时间对齐,可分别获得宽带和超宽带音频的平行数据库. 所有数据在进行处理前,其声音水平需归一化至 -26dBov. 此外,本文从 MPEG 音频质量主观听觉测试数据库中选择了 15 段音频作为测试数据,包含了流行音乐、器乐独奏、交响乐片段以及语音等不同类型. 这些数据长度限制在 10 ~ 20s 范围内,采样率为 32kHz,有效带宽为 14kHz. 该数据通过截止频率为 7kHz 的低通滤波和下采样转换为宽带信号,并将其声音水平归一化到 -26dBov 后作为频带扩展方法的输入. 下面本文分别根据频谱包络估计误差、扩展后音频主客观质量以及计算复杂度对算法性能进行详细分析.

4.2　频谱包络估计误差

为了验证高频频谱包络估计的准确性,本文首先依据重建音频与原始音频高频频谱包络间的均方误差对本文所提方法和 GMM 参考方法进行了对比. 此处,频谱包络估计的均方误差可以定义如下,

$$e_{MS}(i)=\frac{1}{4}\sum_{n=0}^{3}(\boldsymbol{F}_{RMS}(i,n)-\hat{F}_{RMS}(i,n))^2 \qquad (14)$$

其中,$e_{MS}(i)$为第 i 帧音频信号高频频谱包络的均方误差,$\boldsymbol{F}_{RMS}(i,n)$和 $\hat{F}_{RMS}(i,n)$分别为第 i 帧音频第 n 个高频子带的子带均方根真实值和估计值,$n=0,1,2,3$ 为子带序号. 接下来,将每段测试数据上所有帧的均方误差进行平均,其结果作为最终的误差测度.

表 7 给出两种方法对于不同类型音频信号高频频谱包络估计误差的结果. 其中,乡村、爵士和摇滚音乐高频频谱能量明显高于其他类型音频,因此不同估计方法重建高频频谱包络的平均误差较高. 小提琴独奏和交响乐音频频谱成分则多集中在低频,随着频率增加其高频逐渐暗淡,因此这两种音频频谱包络估计的误差相对较低. 而语音中部分清音高频能量较强,其频谱包络的估计值和原始包络间同样存在较大的误差.

表 7　不同扩展方法高频频谱包络估计的误差

数据类型	GMM	ESN
乡村音乐	10.8282	17.8295
爵士音乐	51.6419	26.5364
摇滚音乐	9.9481	9.8263
小提琴独奏	2.7268	2.5441
交响乐	2.7139	3.4748
语音	8.5385	7.2954
平均值	14.3996	11.2511

总体上讲,本文方法能够有效地估计出高频成分的频谱包络,其频谱包络估计误差的平均值较参考算法降低了 3.15 左右. 对于爵士音乐,ESN 方法重建音频高频能量丰富,其频谱包络更接近于原始音频,而 GMM 方法重建高频频谱则相对暗淡,进而造成了较为明显的估计误差. 而对于乡村音乐和交响乐,ESN 方法重建高频频谱整体能量偏高,其频谱包络估计误差高于参考算法.

4.3　客观质量测试

此外,本文进一步利用 LSD、COSH 以及 DLSD 三种测度对不同方法进行客观评价,结果如表 8 所示.

4.3.1　对数谱失真

本文分别对所提方法和参考方法重建音频进行了 LSD 比较,如表 8 所示. 与频谱包络估计误差分析结果相近,ESN 方法 LSD 的平均性能略优于 GMM 方法. 对于摇滚音乐、小提琴独奏、语音信号,两种方法 LSD 值的差异均在 ±0.5dB 范围内. 而两者 LSD 差异较大的是爵士音乐,这种类型音频信号高频能量比较丰富,并且在时域上低音贝斯伴奏使得该音频存在明显的暂态成分. GMM 重建高频频谱比较平坦,而 ESN 方法重建频谱包络更接近于原始音频,因而获得了较低的 LSD. 而对于交响乐和乡村音乐,ESN 重建高频的整体能量略高,尽管主观听感上音频更为明亮,但是其 LSD 高于 GMM 方法重建音频.

表 8 不同扩展方法重建音频的客观失真测试结果

数据类型	LSD	
	GMM	ESN
乡村音乐	6.6592	7.5732
爵士音乐	12.5328	8.4177
摇滚音乐	6.2152	6.2856
小提琴独奏	3.7840	3.5192
交响乐	3.8029	5.0384
语音	6.2044	5.7943
平均值	6.5331	6.1047
数据类型	COSH	
	GMM	ESN
乡村音乐	28.4653	31.9172
爵士音乐	21.1773	17.1231
摇滚音乐	19.0094	20.1251
小提琴独奏	2.5972	2.6134
交响乐	3.0143	3.2091
语音	39.655	37.7401
平均值	18.9864	18.7880
数据类型	DLSD	
	GMM	ESN
乡村音乐	4.3296	3.7866
爵士音乐	7.2664	4.8089
摇滚音乐	4.1076	3.1428
小提琴独奏	2.3920	2.0596
交响乐	2.90145	3.1192
语音	4.0310	3.3972
平均值	4.1713	3.3857

4.3.2 双曲余弦测度

相比于 LSD，Itakura-Saito 失真为音频频谱峰值提供更多的权重，因而与真实的主观听觉质量更为相近. Itakura-Saito 失真 d_{IS}[23] 可以定义如下，

$$d_{IS}(P_i,\hat{P}_i)=\frac{1}{N_{high}-N_{low}+1}\sum_{n=N_{low}}^{N_{high}}\left[\frac{P_i(n)}{\hat{P}_i(n)}-\log_{10}\frac{P_i(n)}{\hat{P}_i(n)}-1\right] \tag{15}$$

作为距离测度，d_{IS}并不具有对称性，因此本文选择了 COSH 测度作为修正失真测度来描述重建音频的感知失真. COSH 测度 d_{COSH}定义如下[23]，

$$d_{\mathrm{COSH}}(i)=\frac{1}{2}[d_{IS}(P_i,\hat{P}_i)+d_{IS}(\hat{P}_i,P_i)] \tag{16}$$

COSH 测度只针对 7 ~ 14kHz 频率范围进行计算，且每段测试数据上所有帧测度的平均值作为最终的 COSH 测度. 两种频带扩展方法重建音频的 COSH 值比较结果如表 8 所示. 在整体上，ESN 重建音频的 COSH 值与 GMM 方法比较接近. ESN 方法在爵士音乐和语音信号上重建音频的客观质量要优于 GMM 方法，而在乡村音乐上则略低于参考方法. 综合以上两项性能对比结果，可以获得结论：ESN 静态客观失真相比 GMM 方法略有提升.

4.3.3 差分对数谱失真

音频频谱帧间的连续性与其频谱重建的准确性具有同样的感知重要性. 本文选择 DLSD 作为动态失真测度来评价扩展后音频信号频谱包络的时间演变平滑程度. 如果 DLSD 值较小，则可认为音频频谱在时间上变化相对缓慢，有益于重建音频的整体主观听觉质量. DLSD 测度[24] 可定义如下，

$$d_{DLSD}(i)=\sqrt{\frac{1}{2(N_{high}-N_{low}+1)}\sum_{n=N_{low}}^{N_{high}}\left[10\log_{10}\frac{P_i(n)}{P_{i-1}(n)}-10\log_{10}\frac{\hat{P}_i(n)}{\hat{P}_{i-1}(n)}\right]^2} \tag{17}$$

式中，P_{i-1}和 $\hat{P}_{i-1}$分别表示前一帧原始超宽带音频和重建音频信号的 DFT 功率谱.

表 8 同样给出了两种方法重建音频 DLSD 的结果. 其中，小提琴独奏、交响乐音频高频成分能量较低，同时高频频谱的时间平滑性较好，因此两种扩展方法重建音频的动态失真差异并不大. 而 ESN 方法对爵士、摇滚、乡村音乐中暂态成分的刻画更接近于原始音频，其 DLSD 数值明显优于 GMM 方法. 对于语音，ESN 方法 DLSD 分数在 3.40dB 左右，较 GMM 提升了 0.7dB 左右.

综上所述，在静态失真方面本文所提出的 ESN 方法平均性能相比 GMM 参考算法略有提升；而除交响乐音频外，ESN 方法所重建大部分音频的动态失真均优于参考算法.

4.4 主观偏爱测试

本文采用主观偏爱测试的方法来评价不同扩展方法的主观质量. 测试过程中邀请了 20 名年龄在 22 ~ 28 岁的被测者来选择两种被测项中较为偏爱的一种，或者选择无偏爱. 主观测试安排在静音室中，并选择了 MPEG 音频数据库中的五句作为测试数据（其中包括乡村音乐、爵士音乐、摇滚音乐、小提琴独奏、交响乐各一句）. 测试音频的顺序采用随机排列的方式. 被测者在做出判断之前可随意重复监听测试数据.

本次主观测试分为三组：ESN 方法与 GMM 方法比较、ESN 方法与原始超宽带音频比较、原始超宽带音频与 GMM 方法比较. 最终的主观测试结果如表 9 所示. 结果可以看出，本文所提 ESN 方法扩展后的音频主观质量比 GMM 方法更接近于原始超宽带音频质量. 尽管 ESN 方法重建交响乐音频的客观质量不及参考算法重建音频，但是由于交响乐音频高频能量较为

暗淡,两种扩展方法重建音频的主观质量差异并不明显.

表 9　主观偏爱测试结果

	偏爱前者	偏爱后者	无偏爱
ESN vs GMM	45%	23%	32%
ESN vs 原始音频	23%	29%	48%
原始音频 vs GMM	42%	20%	38%

4.5　计算复杂度

此外,本文分别对所提方法和参考方法每帧内需要乘法计算的次数进行了统计. 两种方法在特征提取、时频变换、频带复制以及高频成分重建等模块的计算过程完全一致,因此可只针对频谱包络估计模块进行复杂度计算. 对于 ESN 方法,其每帧需要进行 245364 次乘法运算;而 GMM 方法则需要进行 256512 次乘法运算. 由此可见,ESN 方法计算复杂度略低于参考算法.

4.6　讨论

本文所提出的 ESN 方法利用储备池中的递归结构,将音频特征空间的动态演变过程引入到高频频谱包络估计方法中,在不增加计算复杂度的前提下降低了扩展后音频的静态和动态失真. 然而,神经网络的参数训练与样本数据的分布特性直接相关. 如果输入宽带音频包含噪声和混响成分,本文所提方法扩展后音频的主客观质量也会出现一定的降低. 在未来工作中,可以考虑将音频增强和频带扩展相结合,改善 ESN 网络在含噪情况下的鲁棒性,进而提升整体算法的实用性.

5　结束语

本文提出了一种基于 ESN 的音频频带扩展方法. 该方法借助 ESN 储备池中的递归结构描述了特征空间状态的动态更新,并依据线性观测方程对高低频特征参数间的映射关系进行拟合,实现了高频频谱包络的有效估计. 主客观测试结果表明,对于多数测试数据,ESN 方法相比于 GMM 参考方法在静态和动态失真方面均获得了提升,其重建音频更接近于原始超宽带音频的听觉质量.

参考文献

[1] VARY P, MARTIN R. DigitalSpeech Transmission-Enhancement, Coding and Error Concealment[M]. UK: John Wiley & Sons Ltd, 2006.

[2] LARSEN E, AARTS R M. AudioBandwidth Extension-Application of Psychoacoustics, Signal Processing and Loudspeaker Design[M]. UK: John Wiley & Sons Ltd, 2004.

[3] AVENDAÑO C, HERMANSKY H, WAN E A. Beyond Nyquist towards to recovery of broadbandwidth speech from narrow-bandwidth speech[A]. EUROSPEECH[C]. Madrid, Spain: ISCA, 1995. 165 – 168.

[4] CHENG Y M, O SHAUGHNESSY D, MERMELSTEIN P. Statistical recovery of wideband speech from narrowband speech[J]. IEEE Transactions on Speech and Audio Processing, 1994, 2(4): 544 – 548.

[5] CARL H, HEUTE U. Bandwidth enhancement of narrowband speech signals[A]. 7th European Signal Processing Conference (EUSIPCO)[C]. Edinburgh, Scotland: EURASIP, 1994. 1178 – 1181.

[6] EPPS J, HOLMES W H. A new technique for wideband enhancement of coded narrowband speech[A]. IEEE Workshop on Speech Coding Proceedings[C]. Porvoo: IEEE, 1999. 174 – 176.

[7] SOON I Y, CHAI K Y. Bandwidth extension of narrowband speech using soft-decision vector quantization[A]. Fifth International Conference on Information, Communications and Signal Processing[C]. Bangkok: IEEE, 2005. 734 – 738.

[8] KORNAGEL U. Techniques for artificial bandwidth extension of telephone speech[J]. Signal Processing, 2006, 86(6): 1296 – 1306.

[9] PARK KY, KIM HS. Narrowband to wideband conversion of speech using GMM based transformation[A]. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)[C]. ISTANBUL: IEEE, TURKEY, 2000. 1843 – 1846.

[10] BOTINHAO CV, CARLOS BS, CALOBA LP, PETRAGLIA MR. Frequency extension of telephone narrowband speech signal using neural networks[A]. IMACS Multiconference on Computational Engineering in Systems Applications (CESA)[C]. Beijing: IEEE, 2006. 1576 – 1579.

[11] TUAN V P, SCHAEFER F, KUBIN G. A novel implementation of the spectral shaping approach for artificial bandwidth extension[A]. 3rd International Conference on Communications and Electronic[C]. Nha Trang, VIETNAM: IEEE, 2010. 262 – 267.

[12] ISER B, SCHMIDT G. Neural networks versus codebooks in an application for bandwidth extension of speech signals[A]. European Conference on Speech and Language Processing (EUROSPEECH)[C]. Geneva, Switzerland: ISCA, 2003. 565 – 568.

[13] JAX P, VARY P. Wideband extension of telephone speech using a hidden Markov model[A]. 7th IEEE Workshop on Speech Coding[C]. DELAVAN, WI: IEEE, 2000. 133 – 135.

[14] JAX P, VARY P. On artificial bandwidth extension of telephone speech[J]. Signal Processing, 2003, 83(8): 1707 – 1719.

[15] SONG G B, MARTYNOVICH P. A study of HMM-based bandwidth extension of speech signals[J]. Signal Processing, 2009, 89(10): 2036 – 2044.
[16] YAGLI C, TURAN M A T, ERZIN E. Artificial bandwidth extension of spectral envelope along a Viterbi path [J]. Speech Communication, 2013, 55(1): 111 – 118.
[17] LUKOEVIIUS M. APractical Guide to Applying Echo State Networks [M]. MONTAVON G, ORR G B, MÁLLER K R. Neural Networks: Tricks of the Trade, Heidelberg: Springer, 2012. 659 – 686.
[18] LUKOSEVICIUS M, JAEGER H. Reservoir computing approaches to recurrent neural network training[J]. Computer Science Review, 2009, 3(3): 127 – 149.
[19] JAEGER H, LUKOSEVICIUS M, POPOVICI D, SIEWERT U. Optimization and applications of echo state networks with leaky-integrator neurons [J]. Neural Networks, 2007, 20(3): 335 – 352.
[20] JAEGER H, HAAS H. Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication[J]. Science, 2004, 304(5667): 78 – 80.
[21] SCHILLER U D, STEIL JJ. Analyzing the weight dynamics of recurrent learning algorithm [J]. Neucomputing, 2005, (63): 757 – 779.
[22] PULAKKA H, LAAKSONEN L, VAINIO M, POHJALAINEN J, ALKU P. Evaluation of an artificial speech bandwidth extension method in three languages [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(6): 1124 – 1137.
[23] GRAY A H, MARKEL J D. Distance measures for speech processing[J]. IEEE Transactions on Audio, Speech, and Language Processing, 1976, 24(5): 380 – 391.
[24] NORDEN F, ERIKSSON T. Time evolution in LPC spectrum coding[J]. IEEE Transactions on Speech and Audio Processing, 2004, 12(3): 290 – 301.
[25] NILSSON M, GUSTAFSSON H, ANDERSEN SV, KLEIJN W B. Gaussian mixture model based mutual information estimation between frequency bands in speech[A]. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)[C]. Orlando, Florida: IEEE, 2002. I-525-528

作者简介

刘 鑫 男,1986 年生于北京. 北京工业大学博士研究生. 研究方向为语音与音频信号处理.

鲍长春(通信作者) 男,1965 年生于内蒙古赤峰. 北京工业大学电子信息与控制工程学院教授,博士生导师. 研究方向为语音与音频信号处理.
E-mail: chchbao@ bjut. edu. cn