第 42 卷第 3 期　　声　学　学　报　　Vol. 42, No. 3
2017 年 5 月　　ACTA　ACUSTICA　　May, 2017

考虑帧间信息的语音带宽扩展*

王迎雪[1,2]　赵胜辉[1†]　匡镜明[1]

(1　北京理工大学 信息与电子学院　北京　100081)

(2　卡内基梅隆大学 计算机与工程学院　美国匹兹堡　15213)

2015 年 10 月 19 日收到

2016 年 4 月 26 日定稿

摘要　语音带宽扩展是为了提高语音质量，利用语音低频和高频之间的相关性重构语音高频的一种技术。高斯混合模型法是语音带宽技术中被广泛应用的一种方法，但是，该方法的映射函数是分段线性函数，且没有考虑语音前后帧的相关信息。因此，提出了一种基于条件受限玻尔兹曼机的方法。该方法利用条件受限玻尔兹曼机提取了语音信号的帧间信息，同时将语音低频、高频特征参数映射为高阶统计特性，深层发掘和模拟了语音低频和高频之间的非线性关系。客观和主观对比测试结果都表明，该方法性能优于传统的高斯混合模型方法。

PACS 数：43.72, 43.60

Speech bandwidth extension supported by temporal information

WANG Yingxue[1,2]　ZHAO Shenghui[1]　KUANG Jingming[1]

(1　*School of Information and Electronic, Beijing Institute of Technology*　Beijing　100081)

(2　*School of Computer Science, Carnegie Mellon University*　Pittsburgh　15213　US)

Received Oct. 10, 2015

Revised Apr. 26, 2016

Abstract　Speech Bandwidth Extension (BWE) aims to improve the quality of speech by reconstructing the missing High Frequency (HF) components using the correlation that exists between the Low Frequency (LF) and HF of speech. The Gaussian Mixture Model (GMM) based methods are widely used. However, the derived mapping function by GMM is a piece-wise linear transformation and ignores the temporal information of speech. Thus, a novel BWE method is proposed for estimation of the HF parts of speech by exploiting Conditional Restricted Boltzmann Machines (CRBM). The proposed method introduces CRBM to obtain time information and model deep non-linear relationships between the spectral envelope features of LF and HF by building high-order eigen spaces between the LF and HF of the speech signal. The objective and subjective test results show that the proposed method outperforms the conventional GMM based method.

引言

窄带电话语音的带宽有限 (300 Hz~3.4 kHz), 因此，它的语音质量有限，自然度不够，在一些对语音质量要求高的场合，不能满足人们的需求。这些不足可以通过引入宽带语音 (50 Hz~7 kHz) 通信得以改善。然而，宽带语音通信与公共电话网络系统间还存在兼容性等问题，还没有得到实际的应用。因此，可以通过语音带宽扩展技术将窄带语音扩展为宽带语音，从而提高语音的质量。目前，该技术已被应用于多种任务，如语音识别 [1]，多播会议 [2] 等。应用最广泛的语音带宽扩展算法是源滤波器模型法。该方法通过能量调整获得高频激励信号，并将其通过由高频频谱包络转换而来的高频合成滤波器，从而重建高频信号。因此，该方法的两个主要部分是高频谱

* 瑞典爱立信课题资助

† 通讯作者：赵胜辉，shzhao@bit.edu.cn

包络估计和高频激励信号的产生, 其中当前研究的重点是高频频谱包络估计。高频频谱包络估计的方法主要包括码本映射法 [3−4]、线性映射法 [5] 及分段线性映射法 [6]、高斯混合模型 (Gaussian Mixture Model, GMM) 法 [7−8]、隐马尔科夫模型 (Hidden Markov Model, HMM) 法 [9−10] 和人工神经网络法 [11−12] 等等。其中, GMM 方法由于具有良好的带宽扩展效果, 得到了广泛的应用。近年来, 许多学者在传统的 GMM 频谱包络估计方法的基础上, 提出了许多改进方法。例如, Seo 等人 [13] 利用最大化后验概率准则和矢量泰勒级数相结合的方法, 针对被噪声污染的窄带语音进行语音带宽扩展。Liu X 等人 [14] 对比了时域平滑频谱系数和梅尔频谱系数的互信息, 并将互信息较大的时域平滑频谱系数作为 GMM 的输入数据, 以此来估计高频频谱包络。

GMM 法的映射函数是一种分段线性映射函数, 然而, 语音低频部分和高频部分的特征参数之间并不是一种简单的线性关系。因此, 采用分段线性映射函数来模拟语音低频、高频特征参数之间的关系显然是不足的。因此, Li 等人 [15] 提出采用深度神经网络模拟语音低频、高频特征参数之间的关系。他们首先采用受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 进行预训练, 然后在精细调整时采用最小均方误差 (Minimizing the Mean Square Error, MMSE) 准则使估计得到的高频特征参数和原始高频特征参数之间的欧式距离最小。我们在之前的工作中 [16] 采用 RBM 将语音低频、高频的特征参数分别转换为它们的高阶统计特性, 以此来发掘和模拟语音低频和高频的非线性关系。但是, 以上算法都没有考虑语音帧间的相关信息。事实上, 语音帧间具有很强的相关性, 这些相关信息不仅有利于改善带宽扩展效果, 也有利于提高带宽扩展后的语音质量。许多学者也发现了语音帧间信息的重要性。例如, Kim 等人 [8] 采用 GMM 方法重构语音时, 不但采用了语音的频谱信息, 而且利用了语音帧间的时域信息, 并且取得了良好的语音带宽扩展效果。Jax 等人 [9] 采用 HMM 方法吸取语音帧间的相关信息。Nour-Eldin 等人 [17−18] 阐述了语音帧间相关信息的重要性, 并且通过提炼动态特征来提取语音前后帧间的相关信息。

为了更好的挖掘和模拟低频特征参数和高频特征参数之间的线性和非线性关系, 同时也为了提取语音前后帧间的相关信息, 本文提出了一种基于条件受限玻尔兹曼机 (Conditional Restricted Boltzmann Machine, CRBM) 的高频频谱包络估计方法。该方法能够利用 CRBM 提取语音前后帧间相关信息的能力以及强烈模拟语音高、低频频谱包络之间非线性关系的能力。该方法首先利用两个 CRBM 将高、低频语音的特征参数分别映射为它们的高阶统计特征信息, 然后利用前馈神经网络 (Feedforward Neural Network, FNN) 将低频的高阶统计特征映射为高频的高阶统计特性, 接着, 将两个 CRBM 和 FNN 级联成一个深度神经网络, 并利用该深度神经网络将低频特征参数映射为高频特征参数, 最后将通过能量调整后的高频激励信号通过由高频特征参数转换而来的高频合成滤波器, 合成语音高频信号。主观测试和客观测试表明, 相比传统的 GMM 方法和其他神经网络方法, 该方法重构的语音失真更小, 语音质量更高。

2 CRBM

CRBM 由 Taylor[21] 提出, 它由受限制玻尔兹曼机 [19] 演化而来。从图 1 中可以看出, CRBM 是一个具有 3 层结构的随机神经网络。这 3 层结构中, 一个为可见层 $\boldsymbol{v}$, 一个为隐含层 $\boldsymbol{h}$, 另外一个为条件层 $\boldsymbol{u}$。可见层与隐藏层之间全连接, 条件层到可见层之间全连接, 条件层到隐含层之间全连接, 可见层内的节点之间、隐藏层内的节点之间、条件层内的节点之间相互无连接。

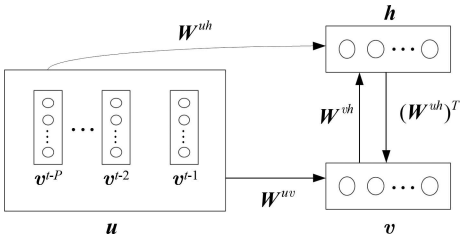


图 1　CRBM 的结构图

给定第 t 帧隐含层的随机变量 $\boldsymbol{h}^t=[h_1^t,\cdots,h_J^t]^{\mathrm{T}}$、可见层的随机变量 $\boldsymbol{v}^t=[v_1^t,\cdots,v_I^t]^{\mathrm{T}}$ 和条件层的随机变量 $\boldsymbol{u}^t=[\boldsymbol{u}^1,\boldsymbol{u}^2,\cdots,\boldsymbol{u}^P]$, $\boldsymbol{u}^p=\boldsymbol{v}^{t-p}=[v_1^{t-p},v_2^{t-p},\cdots,v_I^{t-p}]^{\mathrm{T}}$ (P 是语音帧的数目) 时, 可见层和隐含层变量的联合分布为:

$$p(\boldsymbol{v}^t,\boldsymbol{h}^t|\boldsymbol{u}^t)=\frac{1}{Z}\exp(-E(\boldsymbol{v}^t,\boldsymbol{h}^t|\boldsymbol{u}^t)), \tag{1}$$

$$Z=\sum_h\int\exp(-E(\boldsymbol{v}^t,\boldsymbol{h}^t|\boldsymbol{u}^t))\mathrm{d}\boldsymbol{v}, \tag{2}$$

$$E(\boldsymbol{v}^t,\boldsymbol{h}^t|\boldsymbol{u}^t)=\sum_{i=1}^{I}\frac{(v_i-\widetilde{b}_i)^2}{2\sigma_i^2}-\sum_{j=1}^{J}\widetilde{c}_jh_j-\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{v_i}{\sigma_i}w_{ij}^{vh}h_j, \tag{3}$$

$$\widetilde{b}_i = b_i + \sum_p (\boldsymbol{W}^{u^p v})^{\mathrm{T}} \boldsymbol{u}^p, \tag{4}$$

$$\widetilde{c}_j = c_j + \sum_p (\boldsymbol{W}^{u^p h})^{\mathrm{T}} \boldsymbol{u}^p, \tag{5}$$

其中, I, J 分别表示可见层节点的个数和隐含层节点的个数; $\boldsymbol{W}^{u^p v} \in \mathcal{R}^{I\times I}$ 为条件层节点到可见层节点的连接权重矩阵; $\boldsymbol{W}^{u^p h} \in \mathcal{R}^{I\times J}$ 为条件层节点到隐含层节点的连接权重矩阵; $\boldsymbol{W}^{vh} \in \mathcal{R}^{I\times J}$ 为可见节点和隐含节点的连接权重矩阵; $\boldsymbol{b}=[b_1,\cdots,b_I]^{\mathrm{T}}$ 和 $\boldsymbol{c}=[c_1,\cdots,c_J]^{\mathrm{T}}$ 分别是可见节点和隐含节点的偏移量; σ 是高斯可见节点的标准方差向量, 通常取值为 1。

该模型有 5 种待估参数: $\boldsymbol{W}^{u^p v}$, $\boldsymbol{W}^{u^p h}$, $\boldsymbol{W}^{vh}$, $\boldsymbol{b}$, $\boldsymbol{c}$。CRBM 模型参数的估计一般基于

$$\mathcal{L} = \log \prod_t p(\boldsymbol{v}^t, \boldsymbol{h}^t | \boldsymbol{u}^t)$$

的最大似然估计原理。Hinton[20] 于 2002 年提出的一种快速训练方法 – 对比散度算法 (Contrastive Divergence, CD) 是该模型广泛采用的参数更新算法:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{u^p v}_{i'i}} = \left\langle \frac{v_i^t u_{i'}^p}{\sigma_i^2} \right\rangle_{\text{data}} - \left\langle \frac{v_i^t u_{i'}^p}{\sigma_i^2} \right\rangle_{\text{mod } el}, \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{u^p h}_{i'j}} = \left\langle u_{i'}^p h_j^t \right\rangle_{\text{data}} - \left\langle u_{i'}^p h_j^t \right\rangle_{\text{mod } el}, \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}^{v^t h^t}_{ij}} = \left\langle \frac{v_i^t h_j^t}{\sigma_i^2} \right\rangle_{\text{data}} - \left\langle \frac{v_i^t h_j^t}{\sigma_i^2} \right\rangle_{\text{mod } el}, \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial b_i} = \left\langle \frac{v_i^t}{\sigma_i^2} \right\rangle_{\text{data}} - \left\langle \frac{v_i^t}{\sigma_i^2} \right\rangle_{\text{mod } el}, \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial c_j} = \left\langle h_j^t \right\rangle_{\text{data}} - \left\langle h_j^t \right\rangle_{\text{mod } el}, \tag{10}$$

其中, $\langle\cdot\rangle_{\text{data}}$ 表示在训练集上的期望, $\langle\cdot\rangle_{\text{mod } el}$ 表示在模型分布上的期望。为得到模型分布上的期望, CD 算法根据可见节点和隐藏节点的条件分布对其进行吉布斯采样。

获得模型参数后, 给定随机选择的隐藏层矢量 $\boldsymbol{h}^t$ 和条件层矢量 $\boldsymbol{u}^t$, 重构可见单元 v_i^t 的状态为 $\boldsymbol{v}$ 的条件概率和给定一个随机选择的训练样本 $\boldsymbol{v}^t$ 和 $\boldsymbol{u}^t$, 隐藏单元 h_j^t 的状态为 1 条件概率分别为:

$$p(v_i^t = \boldsymbol{v} | \boldsymbol{h}^t, \boldsymbol{u}^t) = \mathcal{N}\left(\boldsymbol{v}; b_i + \sigma_i \sum_j h_j w_{ij} + \sum_p (\boldsymbol{u}^p)^{\mathrm{T}} \boldsymbol{W}_{:i}^{u^p v}, \sigma_i^2\right), \tag{11}$$

$$p(h_j^t = 1 | \boldsymbol{v}^t, \boldsymbol{u}^t) = s\left(c_j + \sum_i \frac{v_i}{\sigma_i} w_{ij}^{vh} + \sum_p (\boldsymbol{u}^p)^{\mathrm{T}} \boldsymbol{W}_{:j}^{u^p h}\right) \tag{12}$$

其中, s 表示 sigmoid 函数, $s(x) = 1./(1+e^{-x})$; $N(\mu;\sigma^2)$ 表示均值和方差分别为 μ 和 σ^2 的高斯分布。

3 基于 CRBM 的高频频谱包络估计

语音帧间的相关信息具有重要的作用, 为了利用语音帧间的相关性, 同时也为了发掘低频特征参数和高频特征参数的非线性关系, 本文提出了基于 CRBM 的语音带宽扩展方法。

图 2(b) 给出了基于 CRBM 的高频频谱包络估计的网络结构图, 同时, 为了和之前的工作[16]进行对比, 图 2(a) 给出了基于 RBM 的高频频谱包络估计的网络结构图。从图 2 中可以看出, RBM 方法只采用了当前帧的低频特征参数 $\boldsymbol{x}^t = [x_1^t, x_2^t, \cdots, x_M^t]^{\mathrm{T}}$ (M 为维数) 来获取当前帧的高频特征参数 $\boldsymbol{y}^t = [y_1^t, y_2^t, \cdots, y_N^t]^{\mathrm{T}}$ (N 为维数)。CRBM 方法不但采用了当前帧的低频特征参数 $\boldsymbol{x}^t$, 同时也采用了前一帧的低频特征参数 $\boldsymbol{x}^{t-1} = [x_1^{t-1}, x_2^{t-1}, \cdots, x_M^{t-1}]^{\mathrm{T}}$ 和前一帧的高频特征参数 $\boldsymbol{y}^{t-1} = [y_1^{t-1}, y_2^{t-1}, \cdots, y_N^{t-1}]^{\mathrm{T}}$ 来获取当前帧的高频特征参数 $\boldsymbol{y}^t$。这表明相比 RBM 方法, CRBM 方法能够利用语音帧间的相关信息。

从图 2(b) 可以看出, CRBM 方法首先采用低频特征参数 $\boldsymbol{x}^t$, $\boldsymbol{x}^{t-1}$ 获取低频 CRBM, 采用高频特征参数 $\boldsymbol{y}^t$, $\boldsymbol{y}^{t-1}$ 获取高频 CRBM。然后, 采用 FNN 模拟低频高阶统计特性 $\boldsymbol{h}_x^t$ 和高频高阶统计特性 $\boldsymbol{h}_y^t$ 的分布, 其中 $\boldsymbol{h}_x^t$, $\boldsymbol{h}_y^t$ 分别从低频 CRBM 和高频 CRBM 中获取。因此, 该方法的参数集为:

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta}_x, \boldsymbol{\theta}_y, \boldsymbol{\theta}_n\}, \tag{13}$$

其中, $\boldsymbol{\theta}_x = \{\boldsymbol{W}^{x^{t-1}x^t}, \boldsymbol{W}^{x^{t-1}h^t}, \boldsymbol{W}^{x^t h^t}, \boldsymbol{b}_x, \boldsymbol{c}_x\}$ 是低频 CRBM 的参数集, $\boldsymbol{\theta}_y = \{\boldsymbol{W}^{y^{t-1}y^t}, \boldsymbol{W}^{y^{t-1}h^t}, \boldsymbol{W}^{y^t h^t}, \boldsymbol{b}_y, \boldsymbol{c}_y\}$ 是高频 CRBM 的参数集, $\boldsymbol{\theta}_n = \{\boldsymbol{W}^1, \cdots, \boldsymbol{W}^L, \boldsymbol{d}^1, \cdots, \boldsymbol{d}^L\}$ (其中, $L-1 = \{0,1,2\}$ 是隐含层的层数, $\boldsymbol{d}^l$ 是第 l^{th} 层的偏移量, $\boldsymbol{W}^l$ 是 $(l-1)^{th}$ 层到 l^{th} 层的权重矩阵) 是 FNN 的参数集。$\boldsymbol{W}^{x^{t-1}x^t}, \boldsymbol{W}^{x^{t-1}h^t}$, $\boldsymbol{W}^{y^{t-1}y^t}, \boldsymbol{W}^{y^{t-1}h^t}$ 可以获取语音帧间的相关信息。

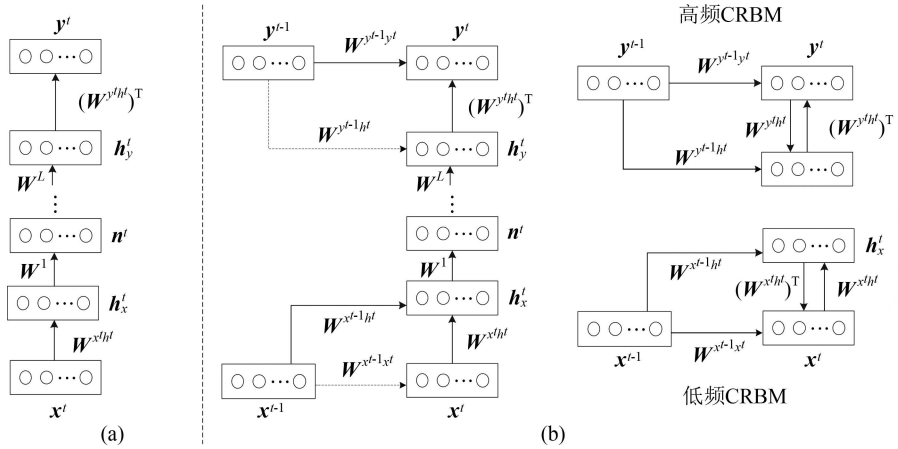


图 2　高频频谱包络估计的网络结构图

在 FNN 的训练阶段，输入向量和目标向量分别为：

$$\boldsymbol{h}_x^t = s\left(\boldsymbol{c}_x + \boldsymbol{W}^{x^t h^t}\boldsymbol{x}^t + \boldsymbol{W}^{x^{t-1}h^t}\boldsymbol{x}^{t-1}\right), \quad (14)$$

$$\boldsymbol{h}_y^t = s\left(\boldsymbol{c}_y + \boldsymbol{W}^{y^t h^t}\boldsymbol{y}^t + \boldsymbol{W}^{y^{t-1}h^t}\boldsymbol{y}^{t-1}\right), \quad (15)$$

估计 FNN 的参数 $\boldsymbol{\theta}_n$ 时，可以通过最小化 FNN 的输出向量 $\widetilde{\boldsymbol{h}}_y^t$ 和目标向量 $\boldsymbol{h}_y^t$ 的误差获得。获取 FNN 的参数后，任意一个输入向量 $\widetilde{\boldsymbol{h}}_x^t$ 转换为：

$$\begin{cases} \widetilde{\boldsymbol{h}}_y^t = \boldsymbol{o}^l = s(\boldsymbol{z}^l), & l \geqslant 1, \\ \boldsymbol{z}^l = \boldsymbol{W}^l\boldsymbol{o}^{l-1} + \boldsymbol{d}^l, & l \geqslant 1, \\ \boldsymbol{o}^0 = \widetilde{\boldsymbol{h}}_x^t, \end{cases} \quad (16)$$

为了将 FNN 的输出向量 $\widetilde{\boldsymbol{h}}_y^t$ 映射为高频特征参数，由式 (11) 得：

$$p\left(\boldsymbol{y}^t|\widetilde{\boldsymbol{h}}_y^t, \boldsymbol{y}^{t-1}\right) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{b}_y + \left(\boldsymbol{W}^{y^t h^t}\right)^{\mathrm{T}}\widetilde{\boldsymbol{h}}_y^t + \boldsymbol{W}^{y^{t-1}y^t}\boldsymbol{y}^{t-1}, \sigma_i^2\right). \quad (17)$$

结合式 (16) 和式 (17)，高带特征参数矢量的估计值为：

$$\boldsymbol{y}^t = s\left(\boldsymbol{b}_y + \left(\boldsymbol{W}^{y^t h^t}\right)^{\mathrm{T}} s(\boldsymbol{z}^l) + \boldsymbol{W}^{y^{t-1}y^t}\boldsymbol{y}^{t-1}\right), \ l \geqslant 1. \quad (18)$$

从式 (14)、式 (16) 和式 (18) 可以看出，当前帧的高频特征参数是低频特征参数的非线性函数，并且，当前帧的高频特征参数由当前帧的低频特征参数、前一帧的低频特征参数、前一帧的高频特征参数共同决定。这表明，该文提出的 CRBM 方法不但发掘和模拟了语音低频特征参数和高频特征参数之间的非线性关系，而且还利用了语音帧间的相关信息。

4　实验与比较

4.1　语音数据库及实验设置

本文采用的数据库包括 NTT-AT 汉语语音数据库 [22] 和北京理工通信技术研究所和爱立信 (Research Center of Digital Communication Technology, RCDCT) 共同录制的汉语语音测试库。NTT-AT 汉语语音数据库和 RCDCT 的汉语语音测试库都包含 96 句男性和女性 (4 名女和 4 名男性) 的发声，每条语句的采样率为 16 kHz, 数据格式为 16 位的 PCM, 每条语句的持续时间为 8 s。本文的训练集共包含从 NTT-AT 汉语语音数据库随机选择的 64 条语句和从 RCDCT 语料库随机选择的 64 条语句。由于训练样本集合与测试样本集不能重复，所以本文的测试集包含 2 个汉语语料库中没有用于训练的所有语音样本。

特征参数提取时，采样率为 16 kHz 的宽带语音首先分别通过高低通滤波器，然后下采样，得到低频语音 (0∼4 kHz) 和高频语音 (4∼8 kHz)。然后对高低频语音分别进行加汉明窗处理，帧长 20 ms, 帧移 10 ms, 预加重的系数为 0.97。高低频声学特征参数分别使用 10 维的 LSF 和 16 维的 LSF。LSF 作为低频 CRBM 和高频 CRBM 的输入数据，在送入 CRBM 训练前，需在整个训练集范围内对其进行归一化，使得每一维特征参数都满足均值为 0、方差为 1 的正态分布。

高频激励信号的产生和能量增益调整也是重构高频语音时比较重要的问题。由于 AMR-WB+ 直接采用低频激励信号作为高频激励信号 [23], 并取得了很好的带宽扩展效果，因此，本文采用低频激励信号直接作为高频激励信号。采用文献 24 中的码本映射法对低频激励信号进行能量增益调整。

表 1 不同的网络结构

不同的网络结构	RBM	CRBM	总层数
Arc.1	16-64-64-10	(16,16)-64-64-(10,10)	4 层
Arc.2	16-64-128-64-10	(16,16)-64-128-64-(10,10)	5 层
Arc.3	16-64-128-128-64-10	(16,16)-64-128-128-64-(10,10)	6 层

为了验证 CRBM 方法的性能, 将 CRBM 方法与传统 GMM 方法[25]、RBM 方法[16] 进行对比。GMM 的低频动态特征参数的维数为 16, 高频动态特征参数的维数为 10, 高斯分量个数取为 128, 模型参数估计采用期望最大算法。CRBM 模型参数估计采用小批量 (minibatch) 的对比散度算法, 每个批量的规模为 64 个训练样本。对比散度算法在梯度下降的过程中采用了十次的吉布斯采样完成对权重的更新, 参数更新的迭代步长设为 0.0001 。对于低频 CRBM 和高频 CRBM, 学习率为 0.0001, 学习轮次 (epoch) 为 500; 对于 FNN, 学习率为 0.0001, 学习轮次为 300 。冲量值 (momentum) 在最初 5 轮设为 0.5, 然后增加至 0.9; 权值衰减因子为 0.0002 。低频 CRBM 和高频 CRBM 的隐含层的节点数设为 64 。RBM 模型参数的设置与 CRBM 模型参数的设置一致。该文对 3 种不同的网络结构 (Arc.1, Arc.2, Arc.3) 进行测试, 这 3 种网络结构具有不同的层数及节点数, 见表 1。

为了评价所提方法的性能, 分别进行了主观评价、客观评价及语谱图比较, 并且给出了不同语音带宽扩展方法的主观测试结果、客观测试结果以及语谱图。主观评价采用平均意见分 (Mean Opinion Score, MOS) 测试。客观评价采用 COSH 距离测度和均方根对数谱距离 (Root mean square log spectral distortion, RMS-LSD) 测度。

4.2 客观评价

COSH 距离测度的定义[25] 为:

$$d_{\mathrm{COSH}}=\sqrt{\frac{1}{2N}\sum_{n=1}^{N}\left[d_{\mathrm{IS}}(A_n(w),\widehat{A}_n(w))+d_{\mathrm{IS}}(\widehat{A}_n(w),A_n(w))\right]}, \tag{19}$$

其中, N 为语音帧的个数; d_{IS} 为板仓距离。

$$d_{\mathrm{IS}}(A_n(w),\widehat{A}_n(w))=\frac{1}{w_2-w_1}\int_{w_1}^{w_2}\left[\frac{g^2\Big/|A_n(w)|^2}{\widehat{g}^2\Big/\left|\widehat{A}_n(w)\right|^2}-\lg\frac{g^2\Big/|A_n(w)|^2}{\widehat{g}^2\Big/\left|\widehat{A}_n(w)\right|^2}-1\right]\mathrm{d}w, \tag{20}$$

其中, $A_n(w)$, $\widehat{A}_n(w)$ 分别表示原始高频和合成高频经过离散傅里叶变换后的谱包络; w_1, w_2 分别表示待测频段的频率下限和频率上限; g, $\widehat{g}$ 分别表示原始高频和合成高频的增益。

均方根对数谱距离 (Root Mean Squared Log Spectral Distance, RMS-LSD):

$$d_{\mathrm{LSD}}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}\int_{w_1}^{w_2}\left[20\lg G_c\frac{A_n(w)}{\widehat{A}_n(w)}\right]^2\mathrm{d}w}, \tag{21}$$

其中, G_c 为增益补偿因子,

$$G_c=\frac{1}{w_2-w_1}\int_{w_1}^{w_2}20\lg\frac{A_n(w)}{\widehat{A}_n(w)}\mathrm{d}w,$$

均方根对数谱距离越小说明系统的性能越好。

本实验对语音高频段 ($w_1=0.25w_s$, $w_2=0.5w_s$, $w_s=16$ kHz) 进行客观测试, w_s 为输入语音的采样频率。表 2 分别给出了不同隐含层数时, 不同方法的 COSH 距离值和均方根对数谱距离值 (所有测试语料的平均值)

表 2 客观测试结果

方法		均方根对数谱距离值 (dB)	COSH 距离值
GMM		4.648	6.93
RBM	Arc.1	4.438	6.76
	Arc.2	4.234	6.42
	Arc.3	4.195	6.27
CRBM	Arc.1	4.279	6.49
	Arc.2	4.053	6.15
	Arc.3	3.881	5.95

从表 2 可以看出，在相同的网络结构下，相比传统的 GMM 方法以及 RBM 方法，提出的语音带宽扩展方法的均方根对数谱距离值和 COSH 距离值最小。在 Arc.3 的网络结构下，CRBM 的均方根对数谱距离值和 COSH 距离值最小。这表明，就客观评价而言，所提算法的性能优于传统的 GMM 方法 RBM 方法，它能够提高重构语音的质量，且在 Arc.3 的网络结构下，重构语音的质量最高。

4.3 主观评价

MOS 即平均意见分，是最常用的主观评价方法之一。该测试是根据听音人对被测系统输出的合成语音主观印象评测，主要是从语音的可懂度、清晰度、自然度等指标去衡量语音质量。该方法将重构语音分成 5 个等级，1 分表示重构后的语音质量最差，而 5 分则代表语音质量最好，可懂度最高。该测试中，一共有 20 位测听人对测试集中的所有语料和用不同方法重构的语音信号进行测听，并给出 MOS 分。表 3 给出了传统的 GMM, RBM, CRBM 合成语音的 MOS 分。

从表 3 可以看出，网络结构相同时，CRBM 获得的平均意见分高于 GMM 方法和 RBM 方法。而且，CRBM 在 Arc.3 时获得了最高的 MOS 分。这表明，就主观测试而言，相比 GMM 方法和 RBM 方法，所提算法性能更加优越，重构后的语音质量更高。

4.3 语谱图比较

图 3 给出了原始宽带语音信号和经不同带宽扩展方法扩展后的语音语谱图。从图 3 可以看出，相比传统的 GMM 方法和 RBM 方法，CRBM 扩展后的语音语谱图更加接近原始宽带语音信号的语谱图。

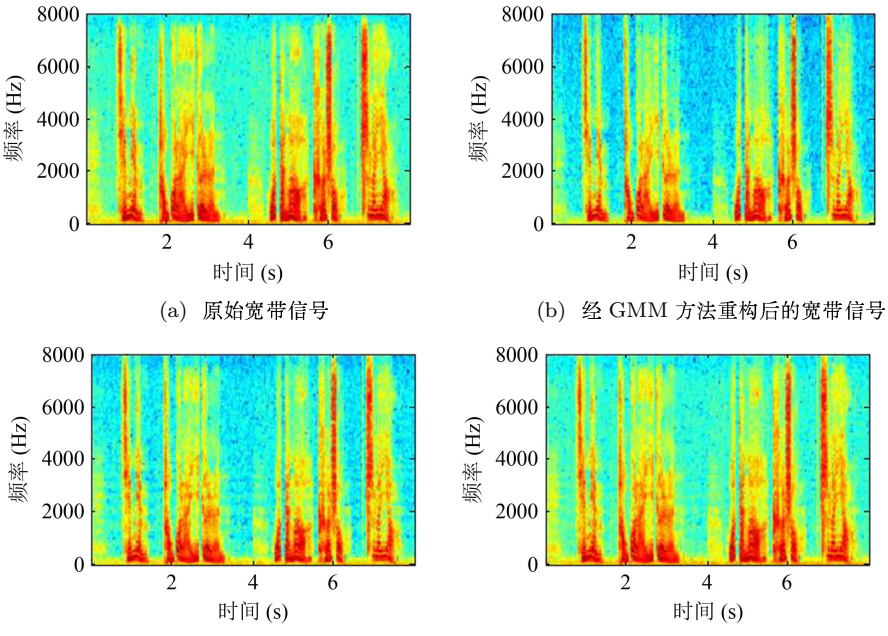
主观测试、客观测试以及语谱图比较结果表明，所提算法的性能优于传统的 GMM 方法和 RBM 方法。这是由于所提算法采用 CRBM 提取了语音信号帧间的相关信息，同时通过 CRBM 将语音低频、高频特征参数映射为它们的高阶统计特性，深层挖掘了低频语音和高频语音之间的非线性关系。

表 3　主观测试结果

方法		平均意见分 (MOS)
GMM		3.348
RBM	Arc.1	3.438
	Arc.2	3.534
	Arc.3	3.695
CRBM	Arc.1	3.579
	Arc.2	3.753
	Arc.3	3.781

5 结束语

本文提出了条件受限制玻尔兹曼机语音带宽扩展方法。该方法采用条件受限制玻尔兹曼机提取语音帧间的相关信息和发掘语音低频和高频特征参数之间的非线性关系。主观测试、客观测试结果都表明，相较于传统的高斯混合模型方法和受限制玻尔兹曼机方法，所提方法的性能更加优越，可以提高重构语音的质量。



(a) 原始宽带信号　(b) 经 GMM 方法重构后的宽带信号

(c) 经 RBM 方法重构后的宽带信号　(d) 经 CEBM 方法重构后的宽带信号

图 3　语谱图比较

致谢

感谢北京理工大学–爱立信国际合作项目、国家留学基金委以及卡内基梅隆大学对本论文的支持。

参 考 文 献

1 Bauer P, Abel J, Fischer V *et al*. Automatic recognition of wideband telephone speech with limited amount of matched training data. Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), 2013: 1232—1236

2 Gandhimathi G, Jayakumar S. Speech enhancement using an artificial bandwidth extension algorithm in multicast conferencing through cloud services. *Information Technology Journal*, 2014; **13**(12): 1953—1956

3 Yoshida Y, Abe M. An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping. Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, 1994: 1591—1594

4 Wang Y X, Zhao S H *et al*. Superwideband extension for AMR-WB using conditional codebooks. Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Florence, Italy, 2014: 3695—3698

5 Nakatoh Y, Tsushima M, Norimatsu T. Generation of broadband speech from narrowband speech using on linear mapping. *Electronics and Communications in Japan, Part 2* (*Electronics*), 2002; **85**(8): 44—53

6 Duy N D, Suzuki M, Minematsu N *et al*. Artificial bandwidth extension based on regularized piecewise linear mapping with discriminative region weighting and long-Span features. Interspeech, 2013: 3453—3457

7 Pulakka H, Remes U, Palomäki K *et al*. Speech bandwidth extension using gaussian mixture model-based estimation of the highband Mel spectrum. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011: 5100—5103

8 Kim K T, Lee M K, Kang H G. Speech bandwidth extension using temporal envelope modeling. *Signal Processing Letters, IEEE*, 2008; **15**: 429—432

9 Jax P, Vary P. Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markov model. Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Hong Kong, Hong kong, 2003: 680—683

10 Bauer P, Abel J, Fingscheidt T. HMM-based artificial bandwidth extension supported by neural networks. 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC), 2014: 1—5

11 Liu H J, Bao C C, Liu X. Spectral envelope estimation used for audio bandwidth extension based on RBF neural network. Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Vancouver, Canada, 2013: 543—547

12 Pourmohammadi S, Vali M, Ghadyani M. Bandwidth extension of narrowband speech in log spectra domain using neural network. *Turkish Journal of Electrical Engineering and Computer Science*, 2015; **23**(2): 433—446

13 Seo H, Kang H G, Soong F. A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2014: 6087—6091

14 Liu X, Bao C C. Audio bandwidth extension based on temporal smoothing cepstral coefficients. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014; **2014**(1): 1—16

15 Li K, Lee C H. A deep neural network approach to speech bandwidth expansion. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015: 4395—4399

16 Wang Y, Zhao S, Liu W *et al*. Speech bandwidth expansion based on deep neural networks. Sixteenth Annual Conference of the International Speech Communication Association, 2015: 2593—2597

17 Nour-Eldin A H, Kabal P. Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech. Interspeech, 2007: 2489—2492

18 Nour-Eldin A H, Kabal P. Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech. Interspeech, 2011: 1185—1188

19 Ackley D H, Hinton G E, Sejnowski T J. A learning algorithm for boltzmann machines. *Cognitive Science*, 1985; **9**(1): 147—169

20 Hinton G E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 2002; **14**(8): 1771—1800

21 Taylor G W, Hinton G E, Roweis S T. Modeling human motion using binary latent variables. In: Advances in Neural Information Processing Systems, 2006: 1345—1352

22 NTT Advanced Technology Corporation. Multi-lingual speech database for telephonometry. Online at http://www.nttat.com/productse/speech, 1994

23 Mäkinen J, Bessette B, Bruhn S *et al*. AMR-WB+: a new audio coding standard for 3rd generation mobile audio services. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005: 1109–1112

24 张勇, 胡瑞敏. 基于高斯混合模型的语音带宽扩展算法的研究. 声学学报, 2009; **34**(5): 471—480

25 Nour-Eldin Arm H, Kabal Peter. Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech. Interspeech, Brisbane, Australia, 2008: 53—56