ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 2,Mar. 2015,pp. 307－318
DOI:10.16337/j.1004-9037.2015.02.008

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

一种基于重构性深度网络的 MELP 语音编码改进算法

张雄伟　吴海佳　张梁梁　邹　霞

(解放军理工大学指挥信息系统学院,南京,210007)

摘　要:为了提高深度模型的编码重构性能,本文为传统对比散度(Contrastive divergence,CD)添加了基于交叉熵的重构误差约束。利用改进后的算法训练了重构性深度自编码机(Reconstructive deep auto-encoder,RDAE),并用 RDAE 替换混合激励线性预测编码(Mixed excitation linear prediction,MELP)语音编码器中 LSF 参数的矢量量化方法。测试结果表明,改进后的算法在损失一定模型似然度的条件下获得了重构性能的提升,当 RDAE 隐藏层结点设为 19 bit 时,本文方法所测得的加权 LSF 距离、重构语音质量、谱失真指标在训练集和测试集上均优于 25 bit 矢量量化方法,即利用本文方法改进的 MELP 编码器,在不降低语音质量的条件下,可将 MELP 编码速率从 2.4 kb/s 降低至 2.1 kb/s,编码速率降低了 12.5%。

关键词:深度学习;深度自编码机;重构性;低速率语音编码;混合激励线性预测

中图分类号:TP181　　**文献标志码**:A

Improved MELP Algorithm Based on Reconstructive Deep Neural Network

Zhang Xiongwei, Wu Haijia, Zhang Liangliang, Zou Xia

(College of Command Information Systems, PLA University of Science and Technology, Nanjing, 210007, China)

Abstract: In order to improve the reconstruction performance of deep models, reconstruction error constraint based on cross entropy is added to traditional contrastive divergence (CD) algorithm. The improved algorithm is used to train reconstructive deep auto-encoder(RDAE), which is used to replace the vector quantization method for LSF in MELP speech coding algorithm. Experimental results show that the improved CD algorithm improves the deep model gain reconstruction performance while costing some likelihood of the model. When the node number of the hidden layer of RDAE is set to 19 bit, the indicators, which include the weighted LSF distance, the performance of reconstructed speech, and the spectrum distortion, perform better in both training set and testing set by the proposed method than by the vector quantization method at 25 bit. That is to say, the coding bitrate of the MELP coder is reduced from 2.5 kb/s to 2.1 kb/s. The reduction rate of the coding bitrate is up to 12.5%, while the speech quality remains.

Key words: deep learning; deep auto-encoder; reconstructive; low bitrate speech coding; mixed excitation linear prediction

---

基金项目:国家自然科学基金(61471394)资助项目;江苏省自然科学基金(BK2012510)资助项目;国家自然科学青年基金(61402519)资助项目;江苏省自然科学青年基金(BK20140071,BK20140074)资助项目。
收稿日期:2014-12-27;**修订日期**:2015-01-30

引 言

语音编码技术是语音通信系统的核心技术,其目标是在控制算法复杂度和通信延迟的前提下,以尽可能低的数码率来传送质量尽可能高的语音[1]。目前,语音编码技术主要可分为3类:波形编码、参数编码以及混合编码。其中,波形编码的优点是语音质量好,缺点是编码速率过高;参数编码基于语音产生模型,提取语音的特征参数,可实现低速率语音编码,但缺点是语音质量较差;混合编码保持了参数编码低速率的优势,同时能保持较好的语音质量。混合编码目前已被广泛应用于移动数字通信领域,其中混合激励线性预测编码(Mixed excitation linear prediction, MELP)[2]就是一种经典的混合编码方案。

MELP以线性预测编码(Linear predictive coding, LPC)参数模型为主框架,采用混合激励的形式,并结合多带的思想,是一种理想的低速率语音编码方案。MELP在对语音进行编码时,首先从语音帧中提取参数,并对这些参数进行量化编码。以线谱频率(Line spectral frequency, LSF)参数为例,编码器从每个语音帧中提取10个LSF参数,并对这10个LSF参数使用四级矢量量化(Vector quantization, VQ),各级的量化位数分别为7,6,6,6。在MELP参数码字分配中,LSF参数共分配了25 bit,占编码总位数将近一半。进一步提高LSF参数的量化压缩比率对降低语音编码速率将起到重要作用。

VQ是一种高效的数据编码压缩方法,而码书设计则是VQ的关键环节。从数学观点看,VQ中的码书设计本质是以系统的失真函数最小作为目标,即寻求将所有训练矢量划分为 N 类的最佳方案,各类的质心则作为码书的码字。然而这是一个非凸优化问题,若通过穷举码书的全部可能来寻找全局最优码书,在现有计算能力下几乎不可能实现。MELP中VQ码书训练算法采用LBG算法[3],其所获得的码书为局部最优或接近全局最优,存在初始码书的选择影响码书训练的收敛速度和最终码书的性能的缺点。

对于非凸优化问题,目前机器学习领域提出了一类新的方法,即深度学习。深度学习首先通过对深度模型进行逐层无监督预训练,获得一个符合训练样本集概率分布的生成模型,然后再通过对深度模型进行有监督微调,可使深度模型具有较好的分类或者回归性能。据文献报道,利用深度信念网(Deep belief network, DBN)构建的语音识别模型,在TIMIT[4]语音库上获得了高达78%的识别正确率[5];利用深度自编码机(Deep auto-encoder, DAE)直接对语音幅度谱进行编码,与VQ方法相比,新方法获得明显的编码性能提升[6]。

深度模型的预训练过程采用对比散度算法(Contrastive divergence, CD)[7],该算法的训练目标是最大化模型边缘概率分布与训练样本集概率分布的似然度,并没有对模型的重构误差添加额外约束。在这样的预训练目标下,模型可以收敛到似然度全局最优值附近,而并非重构性能全局最优值附近。本文尝试对CD算法添加重构误差约束,并用于训练重构性深度自编码机(Reconstructive DAE, RDAE),利用RDAE替换MELP中LSF量化方法,从而在不降低语音质量的条件下,进一步降低编码码率。

1 从DBN到RDAE

如图1所示,RDAE由DBN镜像相连而成。其中,DBN由多层受限玻尔兹曼机(Restricted Boltzmann machine, RBM)[8]堆叠而成。在预训练DBN阶段,采用重构误差约束的CD算法对各层RBM进行逐层无监督训练。利用添加重构误差约束的CD算法训练出的每一层RBM对输入数据重构后的期望误差更小。预训练完毕后,将DBN与其自身镜像对接,从而构成RDAE。对RDAE采用BP算法进行调优训练,BP算法为有监督训练,RDAE的调优训练目标是获得重构输入信号的能力,因此使用输入样本自身作为监督信号。

构成DBN的基本模块是RBM,预训练是针对RBM进行逐层训练。如图2所示为一个典型的

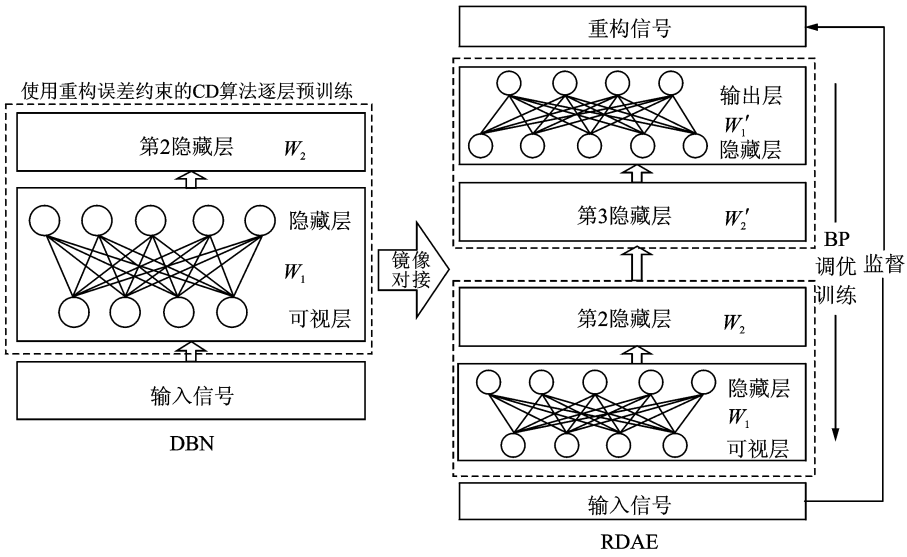


图 1 RDAE 构建过程

Fig. 1 Build process of RDAE

RBM,其由 m 个可视节点和 n 个隐藏结点组成,其中可视节点之间相互独立,仅与隐藏结点相关,同样隐藏结点之间也相互独立,仅与可视节点相关。RBM 的结点为 0/1 随机数。

RBM 中有 3 个参数,$\boldsymbol{W}_{n\times m}$是可视层与隐藏层之间的权重矩阵,$\boldsymbol{b}=(b_1,b_2,\cdots,b_m)$是可视节点的偏移量,$\boldsymbol{c}=(c_1,c_2,\cdots,c_n)$是隐藏结点的偏移量。这 3 个参数决定了 RBM 如何从一个 m 维的样本中提取 n 维的特征。下面阐述利用 RBM 提取样本特征的过程。

对于样本 $\boldsymbol{x}=(x_1,x_2,\cdots,x_m)$,通过 RBM 可提取其特征 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)$。特征提取的规则如下:

(1)利用公式 $p(h_i=1\mid v)=\sigma(\sum_{j=1}^{m}w_{ij}\times v_j+c_i)$ 计算隐藏层第 i 个结点取值为 1 的概率,其中 $v_j=x_j$,$\sigma(x)=1/(1+\mathrm{e}^{-x})$,称为 sigmoid 函数。

(2)产生一个 0～1 之间的随机数,若该随机数小于 $p(h_i=1\mid v)$,则 y_i 的取值就是 1,否则就是 0。反之,若已知样本的特征为 y,通过特征也可重构样本。重构过程如下所示:

①利用公式 $p(v_j=1\mid h)=\sigma(\sum_{i=1}^{n}w_{ij}\times h_i+b_j)$ 计算可视层第 j 个结点取值为 1 的概率,其中 $h_i=y_i$;

②产生一个 0～1 之间的随机数,若该随机数小于 $p(v_j=1\mid h)$,则 x_j 的取值就是 1,否则为 0。

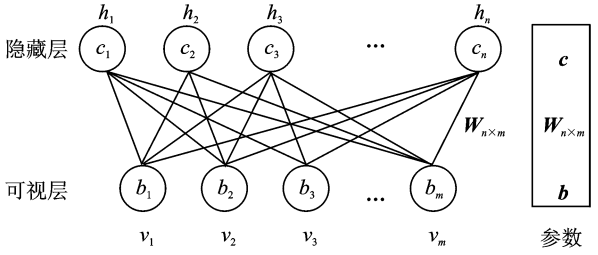


图 2 受限玻尔兹曼机

Fig. 2 Restricted Boltzmann machine

2 重构误差约束的 CD 算法

训练 RBM 的原始目标函数为

$$\underset{\{\boldsymbol{W},\boldsymbol{b},\boldsymbol{c}\}}{\arg\max}\sum_{k=1}^{K}\log p(v^{(k)}) \tag{1}$$

式中：K 为训练样本集容量；$p(v)$为可视层边缘概率。

在原始 CD 算法中，迭代增量可简写为如下形式

$$\begin{aligned}\Delta w_{ij} &= \eta(\langle v_i^+ h_j^+\rangle - \langle v_i^- h_j^-\rangle)\\ \Delta b_j &= \eta(\langle h_j^+\rangle - \langle h_j^-\rangle)\\ \Delta c_i &= \eta(\langle v_i^+\rangle - \langle v_i^-\rangle)\end{aligned} \tag{2}$$

式中：η 为学习速率；v_i^+ 表示训练样本第 i 维分量；h_j^+ 表示训练样本对应的隐藏层状态第 j 维分量；v_i^- 和 h_j^- 分别表示重构的可视层状态分量和其对应的隐藏层状态分量；$\langle\cdot\rangle$表示求取在训练样本集上的平均值。

原始 CD 的训练过程可通过图 3 形象地描述。其中 K 表示样本容量，I 表示可视层结点总数，J 表示隐藏层结点总数，从 $\boldsymbol{V}^+$ 到 $\boldsymbol{H}^+$ 的过程称为上采样，从 $\boldsymbol{H}^+$ 到 $\boldsymbol{V}^-$ 的过程称为下采样。

为了在 CD 训练过程中添加重构误差约束，可通过向目标函数添加正则化项的方法实现。为了精确控制正则化项，将使用一个指派矩阵 $\boldsymbol{Z}\in\mathbf{R}^{I\times K}$，该矩阵中每一个元素 $z_i^{(k)}\in[0,1]$，表示第 k 个训练样本下采样后激活可视层第 j 个结点的概率。指派矩阵的意义在于给下采样后的可视层提供一套行列模板，行域对应于可视层某结点针对不同信号的下采样激活表现，列域对应于可视层针对某信号的下采样激活表现。

该指派矩阵对下采样后可视层的约束可通过交叉熵来体现。文献[9]给出利用交叉熵约束隐藏层的方法，其中交叉熵的定义如下

$$z_j^{(k)}\log h_j^{(k)+} + (1-z_j^{(k)})\log(1-h_j^{(k)+}) \tag{3}$$

类似地，针对可视层的交叉熵可定义如下

$$z_i^{(k)}\log v_i^{(k)-} + (1-z_i^{(k)})\log(1-v_i^{(k)-}) \tag{4}$$

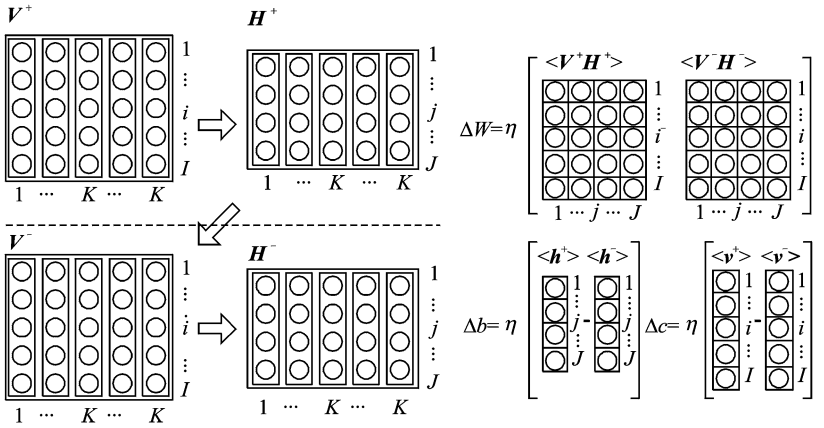


图 3 原始 CD 算法训练过程示意图

Fig. 3 Training process of original CD algorithm

对式(4)进行合并,可得到

$$\log\left[(v_i^{(k)-})^{z_i^{(k)}}(1-v_i^{(k)-})^{1-z_i^{(k)}}\right] \tag{5}$$

从合并后的形式可看出,当可视层第 i 个结点对第 k 个输入样本的一次抽样重构概率 $v_i^{(k)-}$ 与指派矩阵 $\boldsymbol{Z}$ 所指派的 $z_i^{(k)}$ 共振时,交叉熵最大(所谓共振,可理解为矢量方向一致,或解释为两者同时取大值或者同时取小值)。因此,利用该交叉熵作为 CD 训练目标函数的正则化项,可以控制可视层重构信息的表现,达到为可视层重构信息添加约束的目的。带有该交叉熵正则化项的 CD 训练目标函数如下

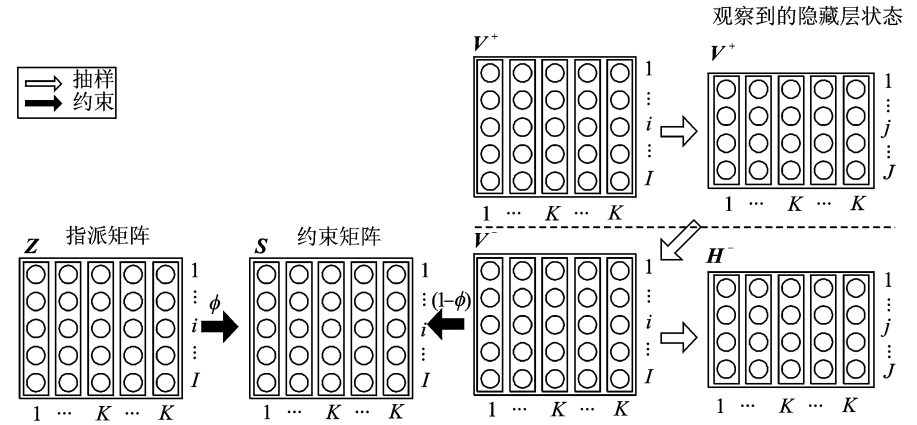
$$\underset{\{\boldsymbol{W},\boldsymbol{b},\boldsymbol{c}\}}{\arg\max}\left\{\sum_{k=1}^{K}\left[\log p(v^{(k)})+\lambda\sum_{j=1}^{J}(z_i^{(k)}\log v_i^{(k)-}+(1-z_i^{(k)})\log(1-v_i^{(k)-}))\right]\right\} \tag{6}$$

式中:λ 为正则项系数。

基于式(6)所得到的迭代如下

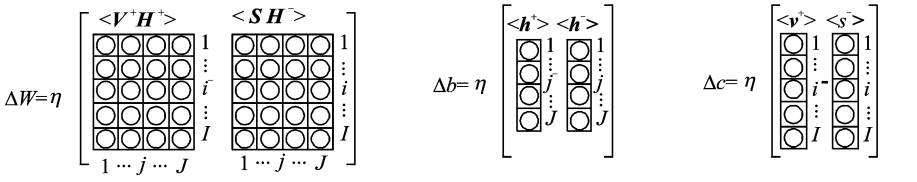
$$\begin{aligned}\Delta w_{ij}&=\eta(\langle v_i^+h_j^+\rangle-\langle v_i^-h_j^-\rangle)+\varepsilon\langle(v_i^--z_i)h_j^-\rangle=\\&\quad\eta(\langle v_i^+h_j^+\rangle-\langle s_ih_j^-\rangle)\\\Delta b_j&=\eta(\langle h_j^+\rangle-\langle h_j^-\rangle)\\\Delta c_i&=\eta(\langle v_i^+\rangle-\langle v_i^-\rangle)+\varepsilon\langle v_i^--z_i\rangle=\\&\quad\eta(\langle v_i^+\rangle-\langle s_i\rangle)\end{aligned} \tag{7}$$

式中:对于第 k 个训练样本第 i 维,$s_i^{(k)}=\varphi z_i^{(k)}+(1-\varphi)v_i^{(k)-}$,在图 4 中,$\boldsymbol{s}$ 为约束矩阵,其各列由 $s_i^{(k)}$ 组成;φ 为指派度,$\varphi=\varepsilon/\eta$。该迭代过程如图 4 所示。



(a) 抽样和生成约束矩阵的过程

(a) Sampling and constraint matrix generating process



(b) 参数更新过程

(b) Parameters updating process

图 4 对可视层添加约束的 CD 训练过程示意图

Fig. 4 CD algorithm with constraint to visible layer

通过设计指派矩阵 $\boldsymbol{Z}$ 可按需调整对可视层下采样重构后的约束，即使 $\boldsymbol{V}^+ - \boldsymbol{V}^-$ 尽可能小。欲实现该约束，只需令指派矩阵 $\boldsymbol{Z}=\boldsymbol{V}^+$。实际上，只有当 $\boldsymbol{V}^- = \boldsymbol{V}^+$ 时，式(5)所表示的交叉熵取最大值，从而达到了使用训练样本引导重构值的目的。

3 基于 RDAE 的改进 MELP 算法

MELP 分为编码器和解码器两部分。编码器首先从语音帧中提取参数，如表 1 所示。其中 LSF 参数、增益、基音周期、残差谐波谱这 4 个参数采用矢量量化进行编码。以 LSF 参数为例，编码器从每个语音帧中提取 10 个 LSF 参数，对这 10 个 LSF 参数使用 4 级矢量量化，各级的量化位数分别为 7，6，6，6。在 MELP 参数比特分配中，LSF 参数共分配了 25 bit，占编码总位数将近一半。进一步提高 LSF 参数的量化压缩比率，对降低语音编码速率起到重要作用。本文将针对 LSF 参数的矢量量化过程进行改进，利用基于 RDAE 构建的 LSF 参数编码方法来替换多级矢量量化方法，以期在不损失语音质量的条件下，进一步降低编码速率。另外 3 个矢量量化的参数通过本文方法可同样进行改进。

表 1 MELP 参数比特分配

Table 1 Bits allocation of MELP parameters

参数	比特位分配	浊音帧总比特位	清音帧总比特位
LSF 参数	7,6,6,6	25	25
增益	3,5	8	8
基音周期	7	7	7
非周期标志	1	1	纠错
残差谐波谱	8	8	纠错
分带清浊音	4	4	纠错
抖动标志	1	1	同步

构建基于 RDAE 的 MELP 编码器分两个步骤进行，首先是模型训练过程，然后是编解码过程。模型训练过程示意图如图 5 所示。RDAE 的下半部分即为 LSF 标准化参数的编码器，上半部分即为 LSF 标准化参数的解码器。编码器由两层 RBM 叠加而成，下层 RBM 的隐藏结点作为上层 RBM 的可视结点，第 1 层 RBM 可视层结点数为 10，第 1 隐藏层结点数为 100，第 2 隐藏层结点数决定了编码率，若设

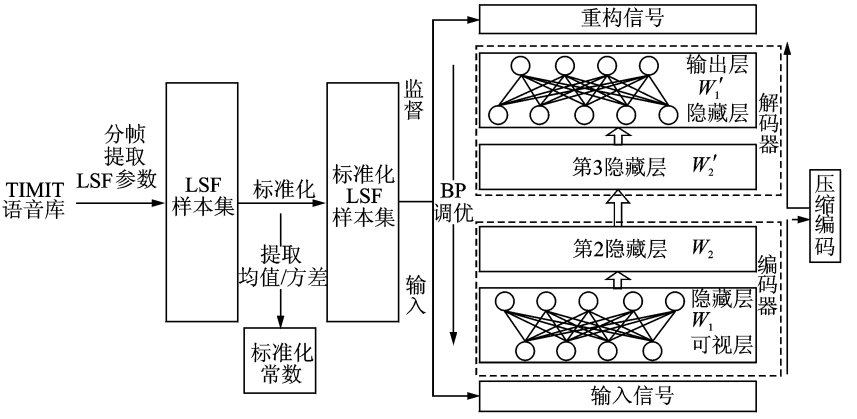


图 5 利用 LSF 样本集训练 RDAE

Fig. 5 RDAE training with LSF sample set

置为 25,则与矢量量化编码的压缩率相同,若低于 25,则获得更低压缩率。解码器的结构与编码器结构相对称。

将样本输入 RDAE 前需要先对样本进行标准化操作。每一帧的 LSF 参数共有 10 个系数,范围为 0～1的实数,且这 10 个系数按照由小到大的顺序排列。如图 6 所示,从 TIMIT 语音库中抽取了 100 句话,并统计了所有帧的 LSF 参数分布情况。可以看出,LSF 参数的 10 个系数各自都近似满足正态分布。因此可尝试通过逐维标准化的方法来对 LSF 参数进行标准化操作,即对各维分别统计均值和方差,并将均值归零,方差归一。通过对 100 句话进行统计,所提取出的 LSF 标准化常数如表 2 所示。该标准化常数在解码时用于恢复 LSF 的统计特性。

在获得了标准化后的 LSF 样本集后,可使用前两节的方法对 RDAE 进行预训练和调优训练。训练完毕后,即可将 RDAE 用于替换 MELP 中 LSF 参数的矢量量化过程。图 7 为基于 RDAE 的 MELP 编解码流程。

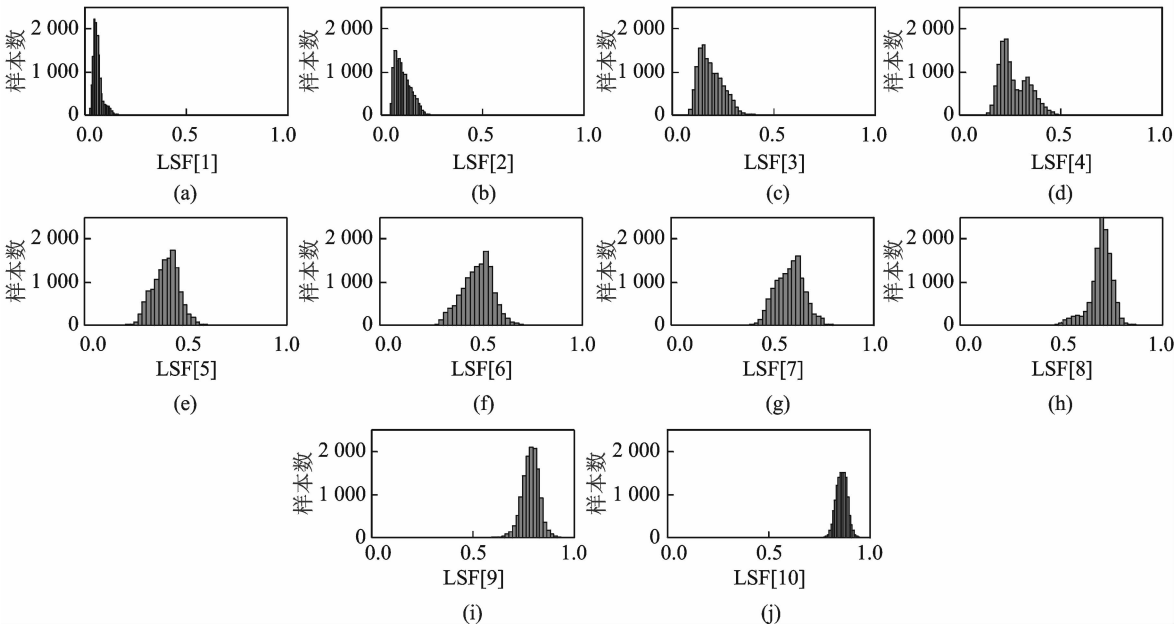


图 6 LSF 各维数据分布统计图

Fig. 6 Statistical graph of each dimension of LSF

表 2 LSF 标准化常数

Table 2 Normalization constant of LSF

LSF 参数序号	均值	标准差
1	0.063 1	0.021 4
2	0.109 7	0.038 0
3	0.189 1	0.056 4
4	0.271 6	0.069 8
5	0.404 1	0.060 5
6	0.484 5	0.071 9
7	0.588 6	0.067 1
8	0.699 4	0.053 4
9	0.787 1	0.041 1
10	0.860 5	0.026 0

原始的 MELP，采用 8 kHz 采样，每帧 180 个样点，用 54 bit 进行编码，码率为 2.4 kb/s，其中25 bit 为 LSF 参数的矢量量化编码。使用 RDAE 替换 MELP 中对 LSF 参数的矢量量化过程，若 RDAE 第 2 隐藏层结点数为 M，则改进后的 MELP 码率为$\frac{M+29}{22.5}$ kb/s。

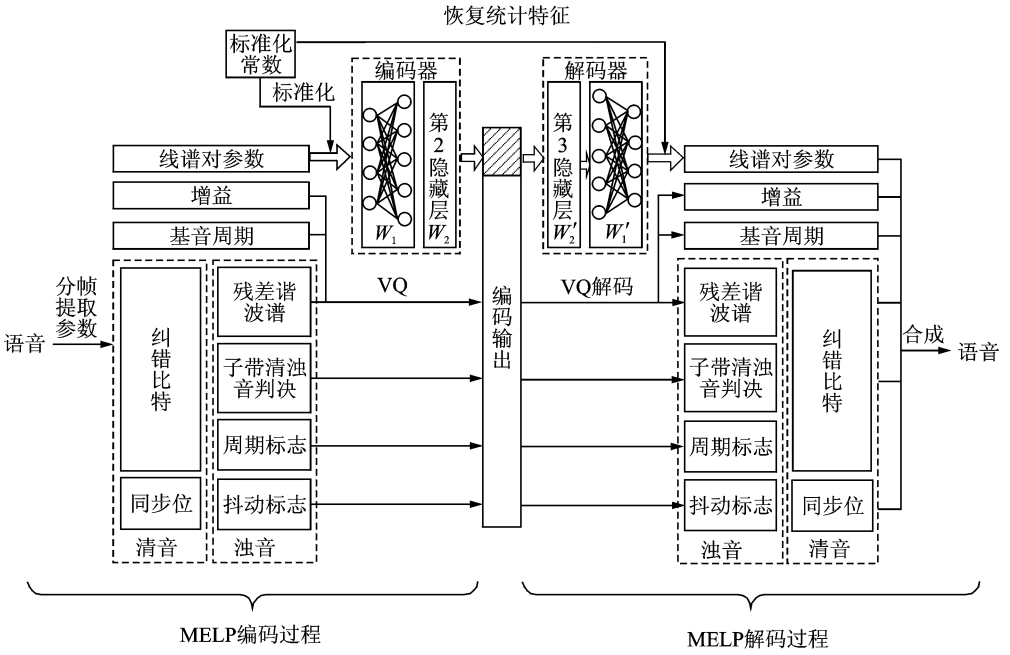


图 7 基于 RDAE 的 MELP 编解码流程

Fig. 7 Coding and decoding process of MELP based on RDAE

4 实验测试

本节对添加重构误差约束的 CD 算法性能以及改进后的 MELP 编码器性能分别进行测试。

4.1 重构误差约束的 CD 算法性能测试

原始 CD 算法的训练目标是最大化模型的边缘概率分布与训练样本概率分布的似然度，添加重构误差这一额外约束后会影响似然度。对算法的性能分析需要考虑重构性能和似然度这两方面。

为了在测试中能精确统计 RBM 的边缘概率分布，考虑到运算量的限制，选择小规模的 RBM 进行遍历统计。实际上统计一个由 v 个可视结点和 h 个隐藏结点构成的 RBM 所有状态，其计算量为 2^{v+h}。测试中，设 $v=8$，$h=6$。训练样本为 8 维的服从某伯努利分布的数据集。该数据集是通过对一个一维的高斯混合密度函数抽样获得。具体方法如下：首先分别生成 10 组服从不同高斯分布的随机数据集，这 10 组数据集的容量与高斯混合函数的混合系数相对应；将这 10 组数据合并成一组，即构成了服从由 10 个高斯分量组成的高斯混合分布的数据集，将数据集中元素二值化，即可获得所需的训练数据集 S。如图 8 所示，为所生成的训练数据集 S 中样本点的概率分布情况。

首先测试重构误差期望与训练样本概率分布的关系。重构误差期望体现了训练所得的 RBM 对符合训练样本概率分布的输入数据的重构能力，重构误差期望越小，则重构性能越好，否则越差。该指标的计算方法如下：对可视节点所有可能的输入 v_i^+ 依次进行编码并重构得到 v_i^-，计算重构误差 $\Delta v_i = |v_i^+ - v_i^-|$；计算该误差在训练样本概率分布下的期望，记为

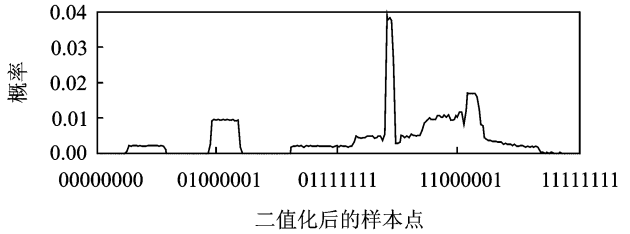


图 8 生成的训练样本集 S 中样本点的概率分布情况

Fig. 8 Probability distribution of generated training sample set S

$$E_{\mathrm{R}} = E_{p(s)}(\Delta v_i) = \sum_i p(v_i) \times \Delta v_i \tag{8}$$

指派度会影响重构误差,当指派度越大时,在训练样本概率分布越大处,数据经编码重构的误差越小,而代价是在训练样本概率分布较小处,数据编码重构的误差会增大。如图 9(a~f)所示。

图 9 中,横坐标为 RBM 的输入值,纵坐标为重构误差。对比图 8 中训练样本的概率分布可以看出,重构误差的趋势与训练样本概率分布趋势相反,训练样本概率分布越大处,重构误差越小。随着指派度 ϕ 的增大,在样本概率分布较大处,重构误差接近 0,但在样本概率分布较小处,重构误差接近 8(对于含 8 个可视结点的 RBM,其重构误差最大即为 8)。表 3 给出了不同 ϕ 下重构误差期望指标 E_{R} 的值。

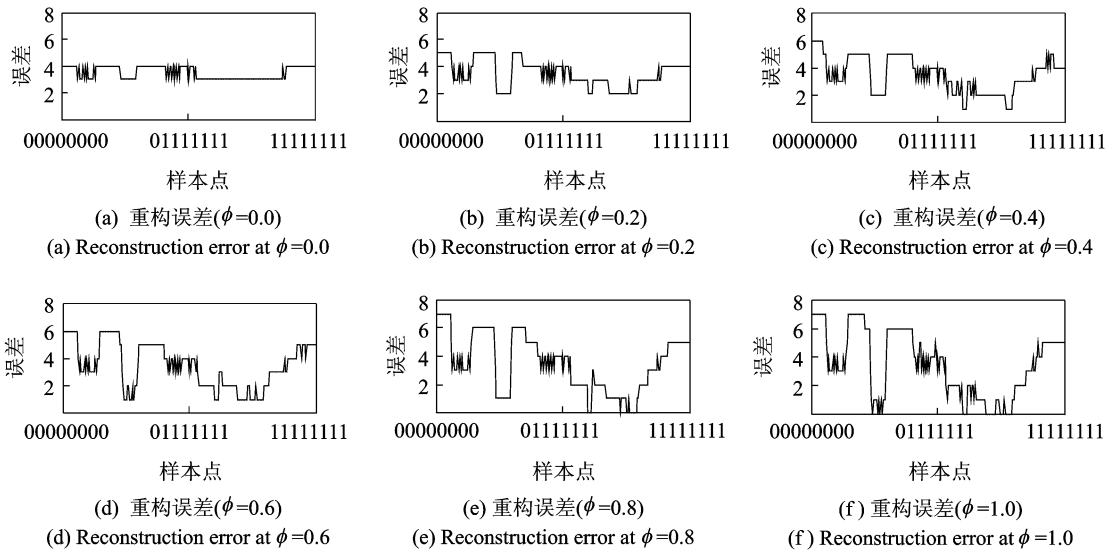


图 9 不同指派度下的重构误差

Fig. 9 Reconstruction error under different assignment degree

表 3 不同指派度下重构误差期望 E_{R}

Table 3 E_{R} under different assignment degree

ϕ	E_{R}	ϕ	E_{R}
0.0	3.25	0.6	2.56
0.1	3.01	0.7	2.47
0.2	2.95	0.8	2.35
0.3	2.89	0.9	2.25
0.4	2.76	1.0	2.19
0.5	2.65		

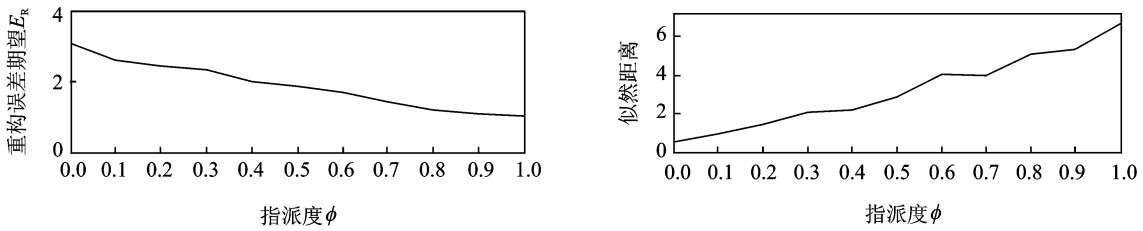
从表 3 可以看出，随着 ϕ 的增大，重构误差期望逐渐减小，$\phi=0$ 为原始 CD 算法。该变化趋势表明，添加了重构误差约束后，RBM 对符合训练样本概率分布的数据具有更好的重构性能。

进一步测试重构误差期望以及似然距离与指派度之间的关系。似然距离体现了训练所得 RBM 关于可视结点的边缘概率分布与训练样本概率分布的拟合程度，似然距离越小，则似然度越高，否则越差。该指标的计算方法如下：遍历 RBM 的 v 和 h，构造关于 v 和 h 的联合分布律 $p(v,h)$，统计关于 v 的边缘分布，利用 KL 距离来评估似然度

$$KL=\sum_{x\in S}p_s(x)\ln\frac{p_s(x)}{p_v(x)} \tag{9}$$

式中：S 为训练样本集；$p_s(x)$ 表示 x 在训练样本分布上的概率；$p_v(x)$ 表示 x 在 RBM 关于可视结点边缘分布上的概率。在实际计算过程中，对 $p_s(x)$ 和 $p_v(x)$ 整体叠加一个足够小的数，避免出现分母为 0 或者对 0 求对数的情况。

测试重构误差约束对 RBM 可视结点概率分布与训练样本概率分布之间似然度的影响。如图 10 所示，重构误差期望 R 以及似然距离指标 KL 随着指派度 ϕ 的变化趋势。



(a) 重构误差与指派度的关系
(a) Relationship between reconstruction error and assignment degree

(b) 似然度与指派度的关系
(b) Relationship between likelihood and assignment degree

图 10 两种指标变化趋势对比

Fig. 10 Variation trends of two indicators

从图 10 可看出，随着指派度 ϕ 的增大，重构误差期望 E_R 逐渐减小，重构性能提高，但似然距离却在逐渐增加，似然度降低。由此可见，添加重构误差约束的 CD 算法可获得编码重构性能的提升，但会损失一定的似然度。

4.2 改进后的 MELP 算法性能测试

针对改进后的 MELP 编码器测试了如下性能指标：(1)分别针对训练语音和测试语音，测试了在不同编码率下，RDAE 对 LSF 参数的编码重构精度，并与 VQ 量化编码重构精度作对比；(2)测试了不同编码率下，改进后的 MELP 编码器重构语音的质量。RDAE 的第 2 隐藏层结点数分别选择了 25,23,21,19,17,15 这 6 种情况。从 TIMIT 语音库中各选择 100 句话分别作为训练语音和测试语音，训练语音总帧数为 12 643 帧，测试语音总帧数为 14 424 帧。指派度 ϕ 取值为 1。

表 4 给出了不同压缩率下，RDAE 对 LSF 参数的编码重构误差(使用加权 LSF 距离[10]度量)和重构语音质量(使用 PESQ 得分[11]度量)。表 5 为不同压缩比下谱失真指标与谱失真异常帧占总帧数的比例。

表 4,5 最后一行都给出 25 bit 矢量量化方法的对比值。从表中记录的测试数据可看出，随着 RDAE 第 2 隐藏层结点数的递减，编码码率在进一步降低，当结点数降低到 19 bit 时， 本文算法所测得

表 4 基于 RDAE 的 MELP 编解码性能

Table 4 Performance of RDAE-based MELP

RDAE 第 2 隐层结点数	压缩率		加权 LSF 距离		重构语音 PESQ	
	LSF 编码压缩率/%	总编码速率/(kb·s^{-1})	训练样本	测试样本	训练样本	测试样本
25	7.8	2.4	0.028 7	0.029 1	3.157 6	3.137 3
23	7.1	2.3	0.030 6	0.031 3	3.123 3	3.108 6
21	6.5	2.2	0.032 4	0.033 5	3.101 5	3.094 0
19	5.9	2.1	0.035 8	0.036 9	3.089 0	2.990 2
17	5.3	2.0	0.044 7	0.049 1	2.875 5	2.757 0
15	4.6	1.9	0.048 5	0.051 4	2.607 8	2.552 1
VQ	7.8	2.4	0.036 7		3.088 9	

表 5 LSF 在不同编码位数下谱失真及异常比例

Table 5 Spectrum distortion and abnormal percentage of LSF under different coding bitrate

RDAE 第 2 隐藏层结点数	谱失真/dB		异常值/dB			
	训练集	测试集	2～4		>4	
			训练集/%	测试集/%	训练集/%	测试集/%
25	1.051 3	1.085 7	3.03	3.72	0	0
23	1.082 8	1.116 2	3.53	4.01	0	0
21	1.111 9	1.136 9	3.76	4.13	0	0
19	1.116 7	1.145 5	3.94	4.79	0	0
17	1.293 8	1.301 2	9.18	9.30	0	0
15	1.410 9	1.444 8	12.54	13.78	0	0.006 9
VQ	1.221 8		5.07		0.015 8	

的加权 LSF 距离、重构语音质量、谱失真指标在训练集和测试集上都比 25 bit 矢量量化方法所测得的性能更优。因此,利用本文方法改进的 MELP 编码器,在不降低语音质量的条件下,可将编码速率从 2.4 kb/s降低至2.1 kb/s,编码速率降低了 12.5%。

5 结束语

本文利用交叉熵对传统 CD 算法添加了重构误差约束,并利用改进后的 CD 算法训练 RDAE,在损失一定的似然度条件下,提升了 RDAE 的重构性能。在此基础上,使用 RDAE 替换 MELP 编码器中对 LSF 参数的矢量量化编码过程,在不降低语音质量的条件下,进一步降低了语音编码速率。本文方法仅针对 MELP 中 LSF 参数的矢量量化过程进行改进,实际上,MELP 编码器所提取的语音参数中,增益参数、基音周期参数和残差谐波谱参数都可以采用同样的改进方法,从而可进一步降低编码率。本文方法仅针对单帧进行处理,若通过多帧联合的方法,预期将能进一步降低编码率。多帧联合处理本质上是通过去除连续多帧之间的冗余信息来降低码率。这将作为下一步的研究工作。

参考文献：

[1] 张雄伟，陈亮，杨吉斌. 现代语音处理技术及应用[M]. 北京：机械工业出版社，2003：112-172.
Zhang Xiongwei, Chen Liang, Yang Jiebin. Modern speech processing technique and application[M]. Beijing: China Machine Press, 2009:112-172.

[2] Department of defense telecommunications systems standard. Mil-std-3005, analog-to-digital conversin of voice by 2 400 bit/second mixed excitation linear predication(MELP) [S]. Washington, USA: Final Committee Draft, 1995.

[3] Gray R M. Vector quantization[J]. IEEE ASSP Magazine, 1984,4:4-29.

[4] Garofolo J S. Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database[R]. NIST Tech Report 1988-552. USA: NIST, 1988.

[5] Mohamed A, Yu D, Deng L. Investigation of full-sequence training of deep belief networks for speech recognition[C]//Proceeding of Interspeech. Chiba, Japan: International Speech Communication Association (ISCA), 2010:57-65.

[6] Deng L, Seltzer M, Yu D, et al. Binary coding of speech spectrograms using a deep auto-encoder[C]// Proceeding of Inter Speech. Chiba, Japan: International Speech Communication Association (ISCA), 2010:23-34.

[7] Carreira-PerpinanM A, Hinton G E. On contrastive divergence (CD) learning[C]// Proceedings of AISTATS. Savannah Hotel, Barbados: Microtome Publishing, 2005:93-101.

[8] Desiardins G, Courville A, Bengio Y. Tempered Markov chain Monte Carlo for training of restricted Boltzmann machine [C]//Proceedings of AISTATS. Sardinia, Italy: Microtome Publishing, 2010:145-152.

[9] Nair V, Hinton G E. 3D object recognition with deep belief nets[C]//Proceedings of NIPS. Vancouver, B C, Canada: Neural Information Processing Systems Foundation, 2009: 326-337.

[10] PaliwalK K, Atal B S. Efficient vector quantization of LPC parameters at 24 bits/frame[C]//International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Toronto, Canada: IEEE Computer Society Press, 1991:661-664.

[11] ITU-T Study Group 12. ITU P 862-2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs[S]. Geneva, Switzerland:ITU, 2001.

作者简介：**张雄伟**(1965-)，男，教授，博士生导师，研究方向：语音信号处理、智能信息处理，E-mail：xwzhang9898@163.com；**吴海佳**(1986-)，男，博士研究生，研究方向：智能信息处理；**张梁梁**(1989-)，女，博士研究生，研究方向：智能信息处理；**邹霞**(1979-)，男，副教授，研究方向：语音信号处理。