基于SOM网络的语音训练可视化方法研究

杨 丹[1],徐 彬[2],王 旭[1],廖富成[1]
YANG Dan[1],XU Bin[2],WANG Xu[1],LIAO Fu-cheng[1]

1.东北大学 信息科学与工程学院,沈阳 110004
2.东北大学 计算中心,沈阳 110004
1.School of Information Science & Engineering,Northeastern University,Shenyang 110004,China
2.Computer Center,Northeastern University,Shenyang 110004,China
E-mail:yangdan@mail.neu.edu.cn

YANG Dan,XU Bin,WANG Xu,et al.Research of speech training visual method using Self-Organizing Map.Computer Engineering and Applications,2008,44(18):15-16.

Abstract: A speech training method using output-layer visualization of Self-Organizing Map(SOM) is proposed.SOM is a neural network model that can transform input data onto two-dimensional plane or curve surface of output layer neurons.The subjects guide their pronunciation through visual feedback from positional information of output layer neurons.In order to improve the clustering of SOM,the authors make strengthen training and discuss how to choose the number of neurons in the output layer.The results show the proposed speech training method is simple and straightforward.It effectively realizes "speak when seeing the picture".

Key words: Self-Organizing Map(SOM);speech training;visual feedback

摘 要:提出了一种利用SOM网络输出层可视化的特点进行语音训练的方法。SOM网络能够将输入向量映射到二维平面或曲面上,受试者通过视觉反馈的位置信息,指导其发音行为。为了提高SOM聚类效果,SOM还进行加强训练;讨论了SOM输出层神经元个数对聚类的影响。实验结果表明,提出的利用SOM语音训练方法,直观简单,能够有效地实现"看图说话"。

关键词:SOM网络;语音训练;视觉反馈
DOI:10.3778/j.issn.1002-8331.2008.18.005 **文章编号:**1002-8331(2008)18-0015-02 **文献标识码:**A **中图分类号:**TN912.3

1 引言

语音训练中的反馈环节是理想输出与实际发音之间的关系,它使受试者意识到发音的声学特点及与标准发音的误差。需要进行语音训练的人(如外语的学习者和有语音功能障碍的人)从反馈信息中调整修正自己的发音模式,直到发音标准为止[1]。目前语音训练中的反馈途径主要有两种:听觉反馈和视觉反馈。其中听觉反馈多是为聋儿语音训练使用,利用他们的残余听力采用助听器或进行人工耳蜗植入进行听觉补偿,实现发音指导;而视觉反馈多是借助视觉通道,将受训者的发音形象的显现出来,从而调节他们的发音行为[2-4]。语谱图是最早的一种"可见语言",它将可听的语言描绘成可见的图形,显示语音的声学特点,受试者经过一定时间的学习训练后能够将不同发音以图形方式区别开,实现"看图说话"。

本文提出一种利用SOM(Self-Organize Map)网络的语音训练可视化方法。在SOM网格上建立起有意义的坐标系,用输出层网格上神经元的位置表示输入模式,提供一种语音可视化表达。利用SOM网络输出层显示的位置信息进行语音训练。

2 语音训练图的建立

在1981年,芬兰的Kohonen教授提出了自组织持征映射网络(SOM)。他认为一个神经网络接受外界输入模式时,能自动地分为不同的区域。各区域对不同输入模式具有不同的响应特征[5]。在学习过程中,只需向网络提供一些学习样本,而无需提供理想的输出,与传统的模式聚类方法相比,它所形成的聚类中心能映射到一个曲面或平面上。并且保持拓扑结构不变。本文利用SOM网络构建一个可视化的语音训练平台。

2.1 建立语音训练的样本空间

在实验中,选取汉语5个单韵母a,i,u,o,e作为语音训练的目标。语音信号的采集是由audio view 1.50v录音软件录制,文件格式为wav音频规格,采样频率为44.1 kHz,16 bit,

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.59977024)。
作者简介:杨丹(1979-),女,博士研究生,主要研究领域为语音信号处理及应用、生物系统建模等;徐彬(1980-),男,博士研究生,主要研究领域为信号检测与分析、神经网络建模等;王旭(1957-),男,博士,教授,博士生导师,主要研究领域为智能控制技术、神经网络建模、生物信息技术等方面等。
收稿日期:2008-02-28 **修回日期:**2008-03-24

mono。建立 40 组共 200 个样本。训练样本和测试样本的个数，根据实验要求选定。

由于语音信号从嘴唇辐射会有 6 dB/oct 的衰减，因此在对语音进行处理之前，按 6 dB/oct 的比例对信号加重。将加重后语音信号分为 14 个语音分帧，分别求取 12 阶 MFCC(Mel Frequency Cepstrum Coefficient，MFCC)系数和差分倒谱系数，然后将其组合成 MFCC 系数[6]。去掉了首尾两帧，构建 MFCC 系数向量为 240 维。得到的 MFCC 系数分布区间大，在神经网络调整权值时会带来难度。所以，将这些系数进行归一化处理，从而得到语音训练的样本。

2.2 SOM 网络的训练

SOM 网络的学习过程包括最佳匹配神经元的选择和权矢量的自适应变化过程两部分[7]。

具体步骤如下：

(1)初始化。对权值向量 $\boldsymbol{w}_{ij}(0)$ 赋予[-1,1]区间内的随机值，其中(i,j)为输出层神经元的坐标。

(2)取样。样本空间为

$$X_t=[X_{t1},X_{t2},\cdots,X_{tk}] \tag{1}$$

式(1)中 t 为当前训练次数，k 为输入样本序数。

(3)相似性匹配。计算输入模式与每个输出神经元节点连接权矢量的距离 d_{ij}

$$d_{ij}=\sqrt{\sum_{m=1}^{M}[X_{tkm}(t)-w_{ijm}(t)]^2} \tag{2}$$

式(2)中，M 表示样本向量维数。选择具有最小距离的输出节点(i^*,j^*)作为获胜节点，如式(3)。

$$d_{i^*j^*}=\min\{d_{ij}\} \tag{3}$$

(4)更新。以获胜神经元(i^*,j^*)为中心，利用式(4)更新权值。

$$w_{ijm}(t+1)=w_{ijm}(t)+\eta(t)\delta(t)[X_{tkm}(t)-w_{ijm}(t)] \tag{4}$$

式(4)中，$\eta(t)$为学习率参数，$\delta(t)$为获胜神经元(i^*,j^*)周围的邻域半径调整函数，如式(5)。

$$\delta(t)=\text{int}[\delta_0(1-t/T)] \tag{5}$$

式(5)中，T 为训练总次数，t 为当前训练次数，δ_0 是邻域半径初始值。

(5)继续。当训练样本集 X 中每个样本都经过一次训练之后，返回步骤(2)，直到 $t>T$。

2.3 SOM 网络的加强训练

由于某些输入样本比较相近，存在某些神经元对两个或者多个输入样本都能产生响应，这类神经元是边界神经元。为了提高 SOM 的聚类性能，对 SOM 网络进行加强训练。加强训练是在 SOM 网络经过第一轮训练完成之后所得权值的基础上进行的。第一轮训练称为“粗调”，加强训练称为”微调”。其学习过程如下：

(1)根据粗调的训练结果，标识输出层神经元模式。

(2)初始权值。微调时的初始权值 $w'_{ij}(0)$为粗调训练后的权值 $w_{ij}(T)$。

(3)取样和相似性匹配。样本空间和匹配策略与粗调时相同。

(4)更新。先进行模式类判别。如果输出的响应模式类与标识的类别一致的话，按式(6)进行正向调整；响应模式类与标识的类别不一样，则按式(7)进行负向调整。λ 为负向调整系数，取[[0,1]之间。

$$w_{ijm}(t+1)=w_{ijm}(t)+\eta(t)\delta(t)[X_{tkm}(t)-w_{ijm}(t)] \tag{6}$$

$$w_{ijm}(t+1)=w_{ijm}(t)-\lambda\eta(t)\delta(t)[X_{tkm}(t)-w_{ijm}(t)] \tag{7}$$

(5)继续。当训练样本集 X 中每个样本训练一次之后，返回步骤(2)，直到 $t>T2$。$T2$ 为微调训练次数。

SOM 网络经微调后，生成输出影射图。受试者根据视觉反馈的信息，指导自己的发声。

3 实验结果

实验中提取 MFCC 系数时，输入层神经元维数为 240；网络输出层神经元个数为 9×9。粗调时的学习效率 $\eta_0=0.4$；粗调学习半径 $\delta_0=7$；微调学习效率函数，如式 8；微调学习半径 $\delta_0=2$。

$$\eta(t)=\eta_0\exp(1-\frac{1.5t}{T}) \tag{8}$$

SOM 网络经粗调训练 1 000 次，微调训练 500 次，生成语音映射图，如图 1。其中的方格表示输出层的神经元；格内的字母表示该神经元所归属的模式类；未标识的神经元为屏蔽神经元，同一方格中含有两个字母标识是边界神经元。

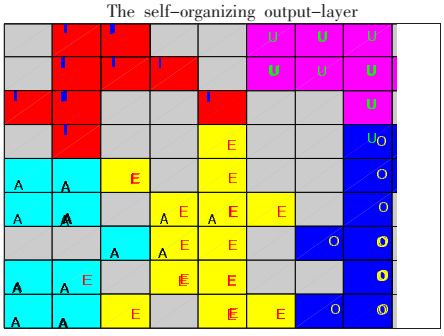

图 1 汉语单韵母 a,i,u,o,e 通过 SOM 生成映射图

在实验中 SOM 网络的输出层神经元个数对语音训练映射图有相当大的影响，如表 1。可以看出，在不改变 SOM 网络的其他参数的条件下，输出层的神经元个数大于类别数的 5 倍以上，SOM 网络的聚类效果一般；在 10~20 倍时，聚类效果理想；当大于 20 倍时，聚类效果基本没有提高，而其计算量无限增加。

表 1 输出层神经元个数与网络训练时间及聚类率关系

输出层神经元个数	3×3	4×4	5×5	6×6	7×7	8×8	9×9	10×10	11×11
训练时间/s	117	149	219	299	402	506	615	732	835
聚类正确率	0.46	0.71	0.83	0.85	0.90	0.91	0.93	0.92	0.92

图 2 为标准成年女的汉语 5 个单韵母发声的语谱图。它反映了发音时长、共振峰纹理、基频位置等一些声学特征，但作为语音训练的可视化图形来说，要求受试者有一定专业的语音背景知识，经过长期训练才能实现看图说话。而利用 SOM 网络在将标准汉语单韵母发音进行训练，生成如图 1 的语音映射图。受试者如发音有误或不标准时，映射在输出层的位置不同，结合标准语音映射图，训练自己的发音，调整发音直到位置相同位置。从而达到进行实现语音训练的目的。

(下转 20 页)