·多媒体技术及应用· 文章编号：1000-3428(2018)09-0256-07 文献标志码：A 中图分类号：TP391

基于神经网络的 AVS-P10 开环模式选择算法优化

崔佰会[1]，高 戈[1]，姜 林[1,2]

（1. 武汉大学 国家多媒体软件工程技术研究中心，武汉 430072；2. 东华理工大学 软件学院，南昌 330013）

摘 要： 现有的开环模式选择算法依赖信号分类的准确率，但多数情况下准确率较低，造成开环模式下编码音质较差。为此，提出一种改进的基于神经网络的开环模式选择算法。使用神经网络替换原开环模式选择的决策树算法，拟合闭环模式选择结果进行训练得到模式选择分类器，按照闭环模式选择的逻辑过程，运用神经网络预测输入的信号，在 ACELP256 和 TVC256 两种编码模式的信噪比取代编码尝试计算得到的信噪比。实验结果表明，与原 AVS-P10 开环选择方法相比，提出的 2 种模式在语音分类准确率上分别提升 5.96% 和 18.07%，在音乐分类准确率上分别提升 3.84% 和 20.29%，其主客观编码音质评测明显提升。

关键词： 神经网络；先进音视频编码；模式选择；特征选择；信号分类；信噪比估计

中文引用格式： 崔佰会，高 戈，姜 林. 基于神经网络的 AVS-P10 开环模式选择算法优化[J]. 计算机工程，2018，44(9)：256-262.

英文引用格式： CUI Baihui，GAO Ge，JIANG Lin. Optimization of AVS-P10 open-loop mode selection algorithm based on neural network[J]. Computer Enginering，2018，44(9)：256-262.

Optimization of AVS-P10 Open-loop Mode Selection Algorithm Based on Neural Network

CUI Baihui[1]，GAO Ge[1]，JIANG Lin[1,2]

（1. National Engineering Research Center for Multimedia Software，Wuhan University，Wuhan 430072，China；
2. Software College，East China University of Technology，Nanchang 330013，China）

【Abstract】 The existing open-loop mode selection algorithm relies on the accuracy of signal classification，but in most cases the accuracy is low，resulting in poor coding quality in open-loop mode. Therefore an improved neural network based open-loop mode selection algorithm is proposed. The decision tree algorithm of the original open-loop mode selection is replaced by the neural network，and the closed-loop mode selection is selected for training to obtain the mode selection classifier. According to the logic process of the closed-loop mode selection，the neural network is used to predict the input signal，and the two codes are ACELP256 and TVC256. The signal to noise ratio of the mode replaces the signal-to-noise ratio that the coding attempt is calculated. Experimental results show that the accuracy of the two methods is 5.96% and 18.07%，respectively，and the accuracy of music classification is 3.84% and 20.29%，the performance of subjective and objective tone has high improvement.

【Key words】 neural network；advanced audio and video coding；mode selection；feature selection；signal classification；Signal to Noise Ratio(SNR) estimation

DOI： 10.19678/j.issn.1000-3428.0047850

0 概述

AVS-P10 是面向新一代移动通信系统的低码率高保真音频编解码技术标准，该标准应用于多媒体通信、无线宽带、互联网宽带流媒体业务、移动通信等多个信息通信领域[1]。AVS-P10 的核心编码器由代码码本激励线性预测 ACELP[2] 和变换域矢量编码 TVC[3] 组成，ACELP 用于编码类语音和瞬态信号，TVC 用于编码类音乐信号。在帧结构上，每 1 024 个连续的采样点组成一个 80 ms 的超帧，每个超帧被分成 256 点、512 点或者 1 024 点的子帧。其中，256 点的子帧采用 ACELP 或者 TVC 编码，512 点和 1 024 点子帧采用 TVC 编码。根据上述帧结构，对不同子帧采用4种不同的编码模式，即

基金项目： 国家自然科学基金“基于上下文相关的音频非盲带宽扩展编码研究”(61762005)。

作者简介： 崔佰会(1991—)，男，硕士研究生，主研方向为音频信号编码；高 戈，副教授；姜 林，副教授、博士研究生。

收稿日期： 2017-07-06 **修回日期：** 2017-08-25 **E-mail：** 903414207@qq.com

ACELP256、TVC256、TVC512 和 TVC1024。由此,在一个超帧上(1 024 样点)会有 26 种不同的编码模式组合,具体采用哪种编码模式取决于信号特征。针对该问题,本文提出了 2 种基于神经网络的低复杂度的 AVS-P10 模式选择方案。AVS-P10 编码可以通过闭环或者开环模式选择来决定。

1　模式选择算法

1.1　闭环模式选择

闭环模式选择对 1 024 样点的超帧中所有可能的模式组合进行编码,然后根据信噪比选择最优的组合。为了避免 26 种编码模式组合中的冗余,降低选择的复杂度,闭环模式首先对 1 024 点超帧中的 4 个256 点帧尝试使用 ACELP256 模式和 TVC256 模式编码,然后分别对前 2 个和后 2 个 256 点帧尝试使用 TVC512 模式编码,最后对一个超帧尝试使用 TVC1024 模式编码。编码详细过程如表 1 所示。

表 1　AVS-P10 闭环模式选择

步骤	第 1 个 256 点帧	第 2 个 256 点帧	第 3 个 256 点帧	第 4 个 256 点帧
1	ACELP256			
2	TVC256			
3		ACELP256		
4		TVC256		
5	TVC512			
6			ACELP256	
7			TVC256	
8				ACELP256
9				TVC256
10			TVC512	
11	TVC1024			

以上模式选择的标准是加权语音 $x(n)$ 和合成加权语音 $\hat{x}_w(n)$ 间的分段信噪比均值。第 i 个子帧的分段 SNR 定义如下:

$$segSNR_i = 20\lg\left(\frac{\sum_{n=0}^{N-1} x_w^2(n)}{\sum_{n=0}^{N-1}\left(x_w(n) - \hat{x}_w(n)\right)^2}\right) \tag{1}$$

其中,$segSNR_i$ 为第 i 个子帧的分段 SNR,N 表示 64 个样点的子帧长度。

分段 SNR 均值定义如下:

$$\overline{segSNR} = \frac{1}{N_{SF}}\sum_{i=0}^{N_{SF}-1} segSNR_i \tag{2}$$

其中,N_{SF}表示帧中子帧的标号。

虽然全复杂度的闭环模式选择能够提供最优的编码质量,但是时间复杂度很高,这给 AVS-P10 的使用和推广带来很大的阻碍。

1.2　开环模式选择

开环模式选择是低复杂度的编码模式选择方法,该方法不是对所有模式进行编码尝试,而是根据信号类型初步决定对 256 点帧选择哪一种编码模式进行编码,接下来的步骤类似于闭环模式,分别对前 2 个和后 2 个 256 点帧尝试使用 TVC512 模式编码,最后对一个超帧尝试使用 TVC1024 模式编码,通过比较编码信噪比确定最终的编码模式。为了进一步降低计算复杂度,开环模式又做了以下优化,如果前 2 个 256 点帧或者后 2 个 256 点帧选择了 ACELP256 编码模式,则放弃对 TVC512 的编码尝试;如果 4 个 256 点帧中选择了 ACELP256 编码模式,则放弃对 TVC1024 的编码尝试。开环模式选择虽然增加了信号信息选择模块,但是相对于对各种模式进行编码尝试,还是大大降低了计算的复杂度。AVS-P10 技术标准中的开环模式选择首先提取输入信号的特征参数,然后根据这些特征参数使用决策树算法将信号初始划分为不确定类、非有用信号、语音和音乐,最后根据当前的语境进行 2 次修正分类,最终确定信号为非有用信号、语音和音乐。对语音信号使用 ACELP256 模式编码,其他信号使用 TVC256 模式编码。开环模式虽然大大降低了计算的复杂度,但是由于原开环模式做的主要工作是解决语音与音乐的分类,这对于最佳模式选择的决策有一定的偏差,再加上语音与音乐分类的准确率不高,导致最终的开环模式选择的准确率特别低。表 2 统计了开环模式选择相对于闭环模式的准确率,测试数据为单声道的语音和音乐信号,比特率为 24 Kb/s,内部采样率为 25.6 kHz,共 657 728 超帧。

表 2　AVS-P10 开环模式选择分类准确率　%

AVS-P10 开环选择	准确率
语音	63.2
音乐	68.5

由表 2 可见,语音和音乐的分类准确率都不是很高,但是低复杂度的开环模式选择对 AVS-P10 技术标准的推广有着至关重要的作用,所以迫切需要找到一个准确率高的开环模式选择算法。针对此类算法,各国学者进行了大量的研究。文献[4]使用子帧能量标准差、高低子带能量比和子帧能量特征,利用神经网络算法来优化 AMR-WB +[5-6]低复杂度的编码模式选择,该方法时间复杂度低,分类准确率较高,与标准的低复杂度 AMR-WB + 模式选择算法相比,音乐平均错误率降低了 20%,语音降低了 15%。文献[7]利用随机森林算法对噪声不敏感、不易过拟合的优点,使用 Filter 方法(一种通过评估数据内部属性给所有特征进行评分的方法)中的 Relief[8]算法

(Relief 算法是基于样本学习的特征权重计算方法,也是一种著名的多变量 Filter 方法的特征选择算法[9])过滤某些最不相关特征和 Wrapper[10]模式(通过对所有特征子集进行训练与测试选出最优特征子集)选取最佳特征子集,最后实现信号的可靠分类,对语音和音乐分类的准确率分别达到了 93.2%、84.6%,但是由于其方法构造的随机森林深度太大,造成时间复杂度太高,而且其测试数据来自网上的流行音乐,比较单一,实验结果不具有鲁棒性。

本文使用原 AVS_P10 开环模式选择的部分特征,包括短时平均过零率、子帧短时能量、子帧子带能量、子帧子带能量比,并提取了频谱质心、带宽、频谱波动3 个特征,利用神经网络算法实现 2 种方案。

2 神经网络模式选择

ACELP 主要用于语音或者瞬态信号编码,TVC 主要用于音乐或者连续信号编码,如果将语音与音乐或者瞬态信号与连续信号差别较大的特征与编码相关的结果建立一个映射关系,就可以利用这些特征来预测编码相关的结果。本文使用神经网络建立一个非线性映射模型,利用音频相关的特征预测编码模式类别或者信噪比。

假设 N 为音频信号特征集,C 为编码类别或者信噪比集合,可以从映射函数集合 $F=\{f:N\mapsto C\}$ 中寻找一个最优映射函数。为了找到一个映射函数,引入损失函数 $Loss(c,y)$,其中,c 为原始编码模型或者信噪比,$y=f(n)$ 为音频信号特征 n 下的预测值。根据统计学习理论[11],预测函数 $f(n)$ 的风险为损失函数 $Loss(c,y)$ 的数学期望:

$$R(f)=E[Loss(c,f(n))]=\int_{C,N}Loss(c,f(n))P(c,n)\mathrm{d}c\mathrm{d}n \tag{3}$$

其中,$p(c,n)$ 为 N 训练集和 C 训练集之间的联合概率分布函数。本文的最终目标是寻找一个最优函数 f_F^*,使 $R(f)$ 最小:

$$f_F^*=\arg\min_{f\in F}R(f)=\arg\min_{f\in F}\int_{C,N}Loss(c,f(n))P(c,n)\mathrm{d}c\mathrm{d}n \tag{4}$$

在学习算法中,联合概率分布函数 $p(c,n)$ 是未知的,所以风险函数 $R(f)$ 不能直接计算得到,但可以通过计算训练集上平均损失函数近似得到 f_F^*,这个值被称为经验风险:

$$R_{\mathrm{emp}}(f)=\frac{1}{N_n}\sum_{i=1}^{N_n}Loss(c_i,f(n_i)) \tag{5}$$

其中,N_n为训练集大小。通过经验风险最小化理论,根据学习算法寻找一个预测函数 $\hat{f}$ 使经验风险最小:

$$\hat{f}=\arg\min_{f\in F}R_{\mathrm{emp}}(f)=\arg\min_{f\in F}\frac{1}{N_n}\sum_{i=1}^{N_n}Loss(c_i,f(n_i)) \tag{6}$$

为了找到预测函数 $\hat{f}$,通常使用绝对差值 $|c-f(n)|$作为损失函数,本文使用神经网络作为预测函数 $f(n)$,那么预测函数可被表示为:

$$f(w,b,n_i)=\sum_{j=1}^{M}wn_{i,j}+b \tag{7}$$

其中,w 为网络权值,b 为偏移值。综上,预测函数 $\hat{f}$ 的最终形式为:

$$\hat{f}(c,n,w,b)=\arg\min_{f\in F}\frac{1}{N_n}\sum_{i=1}^{Nn}\sum_{j=1}^{M}|c_{i,j}-(wn_{i,j}+b)| \tag{8}$$

其中,i 为帧索引,j 为音频特征索引,M 为特征个数,w 和 b 分别为神经网络权值和偏移。

与原开环模式的决策树算法相比,原开环模式的决策树算法只能根据特征的先后顺序选择部分特征进行判断,并且一旦某个特征判断错误,就会造成整个结果的错误;而本文中的神经网络算法会给每个参与预测的特征分配一个权值,所有输入的特征都会参与预测,预测结果有更好的鲁棒性。

本文考虑 2 种方案进行模式选择:第 1 种方案使用神经网络替换原开环模式选择的决策树算法,拟合闭环模式选择进行训练,最后得到模式选择分类器,如图 1 所示;第 2 种方案按照闭环模式选择的逻辑过程,使用神经网络预测输入的信号在 ACELP256 和 TVC256 2 种编码模式的信噪比来取代编码尝试计算得到的信噪比,如图 2 所示。

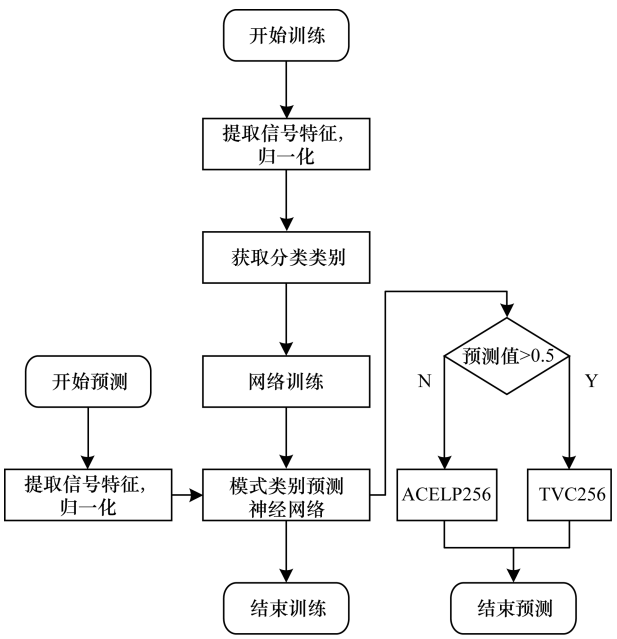

图 1 神经网络预测模式类别(方案 1)

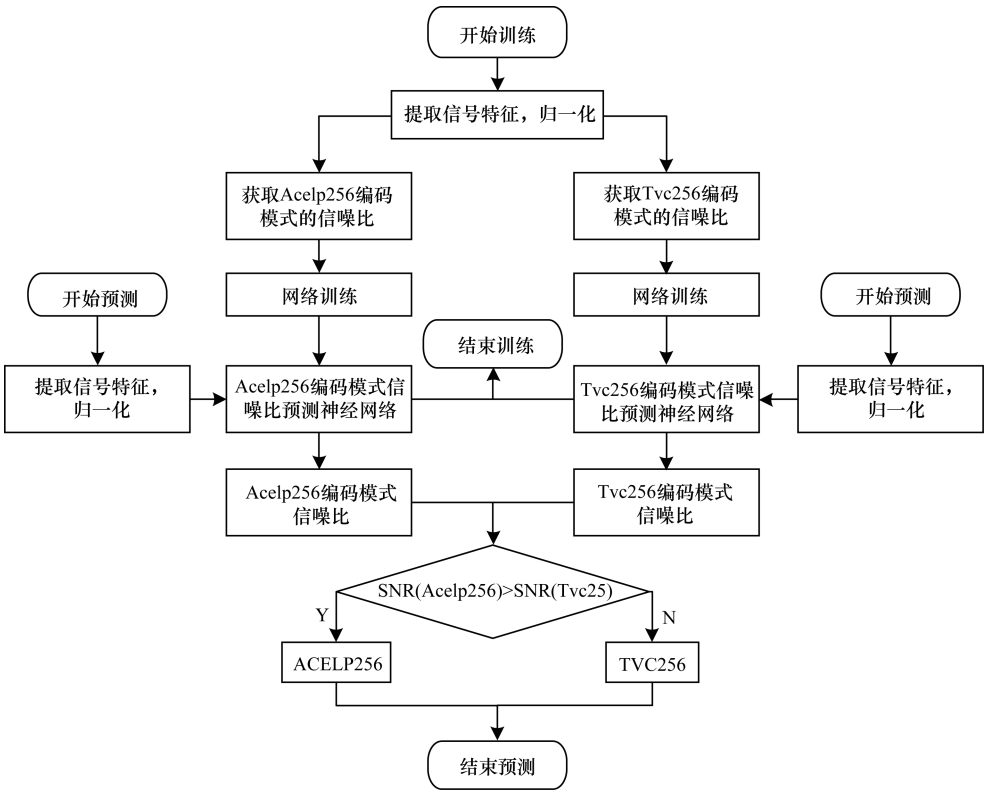

图 2　神经网络预测信噪比(方案 2)

从图 1、图 2 可以看出,方案 1 和方案 2 的过程非常相似,都是利用神经网络对音频特征进行训练与预测。方案 1 直接对分类结果进行训练与预测;方案 2 训练与预测的目标为 ACELP256 和 TVC256 2 种编码模式的信噪比,需要训练得到 2 个网络,预测结果也有 2 个值,通过比较这 2 个值的大小间接预测分类结果。

2.1　特征参数

本文使用频谱质心、带宽、短时平均过零率、子帧短时能量、子帧子带能量、频谱波动、子帧子带能量比 7 种特征,共 29 个特征值,其中子带能量和子带能量比各提供 12 个特征值,其余 5 个特征各提供一个特征值。

2.1.1　频谱质心

频谱质心可以看成是一个频率的均值或期望,它也被称作音频亮度,一般来说音乐的谱质心要高于语音[10],其计算公式如下:

$$W_c=\frac{\int_0^{\omega_0}\omega\ |F(\omega)|^2\mathrm{d}\omega}{\int_0^{\omega_0}|F(\omega)|^2\mathrm{d}\omega} \tag{9}$$

其中,ω 为角频率,$F(\omega)$ 为经过快速傅里叶变换得到的频谱。

2.1.2　带宽

带宽指音频信号中含有的有效成分的频率范围,即最高频信号与最低频信号的差值[12],其计算公式如下:

$$BW=\sqrt{\frac{\int_0^{\omega_0}(\omega-\omega_c)^2\ |F(\omega)|^2\mathrm{d}\omega}{\int_0^{\omega_0}|F(\omega)|^2\mathrm{d}\omega}} \tag{10}$$

其中,ω_c 为频谱质心,$F(\omega)$ 为经过快速傅里叶变换得到的频谱。

2.1.3　短时平均过零率

短时平均过零率可以考察音频信号时域波形通过时间轴的情况,也在一定程度上反映其频谱性质,可以通过信号的短时平均过零率粗略估计该信号的谱特性。一般语音信号幅度变化比较大,其短时平均过零率比音乐信号的要高很多,计算方法如下:

$$zcr=\frac{1}{T}\sum_{i=1}^{T-1}(x(i)x(i-1)<0) \tag{11}$$

其中,当 $x(i)x(i-1)<0$ 取 true 时为 1,取 false 时为 0,T 为一帧内的样点个数。

2.1.4　子帧短时能量

子帧短时能量为一帧样点值的加权平方和,定义 n 时刻某语音信号的短时平均能量 En 为:

$$En=\sum_{m=n-(N-1)}^{n}[x(m)w(n-m)]^2 \tag{12}$$

$$w(n-m)=\begin{cases}0.54-0.46\cos[2\pi(n-m)/(N-1)],\\ m\leqslant n\leqslant N+m\\ 0,\text{其他}\end{cases} \tag{13}$$

其中,N 为窗长,$x(m)$ 为时域信号。

2.1.5 子帧子带能量

在 AVS-P10 中,低频的带宽是 6.4 KHz,考虑心理声学模型的特点,本文采用非均匀划分,共划分为 12 个子带,即 0 Hz ~ 200 Hz、200 Hz ~ 400 Hz、400 Hz ~600 Hz、600 Hz ~800 Hz、800 Hz ~1 200 Hz、1 200 Hz ~ 1 600 Hz、1 600 Hz ~2 000 Hz、2 000 Hz ~ 2 400 Hz、2 400 Hz ~3 200 Hz、3 200 Hz ~4 000 Hz、4 000 Hz ~4 800 Hz、4 800 Hz ~6 400 Hz。子带能量的计算公式如下:

$$P_j = \lg\left(\int_{L_j}^{H_j} |F(\omega)|^2 \mathrm{d}\omega\right) \tag{14}$$

其中,H_j 为子带频率的上界,L_j 为子带频率的下界。

2.1.6 频谱波动

频谱波动反映了一帧内相邻 2 个子帧间的频谱值平均变化情况[13],其计算公式如下:

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\lg(A(n,k)+\delta) - \lg(A(n-1,k)+\delta)]^2 \tag{15}$$

$$A(n,k) = \left|\sum_{m=-\infty}^{\infty} x(m)w(nL-m)\mathrm{e}^{-\mathrm{j}\frac{2\pi}{L}km}\right| \tag{16}$$

其中,$X(m)$ 为离散时域信号,$w(m)$ 为窗函数,L 为窗长度,K 为离散傅里叶变换的个数,δ 是为了避免溢出增加的非常小的值,一般取 10^{-6},N 为子帧个数。

2.1.7 子帧子带能量比

子带能量比指某个子带的能量与所有子带能量和的比值,可以衡量音频信号频域分布的特征,其计算方法如下:

$$p\_radio(j) = \frac{p(j)}{\sum_{k=1}^{N} p(k)} \tag{17}$$

其中,$p(j)$ 表示第 j 个子带能量,N 表示子带的个数。

2.2 神经网络预测模式类别

第 1 种方案使用神经网络替换原开环模式选择的决策树算法,直接预测模式的类别。神经网络有 3 层,每一层分别有 29、8、1 个节点,对于隐含层数的确定,首先,取输入层的三分之一及前后波动的 5 个节点,共 11 种情况进行训练与测试,然后取分类准确率最高的节点个数作为最终的节点个数,最后确定为 8。网络训练的数据包括语音、音乐信号,使用监督学习方式,神经网络所需要的输出是闭环选择的模式。因此,本文目标就是对 256 点子帧,使用 29 个特征参数作为输入,以及拟合闭环模式。网络训练中的测试数据包括语音、音乐信号,用来衡量基于神经网络的开环模式选择的性能。

2.3 神经网络预测信噪比

第 2 种方案使用神经网络首先估计音频信号的信噪比,然后通过比较信噪比大小来间接预测模式的类别。这种方案实际上模仿了 AVS-P10 闭环模式选择的逻辑与结构,与 AVS-P10 闭环模式选择尝试编码得到的信噪比不同的是,本文使用神经网络预测信噪比。更确切地说在表 1 开环模式选择的第 1、2、3、4、6、7、8、9 步被训练的神经网络预测得到的平均信噪比取代,训练的标签为闭环模式选择中 ACELP256 和 TVC256 2 种编码模式的信噪比,神经网络的输入与第 1 种方案相同。其他步骤跟闭环模式相同,对中帧(512 样点)和超帧使用变换域编码模式尝试编码,最后比较各种模式组合的信噪比大小,选择信噪比最大的模式组合作为最佳编码模式。该方案的详细过程如表 3 所示。

表 3 基于神经网络信噪比估计的模式选择

步骤	第 1 个 256 点帧	第 2 个 256 点帧	第 3 个 256 点帧	第 4 个 256 点帧
1	ACELP256 信噪比			
2	TVC256 信噪比			
3		ACELP256 信噪比		
4		TVC256 信噪比		
5	TVC512			
6			ACELP256 信噪比	
7			TVC256 信噪比	
8				ACELP256 信噪比
9				TVC256 信噪比
10			TVC512	
11	TVC1024			

3 实验结果与分析

本文从客观和主观 2 个方面对实验结果进行评价,客观测试包括分类准确率测试、信噪比测试、时间复杂度测试;主观测试主要采用 MUSHRA[14] 评测方法。本文使用的训练集和测试集都包括语音和音乐信号,语音主要包括中文普通话、英语、德语语言对话,音乐包括各种纯乐器、混合乐器演奏。其中训练集时长 3 h39 s,共 657 728 帧,每帧 256 样点,训练中网络的学习速率为 0.01,迭代 10 000 次,目标精度为 0.01,训练的终止条件为达到目标精度或者迭代次数;测试集时长 2 h5 s,共 375 680 帧,每帧 256 样点,测试集与训练集没有交叉。

3.1　客观表现评估

3.1.1　分类准确率测试

表 4 提供了原 AVS-P10 开环选择、神经网络预测模式类别、神经网络预测信噪比 3 种方法相对于原 AVS-P10 闭环选择的分类准确率。由表 4 可以看出,本文提出的 2 种神经网络方法比原 AVS-P10 开环选择方法的分类准确率都有很大程度的提高,语音分类准确率分别提升了 5.96% 和 18.07% ,音乐分类准确率分别提升了 3.84% 和 20.29% 。由表 4 还可以看出,神经网络预测信噪比方法对语音的分类准确率最高,神经网络预测模式类别方法对音乐的分类准确率最高。

表 4　以闭环模式选择为参照的准确率　%

类别	原 AVS-P10 开环	神经网络预测模式类别	神经网络预测信噪比
语音	63.20	69.16	81.27
音乐	68.50	72.34	88.79

3.1.2　信噪比测试

表 5 提供了原 AVS-P10 闭环选择、原 AVS-P10 开环选择、神经网络预测模式类别、神经网络预测信噪比 4 种方法的信噪比,结合表 4 可以看出,算法的信噪比与分类准确率成正比关系,分类准确率越高,信噪比越高,对应的编码质量就越高。神经网络预测模式类别算法的信噪比虽然也有一定的提高,但是与原 AVS-P10 闭环相比也有一定的差距;神经网络预测信噪比算法的信噪比分类准确率高,信噪比与原 AVS-P10 闭环相当。

表 5　信噪比测试结果　dB

类别	原 AVS-P10 闭环	原 AVS-P10 开环	神经网络预测模式类别	神经网络预测信噪比
语音	9.917	7.273	8.389	9.571
音乐	11.795	8.512	9.893	11.624

3.1.3　时间复杂度测试

为了更加精确地计算各种算法的时间复杂度,本文将神经网络训练得到网络集成到 AVS-P10 源代码中,统计各种算法在相同输入的情况下所使用的时间。用于测试的 PC 机的 CPU 为 Intel Pentium G640 处理器,CPU 主频为 2.8 GHz,内存容量为 4 GB,操作系统为 Windows7 x64 家庭版,编译器为 VS2010,测试的数据为 1.59 GB 的单声道音频文件,码率为 24 Kb/s。统计的编码时间如表 6 所示,结果表明,本文方法计算复杂度与 AVS-P10 开环选择方式相当。

表 6　时间复杂度测试结果　s

方式	时间
原 AVS-P10 闭环	2 411.24
原 AVS-P10 开环	1 326.05
神经网络预测模式类别	1 389.32
神经网络预测信噪比	1 492.46

3.2　主观表现评估

MUSHRA 方法能够通过较小的测试规模,对中等质量的音频信号进行评价[15]。MUSHRA 的打分范围为 0 分 ~100 分,分数越高,音质越好。本文的测试序列采用 MPEG 12 个标准测试序列,该测试序列被广泛用于评估编码器的主观质量。每组测试包含闭环选择编码的重建信号、开环选择编码的重建信号和本文所提出的 2 种神经网络方法编码的重建信号,所有编码方法均采用 24 b/s 码率进行编码。参与打分的测试人员均来自实验室音频组有一定测试经验的学生,共 10 名(5 男 5 女),为了排除部分有失偏颇的比较极端的分数,对于每个测试序列,取 95% 的置信区间,然后取平均值得到最终分值。测试结果如图 3 所示。

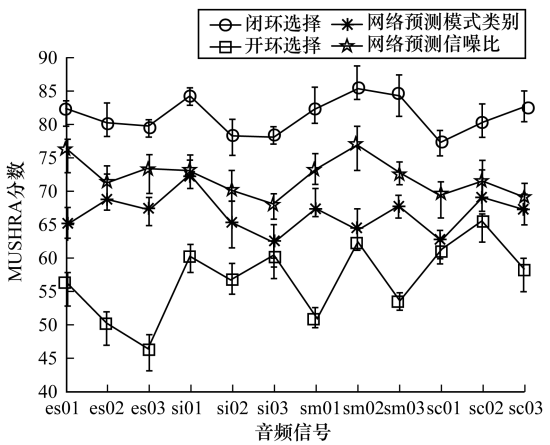

图 3　MUSHRA 主观表现评估测试结果

从图 3 可以看出,本文提出的 2 种神经网络方法较原 AVS-P10 开环模式选择的 MUSHRA 得分大幅提高。

4　结束语

本文提出 2 种基于神经网络的低复杂度的 AVS-P10 模式选择方案。第 1 种方案使用神经网络替换原开环模式选择的决策树算法,拟合闭环模式选择结果进行训练,最后得到模式选择分类器。基于神经网络的开环模式选择(以闭环模式选择为标准)与原标准模式下的开环模式选择相比,语音分类准确率提升了 5.96% ,音乐分类准确率提升了 3.84% 。第 2 种方案

按照闭环模式选择的逻辑过程,使用神经网络预测输入的信号在 ACELP256 和 TVC256 2 种编码模式的信噪比来取代编码尝试计算得到的信噪比。第 2 种方案全复杂度的闭环模式选择降低了 38% 的复杂度,编码质量比原闭环模式有了较大的提高。下一步将预测中帧或超帧的编码模式,并对 AVS-P10 开环模式进行优化,降低时间复杂度。

参考文献

[1] 数字音视频编解码技术标准工作组. GB/T 20090——2013 信息技术先进音视频编码第 10 部分移动语音与音频编码标准[S]. 2013.

[2] BERND G, PETER V. High rate data hiding in ACELP speech codecs [C]//Proceedings of ICASSP'08. Washington D. C., USA: IEEE Press, 2008: 4005-4008.

[3] 蒋三新. 混合音频信号的压缩与重建方法研究[D]. 上海: 上海交通大学, 2015.

[4] LI X, ZHAO L, WEI L, et al. Deep saliency: multi-task deep neural network model for salient object detection[J]. IEEE Transactions on Image Processing, 2016, 25(8): 3919-3930.

[5] LECOMTE J, RICHARD G, LEFEBVIR R. An improved low complexity AMR-WB + encoder using neural networks for mode selection[C]//Proceedings of Audio Engineering Society Conference. Washington D. C., USA: IEEE Press, 2007: 125-132.

[6] LING X U, HUANG B, YANG Z Q. AMR-WB +: a new audio coding standard for 3rd generation mobile audio services[J]. Audio Engineering, 2008, 2(2).

[7] 荣 康, 涂卫平, 姜 林. 基于随机森林的 AVS-P10 开环编码质量优化算法[J]. 计算机工程与应用, 2016, 52(24): 57-61.

[8] KIRA K, RENDELL L A. The feature selection problem: traditional methods and a new algorithm[C]// Proceedings of the 20th National Conference on Artificial Intelligence. [S. 1.]: AAAI Press, 1992: 129-134.

[9] 李晓岚. 基于 Relief 特征选择算法的研究与应用[D]. 大连: 大连理工大学, 2013.

[10] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[11] VAPNI K, VLADIMIR N. The nature of statistical learning theory[M]. Berlin, Germany: Springer, 1995.

[12] 白 亮. 音频分类与分割技术研究[D]. 长沙: 国防科学技术大学, 2004.

[13] LU L, LI S Z, ZHANG H J. Content-based audio segmentation using support vector machines [C]// Proceedings of IEEE International Conference on Multimedia and Expo. Washington D. C., USA: IEEE Press, 2001: 749-752.

[14] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.

[15] BREIMAN L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

编辑 索书志

(上接第 255 页)

[7] HARE S, SAFFARI A, TORR P H S. Struck: structured output tracking with kernels[C]//Proceedings of the 28th International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2011: 201-212.

[8] ZHANG K, ZHANG L, YANG M H. Real-time compressive tracking [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2012: 864-877.

[9] KALAL Z, MIKOLAJCZYK K, MATAS J. Tracking-learning-detection [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(7): 1409-1422.

[10] CHAI C, WANG Y L. Face detection based on extended haar-like features[C]//Proceedings of IEEE Conference on Mechanical and Electronics Engineering. Washington D. C., USA: IEEE Press, 2010: 159-168.

[11] TREFNYJ M J. Extended set of local binary patterns for rapid objection detection[C]//Proceedings of Computer Vision Winter Workshop. Nove Hrady, Czech Republic: Czech Pattern Recognition Society, 2010: 1589-1596.

[12] WANG X, HAN T X, YAN S. An HOG-LBP human detector with partial occlusion handling [C]// Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE Computer Society, 2009: 32-39.

[13] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, USA: [s. n.], 2001: 511-518.

[14] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to Boosting[J]. Journal of Computer and System Science, 1997, 55(1): 119-139.

[15] ZHANG Huaxun. A classifier training method for face detection based on Adaboost [C]//Proceedings of International Conference on Transportation, Mechanical and Electrical Engineering, 2011, 28(7): 240-244.

编辑 索书志