**歌声合成软件Vocaloid的介绍**

Vocaloid是日本乐器制造商雅马哈公司与西班牙庞贝法布拉大学的MTG小组合作开发的电子音乐制作语音合成软件. 它使用频域式歌唱声纹接合与整形技术, 先采集人类声音的标本, 然后在制作歌声数据库, 最后通过输入音调和歌词, 就可以合成为原为人类声音的歌声.

从生成属性上看, Vocaloid合成的音乐属于自然媒体. 该软件合成的音乐存储形式与普通的音频相同.

从感觉器官上分类, 这种媒体属于听觉媒体, 因为它的表现形式为声音.

Vocaloid需要人类声音作为音源库, 所以需要将声音这种模拟信号转换成数字信号输入到计算机中. 声音的数字化需要经历三个阶段: 采样, 量化, 编码. 采样是将时间上的连续模拟信号在时间轴上离散化的过程, 为了不失真, 采样频率需要大于声音最高频率的两倍. 量化将取值的每一个样本转换为离散值表示. 该离散值使用二进制表示. 二进制位数反映度量声音波形幅度的精度. 位数越大, 精度越高, 声音质量越好, 需要的存储空间也就越大. 编码是数字化的最后一步, 为了方便计算机的存储和处理, 需要对量化得到的二进制数字进行编码, 减少数据量.

Vocaloid合成的歌声在计算机中的存储和普通音频文件没有差别. 以wav文件为例介绍它在计算机中的表示. wav文件内容以chunk为最小单位, 一般由3个区块组成: RIFF chunk, Format chunk, Data chunk. 三个区块都以标识和长度作为开始, 标识为字符串, 分别是“RIFF”, “fmt”, “data”. Format区块记录音频格式, 声道数, 采样率, 每秒数据字节数, 数据块对齐, 采样位数. Data区块记录音频数据.

Vocaloid通过三层建模对声音进行合成.

Low Level: Vocaloid制作音源库时使用SMS技术将语音拆分成若干不同频率和幅度的正弦波和气音. 合成时给定控制点, 生成若干正弦波和气音叠加.

Mid Level: EpR语音模型生成谐波和气音的频谱包络和相位. 它把语音的频谱包络视为一条激励曲线和一条共振曲线的和. Vocaloid的音源库包括了大量的EpR参数, 通过在合成中修改这些参数即能实现时间缩放, 音高变换, 发音过渡, 音色修改.

High Level: Vocaloid使用Sonic Space将乐谱转换成EpR参数. 它包含High Level和Mid Level的数据库, 通过算法在这两层的数据库之间进行匹配.

在输出设备上展现歌声的过程可视为声音采样的逆过程, 这个过程将离散的数字信号转换成连续的模拟信号.

参考文章:

Vocaloid维基百科 <https://zh.wikipedia.org/wiki/VOCALOID>

Vocaloid原理揭秘 <http://bbs.ivocaloid.com/thread-126991-1-1.html>

声音信号的数字化 <https://blog.csdn.net/HaoDaWang/article/details/79347562>

wav文件格式详解 <https://www.jianshu.com/p/947528f3dff8>