CVE3.5 算法构想 v0.1

Sleepwalking

2014.1.5

0. 前言

CVE3 引擎在变调和辅音处理方面表现不尽人意，我决定放弃 CVE3，编写下一代引擎，暂定版本号 CVE3.5。

这是 CVE3.5 将使用的算法的初步构想，意味着最终使用的算法可能会和这里提出的不同，或作出一定改进。

这篇文章建构在 Dr. Xavier Serra 的论文基础上：

Serra, X. (1989). A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition. Dept. of Music / CCRMA.

可以在 UPF 网站上下载到：

http://mtg.upf.edu/content/serra-PhD-thesis

这篇文章仅包含了构想的梗概，没有包含实现的细节。在 1、2、3 节中我在 SMS 基础上扩展出一个具有更高还原度并支持辅音的模型和语音重构方法。第 4 节中我使用这个模型对语音进行变调、长度调整、音色调整和音节过渡。

这里提出的构想没有经过完整的测试，所以我下一步是用 Octave 逐步对其进行验证。

1. 背景

CVE3.5 将使用一种改进的 SMS(Spectral Modeling Synthesis)算法。下面简要介绍一下 SMS 使用的 DSM(Deterministic plus Stochastic Model)语音建模方法。

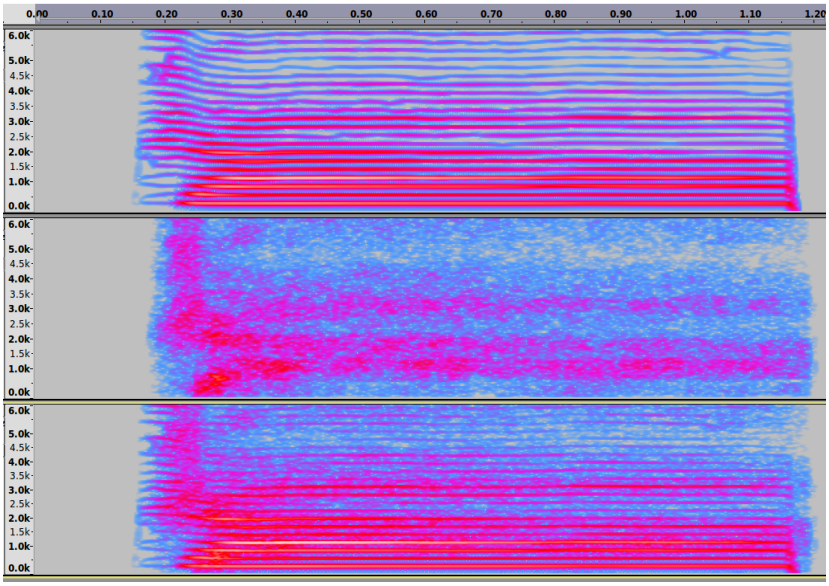
这玩意似乎国内玩的人比较少，暂没找到中文译名，下面穿插的英文可能有些蛋疼。

DSM 将一段语音近似为 Deterministic 部分和 Stochastic 部分的组合。前者由若干正弦波组成；后者是一段通过滤波器的噪音。对于周期性明显的语音信号，Deterministic 部分可视为**谐波**的组合。

Deterministic 和 Stochastic 部分是通过对语音的短时傅立叶分析获得的。

SMS 合成中，Deterministic 部分直接靠一系列余弦函数的和求出；Stochastic 部分可以是经过时变滤波器的白噪音，也可以用 STFT 合成。将两部分混音可得到听觉上接近原始信号，但数据上和原始信号不同的信号。

广义一些，DSM 是一种 Sinusoidal Plus Residual Model，因为 Deterministic 是正弦波构成的，Stochastic 就是除了 Sinusoidal 以外剩下的部分了（**我个人理解**）。



上：Sinusoidal(Deterministic)

中：Residual(Stochastic)

下：Sinusoidal + Residual 的 SMS 重构

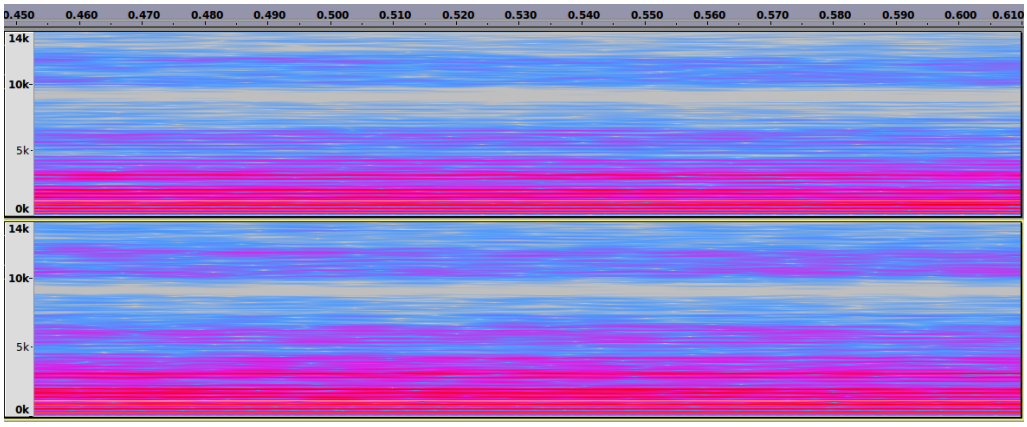
2. 改进的 SMS 算法

请注意 SMS 重构的信号仅在**听觉上**和原始信号**相近**，就是说 SMS 重构的信号和原始信号听起来是有细微差别的。①经测试 SMS 合成的语音有轻微的平淡感，不如原始语音清脆。② SMS 的另一个缺点在于无法对辅音部分以及辅音和元音衔接部分建模。

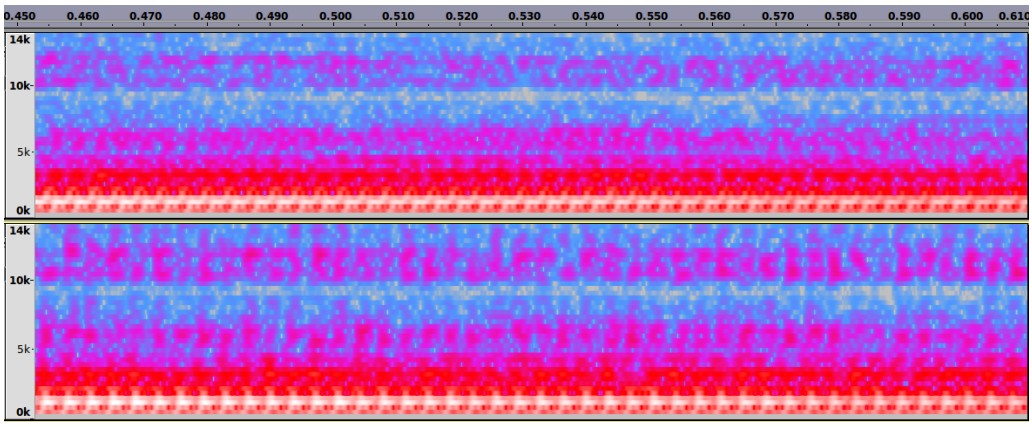
对于问题②，初步想法是把辅音和元音部分分开处理，这会在第三节中详细介绍。

对于问题①，我分析了 SMS 重构信号和原始信号的频谱（不含辅音部分），发现当 FFT 大于 1024 点时两张频谱几乎没有任何区别；只有在 256 点以下时可以看出区别：原始信号的频谱，尤其是 9k-15k 频段有明显的伴随周期的脉冲；而 SMS 重构的信号几乎看不出这种脉冲。这是由于声门张开闭合瞬间的振动引起的短暂信号幅度增强。因为这个增强的持续时间太短，只有靠较小的分析窗才能观察到。

改进措施是人为地随周期调整 Residual 信号的幅度，构建出周期性脉冲。经测试合成自然度有一定提高，受众几乎无法分辨出重构信号和原始信号。



1024 点 FFT 频谱，汉宁窗，上：SMS 重构信号，下：原始信号



128 点 FFT 频谱，汉宁窗，上：SMS 重构信号，下：原始信号

3. 辅音衔接处理

3.0 辅音分类和任务描述

CVE3.5 把辅音分为三类分别处理：塞音，擦音，鼻音。这不是语音学上的分类：

塞音指波形非周期性且短暂的辅音，还包括元音的起始部分。

擦音指波形非周期性且持续的辅音。

鼻音指波形呈周期性的辅音，如汉语拼音"l"、"m"。

CVE3.5 对辅音的处理包括：

改变塞音与元音衔接部分的音高。

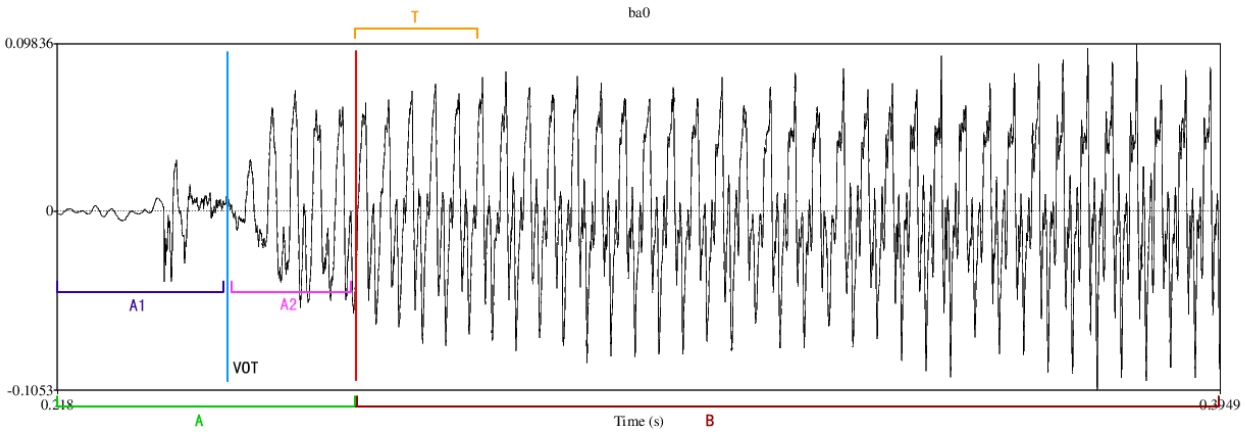
改变擦音的长度、与元音衔接部分的音高。

改变鼻音的长度、鼻音的音高、与元音衔接部分的音高。

另外还有对于各类辅音响度的处理。

3.1 分类：塞音的处理

这里以 Rocaloid_Cyan 音库采样中的"ba"为例：



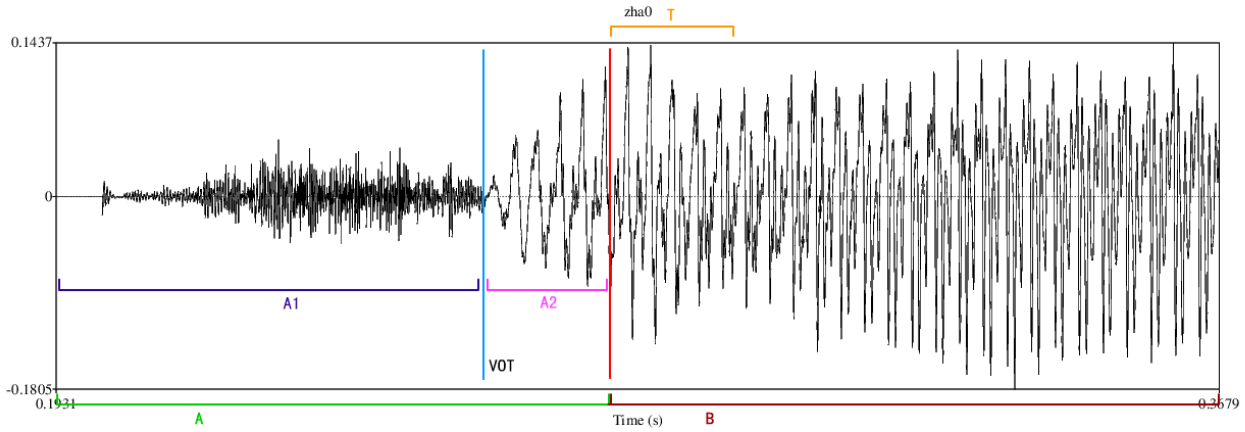
图中辅音与辅元衔接部分标注为 A，元音稳定部分为 B，过渡段为 T。

A 中 VOT(Voice Onset Time)之前的部分为 A1，VOT 后 B 之前的部分为 A2。

CVE3.5 不对 A1 部分做任何处理；A2 部分的音调调整通过 PSOLA 实现；B 部分使用 SMS 合成。在 B 部分开头记录下各谐波相位，作为 SMS 合成的初始相位，以确保过渡段 T 的相位同步。

3.2 分类：擦音的处理

这里以 Rocaloid_Cyan 音库采样中的"zha"为例：



与塞音基本相同，图中辅音与辅元衔接部分标注为 A，元音稳定部分为 B，过渡段为 T。

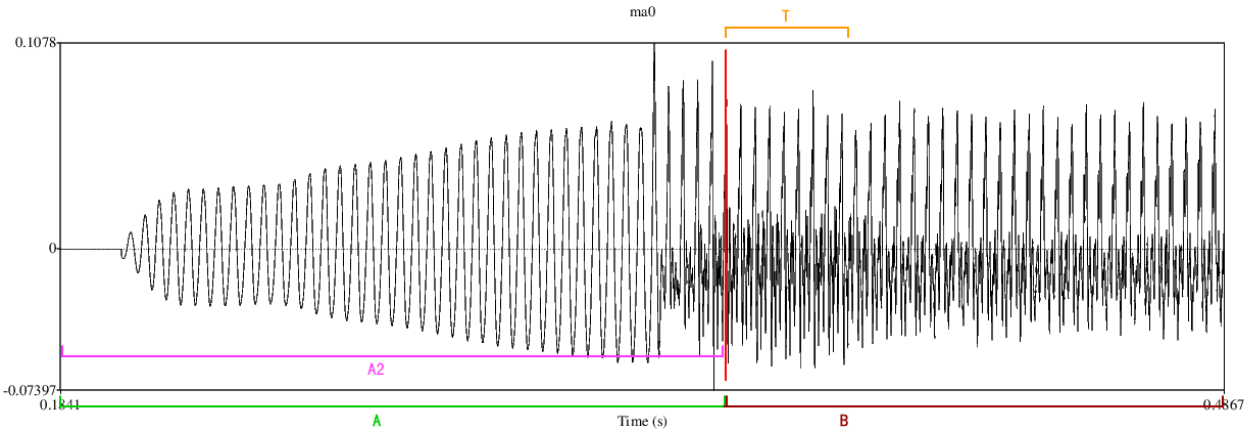
A 中 VOT 之前的部分为 A1，VOT 后 B 之前的部分为 A2。

A1 部分通过重采样实现时长伸缩，然后通过 STFT 恢复重采样后的频谱特性。

A2 部分、B 部分的处理同塞音一致。

3.3 分类：鼻音的处理

这里以 Rocaloid_Cyan 音库采样中的"ma"为例：



图中辅音与辅元衔接部分标注为 A，元音稳定部分为 B，过渡段为 T。

A 中 VOT 之前的部分 A1 段被省略，VOT 只后 B 之前的部分为 A2，占 A 部分全部长度。

A2 部分通过 PSOLA 实现时长、音高调整。

B 部分和 T 部分处理方式与塞音、擦音相同。

4. 使用改进型 SMS 进行合成

至此有了一个支持辅音的改进的 DSM 模型，下一步是将这个模型作为语音信号的中间表示方法，通过修改 DSM 参数实现语音信号的灵活转换。

4.0 音源库文件包含的信息

1. 三种辅音模式均需要完整的 A、T 部分的时域信号。
2. 三种辅音模式均需要对 A2 部分进行 PSOLA 处理，故需要记录 Voice Pulses 的位置。
3. 针对不同辅音模式记录 A1、A2、T、B 时间节点的位置。
4. B 部分中每隔一段 HopSize 记录各谐波频率和响度。
5. B 部分中每隔一段 HopSize 记录 Residual 的频谱包络。

4.1 语音信号的时间伸缩

有两种实现方式：

1. 分别在 B 部分 Sinusoid 数据帧、Residual 数据帧之间插入过渡帧。
2. 将 B 部分的尾部线性过渡至 B 部分开头。

4.2 语音信号的音高变换

A 部分的音高变换见第 3 节。

B 部分的音高变换：

Sinusoidal 部分：按比例将谐波频率缩放，维持谐波频谱包络形状。

Residual 部分：因为频谱包络形状被维持，不需处理。

4.3 语音信号的音色/共振峰调整

CVE3.5 暂不支持 A 部分的音色调整（A 部分持续时间太短，而且对辅元过渡部分作音色调整意义不大）。

所有共振峰调整的共同实现方法：

首先有一个可以产生**近似**频谱包络的函数 E(x, F1, F2, F3..., S1, S2, S3...)

对 Sinusoidal 和 Residual 部分的频谱同时除以原始采样对应位置的共振峰参数的 E(x)，然后乘以目标共振峰参数的 E(x)。

4.3.1 使用 LC-FECSOLA 作为 E 函数

参见 http://bbs.ivocaloid.com/thread-120592-1-1.html

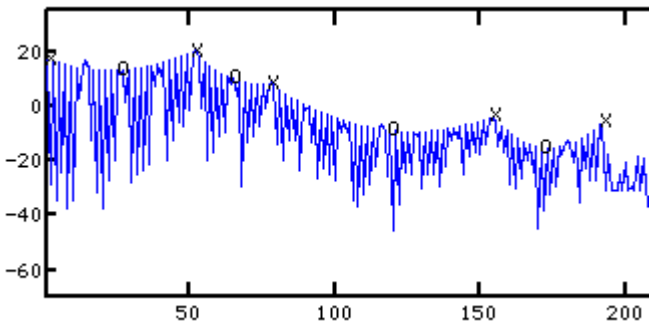
这个方案多出一个步骤：原始采样的频谱作为 E 函数的输入。

4.3.2 使用分段二次函数逼近共振峰包络

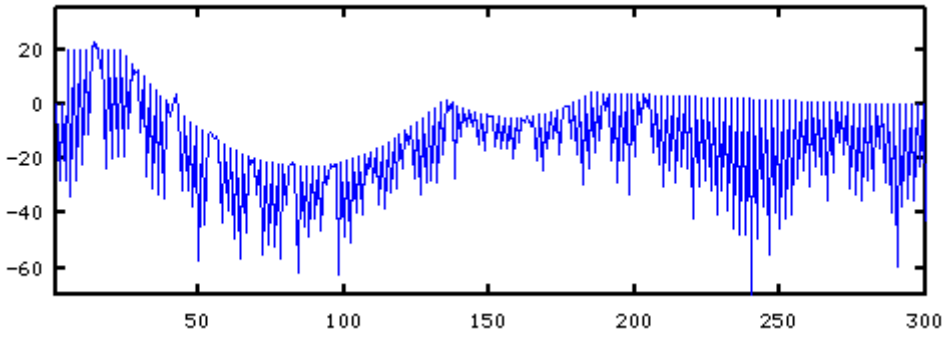
*这里的幅度为分贝频幅谱上的幅度

提供 F1, F2, F3...的频率、幅度以及共振谷的频率、幅度（作为 E 函数的参数）。

可以将频谱包络拟合为分段二次函数。段数即为处理的共振峰数量。



元音"a"



元音"i"

4.4 语音信号的过渡

在 4.3 基础上，对两个语音信号的 Sinusoidal 和 Residual 部分分别做线性插值。