

Composing with phonemes: the 'Phoneme' VST plug-in

Final project summary paper by Sandra Pauletto,
MSc in Music Technology, University of York
Project supervisors: Prof. D. M. Howard and Dr. A. Hunt
Submission date: 31st August 2001

The aim of this project is to develop a software tool for composition, which uses synthesised English phonemes as base sounds. This tool should be usable with a common desktop computer. For this reason it was decided to develop it in the form of a VST (Virtual Studio Technology) plug-in. VST plug-in are sound processes that work inside a host application such as the Steinberg sequencer Nuendo 1.5. The developed plug-in, called 'Phoneme', synthesises ten vowels and allows the user to control their interpolation and various synthesis parameters. Real-time playing and parameter automation is possible. The final musical composition can be saved as an audio file and parameter configurations can be saved as a sequencer project file.

Introduction

A phoneme is a speech sound; for instance vowels and consonants are phonemes. Vocal sounds have always been interesting for composers because they can be at the same time completely abstract sounds or reminiscent of language and the expression of human emotions. A lot of research has been done on speech, both for musical and scientific reasons. Studies on perception, analysis and synthesis of sound in general have benefited a great deal from this research. These studies also allow us to think that a timbre space can be defined with its dimensions and parameters and it can influence musical structures. Phoneme sounds are a part of this space and a very particular one because of their strong relation with human language. Many interesting books have been written on this subject. From *Sound structure in music* by Robert Erickson (1975), which explores how music structure can be derived from the

properties of sounds, to the more recent *Music, language, speech and brain* (Sundberg, 1991), which puts together articles from different scientists and composers on the relation between perception of speech and music composition. There are also many examples of compositions based on vocal sounds (*Speech songs* by C. Dodge, *The vox cycle* by T. Wishart, etc.) that testify for a great interest on the use of these sounds. Wishart writes in his book *On sonic art*:

This [interest] is partly due to the obvious immediate significance of the human voice to the human listener, but also the unique complexity of articulation of the source. The ability to produce a rapid stream of timbrally disjunct entities is uncharacteristic of any other source...The speech stream is thus an archetypal example of complexly evolving timbral morphology. (Wishart, 1985, p. 82)

The acoustic nature of the speech signal

The description of speech production is based on the source/filter paradigm (Fant, 1960). The three basic elements of this paradigm are the power source, the sound source and the sound modifiers. The power source is produced by the compressive action of the lung muscles, which results in the amplitude of the voicing source. The sound source is of two types: the vibration of the vocal folds and the turbulent flow past a narrow constriction (the noise source). These two sources produce voiced speech, voiceless speech and, if combined, mixed excitation. The sound modifiers are the articulators of the structures that enclose and interconnect the airways above the larynx. The phoneme sounds produced by this system have been classified using many labels. First, in English, there are 44

phonemes, which are divided into 20 vowels and 24 consonants. Vowels are characterized by peaks in their spectrum envelope called formants. They further divide in monophthongs and diphthongs (combinations of 2 monophthongs) and are all voiced (i.e. the vibrations of the vocal folds is the sound source). Consonants can be voiced (the sound source is a mixed excitation) or voiceless (the sound source is noise). They can differentiate in various classes depending on where a constriction is made in the vocal tract (place label) or how the noise source is produced (manner label). The manner labels are plosive (e.g. /p/), nasal (e.g. /n/), fricative (e.g. /f/), affricate (e.g. /tʃ/), aspirate (e.g. /h/) and semivowel (e.g. /l/). The place labels are bilabial (e.g. /b/), labio-dental (e.g. /v/), dental (e.g. /t/), alveolar (e.g. /d/), post-alveolar (e.g. /r/), palatoalveolar (e.g. /ʒ/), velar (e.g. /N/), glottal (e.g. /h/).

The synthesiser

To develop a software composition tool requires addressing various problems. First, a decision has to be made on what synthesis technique to use. In fact, scientists developed many different ways to synthesise vocal sounds. Examples of techniques are the FOFs or Formant Wave Functions (developed by X. Rodet at Ircam in 1984), Linear Predictive Coding (LPC) (Atal, 1970; Makhoul, 1975; in Cook, 1996) or formant synthesisers (Rabiner, 1968; Klatt, 1980; in Cook, 1996). For this project the model of the Klatt synthesiser (Klatt, 1980) has been chosen because of the numerous parameters it can allow for user control. This synthesiser models the source/filter paradigm and it combines two different configurations: the cascade and the parallel configuration. Two voicing sources are coded. The voicing source is simulated by a pulse wave, while a pseudonumber generator with a Gaussian distribution simulates the noise source. The voicing source is sent through a cascade of filters or resonators, which adequately shape the voicing source and simulate the formant peaks. The noise source is sent mainly to the parallel configuration (the formant filters are connected in parallel) and it is modulated by the voicing source for voiced consonants. The cascade

configuration is best used for synthesising vowels and it cannot synthesise consonants, with exception of nasals if the noise source is introduced. The parallel configuration can synthesise both consonants and vowels, but more parameters have to be controlled in this configuration for vowel synthesis.

VST: Virtual Studio Technology

Virtual Studio Technology was created by Steinberg in 1996 to produce a framework for programming audio processes that can be used inside a host application. This has been possible because desktop computer processors nowadays can handle hundreds of thousands of samples per second and therefore produce CD quality audio. The language used for programming VST plugins is C++. The VST Software Development Kit (SDK) can be downloaded from the Internet for free and it provides developers with plug-in examples, which illustrate the framework in which a plug-in is coded. The main characteristics of a VST plug-in are that it is an audio process used inside a host application, it uses the computer processor (it does not need a dedicated Digital Signal Processor or DSP), the audio process is platform independent, on the Windows platform a VST plug-in is a Dynamic Link Library (.dll) file. A VST plug-in uses the base classes provided in the SDK called `AudioEffect` and `AudioEffectX`. The programmer does not need to code a graphical user interface to test the audio process because the host application can already produce a default user interface, which represents all the user's parameters with sliders or knobs. If the programmer wishes to create a different user interface, he has to introduce an editor that might make the plug-in dependent on the platform. VST plug-ins can be of two types: effects or instruments. An effect takes samples from an audio file and performs a process, while in an instrument the sound is synthesised by the plugin and it can be played using plug-in automation tracks or via MIDI.

The 'Phoneme' plug-in

The 'Phoneme' plug-in is a VST instrument because the sound is synthesised by the plug-in itself and it can be played in real-time with the mouse or using the plug-in automation tracks. In

the automation tracks, parameter values can be drawn, edited and played back before the final audio file is exported. At this stage of its development, the plug-in can synthesise 10 English vowels. The values for formant frequencies and bandwidths of the correspondent resonators and ranges of fundamental frequency are taken from the article Software for a cascade/parallel formant synthesiser by D. Klatt (1980). The user can interpolate between the vowels in three different ways (linear, sinusoidal or with a complete sinus cycle). The volume of the voicing source can vary. The fundamental frequency of the vowels can change linearly from 50Hz to 500Hz (male speech range) or following a semitone's scale (from F2 to C5). Vibrato can be added and its depth and rate can vary between 0 to 5 semitones and 1Hz to 7Hz respectively. The singing formant can change as well. Singers can enlarge the pharynx cavity and lower the glottis when they sing. The result is that the 4th and 5th formant frequencies get closer together (Dodge and Jerse, 1997). The singing formant parameter varies from getting the 4th and the 5th formant frequency close to the 3rd to linearly part them away. More than one plug-in can be played at the same time so that horizontal and vertical relations between sequences can be controlled. The voicing source is produced by a band limited-pulse generator, where the number of present harmonics stays below the Nyquist frequency (half the sampling rate, which is 44100 samples per second). This is to avoid aliasing distortion. Then the samples are sent through the cascade configuration of the Klatt synthesiser. The first resonator and antiresonator shape the voicing source, and the next five resonators simulate the first five formants. A high pass filter at the end simulates the radiation characteristics, i.e. how the sound is radiated from the head. The plug-in uses the default graphical interface supplied by the host application. Every parameter is displayed as a slider. In certain cases though, the user parameter would be better displayed as a button (for example when choosing between the three different types of interpolation). The solution of this problem has been dividing the slider in three different zones (beginning, middle and end) that identify the three options. If the number of

options becomes much bigger, another solution should be provided.

Conclusions

A VST plug-in has been developed which synthesises ten English vowel sounds interpolates between them and gives to the user control over various synthesis parameters. This plug-in is best used in the Steinberg sequencer Nuendo 1.5. Real-time playing using the mouse and playback through sequencer automation tracks is possible. More than one plug-in can be played at the same time and a final audio file can be exported at the end. The plug-in already allows the production of a complete composition, but in many areas it could be improved. In particular, consonant sounds can be implemented with the relative parameters, a more adequate user interface could be implemented and the use of a MIDI interface could be considered. Finally, the implementation of an analysis stage could give the user the possibility of analysis, modification of parameters and resynthesis, increasing the creative use of this tool for composition.

References

- Cook, P. R.**, 1996, Singing voice synthesis: History, current work, and future directions, *Computer Music Journal*, 20:3, pp. 38-46
- Dodge, C. and Jerse, T. A.**, 1997, *Computer Music: synthesis, composition, and performance*, Schirmer Books
- Erickson, R.**, 1975, *Sound structure in Music*, University of California Press
- Fant, C. G. M.**, 1960, *Acoustic theory of speech production*, The Hague: Mouton
- Klatt, D. H.**, 1980, Software for a cascade/parallel formant synthesizer, *Journal of Acoustical Society of America*, 67:3, pp. 971-995
- Rodet, X.**, 1984, Time Domain Formant-Wave-Function Synthesis, *Computer Music Journal*, 8:3
- Sundberg, J.**, 1991, *Music, Language, Speech and Brain*, Nord and Carlson
- Wishart, T.**, 1985, *On sonic art*, York: Imagenearing Press. Republished by Harwood Academy Publishers in 1996