REAL-TIME SINGING SYNTHESIS USING A PARALLEL PROCESSING SYSTEM

I.S. Gibson, D.M. Howard & A.M. Tyrrell

Department of Electronics, University of York, Heslington, York YO1O 5DD. UK.

Abstract

Existing singing synthesis systems are usually monophonic and implemented using single processor systems. A polyphonic (multi-voice), formant-based synthesiser has been designed which addresses the needs of real-time performance. Singing synthesis, as opposed to speech synthesis, is highly dependent on the control of score parameters such as event onset times and fundamental frequency. The novel system described in this paper is designed to allow control of such parameters using the MIDI protocol and a graphical user interface. The hardware consists of a PC, ADSP-21060 SHARC processors and a Xilinx FPGA chip. This paper presents an introduction to the system, a description of the model used and the method of implementation on a parallel system. The results show processing requirements for polyphonic sound synthesis using the synthesis model.

1 Introduction

Digital singing synthesis techniques are often calculation-intensive. This research aims to produce a real-time polyphonic singing synthesiser. Both synthesis models and user interfaces have been considered.

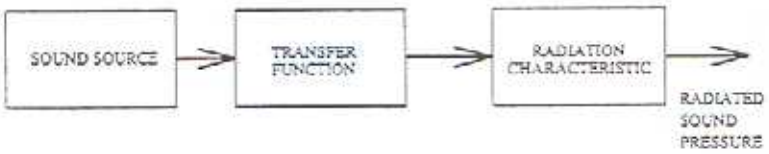

Figure 1: A source/filter model for speech

The synthesis of speech is generally accomplished using a *source/filter* model [Fant G, 1960] (figure 1). The vocal system produces three types of sounds: voiced, unvoiced and a combination of these two. Voiced sounds are produced when the lungs create an airstream which is passed through the glottis either causing the vocal folds to vibrate (for a voiced sound) or causing turbulent airflow through a narrow constriction (for unvoiced sound). The acoustic result of vocal fold vibration is considered to be the *source*. The signal is modified by the vocal tract (the *filter*), in particular the *articulators* (e.g. lips, jaws, tongue, pharynx). These change the volume of the cavities in the vocal tract and the acoustic responses (or *formants*) associated with them.

Existing models of the vocal tract connect the formants in series [Klatt D, 1980] or parallel [Holmes J.N., 1987]. Source models usually involve cycling through a

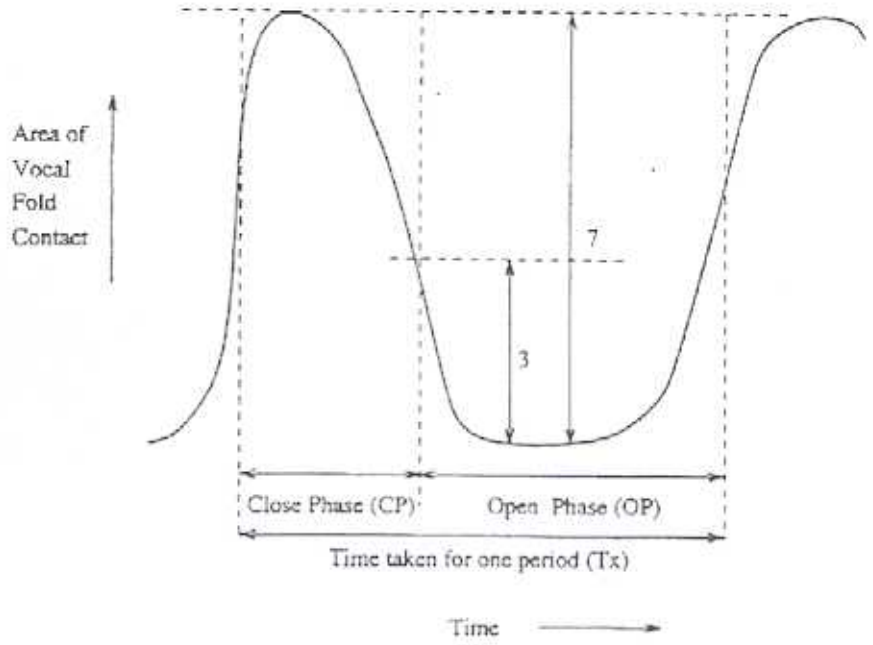

Figure 2: A glottal waveform showing closed phase and open phase portions

wavetable which is pre-defined but which may be varied in frequency and amplitude. A white noise generator is sometimes implemented for unvoiced speech.

The wavetable for voiced speech is usually based upon activity of the vocal folds (figure 2). The closed phase (CP) and open phase (OP) portions of the waveform correspond to periods of vocal fold contact and separation. OP begins at a point where the negative-going lx crosses a level corresponding to $\frac{3}{7}$ ths of that cycle's peak to peak amplitude [Davies et al, 1986]. On output, the fundamental frequency is varied by skipping samples in the wavetable. For example, a wavetable of 256 samples, played at 44 kHz will produce a fundamental frequency of approximately 172 Hz.

Vowel	f1	f2	f3	f4	f5
A	609	1000	2450	2700	3240
E	400	1700	2300	2900	3400
IY	238	1741	2450	2900	4000
O	325	700	2550	2850	3100
OO	360	750	2400	2675	2950
U	415	1400	2200	2800	3300
ER	300	1600	2150	2700	3100
UH	400	1050	2200	2650	3100

Table 1: Five formants frequencies (Hz) of eight vowels sung by a male

The filter section of the synthesis model usually provides a number of resonances with time varying centre frequencies and bandwidths. The remaining formants (usually a total of 6 is common) are fixed. Some typical formant values for the male singing voice are shown in table 1.

Typically, real-time singing synthesis systems offer control over: fundamental fre-

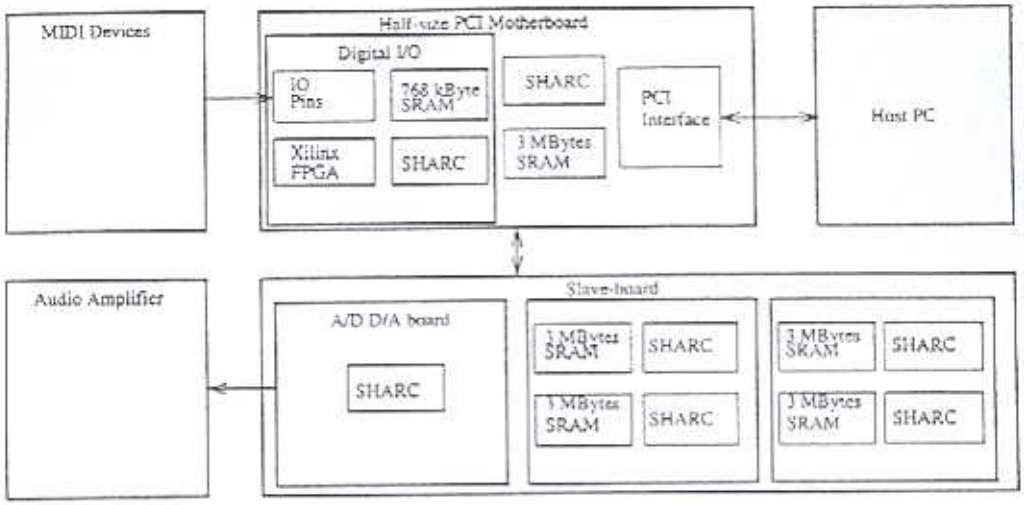

Figure 3: The Singing Synthesis System

quency, voice quality (e.g. Open Quotient), the glottal spectrum, vibrato and formant frequencies. Although existing real-time singing synthesis systems offer considerable control to the performer, generally they do not offer polyphonic output due to processing limitations of single processor systems. Therefore real-time polyphonic singing synthesis is ideally suited to parallel processors, whereby each voice can be implemented on one or more processors.

1.1 The Hardware

The design of the singing synthesis system is shown in figure 3. Data is input to the network using a graphical user interface on the Host PC and through MIDI devices [IMA, 1988] such as keyboards and hardware controllers. Sound data is output through a D/A board to an audio amplifier. ADSP-21060 SHARC processors are used to process the audio data.

The ADSP-21060 SHARC is a 32-bit processor. It has 4 Mbits of dual ported on-chip SRAM (organised as two blocks of 2Mbits each). The memory can be configured as either 128K words of 32-bit data, 256K words of 16-bit data, 80K words of 48-bit instructions, or combinations of word sizes up to 4 megabits. Memory may be accessed with direct and indirect addressing. Up to 6 processors can access each others internal RAM and I/O registers at a rate of up to 240 Mbytes per second.

There is a 25 ns instruction cycle time with an operating time of 40 MIPS. An on-chip instruction cache enables two operands and an instruction to be fetched for every cycle (a three-bus operation). The cache is selective, only storing instructions which conflict with program memory data accesses. A 1024-point complex FFT executes in 0.46 ms. Both fixed and floating point data formats are supported.

Off-chip memory and peripherals may be accessed using the external port. An address space of 4-gigawords is provided in unified address space. The host interface allows connection to 16 and 32-bit microprocessor buses. Four channels of DMA are available for the host interface.

2 Implementation Of The Synthesis Model

This section describes how singing synthesis model has been designed to run on a network of ADSP-21060 SHARC processors and an FPGA processor.

2.1 The Synthesis Model

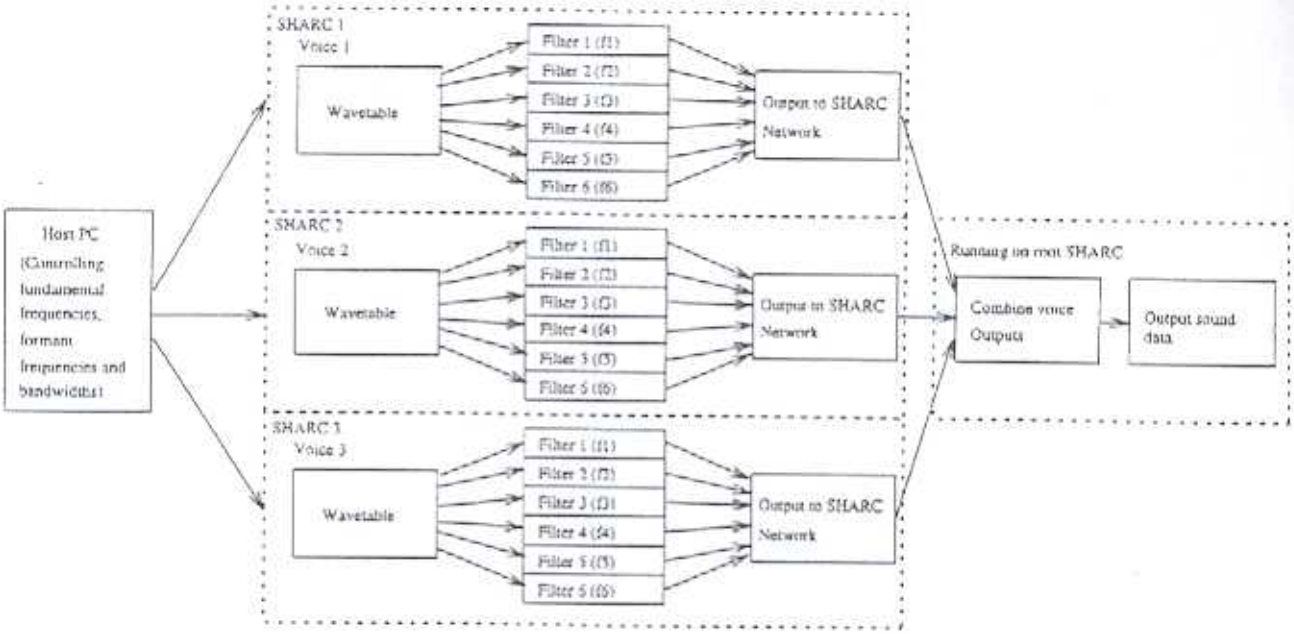

Figure 4: Design for a three voice, six formant singing synthesiser

A parallel formant synthesiser has been implemented in C running on a SHARC network (figure 4). Each voice is generated using a single ADSP-21060 SHARC. The model consists of a wavetable generator followed by a series of bandpass filters arranged in parallel. A wavetable is held in memory and used for the digital oscillator or vocal source.

Each bandpass filter has a centre frequency and bandwidth corresponding to typical formant values for a singing voice. The resulting samples from each of the bandpass filters are combined to produce the output waveform for one voice. The output from all of the voices are combined and sent to the ADSP-21060 root processor. A soundfile is created or sound is played through the soundcard.

2.2 Interfaces To The SHARC network

The AOAC509 Programmable Digital Interface is to be used for digital input and output. This module consists of a ADSP-21060 SHARC and a Xilinx 4K FPGA. The FPGA is mapped to the SHARC with an IO bandwidth of up to 160MBytes per second. There are 110 pins which can be programmed by the user for I/O. External devices are connected using 2mm socket strips.

The FPGA will be used to allow control input to the system using MIDI, the standard protocol for communication between music devices. It will be programmed using Handel-C. The MIDI protocol includes control information for note events,

timbre (e.g. filter cutoff and resonance parameters), real-time control of timbre and downloading of instrument patches.

It is invisiged that either a stereo or multi-channel output will be implemented using the D/A converter. A stereo system would allow the user to pan each individual singing voice to a separate position in the mix. A multi-channel system would send voices to an external mixing desk where further processing of the voices would be undertaken (panning, volume levels, EQ, reverberation etc).

3 Results

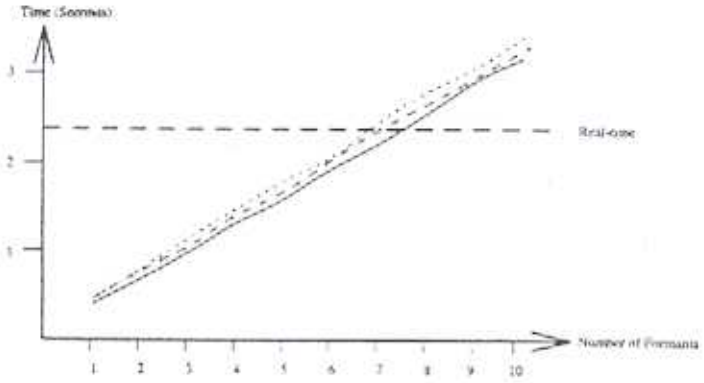

Figure 5: Timings for synthesis of 1 voice (solid line), 2 voices (dashed line) and 3 voices (dotted line)

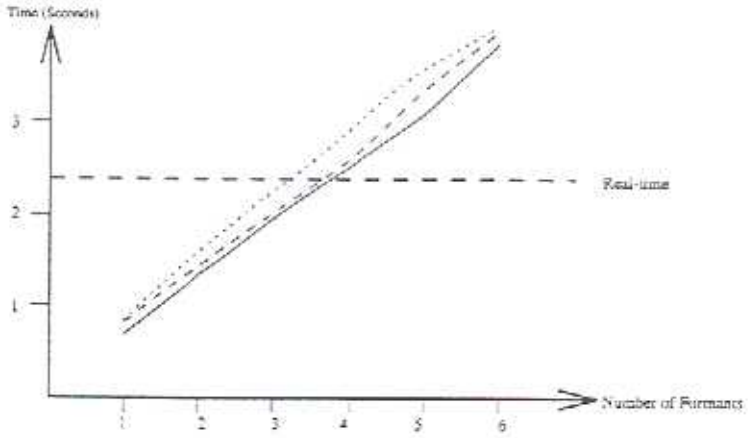

Figure 6: Timings for synthesis of 2 voices on each processor for 1 processor (solid line), 2 processors (dashed line) and 3 processors (dotted line)

The singing synthesiser is capable of synthesising three voices each with six formants in real-time. Real-time is defined as a data output rate of 44.1 kHz (CD quality audio) or more. The results show the calculation times for all sound data to reach the root SHARC node based on output of 102400 samples per voice (this being equal to 100 cycles of the 1024 word wavetable).

Figure 5 shows timings of synthesis for one to three voices. Each voice was synthesised with one to ten formants. Timings were also made running two voices on each SHARC node, each being synthesised with one to six formants. The results are

shown in figure 6. Variables such as formant frequencies, formant bandwidths and wavetable size have been found to have no effect on processing time.

4 Conclusions and Future Developments

A real-time polyphonic sound synthesiser has been successfully designed and implemented using ADSP-21060 SHARC processors and an FPGA. Since the algorithm for each voice is identical, polyphony can easily be increased by adding further nodes to the network.

A comprehensive user interface is to be implemented, the design of which is based upon results from a user survey. A GUI will allow mapping of control parameters to synthesis parameters through a Windows interface. Consideration will be given to designing the interface for use in real-time performance by musicians.

A white noise generator may be implemented in the model and used to generate unvoiced sounds. The synthesiser may require two SHARC processors to synthesise a single voice, one for generating sound sources (for voiced and unvoiced sounds) and the other for processing the filter functions. The focus of current work is to explore the feasibility of implementing a more comprehensive model.

5 Acknowledgements

This project is supported by EPSRC grant GR/K 72490

References

[Cook P, 1993] COOK P.R. (1993). SPASM, A Real-Time Vocal Tract Physical Model Controller; and Singer, the Companion Software Synthesis System. *Computer Music Journal*, 17, (1), 30-43.

[Davies et al, 1986] DAVIES P., LINDSEY G.A., FULLER H., FOURCIN A.J. (1986). Variation of glottal open and closed phases for speakers of English. *Proceedings of the Institute of Acoustics*, 8, (7), 539-546.

[Fant G, 1960] FANT G. (1960). *Acoustic Theory of Speech Production*, The Hague: Mouton.

[Holmes J.N., 1987] HOLMES J.N. (1987) A Parallel Formant Synthesiser For Machine Voice Output. In Fallside, F., and Woods, W.A.(Eds.): Computer Speech Processing, Cambridge: Cambridge University Press.

[IMA, 1988] IMA (1988). MIDI 1.0 Detailed Specification, version 4.0. International MIDI Association.

[Klatt D, 1980] KLATT D. (1980) Software For a Cascade/Parallel Synthesiser, Journal of The Acoustical Society of America, 67, (3), 971-995.