DESIGN AND IMPLEMENTATION OF A MULTI-CHANNEL FORMANT SPEECH SYNTHESIS ASIC

C. D. Summerfield

Syrinx Speech Systems Pty Ltd

ABSTRACT - This paper describes the design and implementation of a single chip multi-channel formant speech synthesis Application Specific Integrated Circuit (ASIC). The aim of the R&D project was the design and implementation a cost effective device capable of synthesising multiple channels of high quality speech. The paper describes the ASIC architecture developed to implement the complete multi-channel synthesis system on a single chip. At the core of the device is a fully synchronous bit-serial signal processing architecture which implements the formant synthesiser function. This is augmented by circuits which implement a multi-channel interactive glottal source function, delay line elements for the filter network and a flexible interface circuit which enables the device to be directly connected to an industry standard 32 bit bidirectional data bus. Using the device it is feasible to implement a complete multi-channel text-to-speech system using just two components, a microprocessor unit to run the text-to-speech algorithm and the multi-channel speech synthesis ASIC to provide the synthetic speech output.

INTRODUCTION

The past decade has seen an explosion in the application of speech synthesis technology. With the rapid expansion in Information Technology projected for the coming decade, speech synthesis is set to become a ubiquitous technology for the dissemination and communication of information. One of the major factors preventing the widespread application of the of speech synthesis at present is the relatively high cost of providing the technology to support high quality speech synthesis services. With this problem in mind, the aim of the project has been to design and implementation of a cost effective means of synthesising large quantities of high quality speech. The outcome of the project is a low cost ASIC device capable of synthesising up to 16 channels of speech at once.

Over the past 40 years, formant speech synthesis has become a well established method for generating synthetic speech output. It has been extensively used in text-to-speech systems and forms the basis for many successful commercial speech synthesis products. One of the attractions of formant speech synthesis is its ability to generate speech from an acoustical description of the speech signal. This description can be readily generated from an phonetic transcription of the speech message using principle developed in acoustic-phonetics. Although the rules governing the mapping from the phonetic transcription to the formant synthesis acoustic parameters are still the subject of much research activity, the phonetic foundations for the rules are well developed and understood. Further, there are several highly developed systems for converting unrestricted text to a phonetic transcription from which synthetic speech can be readily produced using a formant synthesiser model. Consequently, formant synthesis provides a well founded method of synthesis which is compatible with a large number of existing systems.

Most existing formant speech synthesisers are implemented using programmable Digital Signal Processing (DSP) technology. Although there has been an increase in signal processing bandwidth available

from this technology, it is still insufficient for practical implementation multi-channel formant synthesis models. Further, the high price of these devices also make programmable DSP solution expensive for large scale speech response installation envisaged for telecommunications and other services, such as speech response in the banking industry. In some applications many hundreds of high quality speech synthesis channels are required to provide sufficient access to information to make speech synthesis viable. The limited signal processing bandwidth and the relative high cost of DSP technology makes it unsuitable in such applications.

The solution adopted in this project has been to implement a high performance formant speech synthesis algorithm in ASIC form. ASIC implementation offers sufficient signal processing bandwidth to enable multi-channel synthesis operation. When produced in volume, the cost of the ASIC falls dramatically, further reducing the incremental cost of providing speech synthesis in high volume, high value added applications. The engineering R&D concentrated on developing highly efficient bit-serial speech signal processing architecture which efficiently implements the formant speech synthesis algorithm. This was embedded in structures which provide all other functions to implement a complete multi-channel speech synthesis system on a chip. This includes integration of a multi-channel interactive voice source, a real-time microprocessor interface, delay line elements, and clocking circuitry to create a complete formant speech synthesis system on a single ASIC device. The resulting circuit allows complete high quality multi-channel text-to-speech systems to be implemented with as few as two components, a microprocessor system to support the text-to-speech algorithm coupled directly to the new speech synthesis ASIC.

SYNTHESISER STRUCTURE

The structure of the complete synthesis ASIC is shown in figure 1. The structure consists of the core speech synthesis signal processor, supported by five other modules to complete the design. These include: the interface circuit, which provides an interface to an industry standard 32 bit data bus; the multi-channel excitation function generator which provides voiced and fricative excitation to the formant filter network; the delay line model which implements the delays for the resonance and other filters, the output shift register which provides the multiplexed output speech signal in bit-parallel form and the clock generator which provides clocking for the system and orchestrates the multi-channel synthesis function.

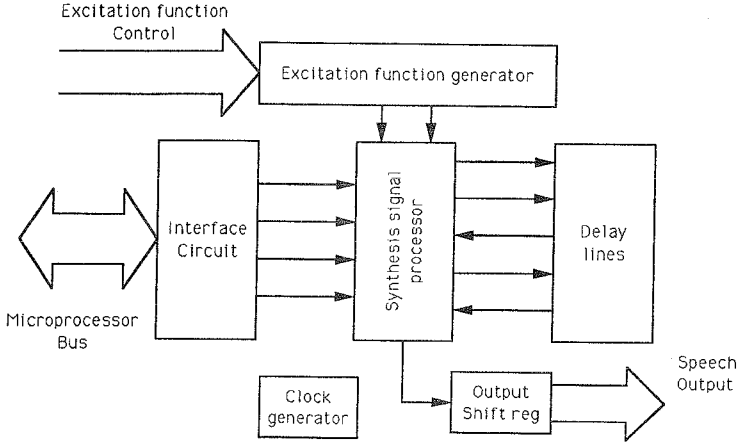

Figure 1. Structure of the speech synthesis ASIC.

Core Formant Speech Synthesiser Signal Processing Structure

The structure of the core speech synthesiser has been well documented (Summerfield & Jabri, 1988, and 1989). The synthesis algorithm is based on a parallel formant filter model originally developed by Holmes (Holmes, 1982) and implemented in DSP form by Quarmby (Quarmby and Holmes, 1984). This algorithm has been shown to produce exceptionally high quality speech synthesis. The ASIC synthesiser uses a fully parallel network of six resonance filters coupled with fixed filters to implement the synthesis function. Excitation to each of the parallel filters is individually controlled by six mixer/gain circuits which explicitly control the degree voiced and fricative excitation applied to each of the filters. The output from the filters are combined in alternate polarity to produce the output synthetic speech signal.

The synthesis algorithm is implemented using a 16/32 bit fully synchronous bit-serial signal processing structure. This consists of a multiplier block, containing two double precision multipliers, controlled by a bank of multiplexers. The multipliers are supplemented with a double precision adder/subtractor network to fully implement the formant filter algorithm. A separate bit-serial shift-and-add structure is used to implement the F1 fixed filter.

Scheduling operations are controlled by a three and a six phase multiplexer clocks. Computation of the filter coefficients, mixer/gain operation and the resonance filter difference equation is controlled by the three phase clock, whilst the six phase clock schedules combination of the parallel formant filters outputs. As this is a fully synchronous bit-serial architecture, the output sample rate from the structure is defined by the bit-serial clocking rate. For a single channel design, the clocking rate is 2.88 MHz (for 10 kHz sample data rate at the output). This is extremely modest for contemporary CMOS fabrication technology and provided ample scope for further multiplexing of the synthesis functions. The target for the design was 16 channels. This requires a clocking rate of 46.08 MHz and is not an unrealistic expectation for current CMOS fabrication technology.

Structure of the Interface Circuit

The structure of the interface circuit is shown in figure 2. This consists of a RAM which stored the input acoustic parameters. This feeds ROM elements which provide the mapping necessary to compute the coefficient for the resonance filters and the conversion from dB to a linear gain scale (Summerfield & Jabri, 1988). The output of the ROM elements are fed to parallel-to-serial shift registers to produce the bit serial input streams to the core speech synthesis signal processor.

Input to the synthesiser is provided through a 32 bit bidirectional bus. Acoustic parameters (frequency, bandwidth, voiced and fricative gain values) are formatted into a 32 bit word and down-loaded, under real-time control, into the interface RAM. During synthesis, the acoustic parameters are sequentially accessed and applied to the ROM mapping functions. The output values from the ROM are latched into the shift registers and applied as bit-serial data to the synthesis signal processing module.

Acoustic parameter down-loading is controlled by a interrupting scheme generated by the synthesiser clocking circuit. The interrupt period is programmable to enable the acoustic parameter update rate to be dynamically controlled during the synthesis operation. The interrupt interval can be controlled from 100 micro-second update rate (continually updating) to 25 ms. The interrupting scheme also includes a programmable look-ahead facility to accommodate interrupt service latencies associated with the host microprocessor.

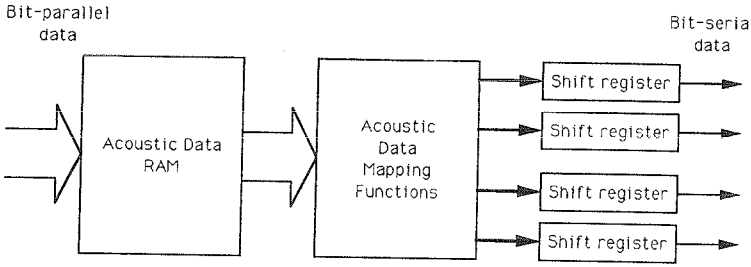

Figure 2. Interface Circuit Structure.

Delay Line Elements

There are three delay line elements in the design, two for the resonance filter implementation and an additional delay line to implement the F1 fixed filter. The delay line length (in bit serial terms) is controlled by the number of channels to be synthesised, minus the computational latency of the core processor. Delay line lengths for a 16 channels system are approximately 4,600 stages, each. It would be impractical to implement these directly using shift register elements.

The solution has been implemented using three separate RAM blocks addresses by separate recirculating counters. Interfacing to the bit-serial processor is provided by 16 bit serial-to-parallel (input) and parallel-to-serial shift registers (output). Reading and writing the 16 bit data into the RAM provides for the necessary delay period. As the delay function is only required for the formant filter data it is possible to reduce the necessary RAM size by two thirds. In this configuration, the clocking to the shift registers is gated to access only those bits associated with the resonance filter operation. In this configuration, the RAM sizes can be reduced by two thirds. This not only reduces the size of RAM, it also significantly reduced the restrictions on access time. Further, to negate the necessity to incorporate excessively long incremental shift register elements at the output of the delay lines needed to trim the bit-serial length of the delay line, a recursive shift register structure is used. This repeatedly applies the same bit-serial word to the signal-processor, allowing a smaller trim delay to be inserted as required.

Excitation Function Generators

Two excitation functions are generated, a fricative function and a voiced excitation function.

The fricative excitation generator is implemented using a 16 bit shift register which is recursively connected through exclusive OR gates to produce a pseudo-random sequence generator. The output is latched in the design to scale the fricative source making control of the unvoiced component simpler.

The structure of the voice source function generator is shown in figure 3. This is designed to produce 16 independently controllable interactive voiced source excitation functions. The structure consists of a ROM which contains profiles for a number of the differential glottal volume-velocity functions. This is addressed by a multiplexed programmable counter circuit. The RAM at the input to the counter stores the pitch period values which control the period of the pitch counter operation. This RAM also holds the glottal pulse shape page address which allows the user to dynamically select various glottal pulse shapes.

Access to the multi-channel voice source is provided through a second RAM element located on the ASIC. This allows asynchronous down-loading of the glottal source parameters and prevents the possibility of causing glitches during the glottal pulse generation. By also incorporating a status register

accessible by the external processor (via the bidirectional data bus), voice parameter down-loading can be synchronised with the the pitch to allow pitch synchronous updating of the glottal source functions. This is useful for synthesising non-periodic phenomena associated with the voiced excitation, such as pitch jitter and vocal fry.

The ROM address of data buses are made available externally to allow users to connect their own glottal ROM to the device. This adds further flexibility to the synthesis ASIC to allow users to modify the voicing function and the output speech quality.

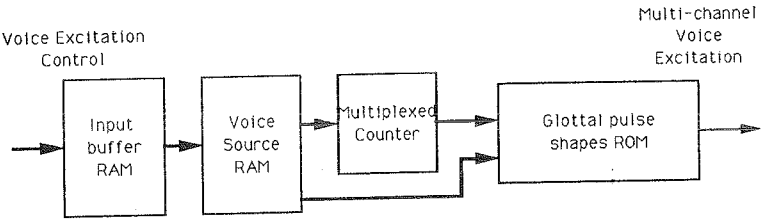

Figure 3. Structure of the multi-channel interactive glottal source function generator.

Clock Generator

Clocking signals for the complete synthesis ASIC are derived from an external crystal oscillator. As the ASIC is fully synchronous, the frequency of the clock determines the number of synthesis channels and is specified by $2.88 \times (No.of channels)$ (MHz). A 16 channel synthesis device, therefore, requires a 46.08 MHz crystal oscillator.

The clocking circuit is responsible for producing all the clocks using in the synthesis ASIC. The circuit consists of a network of synchronous counters which generate the control pulses for the bit-serial signal processor, three multiplexer clocks for multiplier scheduling, combination of the formant filter outputs, and multiplexing of the speech synthesis channels. A separate counter is used to generate the update interrupt signals. The circuit also provided clocking signals for the delay lines, the interface circuit and the glottal source function generator.

Output Shift Register

The output shift register converts the bit serial output from the synthesis signal processor and produces the multi-channel synthetic speech waveform as a time multiplexed 16 bit data stream in bit-parallel format. Speech sample values are reproduced synchronously with a four bit channel address and is used to demultiplex the multiple speech channels. Demultiplexing can be performed either in digital or analogue domains.

DESIGN COMPILATION AND ASIC FLOORPLANNING

The complete speech synthesis ASIC has been designed with the European Structures' (ES2) SOLO 1400 design tool. This is an upgraded version of the SOLO 1000 tool used to implement the original proof-of-concept device (Summerfield & Jabri, 1989). The SOLO 1400 tool provides options to allow the

users to define compilable RAM and ROM mega-cells which can be directly integrated into the ASIC design.

Design of the complete device was controlled by a strict hierarchical procedure. The design for each of the modules were entered separately via the the tools schematic capture facility. Following this each module was rigorously tested to ensure it conformed to both speed and functional specifications. The bit-serial signal processor (originally implemented using the SOLO 1000 tool) was recompiled from its original Hardware Description Language (HDL) MODEL code definition using the new 1.5 micron library parts supplied as part of the software tool. During this process the two phase clocking scheme used in the original design was modified to a single phase system. Although this makes the design more process dependent, it reduced the size of the macro and allowed an increase in the maximum clocking speed to be achieved. (A spin-off of this modification was also a significant speed-up in simulation run time.)

In large ASIC design, such as this device, it is important to carefully controlled to optimise chip area. As the design is "core bound" (ie. the size of the chip is determined by the size or the core circuitry as opposed to the size of the pad ring), there are significant advantages in concentrating on optimising layout and floorplanning. One of the advantages of the design is the lack of large global buses. This minimised the amount of metal inter-connects and produces an extremely compact layout. The design is made even more efficient by the extensive use of bit-serial signal processing techniques which minimise the data communications overhead within the signal processing architecture. This is an important consideration in ASIC design environments as it minimises the demands made on the automatic cell placement and route algorithms used by the ASIC design tools and ensures that a highly efficient chip design is achieved.

The total design is estimated to contain approximately 250,000 transistors and is less than 100 sq mm (at 1.5 micron geometries), including the pad ring.

CONCLUSION

This paper has described the design and implementation of a multi-channel formant speech synthesis ASIC for applications in Telecommunications and Information Technology. The device provides an extremely cost effective, single chip solution to the provision of high quality speech synthesis in multi-channel applications. These include high value added applications in telecommunications enhanced services, the provision of speech synthesis in banking and financial services markets and speech response in information services.

ACKNOWLEDGEMENTS

This work is supported by OTC Ltd. ASIC design and verification services were provided by Mr. Toby Cross and Mr. Graig Hepworth of AS2 Pty Ltd.

REFERENCES

Summerfield, C.D. & Jabri, M.A., (1988) "A formant speech synthesiser ASIC: Functional design", Second Australian Conference on Speech Science and Technology, pp. 8-1.

Summerfield, C.D. & Jabri, M.A., (1989) "Design and implementation of a formant speech synthesiser ASIC", Proc. Int Conf on Acoustic, Speech & Signal Processing (ICASSP), Glasgow, Scotland.

Holmes, J.N., "Formant synthesizers: cascade or parallel?", (1982) JSRU Research Report 1017.

Quarmby D & Holmes, J.N., "Implementation of a Parallel-Formant Speech Synthesiser Using a Single-Chip Programmable Signal Processor",(1984) Proc. IEE Vol 131 (Part F.) No. 6 pp 563-569.