AN OPTIMISED PARALLEL FORMANT SPEECH SYNTHESIZER

J R Andrews and K M Curtis
Department of Electrical and Electronic Engineering, University of Nottingham, UK

ABSTRACT

The problem that is addressed in this paper is the production of high quality natural-sounding multi-gender speech for different languages. The development of the Nottingham Parallel Seven Formant Speech Synthesizer is described, presenting the reasons for the choice of structure, component parts and demisyllable synthesis units. Results of the synthesis of German speech sounds and an utterance are given.

HIGH QUALITY SPEECH

Speech is regarded as the most natural form of communication known to man. The human-computer interface would be drastically revolutionised if speech was the medium chosen for the transfer of information. For example, the computer keyboard could be replaced by commands being issued and received by speech. This requires high quality speech synthesis and recognition systems, which are reliable, robust and can produce naturally sounding speech and recognise normal human speech.

The acoustic theory of speech production [1] provides a sound basis from which this problem can be tackled, in the frequency domain using formants.

In order to produce high quality speech, a flexible well-designed synthesis model must be sought. The most significant models are the five formant cascade and parallel KLSYN88 [2], developed by Klatt *et al* and the five formant parallel synthesizer [3] of Holmes *et al*. Both produce male and female speech and have been used in commercially. The CNET PSOLA synthesis system [4] uses time waveforms/windowing techniques and produces natural-sounding speech.

Diphones and demisyllables are the most widely used units for synthesis. PSOLA uses diphones for its synthesis unit. The German "HADIFIX" synthesis system [5] is based on a combination of diphones and demisyllables. These units have not yet been applied to a high quality formant speech synthesizer.

The synthesizer must have the ability to produce sounds present in most european languages and be flexible enough to produce male, female and child's speech. Therefore, it must incorporate a high level of control and be driven by highly accurate formant and excitation source data.

The synthesizer presented here forms part of a complete synthesis system [6]. It is a highly-programmable parallel formant synthesizer designed to synthesize multi-gender speech sounds of a high quality. A parallel structure provides the most flexible structure for synthesizing all types of sounds. The mapping of the synthesizer onto a network of processors is performed using a novel data/control flow methodology, thus enabling maximised processor utilisation and real-time performance. The synthesizer can model up to seven formants, has a maximum bandwidth of approximately 10 kHz and can use natural voiced sources.

NOTTINGHAM SYNTHESIZER

The synthesizer developed here consists of seven resonators connected in parallel. A novel parallel configuration was chosen for modelling the vocal tract, which is the most flexible design for synthesising all types of sounds. The choice of a solely parallel structure also reduces the complexity of the design. It has been shown [2] that a combination of a cascade and a parallel structure can be used for the production of all sounds.

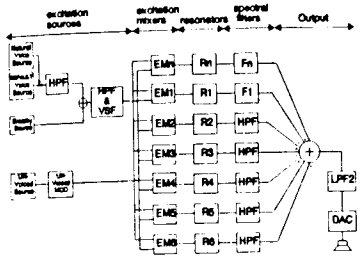

Figure 1. The Nottingham Parallel Seven Formant Speech Synthesizer.

This is because each resonator structure is suited to the production of a certain set of sounds. However, with accurate control of the resonator parameters, high quality sounds can be produced by a parallel branch only. This level of control is achieved by a constant update of all the synthesizer's parameters from sample to sample.

The aims of the design were to provide accurate control of all synthesizer parameters and to keep all specifications as dynamic as possible, in order to provide a highly adaptable synthesizer and thus to produce high quality natural sounding speech. Figure 1 shows the final components of the Nottingham Parallel Seven Formant Speech Synthesizer. The three main components are the excitation sources, the bank of resonators (representing vocal tract transfer function), and the output filter. In developing the final structure shown below other synthesizer designs were researched and tested [7].

The most significant parts of the synthesizer design are the default voice source, the novel use of individual excitation mixers and resonators. The choice of demisyllable synthesis unit will also be described.

Default Voice Source

The default model of the voiced source in the synthesizer is a complex design incorporating a number of important spectral features. The equation used in this model was based on the voiced source model employed in the KLSYN88 formant synthesizer [2].

The voiced source contains provision for dynamic variation of; amplitude of voicing (AV in dB); fundamental frequency (F0 in Hertz); open quotient, OQ in % of the total pitch period T0, this is the ratio of open glottis time to its pitch period.; period-to-period flutter (or jitter), FL in % of the fundamental frequency F0 value; and Shimmer SH, the period-to-period variation in the amplitude of voicing, A.

The first three parameters are essential for a high quality voice and the last two provide a degree of naturalness. The basic equation which describes the one pitch period of the waveform is of the form shown below.

$$v(nT) = a\left((nT)^2 - \frac{(nT)^3}{(\frac{OQ}{100}) * T0}\right) - DC; \quad 0 \le nT < (\frac{OQ}{100})T0$$

$$= -DC \qquad ;(\frac{OQ}{100})T0 \le nT < T0$$

where a and DC are constants determined by AV, OQ and T0.

Excitation Mixers

Figure 1 depicts each resonator branch as a combination of three distinct parts, the excitation mixer (EM_n, EM_1...EM_6), a formant resonator (R_n, R1,..., R6) and a spectral weighting filter (FN, F1 or HPF). Each resonator branch provides one formant resonance. R_n provides the low frequency component and nasal formant. Resonators R1 to R6 can each provide a formant, with R6 providing the highest formant (up to the bandwidth of the synthesizer's current implementation).

The excitation mixer combines both the voiced and unvoiced excitation for the resonator in its branch. It allows for individual dynamic control of both types of excitation for each resonator. This is unlike other common fixed synthesizers (Holmes and Klatt) and provides an extra degree of freedom in the amount of excitation in each formant.

The excitation mixer requires two parameters A_{Vi} (dB) and A_{Ni} (dB) per speech segment, for specifying voiced and unvoiced source amplitudes. The output of the EM module is defined below.

$$EM_i(nT) = a'\left\{A_{Vi}(S_{gi}(nT)) + A_{Ni}(S_{ni}(nT))\right\}$$

where $EM_i(n)$, $S_{gi}(n)$ and $S_{ni}(n)$ are the outputs of the excitation mixer, voiced source branch and unvoiced source branch respectively. a' is a factor derived from the formant frequency and bandwidth in the resonator equation.

Resonator

The synthesizer utilises a 2nd order resonator to generate the formant. The basic resonator (Rn, R1-R6) requires three parameters to specify its acoustic properties, the formant frequency, F_i Hz, amplitude and bandwidth, BW_i Hz. The formant amplitude is controlled in the excitation mixer. The resonator only requires the F_i and BW_i.

The output of the resonator is:

$$R_i(nT) = EM_i(nT) + B_i\{EM_i(nT-T)\} + C_i\{EM_i(nT-2T)\}$$

where n is the sample number, $EM_i(n)$ $EM_i(n-1)$ and $EM_i(n-2)$ represent the excitation's output at sample (n), (n-1) and (n-2). The coefficients B_i and C_i are calculated directly from F_i and BW_i.

Demisyllable Synthesis Units

The demisyllable was chosen as the speech synthesis unit to ensure synthesis with a very high degree of naturalness. A systematic approach [6] is used to define a demisyllable database. Synthesis is based on three types of demisyllables; (consonant-vowel cluster) and final demisyllables (vowel-consonant cluster) and vowel-vowel clusters. These were required to adequately cover all the coarticulation effects present in european languages. The approach adopted ensures that high quality and uniformity is maintained in the synthesis elements. The synthesis units are formant coded using the analysis system [6].

The temporal minima of the coarticulation effects coincide with the boundary of the demisyllable, therefore, only a small number of relatively simple concatenation rules are required. Each demisyllable is regarded as a number of formant-coded segments. The segments are concatenated by simple linear interpolation.

SYNTHESIZER IMPLEMENTATION

The implementation aspect of the synthesizer design is just as important as the structure of the synthesizer itself. Not only is a high quality natural-sounding speech synthesizer required, but the ability to produce sounds in real-time and as efficiently as possible is also needed.

The choice of both synthesizer design and implementation lead to the choice of the transputer parallel processor and processing language Occam [7].

As the synthesizer design is a complex algorithmic problem, a novel technique [8] of identifying and modelling both the flows of data, and flow of control variables was developed.

The synthesizer solution lead to the construction of a multi-transputer architecture to give a highly versatile and novel speech synthesizer, capable of the easy introduction of further formant generators. see figure 2 below.

The flow of control parameters is a pipeline originating in the "monitor" transputer which calculates the parameters from user input data.

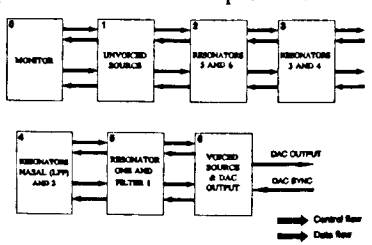

Figure 2. Synthesizer implementation.

The results of the synthesis of an utterance spoken by a male German are now presented.

RESULTS

An utterance was selected for synthesis using the German database of demisyllables. The utterance was "Das Stoßgebëf ist ohne Sinn". The utterance consists of the following demisyllables; /da/ /as/ /_Sto:/ /o:s/ /g@/ /_@/ /@_/ /bE:/ /E:f/ /_I/ /Ist/ /_o:/ /o:_/ /n@/ @_/ /zI/ /In/.

All the demisyllables underwent formant analysis as described in reference [6]. The each demisyllable is described by a number of sets of formant parameters.

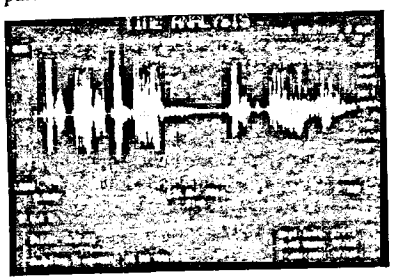

Figure 3. Time waveform of natural sentence /das Sto:sg@bE:f Ist o:n@ zIn/.

CONCLUSIONS

The synthesized utterance contained all types of speech sound. The waveform shows the advantage of demisyllables as units for speech synthesis. The demisyllables were synthesized from sets of formant coded segments and linear interpolation was carried out between the sets of parameters. This produced smooth transitions from segment to segment. However, there were some problems encountered in the production of the synthesized waveform. This was in the control of the unvoiced source model. As can be seen from the synthesized waveform, the amplitude settings for the unvoiced/mixed voiced segments were too high.

Work is under way in developing a more precise strategy for the control of the parameters fed to the synthesizer. Further work is also being carried out to develop a more natural sounding voice source which can be easily implemented into the synthesizer's structure. This should allow the modelling of different voices, accents, and manners of speaking. The synthesizer has shown greatest flexibility in the description of demisyllables.

REFERENCES

[1]Fant G. *Speech Sounds and Features*. MIT Press 1973.

[3]Klatt D H and Klatt L C, "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers", J.Acoust.Soc.Am Vol 87 No.2, Feb 1990, pp 820-856.

[4]Holmes J N, "A Parallel formant synthesizer for machine voice output", in *Computer Speech Processing*, Prentice Hall UK, 1985 pp 163-187.

[5]Hamon C, Moulines & Charpentier F, "A Diphone Synthesis System Based on Time-Domain Prosodic Modifications of Speech", ICASSP'89, pp 8-11.

[6]Portele T, et al, "Hadifix: A Speech synthesis system for German", ICSLP'92, Banff, pp 1227-30.

[7]Andrews J R and Curtis K M, "A Comprehensive Analysis and Synthesis System, and Synthesis Methodology for the Production of High Quality Speech", International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, April 1994, pp 591-594.

[9]Curtis K M, Asher G M, Pack S E & Andrews J, "A Highly Programmable Formant Speech Synthesizer Utilising Parallel Processors", ICSLP'90, Kobe, Japan, 19.14.1-19.14.4.

[11]Curtis K M and Andrews J R, "Control Flow: A Technique for Optimising Processing Architectures", Proc of 9th IASTED, Austria'91, p322-5.