Proceedings of AVIOS '95, the 14th Annual International Voice Technologies Applications Conference of the American Voice I/O Society, San José, Sept. 11-14 1995, AVIOS*: San José, pp 27-44

Real-time articulatory speech-synthesis-by-rules

David Hill‡, Leonard Manzara, Craig Schock

Note that the software developed under this project is available under a General Public Licence from the Free Software Foundation website It runs under NeXTSTEP 3.3 (black hardware, or the Intel version. Major parts have been ported to the Macintosh under OS X, and a port to GNU/Linux under GnuStep is in progress. If you want to help, please contact the project at http://savannah.gnu.org/projects/gnuspeech.

* 4010 Moorpark Ave, Ste 105M, San José, CA 95117, USA, Ph:408-248-1353, Fx: 408-248-0251, email: avios@pilot.net

‡ Also Professor Emeritus of Computer Science & Faculty Professor of Psychology, U of Calgary, Alberta, Canada T2N 1N4

Table of Contents

Appendices

Links to Male-Female-Child “Hello”synthesis comparison (sound files all ~3 MB)
(Composed by Leonard Manzara as a demonstration)

.aiff file	.au file	.snd file	.wav file

Links to additional examples illustrating full articulatory text-to-speech synthesis for male voice
(examples chosen for interest rather than flattery)

Pat-a-pan	Lumberjack	The Chaos

Introduction

An historical perspective

Articulatory synthesis has a natural appeal to those considering machine synthesis of speech, and has been a goal for speech researchers from the earliest days. The earliest documented example of physical modelling was due to Kratzenstein in 1779. His entry for a prize offered to encourage understanding of the nature of the five basic vowels in English, used fixed resonators (Paget 1930). By 1791, Wolfgang von Kempelen had a more complete and elaborate model that used hand control of a leather "vocal tract" to vary the sounds produced, with a bellows for lungs, auxiliary holes for nostrils, and reeds and other mechanisms for the voiced and fricative (sibilant) energy required (Dudley & Tarnoczy 1950). Alexander Graham Bell produced his own physical articulatory model, based on skull castings and internal parts that was able to generate vowels, nasals and a few simple utterances (Flanagan 1972). Riesz produced a mechanical articulatory model in 1927 (Flanagan 1972). A more sophisticated mechanism for controlling the vocal tract shape allowed a skilled person to produce some connected speech.

Other early attempts at synthesis used various forms of spectral reconstruction, starting with Helmholtz and his tuning forks, and progressing through Homer Dudley's "Voder" and the Haskins Laboratory "Pattern Playback" machines, to Walter Lawrence's "Parametric Artificial Talker" (PAT)—the first fully successful synthesiser to parameterise the reconstruction process in terms of vocal tract resonances instead of spectral characteristics (Lawrence 1953).

It was the advent of electronics that accelerated the quest for machine generated speech, but—until now—the most successful synthesisers have used some form of spectral reconstruction, rather than vocal tract modelling. Systems based on spectral reconstruction techniques are often referred to as "terminal analogues", in contrast to approaches based on artificial vocal tracts.

Origins of machine synthesis in Electrical Engineering & Computer Science

Techniques of spectral reconstruction have dominated speech synthesis for the past 40 years for a variety of reasons. Most importantly, they directly mimic what is seen in spectral analyses of speech, and are therefore subject to control on the basis of easily derived and manipulated data, and can be understood more easily by those working with them. The Sound Spectrograph, invented at Bell Laboratories (Koenig, Dunn & Lacy 1946) and manufactured for many years by the Kay Electric Company (a family owned business that still trades in modern audio analysis equipment under the name Kay Elematics), provided such a dramatic picture of speech energy distribution in time and frequency that, at first, people thought all the problems of speech recognition and synthesis had been solved. The form of presentation revealed not only the variation with time of the energy in various frequency bands, but also (to the human eye) vivid depictions of the varying resonances and energy sources involved. In reality, two years training were needed for even the most talented humans to learn to "read" spectrograms for purposes of recognition, and attempts at synthesis were successful only to the extent that spectrographic information could be copied from real speech and turned back into sound (for example, by apparatus such as Pattern Playback at the Haskins Laboratories). It was not until the systematic experiments relating spectral cues to perception were carried out at the Haskins Laboratories that truly synthetic speech became possible (Liberman, Ingemann, Lisker, Delattre & Cooper 1959)—so called "speech-synthesis-by-rules".

Other reasons for the dominance of spectral approaches to defining speech structure include the availability of a wide variety of mathematical models capable of revealing frequency-energy

variations in time, and the considerable expertise in filter design and digital signal processing available from electrical engineering and, later, computer science. This expertise is what allowed the successes of Pattern Playback, PAT and their successors.

Modern articulatory synthesis.
Articulatory approaches to speech synthesis also derived their modern form of implementation from electrical engineering and computer science. Sound propagation in an acoustic tube is modelled algorithmically (as opposed to physically) by the same techniques as used for modelling high-speed pulse transmission-lines[1]. The length of the acoustic tube (vocal tract) can be divided into a series of sections (typically around 40) and each is represented by elements that create the appropriate "impedance" corresponding to the physical reality. At first, these elements were electrical inductors and capacitors (Dunn 1950), and systems tended to be unstable. More recently, digital computer emulation has taken over, and each section comprises a forward energy path, a reverse energy path and reflection paths which connect the forward and reverse paths (e.g. Cook 1989, 1991). Whichever approach is adopted, and even ignoring possible instabilities, problems arise in controlling the models: partly due to the number of sections involved, and the many-to-one relationship between configurations and spectral output; partly because of the number of controls needed; and partly because of the lack of data from real vocal tracts. All of these factors have been addressed by recent advances.

A tube model
In electrical engineering, a transmission line or waveguide is designed to have uniform impedance and avoid reflections (which would constitute noise and/or energy loss when trying to transmit digital pulses from one point to another or to radiate energy). Variations in physical characteristics along the waveguide (including unmatched terminating impedances) cause reflections, providing the basis for resonant behaviour.

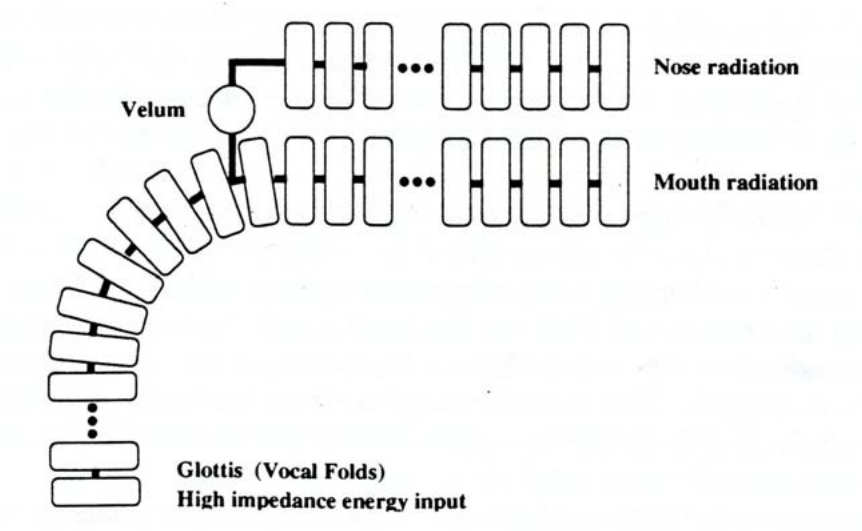

Figure 1: Multi-section physiological model of the human vocal tract

[1] A transmission-line analogue of the acoustic tube which models the sound propagation in the vocal tract may also be correctly considered as a spatial or lattice filter, or as a waveguide (Kelly & Lochbaum 1962; Markel & Gray 1976; Cook 1989, 1991. We prefer the more direct (but less general) term "tube model"

A human vocal tract is an acoustic tube that varies in cross-sectional area. Energy input from the vibrating glottis driven by air pressure from the lungs causes sound waves to propagate within the tube. Reflections due to area changes and end effects, lead to resonances characteristic of the tube shape. It is important to note that only a portion of the energy that reaches the mouth is radiated into the surrounding air. The rest is reflected back into the vocal tract. Modelling this radiation characteristic is one of the problems of articulatory synthesis. The nasal branch behaves in a similar way to the oral branch, and is connected part way along the oral branch by an equivalent of the velum (allowing energy transfer between the nasal branch and the oral branch at that point). Unlike the oral branch, the nasal branch's shape is fixed, for a given speaker. Figure 1 illustrates a conventional articulatory synthesiser (tube model) implementation. The section parameters can be set to emulate varying cross-sectional areas.

Using a tube model

A tube model is thus driven by energy input (glottal excitation, glottal pulses, the voicing waveform). The energy represents air being forced through the vocal folds from the lungs, causing them to vibrate. In addition there is random energy associated with forcing air through narrow, non-vibrating orifices (such as between the wide open vocal folds in an /h/ soundÐaspiration, or between the tongue and the back of the tooth ridge when making an /s/ sound—frication). Ideally one would model the physics of these energy sources, just as the waveguide and radiation characteristics model the physics of sound propagation in the vocal tract. However, good models are not yet available and, even if they were, the amount of computation required would likely preclude practical real-time operation using current signal processing chips.

A satisfactory approach is to use a hybrid model. The oral and nasal tubes are modelled as waveguides, but the energy sources are provided in ready-made form. The glottal waveform is derived from a wavetable (essentially table look-up of the time waveform of the glottal pulses) and the noise sources are provided by suitably filtered white noise.

Various refinements are possible. For example, the shape of the glottal waveform varies as the voicing effort varies; also its shape, onset and offset may differ for different speakers and conditions. Provision can be made to vary these details automatically, as well as to provide alternative glottal waveforms. When voiced fricatives are articulated, the fricative noise is usually amplitude-modulated at the glottal frequency. This effect can be reproduced, and it enhances the listener's ability to perceive one complex speech sound, rather than a collection of unrelated noises[2]. Some energy is radiated through the soft tissues of the throat. This "throat transmission" characteristic can be allowed for in the combined output. Finally, a certain amount of random energy ("breathiness") may be generated along with the voicing energy (glottal pulses), typically due to the vocal folds not closing along all their length during phonation, but perhaps due to roughness in the contact edges of the vocal folds due to the effects of smoking or other pathological changes. This "breathy voice" is more noticeable with females than with males, as the female glottal excitation is often characterised by incomplete closure (the so-called "posterior glottal chink" (Södersten, Hertegård, & Hammarberg 1995.)).

Other characteristics that distinguish females from males include a shorter vocal tract length (say

[2] The melding of various sound components into a single complex sound is crucial to good speech synthesis. The energy present at various frequencies during voicing (sounds where the vocal folds are vibrating) must bear a common amplitude modulation (Broadbent 1962). Timing and appropriateness are also important. Non-speech events do not meld into the time pattern that is speech (Broadbent & Ladefoged 1960). Non-speech events may include events like noise bursts that are inappropriately timed as well as sounds that are inappropriate in other ways. Such events are hard to place in their time relationship to the speech(Ladefoged & Broadbent 1960).

15 or 16 cm instead of 17 to 20 cm), and a higher mean pitch (because the female larynx is smaller than the male larynx for mature males and females). Not all characteristics that distinguish males and females have been fully identified and researched. Producing convincing female voices still causes problems, even for excellent synthesis systems.

The control and computation problems

Two approaches to controlling a physiological analogue of the vocal tract have been taken in the past. One approach is to set up the cross-sectional areas so that, in total, the overall ªshapeº of the vocal tract is approximated. This allows steady-state sounds to be produced, but there are problems (a) in obtaining the data needed, and (b) in moving from one posture to another (what interpolation function should be used, and should it be the same for all sections). The second approach imposes an articulatory framework as a means of constraining the tract cross-sections. This is not as easy as it might seem, partly because of inadequate data, and partly because it is not easy to specify the behaviour of the sections in relation to each other, or in an absolute sense, as a dynamic system with springiness, inertia, deformableness, volume constraints, and the like. The first approach leads to systems termed ªphysiological modelsº and the latter to "articulatory models". Both are based ultimately on the tube model. Other related work on articulatory synthesis uses some form of analytic model based on tube acoustics to determine resonances, but then synthesises speech using a terminal analogue. This is not articulatory synthesis in terms of acoustic realisation, but does tackle the problem of modelling speech knowledge.

As well as the data and manipulation problems, there is also a problem providing enough computational power to manage all the calculations needed for running such multi-section tube models in real-time, in the context of the requirements for vocal tract simulation.

In his survey of speech synthesis models, Carlson (1993) comments that "Articulatory models have been under continuous development, but so far this field has not been exposed to commercial applications due to incomplete models and high processing costs."

Such difficulties dissuaded us from attempting approaches to speech synthesis based on tube models until quite recently. However, it transpires that the control problem can be greatly simplified, making both physiological and articulatory models much more practical and much closer together and reducing the computational problem. As a result, the first commercial application of articulatory-synthesis-based text-to-speech conversion is now available.

Distinctive Regions and Modes: the Distinctive Region Model (DRM)

Acoustical basis

In 1974, Fant and Pauli published some results of their research on the spatial characteristics of vocal tract resonance modes (Fant & Pauli 1974). The work can be viewed as a sensitivity analysis of formant frequency dependence on vocal tract constriction, for various places of constriction, and cites earlier work by Schroeder (1967) and Heinz (1967).

In a resonating vocal tract of roughly uniform cross-section, terminated by a high-impedance glottal source at one end and an open mouth at the other, each formant is represented by a volume velocity standing wave with a node at the glottis, and further nodes and anti-nodes dispersed down the vocal tract, leading to an anti-node at the lips. Figure 2 shows the situation for formant 2. Note the use of two scales, one for volume velocity from positive to negative flow and another for kinetic energy from zero to the value K. Figure 3 also embodies two scales.

[3] Fant, who heads the Speech Technology Laboratory at the Royal Institute of Technology in Stockholm, wrote the classic text on the acoustic theory of speech production, based in part on his Ph.D. research (Fant 1960).

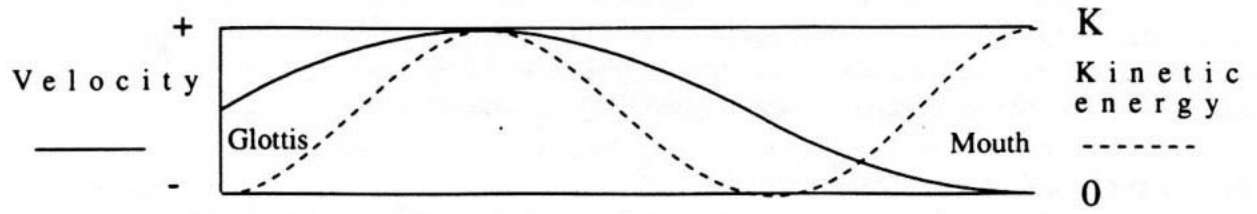

Figure 2: Volume velocity standing wave and kinetic energy for Formant 2 in a uniform tube, open at one end

The figure provides a graph of the range of velocity fluctuation and kinetic energy function for the second formant (F2), velocity being maximum at anti-nodes (one anti-node is located at the mouth), and minimum at nodes (one node is located at the glottis). Figure 3 shows the pressure standing wave and potential energy function for the same formant.

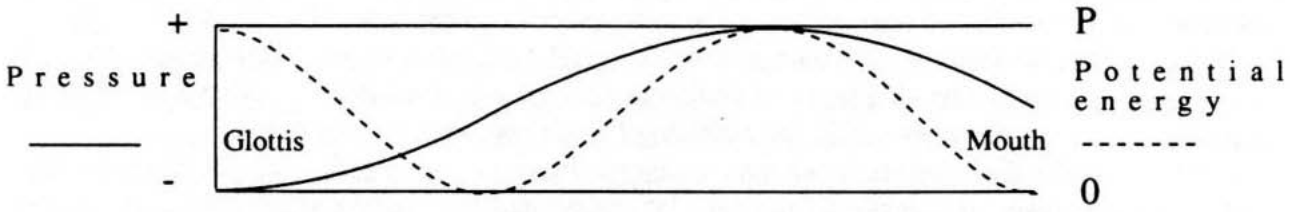

Figure 3: Pressure standing wave and potential energy for Formant 2 in a uniform tube open at one end

The sensitivity function for the same formant relates the kinetic and potential energies by providing a graph of the arithmetical difference between potential and kinetic energies (k - p). Figure 4 shows the (sinusoidal) sensitivity function for F2.

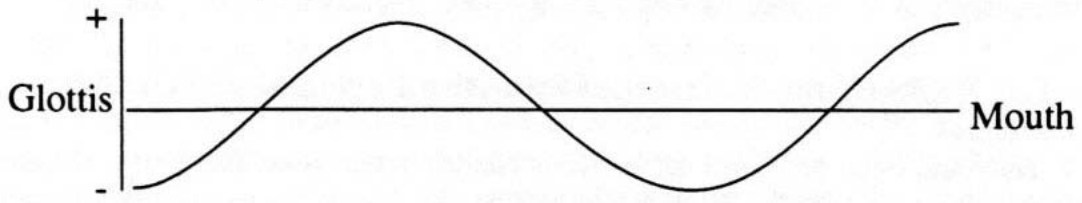

Figure 4: Sensitivity function (f = k - p) for Formant 2 in a uniform tube open at one end

When the graph is negative, potential energy exceeds kinetic. When it is positive, kinetic energy exceeds potential. When it is zero, the kinetic and potential energies are equal. Increasing the area of the tube at a zero sensitivity point has no effect on the frequency of the formant. Increasing it where the function is positive causes the frequency to rise proportionately to the value of the function since the impedance is thereby decreased, so that the pressure decreases, and the volume velocity increases

in that region. Constricting it has the opposite effect. The effects are reversed in regions where the sensitivity function is negative, since potential energy effects dominate.

Formant 2 divides the tube into four distinct regions. In two of the regions, constriction causes the formant 2 frequency to rise, in the other two regions constriction causes the formant 2 frequency to fall—complementary relationship with the first two regions. Related functions can be drawn for formants 1 and 3. Formant 1 produces two regions having complementary effects; formant 3 produces six regions having complementary effects. When the regions for these lowest three formants are combined, the tube is divided into a total of eight regions by the zero crossings of the sensitivity functions, as shown in figure 5.

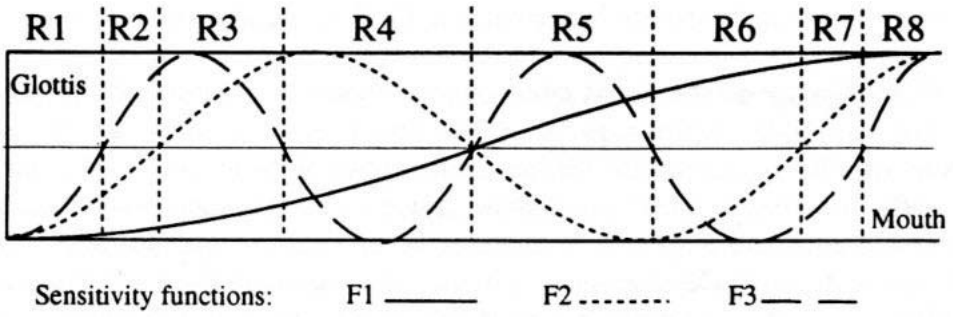

Figure 5: Distinctive regions based on normalised sensitivity function zero-crossings

Figure 6 shows the direction in which a given formant will move for a constriction in the region specified. Each will rise if the corresponding sensitivity function is negative (potential energy greater than kinetic) and fall if it is positive (kinetic energy greater than potential).

	R1	R2	R3	R4	R5	R6	R7	R8
F1	↗	↗	↗	↗	↘	↘	↘	↘
F2	↗	↗	↘	↘	↗	↗	↘	↘
F3	↗	↘	↘	↗	↘	↗	↗	↘

Figure 6: Formant movement directions for constriction in various regions

As the formant frequencies change with change in tube shape, the functions change due to the modified distribution of kinetic and potential energy, but, for relatively small changes of area, away from the boundary regions of the sensitivity functions, the relationship allows the effects of changes in cross-sectional area on formant frequencies to be predicted over a reasonable range of area change.

A DRM-based control system and its advantages

Carré and his co-workers used Fant and Pauli's analysis as the starting point for developing what they call the Distinctive Regions and Modes system, or Distinctive Region Model (DRM) capable of ªproducing optimal formant frequency modulationº. The DRM has some very interesting properties in relation to physiology and language structure, which are pursued in depth in other papers (Carré 1992; Carré, Chennoukh &Mrayati 1992; Carré & Chennoukh 1993; Carré 1994; and Carré &

Mrayati 1994; for example). Our focus is on using Carré and Mrayati's system as a basis for simplifying the control of a tube model, but the derivative results and hypotheses suggest that humans partly evolved and partly learned to use the DRM system in much the same way for optimal control of their vocal tracts, in order to produce human speech. For one thing, the disposition of the articulators coincides very well with the need to control the regions of the DRM system independently, and hence optimise the control of the lowest three formant movements. To control more than three formants independently would require a greater number of much more selective, finely placed articulators. At the same time, the higher formants do not assume arbitrary values. Their values are dictated by the same kind of functions as the lower formants. They simply cannot be manipulated independently, because of the absence of the necessary controls.

The main advantages of the DRM approach to control are threefold. First, it is only necessary to consider eight control regions, or vocal-tract sections, which are closely related to human articulators. Thus the differences between a physiological model and an articulatory model are greatly diminished. It depends somewhat on what the researcher considers are the articulators. If it is acceptable to work in terms of lip opening, tongue position, tongue height, and place of closure, then the DRM model is well suited to representing these parameters almost directly, making allowance for vocal tract continuity constraints and a constant-volume tongue body. If it is necessary to add jaw rotation and perhaps other parameters (as may be needed for facial animation), then a little extra computation is required, based on suitably quantified relations. Secondly, with only eight vocal tract sections to compute, the amount of computation is greatly reduced. This puts real-time articulatory synthesis within the reach of current Digital Signal Processor (DSP) technology. Thirdly, the type of data needed to drive the model is simplified and more accessible, and its amount is greatly reduced.

Characteristic	Spectral reconstruction	Articulatory synthesis
1. Spectral fidelity: formant shapes.	Formant shapes determined by properties of filters used for resonance approximation.	Formant shapes accurately modelled according to physical reality.
2. Spectral fidelity: formant frequencies.	Lowest three or four formant frequencies explicitly defined, remainder usually approximated by lumped fixed filter.	Varying frequencies of an arbitrary number of formants are correctly reproduced on the basis of controlling the lowest three.
3. Dynamic spectral fidelity.	Interpolation between speech postures depends on approximate formant interpolation, independent of the underlying articulatory constraints.	Interpolation between speech postures is modelled directly, based on the changing shape of the emulated vocal tract.
4. Nasalisation.	Nasalisation approximated by various "kludges", using additional poles and zeroes. A notch filter is usually used to "split" F1—a characteristic of nasalisation.	Nasalisation is modelled directly by opening a velar connection to a parallel nasal tube that exchanges energy correctly with the pharyngeal-oral tube. No "kludges" are needed.
5. Fricative spectra and formant transitions.	Fricatives are approximated by filtering noise, according to the type of occlusion of the vocal tract. Fricative formant transitions are not modelled.	Fricatives are approximated by filtering noise, according to the type of occlusion, and are injected at the appropriate place along the length of the vocal tract, producing associated fricative formant transitions.
6. Throat transmission	There is no easy equivalent of throat transmission in spectral reconstruction approaches. Usually not attempted.	Appropriate energy representing throat transmission may be extracted and fed to the output.

Table 1: Comparison of spectral modelling with articulatory synthesis

Since the DRM approach involves articulatory modelling, further advantages are gained. Table 1 illustrates the advantages of an articulatory model compared to a spectral reconstruction model (such as a formant or spectral synthesiser).

It is worth noting that if the DRM control system is deficient in any sense (for example, because the regions shift as constrictions are applied) then these deficiencies are almost certainly reflected in the human's control of the human vocal tract. There is not a lot of margin for adjustment in given human places of articulation, and the goal of human articulation does seem to focus on the lowest three formants. The DRM model appears to be fundamental to human articulation and language structure, a point that is made very convincingly in the papers by Carré et al. already cited.

Building a real-time articulatory-synthesis-based text-to-speech system

Basic tasks

Having become convinced of the merits of the DRM system for controlling a tube model, we elected to modify our existing spectral-reconstruction-based real-time text-to-speech system to use an articulatory synthesiser. Five main problems had to be solved in order to do this.

1. Create a complete, hybrid sound synthesiser built around a tube model of the vocal tract;
2. Port the synthesiser onto DSP hardware and optimise it to allow real-time operation;
3. Build a database framework and editing system for the rules and data needed to construct the control parameters for the tube model;
4. Collect and/or create the data and rules needed to specify vocal tract shapes (speech postures) corresponding to the sounds of the initial target language (English);
5. Integrate the new speech synthesiser into the existing text-to-speech system, replacing the previous spectral reconstruction synthesiser.

Numerous accessory components were also needed, including graphical interfaces to the synthesiser and the rule data-base editor facilities, together with sound analysis and manipulation utilities, also with associated graphical user interfaces.

The synthesiser

Conceptually, the synthesiser is straightforward, and the implementation is quite similar to the waveguide models described by Perry Cook (1989; 1991). In practice there were many details to be worked out, some of which depended on new research. Achieving real-time operation was a tedious exercise in DSP assembler code optimisationÐan exercise that was not completely successful in the context of the 25MHz 56001 DSP that we used, limiting real-time operation to approximately a 16 cm minimum tube length. A Turtle Beach Multisound board, with a 40 MHz DSP, was also usedÐas the plug-in required for our PC implementation, but this "best" off-the-shelf board available used 3-wait-state memory, which offset the advantage of the faster DSP. We expect the problem will be overcome once we move to a more consistently designed 56002-based plug-in, running at 60 MHz. We have considered building our own board, complete with host processor. This would offer other advantages (by offloading the main host, and speeding speech-host/DSP/DAC communication). A preliminary design is complete.

The synthesiser is best illustrated by showing some of the graphical interfaces used to communicate with the synthesiser facilities. When used as part of a text-to-speech system, these graphical interfaces are, of course, irrelevant. The same control data still have to be supplied, though, and this is the task of the parameter construction system based on a subset of the rule database editor.

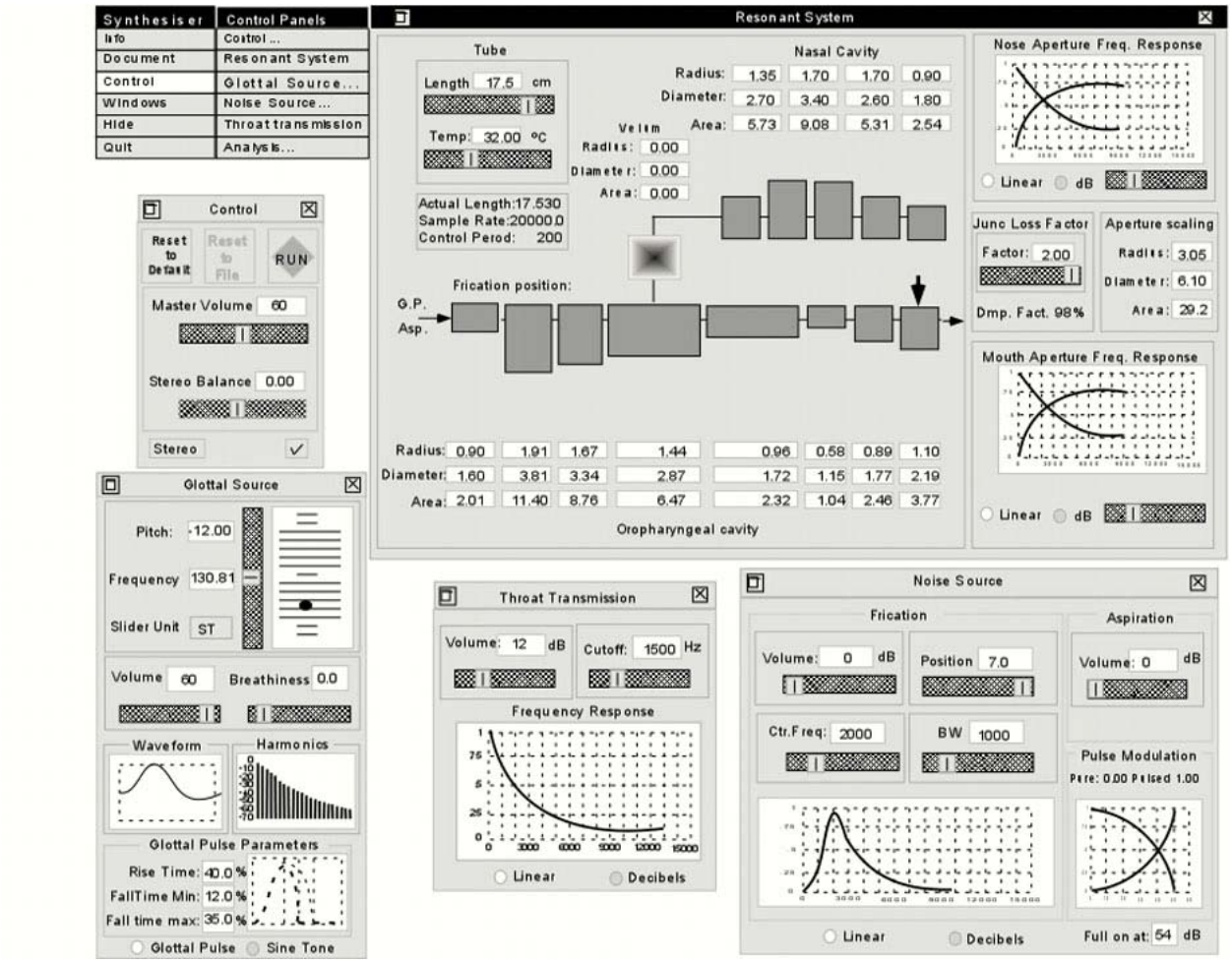

Figure 7: Tube-model-based synthesiser GUI

It will be noticed that many controls are provided. Some controls are varied only between utterances. These so-called "utterance-rate parameters" reflect speaker characteristics. They include volume, stereo balance, glottal waveform characteristics, throat transmission, mouth and nose radiation characteristics, vocal tract length, and nasal branch dimensions. The remaining controls are varied at a speech posture rate. They include the DRM cross-sectional areas and the pitch. Of course, the mean pitch will depend on the speaker, so that pitch is really in both sets of control data. Figure 7 shows all panels together, but it is only necessary to invoke the ones to be used at any given time.

The GUI provides a graphical depiction of the pharyngeal, oral and nasal branches, plus the velar connection. The regions may be varied by direct manipulation, or by typing the radius, diameter, or area into a text field explicitly.

A real-time analysis window can also be called up, whilst running the synthesiser, so that the spectrum of the output sound can be graphed during testing and synthesiser development, as well as providing part of the basis for determining appropriate cross-sectional area data for different speech postures during rule and posture data development.

Determining the best basis for the mouth and nose radiation characteristics proved problematical. A number of models are proposed in the literature, including a circular piston in an infinite baffle, a circular piston in a spherical baffle, and a pulsating sphere. The former was chosen as it seems to provide the best approximation for our purposes (Flanagan 1972, pages 38-40). There is also some difficulty in dealing with the end effect at the mouth in terms of the energy radiated versus the energy reflected, and the final configuration was empirically based.

The fricative energy is supplied by filtering a white noise source and injecting the result into the tube at a variable location (chosen according to the place of tract occlusion). Provision is made to allow the noise to be amplitude modulated at the pitch rate, to a degree determined by the overall amplitude.

Parameter target and rule data management—MONET

Figure 8 illustrates some components of the MONET editing system, used to set up speech posture targets, interpolation rules, and timing data.

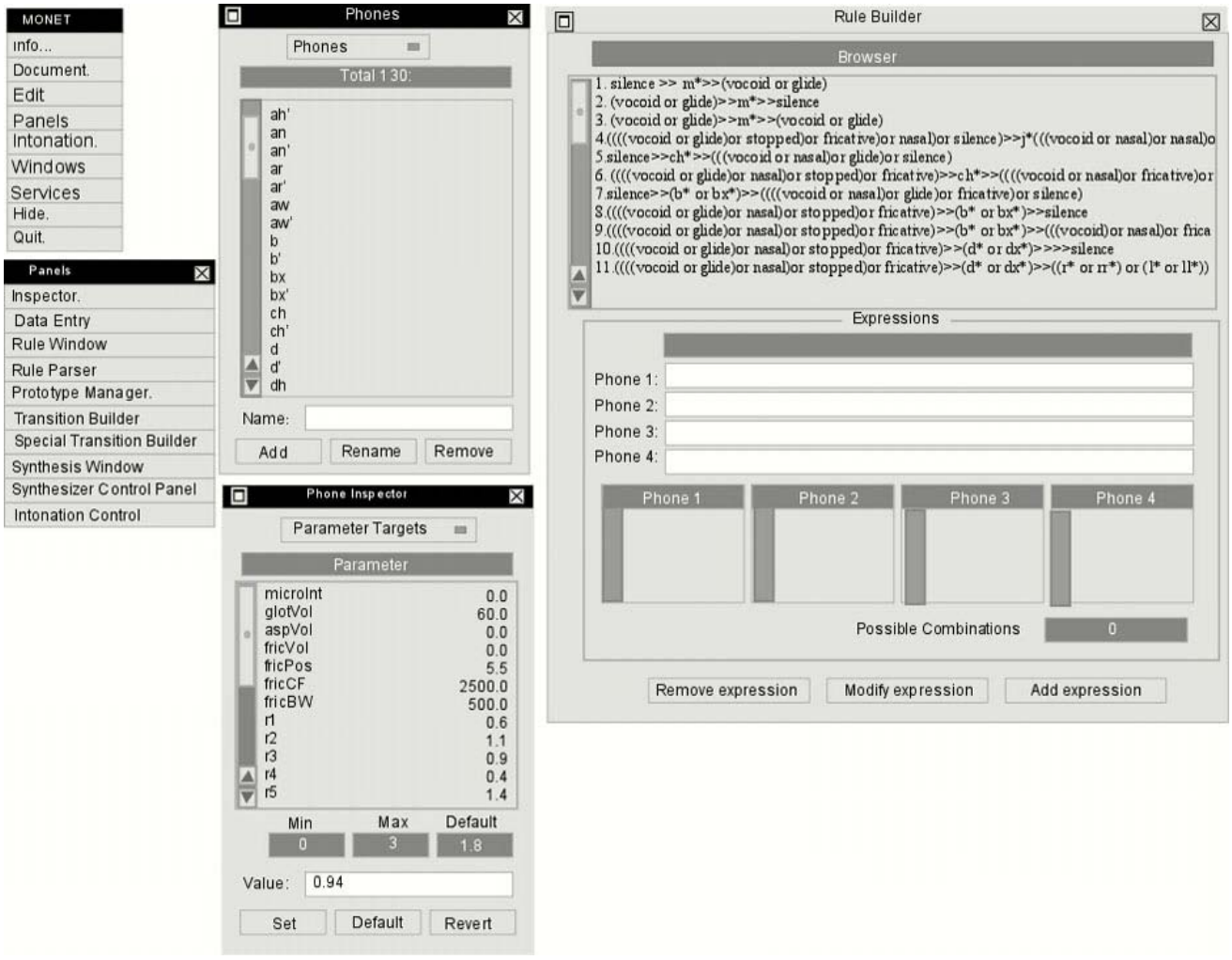

Figure 8: MONET: Target data entry, Inspector, Rule Window and menus

The symbols for speech postures (also referred to somewhat incorrectly as “phones”) are entered in the top left window. The Phone Inspector below then shows default data for that posture, which may be edited. In the figure, the value for R3 is selected, and could be changed. Other options on the data entry window allow the definition of: “Categories” (into which postures may be categorised in order to apply rules to particular categories); “Parameters” (to provide for additions to the synthesiser, or a change to a different synthesiser); “Meta Parameters” (to allow higher level parameters to be defined in terms of lower level parameters thus JawRotation could be a meta-parameter, and it value would be one thing affecting some of the distinctive region radii); and Formula Symbols to provide symbols associated with each posture—such as duration—capable of being used in rule formulae in order to compute needed data for parameter interpolation. The Parameters defined for the current database are partially visible in the Phone Inspector window, and the actual values shown are those associated with the posture “aw”, which is IPA /o:/ as in British English.

The Rule Builder window, which is also visible, allows rules to be entered. A given rule selects a combination of the basic postures which will be treated in a particular way. The simplest rules handle diphone combinations, but triphone or tetraphone combinations may also be handled in order to manage the dynamic, context-dependent aspects of parameter construction. To create a new rule, the components (up to four) that invoke the rule are entered, and then the sub-rules for assigning parameter transition profiles, timing data, and any special events are hooked into the rule using the relevant GUI components.

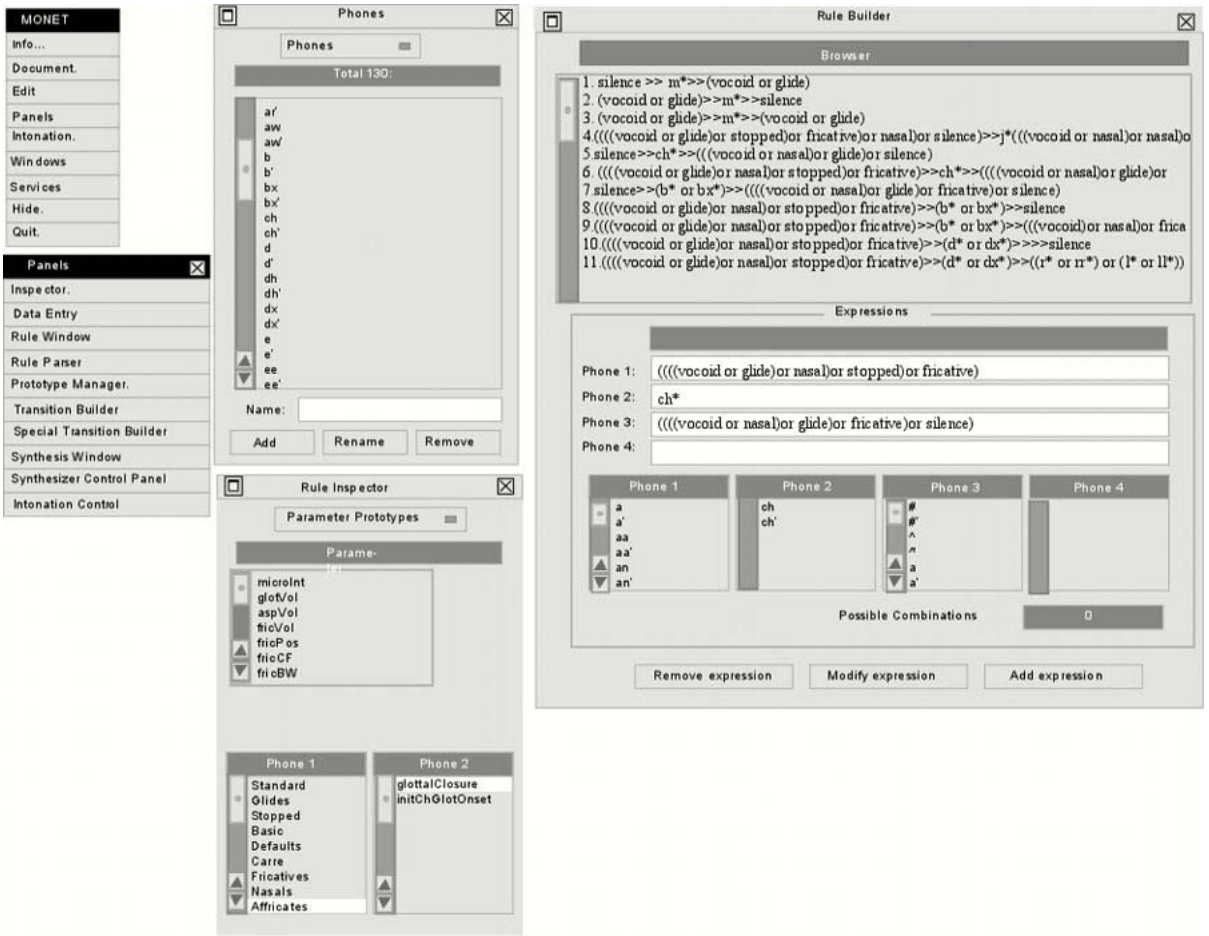

Figure 9: MONET: rule selected. Inspector now able to show transition profiles for parameters

When a rule is selected, the components are displayed separately below the browser and can be edited. Some elements in the rules are categories or Boolean combinations of items, so a display of all the elements covered by a rule component is also provided in scrolling lists, as shown in Figure 9. Once the rule is selected, the Rule Inspector allows a particular parameter to be selected and displays the parameter prototype associated with that parameter, for that rule. By clicking on a different parameter transition profile, a new profile can be assigned to any given parameter via the Inspector. Since new prototypes may be created, the system allows detailed specification of exactly how any parameter varies for a given combination of speech postures. A parallel mechanism allows special transition profiles to be created and allocated. These are used in combination with the normal transitions to allow for specific parameter deviations such as noise bursts. The timing of all points on any profile are defined by formulae, set up using the Equation Prototype Manager. Timing equations are computed based on the Formula Symbols entered for each posture, according to the Formula Symbols defined in the original set-up.

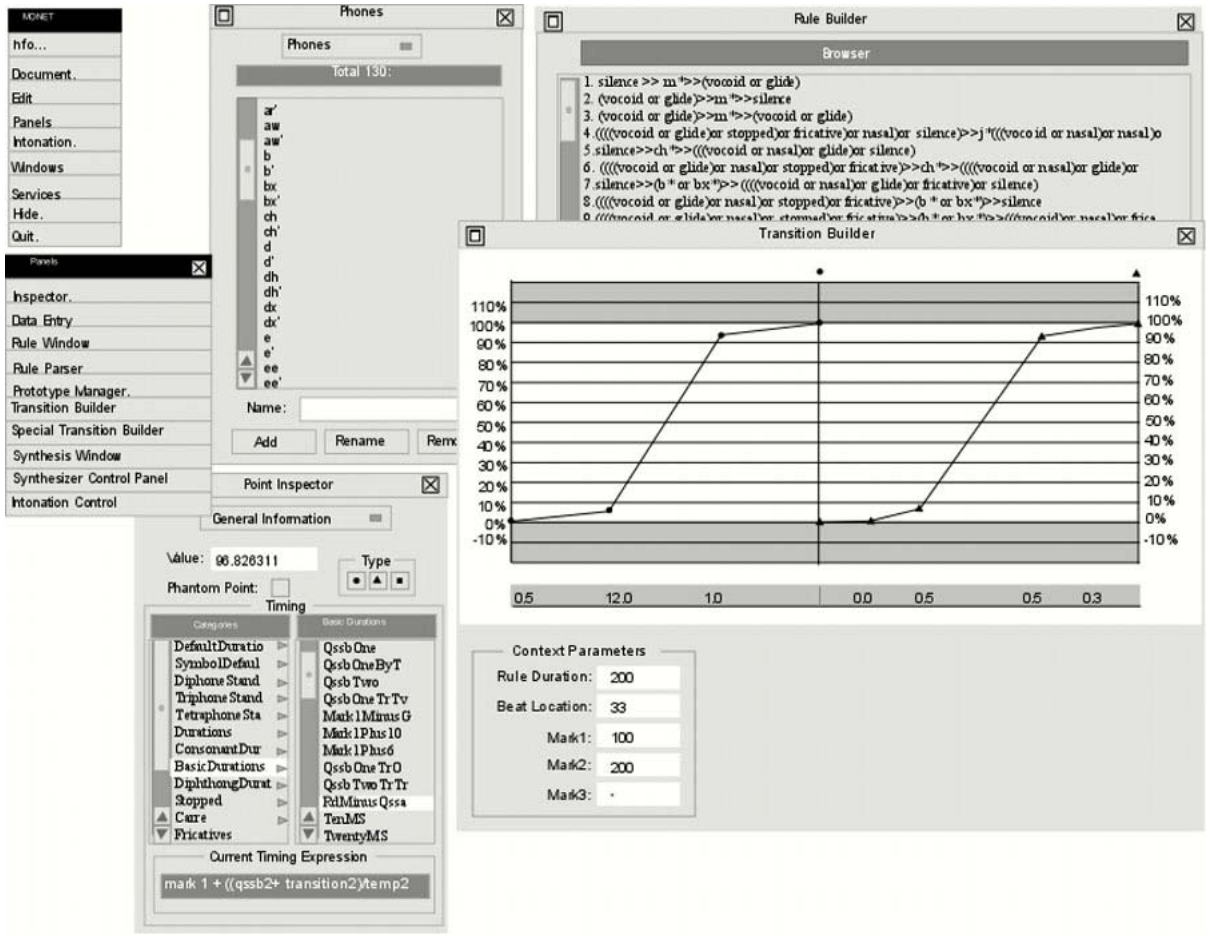

Figure 10: MONET: Double click the transition profile in the browser (Fig. 9) to bring up the transition profile graph and add points. Select any point to see and edit point characteristics.

Double clicking on the Transition Profile name in the Rule Inspector window brings up an actual graph of the transition profile selected, as shown in Figure 10. It may then be edited and refined as required by adding and deleting points, or changing their values and timing. Selecting a particular point changes the Inspector into a Point Inspector, as shown. The timing value is specified in the browser

(by a named timing value previously defined by equation), and the value is defined by a text field. Parameter transition profiles are somewhat more complex than this simple account suggests, but it gives the basic idea. It is very easy to manage relative timing of events between different parameters, based on the equations, and profiles may be specified to an almost arbitrary degree of detail.

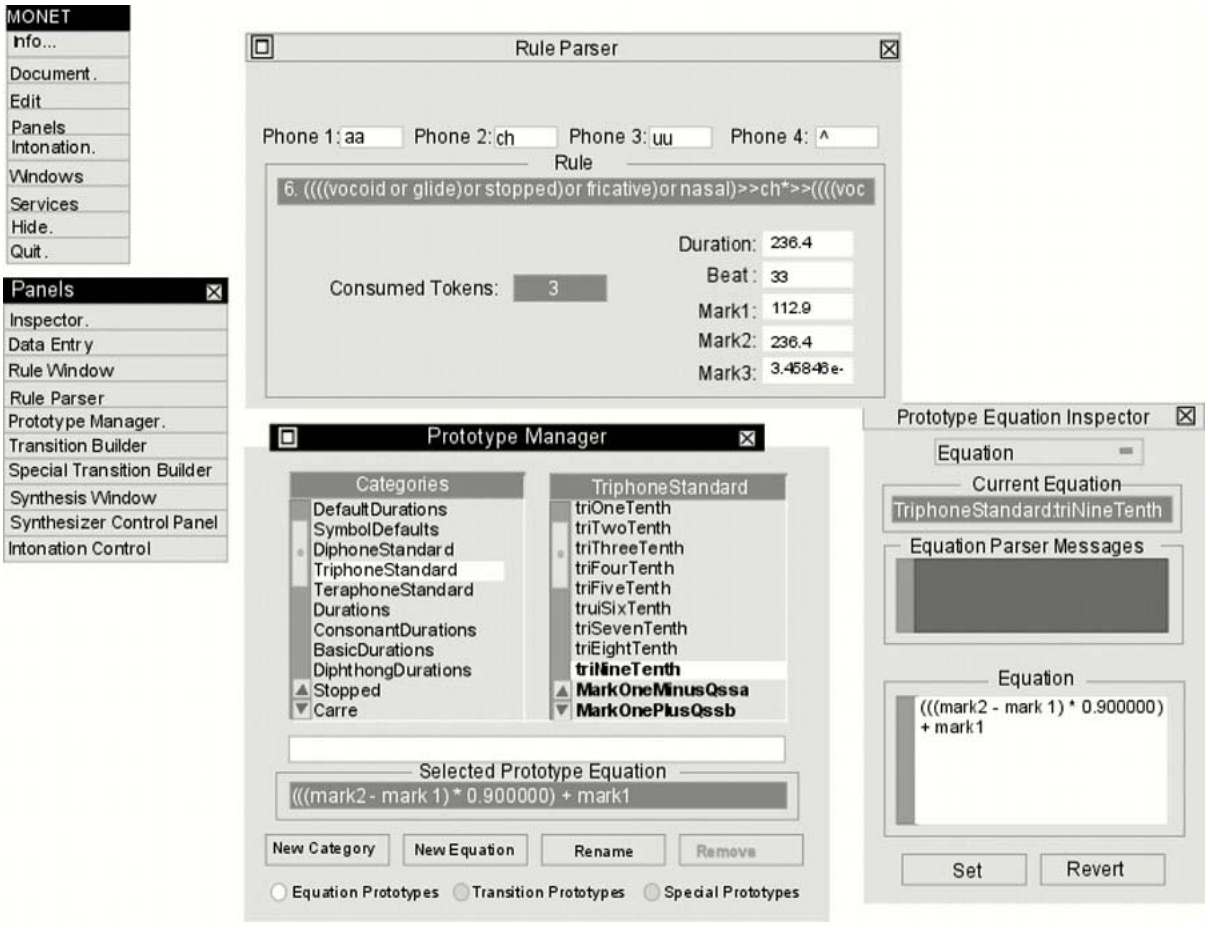

Figure 11: MONET: Rule parser and prototype manager

Figure 11 shows the rule parser, used to check which rule will apply to any given input string, and how many input tokens will be consumed in applying the rule. It also shows the Prototype Manager window, which provides an alternative means of accessing the Transition Profiles directly, together with the Equation Prototypes. The window is used to name and define new prototypes. As already noted, equations are defined in terms of the Formula Symbols entered as part of the Posture Data. The Prototype Equation Inspector in Figure 11 shows the actual equation selected in the Manager's browser.

Building the databases

Faced with the task of building the necessary data to construct articulatory parameters for text-to-speech conversion, we had two choices. Analyse speech, extract the necessary data, and formalise it; or synthesise speech and edit the data and rules to produce the required results. The synthetic approach was chosen because it is more economical, and leads to an easier basis for abstracting

general principles. It is easier to start from simple principles and refine them, as an approach to managing complexity, than to try and see general patterns in a sea of data (even assuming the data collection can be completed in a systematic and useful manner). Subsequent experience more than justified our approach.

MONET was the system we used to create our database of posture targets, rules, profiles and timing equations in order to control the parameter construction needed to drive the synthesiser. The trickiest part, once MONET had been designed and implemented (no easy task), was creating correct vocal tract shapes to supply the critical initial posture target data. Once that was done, the rules were developed from the general to the specific, refining as needed to produce particular effects associated with the dynamics of the various posture combinations. As well as critical listening tests, analyses of the speech output from the system were compared with analyses of natural speech, in order to check the effect of the various components involved in the parameter construction, for each combination. Occam's razor was applied with enthusiasm. Our initial goal was to produce an adequate set of data, rather than a complete set, using the same principle of successive refinement that guided our original approach (Manzara & Hill 1992).

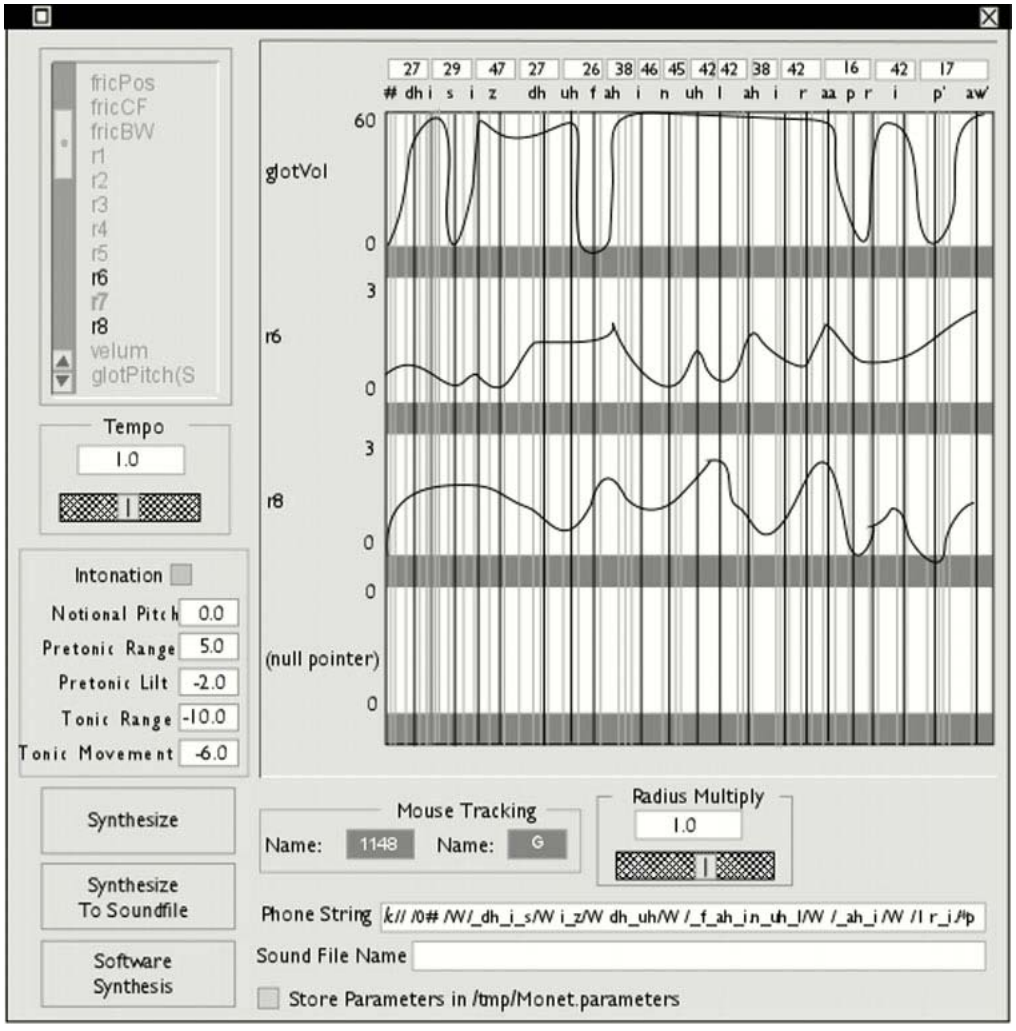

Figure 12: MONET: the synthesis window, showing computed parameter tracks

The complete system

Figure 12 shows the synthesis window associated with MONET. All constructed parameter variations may be inspected and related to each other. The particular set on view is selected using the scrolling list at the top left-hand side, and a zoom control is provided. The sound output may also be listened to in real-time, or a sound file may be created. In addition, the original non-real-time software synthesiser (written in `C', and used in earlier development) may be invoked, to allow the DSP output to be verified against the original algorithm. Another window may also be invoked to show the intonation contour separately, and allow it to be edited. The intonation is normally generated algorithmically, and smoothed. It is useful, however, to be able to experiment with intonation contours for psychophysical experimental purposes, in order to test changes and improvements to the algorithms, and to carry out basic research.

MONET, along with the synthesiser and database, was incorporated into the final text±to±speech system, stripped of all its graphical interaction facilities. In this way we avoided trying to manage multiple databases with the same information, or trying to reconcile two different synthesis systems to produce the same results. The framework from our original spectral±reconstruction±based system provided the input parsing, pronunciation look-up, dictionary management tools, and various utilities. It was comforting to discover that few problems arose from our change of synthesis engine, and that the only changes relevant to pronunciation were very much in the direction of making better phonetic sense.

The text-to-speech system is available commercially in two forms: as a user kit, to allow end users to listen to files, or highlighted text (including items looked up in dictionaries, thus providing a practical pronunciation guide); or as a developer kit, which allows programmers to create new applications (email speakers, telephone database access programs, computerised spoken instruction providers, and the like).

Discussion and future work

As a result of the work we have done, it has become very obvious that our articulatory synthesiser, together with the tools and components we created in order to develop the text-to-speech system, comprise a radically new departure in speech synthesis that can be of great value to speech researchers. The complete system (with tools) has a number of potential uses including:

1. Provides a basis for psychophysical experiments on speech perception. We are already working with colleagues at the University of Calgary (notably Dr. Elzbieta Slawinska), and others elsewhere, to create very precisely controlled stimuli (timing, values, etc.).
2. Provides a systems for intonation research, since hand-tailored or algorithmically determined intonation contours may be applied to arbitrary utterances, on a framework of high-quality synthetic (therefore well-controlled) speech.
3. Provides a basis for developing new rule sets for arbitrary languages and synthesisers.
4. Provides the means of testing hypotheses about the nature of speech production. For example, during the course of our development we naturally considered Carré's hypothesis (Carré 1992) that intervocalic stop sounds are best emulated as a transition of DRM parameters from the preceding vowel to the following vowel, with a closure of the section relevant to the stop superimposed. Although this seems somewhat applicable to bilabial stops (for which the lips are able to move relatively independently of the jaw and tongue), we had no success applying the approach to other stop sounds, as judged both aurally and spectrographically. Our rules vary all the DRM sections for other stops.

The most striking aspect of the synthetic speech produced by our new articulatory synthesiser is

the natural appearance of spectrograms made from samples of it, and the clarity of articulation. Several expert listeners have commented that it almost sounds over-articulated. This is a pleasing "fault" to have, compared to the cotton-wool-in-mouth tendency of some earlier synthesisers. Another important aspect of the synthesis comprises intonation and rhythm, which provide much easier listening than other contemporary synthesis systems (less tiring to listen to, according to those who have spent some time using it). The easier listening may also have something to do with the quality of the speech sounds themselves. The origins of the intonation and rhythm systems are described elsewhere (Taube-Schock 1993; Hill, Schock & Manzara 1992), but considerable refinement has taken place in the intonation system, and this must be the subject of a further paper.

Future work on the synthesis system itself will include further development of the rhythm and intonation, and further refinement of the rules used to create the parameters for synthesis. In addition, we shall be using the system for the kind of basic research outlined at the start of this section, and expect that the results of this work will also feed back into the synthesis system and help us to improve it.

Because of our experience in using the system complete with tools, we also expect to offer a complete system for researchers to use for the same kind of work, and would value any feedback about what facilities are desirable.

Acknowledgements

The authors acknowledge with gratitude the co-operation of Dr. René Carré, of the Laboratoire de Traitment et Communication de l'Information (LCTI) within Ecole nationale supérieur des télécommunications (ENST) in Paris, in Paris, whose published work and support have been an inspiration. The authors also wish to thank Dr. Gerard Chollet, a member of the same laboratory for his long term interest in the ongoing speech research, both at the University of Calgary and at Trillium Sound Research Incorporated. Dr. Chollet made the crucial introduction to Dr. Carré at the 2nd International Conference on Spoken Language Processing, in Banff, in October 1992.

The authors also acknowledge with gratitude the financial support of the National Research Council of Canada under the IRAP RDA program during a major part of this research. Last but by no means least, the authors wish to thank Dr. Clifford Anger, Richard Esau, Michael Forbes, Emil Rem, and Michael Whitt for their confidence in the work which led them to provide financial and business support during the development of the software, and agreed to make the software available under a General Public Licence when the company we all founded (Trillium Sound Research Inc.) was wound up.

References

BROADBENT, D.E. (1962) Attention and perception of speech. *Scientific American*, April

BROADBENT, D.E. & LADEFOGED, P. (1960) Vowel judgements and adaptation level. *Proc. Royal Soc. Series B (Biological Science)* **151** (944), Feb.

CARLSON, R. (1993) Models of speech synthesis. *Speech Technology Lab QPSR*, Royal Institute of Technology: Stockholm, 1-14.

CARRE, R. (1992) Distinctive regions in acoustic tubes. Speech production modelling. *Journal d'Acoustique*, **5** 141 to 159.

CARRE, R. (1994) 'Speaker' and 'speech' characteristics: a deductive approach. *Phonetica*, **51**

CARRE, R., CHENNOUKH, S. & MRAYATI, M. (1992) Vowel-consonant-vowel transitions: analysis, modeling, and synthesis. Proc. 1992 *Int. Conf. on Spoken Language Processing*, Banff, Alberta, Oct. 12-16 1992, 819-822

CARRE, R. & CHENNOUKH, S. (1993) Vowel-consonant-vowel modelling by superposition of consonant closure on vowel-to-vowel gestures. *3rd Seminar of Speech Production: Models and Data*, Saybrook Point Inn May 11-13 (submitted to the *Journal of Phonetics*)
CARRE, R. & MRAYATI, M. (1994) Vowel transitions, vowel systems, and the Distinctive Region Model. in *Levels in Speech Communication: Relations and Interactions*. Elsevier: New York
COOK, P.R. (1989) Synthesis of the singing voice using a physically parameterised model of the human vocal tract. *International Computer Music Conference*, Columbus Ohio
COOK, P.R. (1991) TBone: an interactive waveguide brass instrument synthesis workbench for the NeXT machine. *International Computer Music Conference*, Montreal, PQ, Canada
DUDLEY, H & TARNOCZY, T.H. (1950) The speaking machine of Wolfgang von Kempelen. *J. Acoust. Soc. Amer.* **22** (1), 151-166
DUNN, H.K. (1950) The calculation of vowel resonances and an electrical vocal tract. *J. Acoust. Soc. Amer.* **22** 740-753, Nov
FANT, G. (1960) *Acoustic Theory of Speech Production.* Mouton: The Hague328 pp, (Second Printing 1970)
FANT, G. & PAULI, S. (1974) Spatial characteristics of vocal tract resonance models. *Proceedings of the Stockholm Speech Communication Seminar*, KTH, Stockholm, Sweden
FLANAGAN, J.L. (1972) *Speech Analysis, Synthesis and Perception.* Springer-Verlag:New York (2nd Edition)
HEINZ, J.M. (1967) Perturbation functions for the determination of vocal-tract area functions from vocal-tract eigenvalues. *Speech Technology Lab Qrtly. Progress & Status Report* 2-3 1-14
HILL, D.R., SCHOCK, C-R. & MANZARA, L. (1992) Unrestricted text-to-speech revisited: rhythm and intonation. *Proc. 2nd. Int. Conf. on Spoken Language Processing,* Banff, Alberta, Canada, October 12th.-16th., 1219-1222
KELLY, J.L. & LOCHBAUM, C.C. (1962) Speech synthesis. *Proc. 4th International Congress on Acoustics,* Paper G42: 1-4
KOENIG, W., DUNN, H.K. & LACY, L.Y. (1946) The sound spectrograph. *J. Acoust. Soc. Amer.* **18**, 19-40, July
LADEFOGED, P. & BROADBENT, D.E. (1960) Perception of sequence in auditory events. *Qrtly. J Experimental Psychology* **XII** (3), August
LAWRENCE, W. (1953) The synthesis of speech from signals which have a low information rate. In *Communication Theory* Butterworth: London, 460-469
LIBERMAN, A.M., INGEMANN, F., LISKER, L., DELATTRE, P. 7 COOPER, F.S. (1959) Minimal rules for synthesising speech. *J. Acoust. Soc. Amer.* **31** (11), 1490-1499, Nov
MANZARA, L. & HILL, D.R. (1992) DEGAS: A system for rule-based diphone synthesis. *Proc. 2nd. Int. Conf. on Spoken Language Processing*, Banff, Alberta, Canada, October 12th.-16th., 117-120
MARKEL, J.D. & GRAY, A.H. (1976) *Linear Prediction of Speech.* Springer Verlag: New York, 288pp
PAGET, R. (1930) *Human Speech*. Harcourt: NewYork
SCHROEDER, M.R. (1967) Determination of the geometry of the human vocal tract by acoustic measurements. *J. Acoust. Soc. Amer.* **41**, 1002-1010
SODERSTEN, M., HERTEGARD, S., & HAMMARBERG, B. (1995) Glottal closure, transglottal airflow, and voice quality in healthy middle-aged women. J. Voice. June 2 182-97
TAUBE-SCHOCK, C-R. (1993) *Synthesizing Intonation for Computer Speech Output.* M.Sc. Thesis, Department of Computer Science, University of Calgary, Calgary, Alberta T2N 1N4

This page is intentionally blank