Part 2: RHYTHM – DURATION AND TIMING

Część 2: RYTM – ILOCZAS I WZORCE CZASOWE

From research to application: creating and applying models of British RP English rhythm and intonation

Od badań do aplikacji: tworzenie i zastosowanie modeli rytmu i intonacji języka angielskiego brytyjskiego

David R. Hill

Department of Computer Science
The University of Calgary, Alberta, Canada
hilld@ucalgary.ca

ABSTRACT

Wiktor Jassem's contributions and suggestions for further work have been an essential influence on my own work. In 1977, Wiktor agreed to come to my Human-Computer Interaction Laboratory at the University of Calgary to collaborate on problems associated with British English rhythm and intonation in computer speech synthesis from text. The cooperation resulted in innovative models which were used in implementing the world's first completely functional and still operational real-time system for articulatory speech synthesis by rule from plain text. The package includes the software tools needed for developing the databases required for synthesising other languages, as well as providing stimuli for psychophysical experiments in phonetics and phonology.

STRESZCZENIE

Publikacje Wiktora Jassema i wskazówki do dalszej pracy w sposób istotny wpłynęły na moją własną pracę. W 1997 roku Wiktor zgodził się na przyjazd do mojego Laboratorium Interakcji Człowiek-Komputer na Uniwersytecie w Calgary aby wspólnie zająć się problemami związanymi z rytmem w brytyjskim angielskim oraz intonacją w syntezie mowy z tekstu. Współpraca zaowocowała innowacyjnymi modelami, które zostały wykorzystane do wdrożenia pierwszego na świecie w pełni działającego i nadal funkcjonującego w czasie rzeczywistym systemu artykulacyjnej syntezy mowy z tekstu na podstawie reguł. System ten posiada moduł pozwalający na tworzenie baz danych potrzebnych do syntezy innych języków jak i usprawnienia umożliwiające prowadzenie badań psychofizycznych w dziedzinie fonetyki i fonologii.

1. Background

After spending time as an RAF pilot, attending Cambridge University and becoming a flight-test engineer, I returned to graduate school to follow the newly emerging disciplines of human factors and artificial intelligence.

My first post with ITT Europe (Standard Telecommunication Laboratories – STL – in Harlow, Essex, UK) involved an attempt to apply learning machines to the problem of automatic speech recognition. Amongst other things, this required considerable knowledge of speech structure, from the acoustic signal, through acoustic phonetics, phonology, prosody, and grammar. This was a hot topic at the time and there was a wealth of

research in many countries. Particularly notable centres of experimental phonetics included Gunnar Fant's Speech Technology Laboratory at the Royal Institute of Technology in Stockholm, the Haskins Laboratories (Frank Cooper, Pierre DeLattre, Alvin Liberman, and others), the Research Laboratory of Electronics at MIT, Bell Laboratories, the Speech Communication Research Laboratory in Santa Barbara (John Markel and June Shoup), and Peter Ladefoged's phonetics laboratory at UCLA.

Peter had completed his thesis under supervision from David Abercrombie in the Department of Phonetics at Edinburgh University in the UK and in collaboration with the recently-retired-from-University-College Daniel Jones, the leading phonetician at the time (who had worked with Henry Sweet, the grandfather of phonetics research). PL's thesis examination included Walter Lawrence, the inventor of PAT (the *Parametric Artificial Talker*, the first fully functioning parametric formant-based speech synthesising machine), as outside examiner. Walter was also the "Third Opponent" in Gunnar Fant's thesis defence. The third opponent, Walter said, was included as an examiner to provide light relief for a long public defence carried out in the presence of the King of Sweden. For the occasion Walter created a duet with OVE (Fant's synthesiser) and PAT. The synthesised song – "Daisy, Daisy ..." – later became famous during the shutdown of the computer, *HAL*, in Kubrick's film: *2001: a Space Odyssey* symbolising HAL's regression to its original untrained state.

I was privileged to meet all of these outstanding scholars, and more, during my early research into speech structure. They included Betsy Uldall and Laurie Iles at Edinburgh where I spent several working visits learning more about speech, phonetics and *PAT*.

One notable figure in linguistics research that I soon met, at the Fifth International Congress of Phonetic Sciences in Münster, was Wiktor Jassem. Wiktor, at the time, was working at the Polish "Instytut Podstawowych Problemów Techniki". The pictures in Figure 1 were taken during the 1964 conference.

Readers will know that there are many varieties of English in the world today with significant variation of accent and dialect. However, the gold standard was, for many years, the so-called 'Received Pronunciation' or 'RP' accent of English as heard by listeners to the British Broadcasting Corporation – the BBC or "Beeb" – before regional

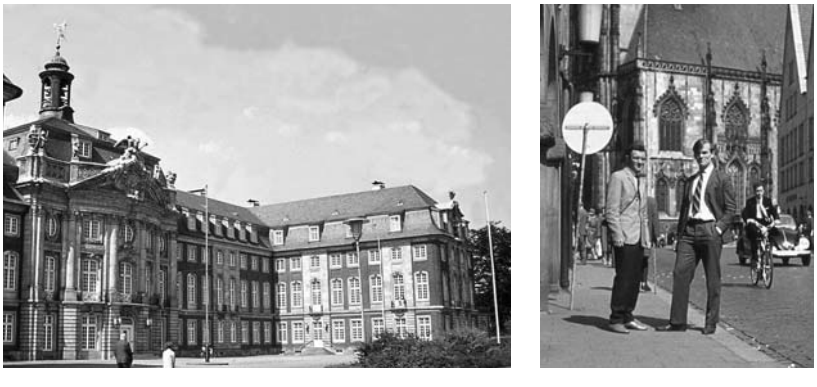

Figure 1: 5th ICPhS, Münster 1964; Wiktor Jassem and David Hill.

accents became acceptable. Listening to the BBC was part of how Wiktor Jassem learned to speak English perfectly, with a wonderful RP accent. I was amazed at this achievement. He spent time at University College, London, working with Peter Denes, A. C. Gimson, and J. D. O'Connor, and being influenced by the published works of the department's founder, and father of English phonetics, Daniel Jones. Windsor Lewis [1] provides a good summary of Wiktor's career (see also the *Biography* section of this volume).

I was overwhelmed by the combination of scholarship and charm as Wiktor took me under his wing for what was my first major international gathering a few short months after I had joined STL and begun my speech research in earnest. It made a big difference to my being able to follow what was said in this multilingual conference, as he explained difficult points in his perfect English. It was amazing to realise that he had not learned English growing up in England.

I persuaded Walter Lawrence to donate a working *PAT* to my project, as a result of my cooperation with the Phonetics Department at Edinburgh University, but it was not possible, for company reasons, to connect it with the PDP-8 computer I had managed to obtain (the first PDP-8 in Britain). Knowing it was essential to establish a sophisticated synthetic speech research facility in order to gain the data needed for speech recognition, I left STL and returned to Canada, which I had come to love after experiencing it during my RAF training. I accepted a post as assistant professor at the new University of Calgary. It turned out that I now had more freedom, but a lot less time. It seemed to be an acceptable trade-off!

It seemed obvious to me that, in order to make progress in comprehensive speech recognition, tackling the problem only at the acoustic-phonetic level was not enough. Ultimately one had to understand what was being said, at a fairly sophisticated level. This would involve rhythm and intonation, grammar, and usable real-world knowledge (with the ability to relate that knowledge to the speech being recognised). This still holds true.

However, I also knew full well that such a system was well beyond the limited resources at my disposal. I was also convinced that high-quality speech synthesis required similar knowledge resources. In order to speak an arbitrary utterance convincingly, and automatically, based on plain (i.e. not specially marked) textual input, the same knowledge of prosody, grammar, and real-world understanding would be necessary.

As a simple illustration of this obvious fact, consider a couple of simple examples. Suppose I am discussing with a friend a meeting later in the day. I say "OK, let's meet at 5 o'clock then." My friend replies: "No earlier!", or did she say "No, earlier!"? These apparently identical responses carry exactly opposite meanings, as determined by the rhythm and intonation that are poorly represented by the lone comma.

A more complex example involves seriously black humour. Suppose I arrive home where my mother and some friends are waiting. I ask: "What are we having for dinner mother?" – polite enquiry. Or I could say "What are we having for dinner ... mother?" (suggesting that mother may have failed to get us any dinner; there's a fall, a pause, and a different, more emphatic, rising intonation on "mother"). Or a very disturbing possibility would be "What are we having for dinner? Mother?" with pause and a rise-fall-rise intonation on "mother" which conveys a plan for cannibalism. The difficulty of conveying these possibilities in punctuated text, even with metacommentary, is indicative of the complexities of rhythm and intonation in conveying the different

possibilities. Such resources are used differently in different dialects (think of modern "up-talk").

Knowledge of the real world, and an ability to apply it involves a complex of information and abilities beyond mere dictionaries, grammar books and punctuation. Intonation and rhythm, in this context, can completely change the meaning of the words spoken. Our knowledge of how to do so is rudimentary, and dialect-specific, whilst our ability to link such changes automatically, to the real world, is currently well beyond the state of artificial intelligence.

Work on excellent speech-synthesis-from-text necessarily covers the gamut of what is required for comprehensive speech recognition, and also provides a much more visible and divisible research path – that is to say, one that is more easily subject to recognisable milestones and control. Experiments with synthesised speech, coupled with careful analyses of real speech, have allowed us to unravel important aspects of speech structure at many levels, building on the seminal work of the pioneers carried out in the early years, and particularly researchers in the 50s at the institutions partially listed above.

Implementing a basic system along these lines, using the technology and knowledge available at the time, occupied the period from when I started at the University of Calgary until the mid-70s, and included work by my post-doctoral fellow, Dr. Miroslav Preučil from the Technical University, Prague on the interpolating interface needed for the more modern, transistorised *PAT* I had been fortunate to obtain from Walter Lawrence, who also visited Calgary. Thus I finally managed to connect a modern version of *PAT* to my equally updated PDP-8/I computer. This step allowed me to develop a first system for computer-controlled speech synthesis using *PAT* nearly a decade after my thwarted attempt at STL. I used parameter generation rules developed as a result of my earlier work at Edinburgh and many scholarly papers. The work was reported at the 7th AICA Congress in Prague [3].

An amusing (and still inexplicable) story is associated with that success. Shortly after the successful hook-up, I was demonstrating the synthesis to a small group in my lab (see Figure 2). When I finished, I terminated the program. To my surprise, there was a sound of heavy breathing. So I shut down the computer. The heavy breathing continued. It was only when I removed the power supply plug from the wall that the heavy breathing was silenced.

Another continuing collaborator in the work during these early days was Ian Witten, formerly a graduate student of mine, but subsequently a lecturer at the University of Essex in the UK, and eventually a professor back at Calgary before his move to Waikato University in New Zealand. I made lengthy visits to Essex University in the 1970s to collaborate with Ian on the work[1]. Ian visited me in Calgary. We developed a simple version of the Halliday intonation model [4]. The synthesis program set (SEGSYN which computed the parameter events needed to generate the synthesiser parameter

[1] I designed and built a newer yet version of *PAT* + computer interface that fitted into a PDP-8/I rack module as a series of standard plug-in boards. One was connected to the departments's TSS-8 time-sharing system in Calgary, making it available for student projects, including course-work in the Human-Computer Interface course – CPSC 481. Another copy was installed on the PDP-11 system at Essex where it provided, as well as a research tool, the 'voice' for a game of 'Moo'.

Figure 2: David Hill's research group in 1970 with PDP-8/I and invisible *PAT*. *Clock-wise from left front: David Hill, Ian Witten, Bruce Akitt, Miroslav Preucil (all University of Calgary), Anton Rosypal (University of Alberta).*

variations, and RTSO, the real-time interrupt handler that produced the parameter tracks as required based on the event list from SEGSYN) was modelled on new work on computer operating system simulation. It solved a host of problems arising in managing the synthesis, allowing separation of segmental voicing control and intonation control, avoiding many timing problems, and allowing great flexibility in departure from a simplistic view of segment boundaries. For example, formant articulator movements do not all begin and end simultaneously. Such effects are important for high quality synthesis. The work is summarised by Hill [5]. The structure we developed has heavily influenced the synthesis work since. The collaboration prepared the ground for a most rewarding visit by Wiktor.

2. The rhythm work with Wiktor Jassem

I was delighted that, in 1976, I was able to persuade Wiktor Jassem to come to my laboratory at the University of Calgary to collaborate on a study of British English rhythm and intonation – an area in which Wiktor had considerable experience and knowledge. He was able to spare 6 months in Calgary – September 1976 to March 1977 – before travelling on to work with Harry Hollien in Florida.

As part of the preliminary work sketched above, I had carried out phonetic analyses and extracted the intonation patterns of utterances provided for two of Michael Halliday's study units in his book on intonation [2] (Study Units 30 and 39 – the latter being conversation between two people rather than separate utterances). I had also carried out some preliminary experiments to investigate the effect of the timing of an intonation rise

relative to the syllable boundaries in a nonsense word "mamama" for a generally declining contour [6]. The interesting result from that work was that the time of intonation rise is perceived categorically with respect to the syllable boundaries.

This was the starting point when Wiktor arrived in Calgary at a time when I had refined the program structure for speech synthesis [5] and was looking for better data with which to drive it.

Wiktor independently re-analysed both the Halliday Study Unit utterances to double check the phones I had analysed – their identity and their duration, and thus to provide solid data for initial model building and testing. The agreement was satisfyingly close with durations within ±2.5 milliseconds (generally regarded as the expected measurement error).

Our intent was to determine which of two models of British English rhythm – the Abercrombie/Halliday (AH) model or the Jassem (WJ) model – provided the better fit to our data. The two models were related. Both involved metrical rhythmic units with a "tendency towards isochrony" (a stress-timed rhythm as opposed to a syllable-timed rhythm like French). That is, the metrical units (rhythmic units) tended to be more equal in length than one would expect from the number of phonetic elements in the units. The difference between the two models lay in Wiktor's innovation of "anacruses" (ANAs), analogous to anacruses in music, notes whose quantity is extra to the strict rhythm of the bars and therefore discounted in assessing the musical rhythm. He considered that British English rhythm comprised two kinds of 'rhythm units': (a) *Narrow Rhythm Units* (*NRU*s) and (b) *Anacruses* (*ANA*s); the latter having no accented syllable, only unaccented syllables, spoken as quickly as possible, consistent with reasonable clarity; and that a *Total Rhythm Unit* (*TRU*) comprised an *NRU* possibly associated with an *ANA*.

In either model, there is a rhythmic unit that, like a musical bar, begins with a stressed syllable, or beat. Such syllables provide the basis for determining initial rhythmic unit boundaries (a necessity for automation of the division). The unstressed syllables associated with a rhythmic unit could be either enclitic (grammatically tied to the syllable/word at the beginning of the rhythmic unit) or proclitic (grammatically tied to the syllable/word at the beginning of the following rhythmic unit). Wiktor's model identified the proclitic syllables as *ANA*s. This means that any *ANA* always *precedes* an *NRU*, to which it belongs grammatically. The crucial difference between the two models of rhythm for spoken British English is the treatment of the unaccented syllables, which – according to the AH model – always follow the accented syllable as part of the same rhythmic unit (a *foot* in the AH model); but according to the WJ model, they are either preaccentual (preictic, that is, *ANA*s) or postaccentual (postictic), the two categories being subject to fundamentally different rhythmic patterning. In particular, the length of *ANA*s would tend to vary directly as the number of phones, showing little, if any tendency for the phones to shorten as the number increased. Clearly this difference will significantly affect any assessment of a tendency towards isochrony in the rhythmic unit durations (*feet* or *NRU*s).

Some linguists reject outright any notion of isochrony in spoken English; others simply express doubt; whilst yet others consider it a given. Regardless, many believe that the effect is at least partially perceptual, rather than objective. Certainly in our measures of foot durations (AH model), we found variation in foot lengths of up to 6 to 1. However, the question is not whether *feet* or *narrow rhythm units* have the *same* duration, regardless

of the number of phonetic elements, but whether a rhythmic unit is shorter than would be expected from knowing the number and type of the phonetic elements in the rhythmic unit.

One quibble I have with Wiktor is worth mentioning. Jassem et al. [7], page 207, make it clear that the WJ model assumes that the rhythm of English speech is a purely *phonetic* phenomenon, with no recourse to any other level of analysis-synthesis, such as grammar or semantics. I regard this as a rather dubious assumption, as the examples I gave at the start may suggest. Perhaps it depends on precisely what is meant by "purely *phonetic*".

Ian Witten, who was also visiting my lab at the time, took the lead in the statistical analysis of the data using the SPSS package, an analysis that involved considerable discussion. The results appeared as a departmental report [8] (the raw data is also available, by request).

A number of papers resulted from the work, the most pertinent of which was undoubtedly Jassem, Hill & Witten [7]. The paper opens with a set of questions on isochronicity and measurement methods to be asked in the context of British English speech rhythm. I will summarise some of main results to which Wiktor Jassem contributed and show how we applied these to a text-to-speech system that was originally implemented on a NeXT computer and has subsequently been ported to both Macintosh under OS X, and GNU/Linux under GNUStep[2]. The complete system not only allows normally punctuated English text to be spoken by computer as a service, but also provides the tools needed to create the databases for other languages, to modify the existing databases, or to produce stimuli for psychophysical experiments on speech.

3. The results

Statistical analysis of the segment duration data found that there was a small number of essentially independent factors determining rhythm from the many that were examined. The basic question asked was: "What determines the duration of a given phone or other unit in connected speech?" – that is, what were the sources of variation in element duration, since it is these durations and variations that provide the substance of rhythm (whilst intonation is partly responsible for stress or accent, which provide the anchor for both models). The results from the statistical analysis are reported in [8].

There were three main levels of analysis: the segment (or individual speech sound) level; the syllable level; and the rhythmic unit level (*feet* or *NRU*s/*ANA*s). At the segment level, we were interested to know what factors contributed to setting the durations of individual speech sounds, and in what proportion. Thus the basis of this analysis was *contribution to variance of mean segment duration*. At the syllable level, we wished to know what factors contributed to variance in mean syllable duration, and in what proportion. Finally, at the level of rhythmic units, we wished to know what

[2] All the software is available free under a General Public Licence at http://savannah.gnu.org/projects/gnuspeech in the SVN repository accessible from that page. Sources are provided as well as builds and links to papers, etc.

factors contributed to the variance of mean duration of rhythmic units (both Halliday's *feet* and Jassem's *rhythm units*), and in what proportion. From these processed data we hoped to deduce other things.

The segment level of analysis was of prime importance, since, as noted, specification of segment duration during synthesis was fundamental to determining the rhythmic timing. An important part of the two higher levels of analysis was to see how well the durations of syllables and rhythmic units were modelled on the basis of data gathered at the segment level, taking account of increasing numbers of factors.

The primary determinant of speech sound duration, as judged by contribution to variance in mean segment duration, was *type*. *Phoneme type* (47 types total) accounted for 45% of the variance in mean segment duration (*only* 45% which might be considered surprising).

Not all the various factors examined were independent. For example, the effect of syllable type on mean segment duration was not independent of the effect of phoneme type on mean segment duration, since phoneme type partially distinguished syllable type. Thus, although the type of syllable into which a segment fell was found to account for 14% of the variance in mean segment duration, it was not found necessary to take this into account as an additional, independent factor.

Secondly, the kind of rhythmic unit into which a segment fell – *tonic*, *final*, *final-tonic* and *unmarked* was reasonably independent and accounted for about 15% of the variance in mean segment duration. For most purposes, it was sufficient to consider only the two classes marked (the first three types) and unmarked.

Thirdly, there was significant reduction in segment duration proportional to the number of segments in a *foot* or *NRU* – the controversial "tendency towards isochrony". This accounted for about 10% of the variation in segment duration.

To summarise, based on Jassem et al. [7], two theories of English rhythm were tested as a basis for modelling British English rhythm for purposes of speech synthesis by computer: the AH foot-based model which postulates a single type of quasi-isochronous rhythm unit; and the WJ model which postulates two types of rhythmic unit, *anacruses* with no isochrony, and *narrow rhythm units* which tend towards isochrony. The results of regression analysis show that the tendency towards isochrony is minimal in *anacruses* and quite distinct, if not very strong, in *narrow rhythm units* and *feet*, but, of course, the foot (AH model) averages out and obliterates the distinction between *anacruses* and *narrow rhythm units* which was shown to be statistically highly significant.

Using an algorithm based on rules appropriate to the WJ model, the temporal organization of speech may be generated from a transcription indicating the occurrence of accent and thereby the start of *TRU*s. Separating the *ANA*s presents a problem as it requires some degree of grammatical analysis. A table of mean phone durations is also required and the durations for synthesis can then be modified by the rules invoked. The needed transcription can be produced using a comprehensive dictionary of word pronunciations which also indicates word stress placement. Grammatical disambiguation for words like "lead" /lɛd/, the metal, and "lead" /li:d/ the act of leading (e.g. a group) is required, though the problems are not all that simple. A simpler program can use foot boundaries, but then loses the information concerning *anacruses* which are relatively unaffected by any isochrony effect, but the approach does avoid the need for grammatical analysis.

Following the study, a subjective experiment was carried out to assess the naturalness of different intonation models – rather than rhythm – for synthetic speech [9]. Because the next step was to use the findings for *automatic* synthesis of spoken utterances from ordinary text, it was decided to work with the AH *foot*-based model to avoid the need for grammatical analysis – a possibility we reluctantly had to postpone to the future, for lack of resources. We would like to have run a rhythm model comparison, with copied natural intonation, but again our resources were limited. These are important experiments that we have not yet performed.

Dictionary look-up specified the segment identities, and identified syllables capable of taking stress – thereby automatically dividing the speech up into AH feet. Segment durations were then assigned, based on the model, the isochrony effect taking the form of a linear regression, whilst the other factors (marked versus unmarked) were covered by table look-up. The model accounted for ~71% of the segment durations found in our real speech analysis. None of the subjects mentioned any problems with the rhythm.

Three versions of each of the first ten sentences from SU 30 [2] were synthesised. One used a copy of the original intonation contour to control the pitch. The second used a synthetic contour provided by a simplified version of the IPO rules for synthesising British English intonation [10]. The third used a synthetic contour provided by a simplified version of Halliday's intonation specification, based on an algorithm implemented by Witten [11:pp201–207], but extended somewhat to include needed contours missing from the original Witten implementation. Subjects were asked to judge the utterances on a 6-point scale from 1 ('very unnatural') to 6 ('very natural'). Each of fifteen subjects heard each of the ten utterances three times in randomised order, but could not repeat an utterance they were currently judging.

Compared to the Witten implementation of Halliday's intonation system, as extended, the IPO approach is much more complete system, backed by considerable in-depth perceptual research. However, the IPO system achieved only slightly better mean scores whilst being considerably more complicated to implement, and showing extreme variability. The apparent perceptual advantage over the Witten system was not statistically significant, and the Witten system produced more uniform naturalness scores as may be seen in Figure 3.

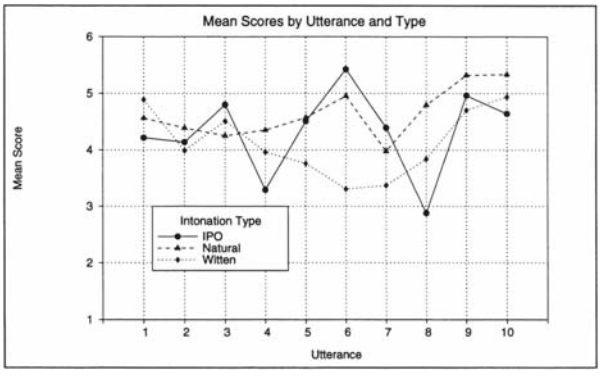

Figure 3: Mean Scores by Utterance and Intonation Type.

The synthesis program used had evolved from the earlier collaboration described above. At this point, in 1992, we had a new version of the *SEGSYN/RTSO* program ported to a NeXT computer, written by Leonard Manzara, a Ph.D. student in music working in my lab, and Craig Schock, one of my M.Sc. students [12]. The NeXT was chosen because it allowed the synthesiser itself to be simulated on the Digital Signal Processor (DSP) chip that was standard on the NeXT, avoiding the need for a separate hardware synthesiser. The whole text-to-speech system could run in real time without special hardware. Leonard suggested that we should form a spin-off company to make and sell our synthesis software, and so we set up *Trillium Sound Research Inc*.

At the *Second International Conference on Spoken Language Processing*, where we presented the work just described, we met up with Gérard Chollet (a one time special Ph.D. student of mine now at the Département Signal, École Nationale Supérieure des Télécommunications). He told us about work by René Carré in his institute on a *Distinctive Region Model* (*DRM*) of the vocal tract which allowed control of a waveguide-simulated vocal tract by just eight parameters [13], based on the results of earlier formant sensitivity analysis by Fant and Pauli [14]. Shortly after obtaining this exciting information (previous waveguide simulations had involved 40 parameters and a problem with instability) NeXT Computer Inc. announced that the NeXT computer hardware was to be discontinued, but the software would be ported to the Intel PC hardware. This was a blow, because such hardware did not come with a DSP as a standard co-processor.

Realising we faced yet another software port, but excited by the news from Gérard, we decided that the port would use the *DRM* vocal tract instead of the formant synthesiser we had been using from the early days, in various forms. This meant that our databases would have to be recreated, since formant data was only indirectly linked to vocal tract shapes through the *DRM* controls.

Seven months of concentrated programming and acoustic analysis followed. Leonard programmed the *DRM*-controllable acoustic tube model, *tube*[3], complete with a nasal branch, and an interface application, *Synthesizer*, to allow direct experiments with *tube*, while Craig developed the editing program needed to build the databases – "My Own Nifty Editing Tool", or *MONET* for short, later rendered as *Monet*. Leonard and I started work with *Monet* whilst it was still under development, providing feedback, which helped to ensure it included the features we needed and that the bugs were removed. We had a Kay *Sona-Graf* to perform the acoustic analyses, and the high quality sound on the NeXT for critical listening, plus *Synthesizer*.

[3] The *DRM* control model for *tube* controls eight tube sections of unequal length determined by the nodes of the format resonances in a uniform tube. To get sections that are of exactly the right length would require a tube model with 30 sections, pooled into eight supersections of varying length corresponding to the *DRM* regions. The problem with this approach is that the shorter the waveguide sections are, the higher the computation rate has to be. On the original NeXT implementation, the DSP was not fast enough to do the necessary computations in real time for a 30 element version. Fortunately a division into ten sections, with the four centre sections ganged in two pairs, together with the surrounding 6 sections, produces a fairly close approximation to the required eight super-sections. This was a vital insight to making the tube model perform in real time, given the relatively limited computational power. With faster CPUs, the exact division model is now possible.

We used the NeXT hardware because we knew the NeXTSTEP development environment would be available on the new hardware. We were able to spec a Turtle Beach *Multisound* plug-in board to provide a DSP. The procedure for creating the databases was pragmatic as no such databases existed. We were, of necessity, pioneers.

We had previously defined successive speech postures (loosely *phonemes*) in terms of formant values, accompanied by information about voicing, voice onset time, aspiration, sibilant noise properties, and possible noise bursts or other 'special events'. Some of this information was applicable to the new articulatory approach (since we would generate the voicing and noises directly, rather than attempting to generate them by acoustic simulation of the vocal folds and airflow). However, we needed the *DRM* radius parameters required to represent the vocal tract configurations for the postures. This was achieved by experiment with *Synthesizer* to determine the tube radii that gave the correct formant values for each posture, using knowledge of the real articulation for each as a constraint. This was based on the seminal work in the 50s that did so much to enlighten us about the acoustic-phonetics of speech. It also used knowledge of the effect that changes in the tube radii in the eight regions had raising or lowering the three main formant values. These effects are shown in Figure 4. Note that important advantages of the articulatory synthesis using the tube model included an accurate simulation of the nasal branch and related energy exchange with the main oral tube, plus correctly produced dynamic higher formants, avoiding any need for a lumped filter to simulate these formants, as required for the earlier formant synthesiser [15].

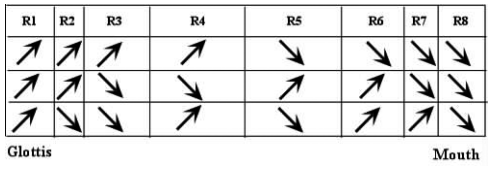

Figure 4: Configurations for articulatory synthesis.

The framework for posture (segment) parameter construction evolved from the simpler version so as to allow *diphones*, *triphones* and *tetraphones*, in order that items like consonant clusters and co-articulation could be handled more appropriately. *Monet* allows the creation of rules for particular combinations of postures (which can be specified in terms of specific postures, or in more general terms related to posture classes). The rules invoke transition profiles specifying the shape and timing for each of the parameters independently, but in time relation. This also allows such features as vowel reduction. Special transition profiles serve to allow arbitrary noise bursts to be added and, because the original concept of linear superposition was preserved, all effects for a particular parameter, as dictated by the rules and profiles, can simply be added to obtain the complete parameter variation. Thus micro-intonation associated with constriction in the vocal tract is simply added to whatever pitch value is called for by the overall intonation pitch variation, whilst voice onset time is simply determined by an appropriate normal transition profile.

The rules are arranged in a hierarchy from most general to most specific. Any posture combination that is not caught by a specific rule eventually defaults to the most general rule and profile – *phone-to-phone*.

4. Epilogue

Not long after *Trillium Sound Research Inc.* had completed, and started selling the new real-time articulatory based synthesis system for NEXSTEP-on-Intel-Processor hardware, Turtle Beach stopped making the DSP card and Steve Jobs killed NeXT, rejoining Apple Computer in the process and taking the "best software on the planet" with him. It all made perfect sense, but caused the demise of many companies that had built their business on the existence of NeXT, including ours. We therefore transferred our entire proprietary software, including the tools we had used, and databases we had created, to the Free Software Foundation as a GNU Project "gnuspeech". The package has largely been ported to the Macintosh under OS X, and somewhat less to GNU/Linux under GNUStep, following the hiatus due to the closure of NeXT. On OS X, only the database creation modules, the dictionary editor, and some of the graphs and bug removal for *Synthesizer* remain to be dealt with. The *Monet* displays, and speech synthesiser daemons are complete and produce speech according to the models and data described.

The software, including sources, is available for download and use under a General Public Licence (GPL).[4] There are three branches. The original NeXT version, the Macintosh OS X version, and the GNU/Linux version. Anyone who can help with continued development of this versatile software suite, or tries it out and has questions, is encouraged to contact the author. The system is not only valuable for text-to-speech conversion, producing high quality speech, but, when complete (as the NeXT version is) is also a tool for creating the databases for languages other than English; for experimental phonetics; for speech aids for the visually impaired [16]; and many other applications. Manuals and related scientific papers are available[5] along with samples of the speech produced from ordinary punctuated English text together with the most recent description of the work [17].

In closing, it must be said that our work on automatic speech synthesis and recognition owes a significant debt to Wiktor Jassem, even though, from point of view of his distinguished career, it was only one small element in his lifetime of scholarship and research. We acknowledge our debt, and are grateful for the opportunity to know and work with him.

References[6]

[1] Windsor Lewis, J. 2003. The contribution to English phonetic studies of Professor Wiktor Jassem. *The Phonetician*. 87, 19–21. (See also Biography, this volume.)

[2] Halliday, M. A. K. 1970. *A Course in Spoken English: Intonation*. London: Oxford University Press. 134 pp + tape recordings.

[3] Hill, D. R. 1973. Control of an analogue speech synthesiser by a time-shared digital computer. *Proc. 7th. AICA Cong. (Hybrid Computing), Prague*, Aug 27–31, 249–252.

[4] http://savannah.gnu.org/projects/gnuspeech (Subversion repository)

[5] http://pages.cpsc.ucalgary.ca/~hill

[6] Many of the papers are available at http://pages.cpsc.ucalgary.ca/~hill

[4] Hill, D. R. (1975) Computer models for synthesising British English rhythm and intonation. *Proc. 8th. Int. Cong. of Phonetic Sciences*, Leeds, UK, paper 129. 17–23.

[5] Hill, D. R. 1978. A program structure for event-based speech synthesis by rules within a flexible segmental framework. *Int. J. Man-Machine Studies* 10 (3), 285–294.

[6] Hill, D. R. and N. A. Reid N. A. 1977. An experiment on the perception of intonational features. *Int. J. Man-Machine Studies* 9 (2), 337–347.

[7] Jassem, W., D. R. Hill and I. H. Witten, I. H. 1984. Isochrony in English speech: its statistical validity and linguistic relevance. In: D. Gibbon and H. Richter, eds., Intonation, Accent and Rhythm: *Pattern, Process and Function in Discourse Phonology*. Berlin: Mouton de Gruyter. 203–225.

[8] Hill, D. R. 1977. Some results from a preliminary study of British English speech rhythm. *94th. Meeting of the Acoustical Society of America*, Miami (Research Report 78/26/5, 28 pp). 12–16.

[9] Hill, D. R., C.-R. Schock and L. Manzara. 1992. Unrestricted text-to-speech revisited: rhythm and intonation. *Proc. 2nd. Int. Conf. on Spoken Language Processing*, Banff, Alberta, Canada. 1219–1222.

[10] Willems, N., R. Collier and J. 't Hart. 1988. A Synthesis Scheme for British English intonation. *J. Acoustical Soc. America* 84, 1250–1261.

[11] Witten, I.H. 1982. *Principles of Computer Speech*. Academic Press: London.

[12] Manzara, L. and D. R. Hill. 1992. DEGAS: A system for rule-based diphone synthesis. *Proc. 2nd. Int. Conf. on Spoken Language Processing*, Banff, Alberta, Canada. 117–120.

[13] Carré, R. 1992. Distinctive regions in acoustic tubes. Speech production modelling. *Journal d'Acoustique*, 5 141–159.

[14] Fant, G. and S. Pauli, S. 1974. Spatial characteristics of vocal tract resonance models. *Proc. Stockholm Speech Communication Seminar*, KTH, Stockholm, Sweden.

[15] Hill, D. R., L. Manzara, and C.-R. Taube-Schock. 1995. Real-time articulatory speech-synthesis-by-rules. *Proc. AVIOS 95' 14th Annual International Voice Technologies Conf*, San Jose, 27–44.

[16] Hill, D. R. and C. Grieb. 1988. Substitution for a restricted visual channel in multi-modal computer-human dialogue. *IEEE Transactions on Systems, Man & Cybernetics* 18 (2), 285–304.

[17] Hill, D. R. In press. A GNU approach to speech database creation for articulatory speech synthesis by computers. *Proceedings of Interspeech 2012, Portland, Oregon*.