Talking with computers.

Simon Lavington. *July 2023.*
lavis@essex.ac.uk

Two curious devices were connected to the Ferranti Atlas computer at Manchester University in the 1960s. One was an on-line X-Ray Diffractometer; the other was simply known as the Speech Converter. This is the story of the latter.

Modest research into optical character recognition had been carried out in the mid-1950s in Manchester. By the autumn of 1962, when the first Atlas was nearing completion, it was decided to explore a more ambitious class of pattern recognition, namely speech. Could the new Supercomputer be made to respond to spoken commands? Furthermore, it would be fun if the machine could respond by uttering synthesised speech. So I, an innocent research student, was recruited to start the ball rolling.

As was customary, the first step was to simulate candidate recognition and synthesising schemes in software – meaning that realistic samples of speech waveforms had to be captured and digitised. This, in turn, required an on-line Analogue-Digital-Analogue converter – known locally as the *Speech Converter*. This equipment was unusual at the time (1963) because it handled DC to 10 KHz with 8-bit accuracy in order to give reasonable-quality speech. It took an 8-bit sample of an input waveform every 50 microseconds, so its potential data rate for input or output (20 Kbytes/sec) was at the time impressive. For comparison, the transfer rate for each Atlas high-performance magnetic tape decks was 48 Kbytes/sec. By the way, the idea of an 8-bit byte as a standard unit of information was not yet commonplace when Atlas was being designed. Atlas used 6-bit characters and 48-bit words.

The Speech Converter was a real-time device with a *crisis time* – meaning that the Atlas CPU must guarantee to service a Speech Converter's interrupt within a critical time if continuous speech of arbitrary duration was to be accommodated. Without buffering, the critical time was very short - 50 microseconds. It was decided that the Speech Converter should therefore assemble six eight-bit samples and issue a request for a 48-bit Atlas word transfer every 300 microseconds. Double-buffering was employed, so the Speech Converter contained two 48-bit flip-flop buffer registers. The total time spent by the Atlas Supervisor (ie Operating System) dealing with each Speech Converter interrupt was about 40 microseconds – once higher-priority interrupts had been dealt with. Recall that the Atlas CPU shared its time between potentially many asynchronous activities.

As it was, one did not lightly run the Speech Converter for more than a few tens of seconds at a time. On Atlas, this meant that incoming data had to be rapidly moved through Virtual Memory from core to drum to magnetic tape in real time. When I

used the Speech Converter for input, the computer was usually entirely devoted to my endeavours. I remember Frank Sumner, my Ph.D. Supervisor, and sundry maintenance engineers and operators all standing around with a look of bemused amazement (or was it anxiety?) on their faces whilst this young research student tried to bring the most powerful machine in the world to its knees! I was not too popular, since most other user-jobs ceased whilst I did my stuff. I remember Frank's wry smile as I stood there nervously counting each 3 Kbyte block of data that was being written up to tape – you could hear the deck jerk forward as each block was written. He remarked that here was I, surrounded by the world's most powerful machine, but counting on my fingers!

Once the Speech Converter's Supervisor routines had been written in machine code, a library of eight signal-processing routines plus an initialising program were written by two research students (myself and Lynn Rosenthal) using the Compiler Special facilities in Atlas Autocode, an Algol-like language. The library, called SPP1, was stored on the same magnetic tape that contained samples of connected speech. SPP1 could be called down by applications programmers.

The SPP1 software.

From the programmer's viewpoint, the Speech Converter simply produced a very large one-dimensional array of fixed-point numbers in the range -127 to +128, representing waveform amplitude-values every 50 microseconds. If this was the result of continuous speech, the first problem was to distinguish between background noise and meaningful utterance, then to segment the speech roughly into words and finally to try and identify phonemes within words. The first six routines of SPP1 were used to get an overall picture of the total record, using statistical measurements derived experimentally. The final two routines performed detailed spectral analysis on chosen sections of the record. In all cases, the user provided each routine with several parameters, some of which related to the time-intervals of the waveform record being inspected and others of which controlled the detailed algorithms contained withing the body of a routine. Without going into too much detail, here is a summary of the eight routines in SPP1.

● r1: prints an amplitude versus time graph of all the digitised samples between a start- and end-address, starting with the first-occurring value greater than a specified threshold. Additionally, certain statistics are printed based on three measurements: the number of zero-crossings (axis-crossings), z; the number of positions of zero time-derivative ('turn-arounds'), t; the function (t – z). *Use of r1*: gives an overall view of the waveform being inspected.

● r2: amplitude bar graph, giving maximum amplitude value during each period of 128 samples (equivalent to 6.4 milliseconds) in a longer section. *Use of r2*: overall view of the likely start- and finish- times of each spoken word in continuous speech.

●r3: Produces counts of z, t, (t – z) and amplitude over successive 600-sample (30 millisec) intervals. Values of these statistics enable a rough classification into sections of the audio record that are either sustained high frequency, turbulence, voiced or noise. *Use of r3*: gives a rough guide to utterance identification.

● r4: same as r3 but on a more precise timescale. *Use of r4*: defining rapid transitions between phonemes and quasi-steady state periods within diphthongs.

● r5: more detailed statistical analysis of zero-crossing data. *Use of r5*: aid to word segmentation in continuous speech.

● r6: fundamental frequency (voice pitch) detection, using a combination of autocorrelation analysis and amplitude peak measurements. The analysis is carried out on consecutive 33 millisec sections of speech, overlapping by 50%. *Use of r6*: as a parameter in speech synthesis systems; also, for pitch-synchronous analysis as in r7.

● r7: Spectral analysis by Fourier analysis. This took frequency spectral cross sections of portions of the speech waveform by forming the Fourier transform of the autocorrelation function. It effectively passed speech consisting of repetitions of the record to be analysed through a series of filters whose bandpass characteristics were pre-defined.

● r8: Spectral analysis by continuous filtering. This routine passed the speech waveform continuously through a set of filters formed from 2nd-order difference equations. This faster routine differs from r7 in that the waveform is continuously passed through filters, rather than discrete portions of record being analysed. It is thus possible to read the filter outputs and produce cross sections at frequent intervals without greatly affecting the computing time, except for time consumed in actually printing results if printing is needed. In effect, advantage is being taken of the computations already performed in previous cross sections. Pitch synchronous analysis is not provided with this routine.

Applications of the Speech Converter and SPP1.

My particular application was automatic speech recognition. At the time, other researchers elsewhere were mostly using analogue equipment. They typically reported accuracies of 95% when dealing with very limited vocabularies – usually the spoken digits zero to nine. Details such as any prior speaker-training or number of different people able to use their equipment were curiously omitted from the reports. In short, this was indeed the very early days of speech recognition.

My recognition scores were similarly pathetic, though I strove *ab initio* to deal with speaker independence by studying a wide range of male and female accents. I developed algorithms that were mathematically simple, fast, and capable of being replicated in external circuitry, thus greatly reducing the incoming data rate to the computer. But basically all my own Ph.D. thesis showed was that the recognition problem was difficult!

My ideas were later taken up by another Manchester research student, Brian Carpenter, who implemented a real-time speech recognition system for a PDP8 computer. To assess performance he used two vocabularies: one was the spoken digits zero to nine; the other was a five-word vocabulary with direct visual feedback. This gave voice control (*up, down, left, right, stop*) of the movement of a spot on the PDP8's display screen. I remember some users frantically shouting "Stop!" if the spot

was moving inexorably off the edge of the screen! They were ignoring Brian's advice to 'speak calmly' – thus spawning a further line of investigation into so-called *human factors.*

In the speech synthesis area, two fellow research students, Ron Mathers and Lynn Rosenthal, had more luck. Their starting-point was a vocoder-type analogue system called *PAT* – (Parametric Artificial Talker) that had recently been demonstrated at the University of Edinburgh. This apparatus was a marvellous assembly of buzzers, hissers and band-pass filters, controlled by eight parameters that were periodically updated. As part of unrelated research, Ron and Lynn wrote a software simulation of PAT using eight 3-bit parameters, updated every fortieth of a second if I remember correctly. To judge the quality of the synthetically-generated speech, the data was streamed to the outside world via the Speech Converter. We stood around delighted by the sounds! Adaptations of the PAT model were investigated and the voice quality improved. Two more Ph.D.s in the bag!

At about this time we were interviewed by a puzzled BBC reporter, investigating 'talking computers'. At the end of the chat, we were asked whether we could synthesise a short BBC announcement in a rich northern accent for their early-morning news programme. We laboured through the night, recording our sample BBC announcement spoken by one of the Electrical Engineering Department's porters who had a distinctive Hilly Lancashire accent in the style of Fred Dibnah. We fed his voice in through the Speech Converter, analysed the waveform to produce sets of the eight parameters required by the PAT model and then fed the synthesised version of Fred's announcement as a string of 8-bit integers representing 50-microsecond samples out through the Speech Converter. In the early hours, a courier took the audio tape to the BBC's Manchester studio. Sure enough, over the radio next morning came not the usual posh south-of-England accent but 'Fred', saying something like: "The time is coming up to seven thirty and you're listening to BBC Look North". Several senior colleagues arrived at work that morning to describe how, unbelievable though it seemed, they swore they had heard 'Fred' the porter make a radio announcement!

What did the equipment look like?

The picture on the next page shows the Speech Converter's performance being checked by me in about 1964. The photo was taken in one corner of the crowded Atlas machine room at Manchester, where the device was installed once Tom Kilburn, everyone's boss and inspiration, had been assured that a prototype had successfully operated in another lab. The thought of some maverick device tripping the power supply of a £2 million installation could not be contemplated! Interestingly, the prototype was built using a 50-volt Westat power unit and the steel framework of a 19-inch Post Office rack left over from the 1948 Baby computer project. The final product was much smarter: it used standard Atlas style of cabinetry and racking for the printed-circuit boards on which discrete components (germanium transistors in those days) were mounted. From memory, the unit in the photo measured approximately 120cm high, 55 cm wide and 45 cm deep.

Looking at the photo, the Speech Converter's front panel included:

- Input and Output amplifier gain control and a volume meter;
- Provision for remote control of an external audio tape recorder or other instrument, by arranging a suitable pause for the tape to reach a steady speed before commencing input or output transfers;
- a 'wait' button that allowed temporary manual halting of input when using a microphone for spoken commands;
- Speed selection: if the full 10 kHz bandwidth was unnecessary for a particular application, the sampling rate and hence bandwidth could be halved so as to give I/O transfers every 600 (not 300) microseconds.

What happened in the end?

In addition to voice studies in Manchester, the speech converter was also used briefly by the Admiralty for a classified signal-processing task. This contract was handled by Ferranti and I was not involved. My guess is that it was concerned with underwater sounds – from submarines or ship's propellers? The Joint Speech Research Unit (an amalgamation of the speech research interests of several UK government departments) was also interested in using the Speech Converter. They came up and watched a demo but I'm not sure whether they actually used it for their speech investigations. Lowly research students were not consulted! STL Harlow did use the device and our software routines for some work on automatic speech recognition. I know they complained bitterly of the cost that Ferranti charged them!

When the Manchester Atlas was shut down in September 1971, the Speech Converter was finally switched off. It was preserved in storage, along with other Atlas bits, and certainly still existed when I left the University in 1986. It has since vanished.

Was speech research continued at Manchester? No. Ron Mathers, whose main interest lay in road traffic simulation and control, went back to New Zealand. Lynn Rosenthal, whose interest had turned to simulating neural nets, went back to America, then to Nigeria, and finally to open an underwater photography and scuba diving school on a Caribbean island. Brian Carpenter left to join CERN. By 1968 most of the efforts of the computing group, including myself and Frank Sumner, were focussed on the design of a new computer, MU5. The attraction of being able to talk to computers had faded.

Of course, nowadays what was formerly a quaint and unreliable party trick has become the robust and useful tool of automatic speech recognition. I readily admit that I am amazed and delighted every time my smart phone responds with complete accuracy to my spoken messages.

More information on the Speech Converter and SPP1 can be found here:
Some facilities for speech processing by computer. S. H. Lavington and L. E. Rosenthal. The Computer Journal, Volume 9, Issue 4, February 1967, pages 330 – 339. See: https://academic.oup.com/comjnl/article/9/4/330/390149