Proceedings of the Institute of Acoustics

A NEW COMPUTER-BASED SPEECH THERAPY TUTOR OFFERING IMMEDIATE AND DEFERRED VISUAL FEEDBACK

J.M. Turnbull (1) A.T. Sapeluk (1) R.I. Damper (2)

(1) Dundee Institute of Technology, Dundee, Scotland
(2) University of Southampton, Southampton, England

Abstract

This paper discusses a computer-based speech therapy tutor which exploits recent developments in high-speed processor technology to display an assessment of vowel quality as an utterance is being made. The visual feedback methods, the analysis algorithms used, the system hardware, and the results of early clinical trials are covered.

1 Introduction

Computer-based speech therapy tutors have been in existence for a number of years, aimed primarily at the hearing impaired[1,2,3]. The majority of such tutors are limited in that they do not continuously display information about the quality of the speech sound during the utterance production. By contrast, our system exploits recent developments in high-speed processor technology which enables it to display an assessment of vowel quality as an utterance is being made.

In its present form the system contains two visual feedback modes, namely, the *immediate feedback* and the *deferred feedback* modes. Using the immediate feedback mode, patients can experiment with moving their tongue, jaw, and lips, while watching the display to see if their production is close to the target sound or not. This allows them to learn the correct vocal-tract shape for a particular sound experimentally. If the therapist requires that the vowel appear in a consonant-vowel or consonant-vowel-consonant context the other feedback mode is used. In this case, the vowel portion must first be located and so feedback is deferred until the whole utterance has been produced.

As the patient is being trained by the therapy tutor, the vowel portion of his or her monosyllabic utterance is matched against a set of prototypes which have been chosen by the therapist. Ideally, these prototypes would be obtained by the therapist and patient working with the tutor in learn mode, the therapist indicating to the computer when a good example has been spoken. Alternatively the prototype can be obtained from a tape recording of the therapist and patient working together and the identified satisfactory utterance used as the model.

The analysis is based on a novel pole-tracking algorithm in which z-plane pole-pairs of the linear prediction model[4] are represented as a single point on a real (rs) plane[5]. Movements of the pole-pairs are tracked in order to identify steady-state regions and, in deferred mode, boundaries delimitting the vowel portion. Subsequently, the patient's utterance is matched against a target, or prototype, speech pattern. We will discuss the methods of obtaining the steady-state regions and boundaries as well as the methods of pattern matching adopted.

The system hardware is based on a portable personal computer with additional computing power in the form of a digital signal processor (TMS32020) and a single floating-point transputer (T800). These high-speed processors reside on plug-in cards so that all hardware is contained within one single unit. The operation of the hardware from microphone input to visual output, the implementation of the analysis and recognition algorithms, and the communications between the concurrent processes will be discussed.

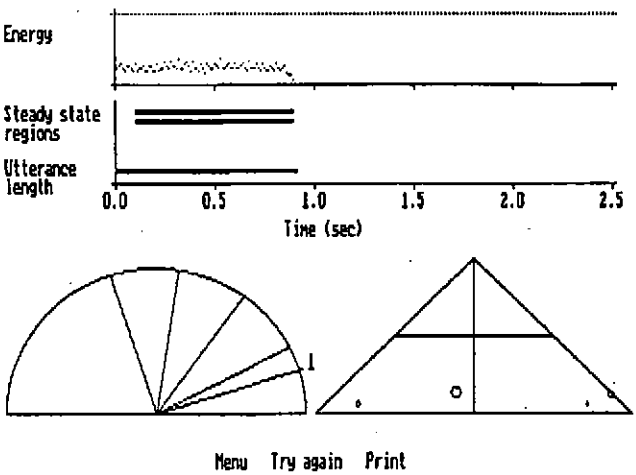

Figure 1: the deferred feedback mode display.

2 The Feedback Modes

In its present form the system contains two modes of operation, namely the *deferred feedback* and the *immediate feedback* modes. The deferred feedback mode is used when the therapist requires that the vowel appear in a consonant-vowel or consonant-vowel-consonant context. In this case, the vowel portion must first be located and so feedback is deferred until the whole utterance has been produced (with the template matching suppressed, this mode is also used when generating prototypes).

Figure 1 shows a typical screen display after the analysis of an utterance using this mode. This display comprises five sections. At the top is a graph against time of the energy (or loudness) of the sound. Below this is graph against time which shows two things, namely the length of any steady-state regions of the utterance, and the length of the utterance itself. Underneath these two graphs are a 'dial' type display, for indicating the closeness of the utterance to a prototype, and a triangular display, for displaying the more technical information (see Section). The remaining section is the horizontal list of options at the bottom of the screen. As can be seen, the therapist can choose to exit this mode (return to the *Menu*), to continue in this mode and have the patient *Try again*, or to get a hard-copy *Print* of the current screen display.

Prior to analysing the utterance, the Tutor will display these four plots devoid of any information and will bleep once to indicate that is waiting to analyse an utterance. As the patient produces the utterance the energy plot is updated. This gives the patient an indication as to the length of his or her utterance so far. When the patient stops producing the utterance the remaining information is displayed with no noticeable delay.

The lower trace against time will show how steady the produced sound was. Ideally for a monosyllabic sound, with the vowel extended, this should be a continuous line which is slightly shorter than the length of the utterance. If the sound was not particularly steady then this line will be fragmented (or not present if the sound was transient throughout). The longest steady state region is emphasised by being displayed as a double bar, and this is the region that is taken for template matching.

The dial display gives an indication of how close the utterance was to the one of the prototypes. The

A SPEECH THERAPY TUTOR

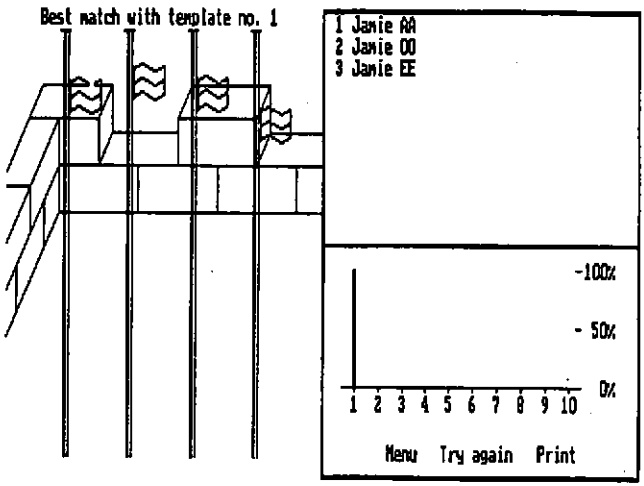

Figure 2: immediate feedback mode display.

pointer moves clockwise from the left; the left most position indicates a very poor match, and the rightmost position indicates a perfect match.

The other mode, the immediate feedback mode, displays results as the speech sound is being produced. Using this mode, patients can experiment with moving their tongue, jaw, and lips, while watching the display to see if their production is close to the target sound or not. This allows them to learn the correct vocal-tract shape for a particular sound experimentally.

Figure 2 shows a typical screen shot for the immediate feedback mode. Again there are five regions of displayed information. The largest region is that containing four flags on flag-poles which are situated on top of a castle tower. Above these flags is a statement of which prototype template the sustained utterance got closest to. The box in the top right of the screen gives a list of the name or identification of each entry in the database (in this case there are three entries). The box in the lower right of the screen displays a histogram of the performance throughout the utterance. This box also contains a list of options that the therapist has, which is identical to the list for the deferred feedback mode.

Prior to analysing the utterance, the screen is drawn as shown but with the flags at the bottom of each pole, no statement of best match, and no bars on the histogram. The Tutor then bleeps once to indicate that it is waiting to analyse an utterance.

As the sound is produced it is analysed and matched against *all* templates in the current database. The heights of the flags correspond to the difference between the utterance and the closest prototype template.

In the early days of the Tutor, the flags at the end of the utterance represented the last frame analysed. However, as the utterance is being finished, control over the articulators is relaxed, and consequently the flags would drop. The final static result displayed would therefore not be representative of the patients overall performance. Therefore the best (highest) position of the flags is recorded and displayed at the end, giving a good indication of the patients performance.

Once the flags have been repositioned, the statement is made as to which prototype the utterance came closest to, and the histogram is drawn. The histogram shows the percentage of closest matches to the prototypes throughout the utterance duration.

3 The Analysis Methods

The Tutor described in the previous section offers two modes of feedback, namely the *deferred mode* and the *immediate mode*. The analysis methods of the deferred mode must be capable of two main tasks: in the worst case an extended vowel must automatically be extracted from a monosyllabic consonant-vowel-consonant utterance, and a template of parameters representing this segment of speech must be compared, or matched, with the templates representing the prototypes, or target sounds.

For the immediate mode there is no need to extract a specific segment of the utterance. Templates representing short contiguous segments are matched against the templates of the prototypes. The analysis methods for the immediate feedback mode is therefore a subset of the methods employed in the deferred feedback mode.

To extract the extended vowel from the CVC utterance a set of analysis parameters must be tracked. The parameters will be rapidly changing during the analysis of the dynamic consonants either side of the vowel and will be static, or at least slowly varying, during the analysis of the vowel. Hence if the deviation of the parameters is monitored, it will show peaks for the consonants and a trough for the vowel, and hence the vowel can be automatically extracted from the utterance as the segment of the speech corresponding to this trough.

The analysis parameters used in this particular application are the poles of the Linear Prediction (LP) model for speech production. A method of determining the poles of the LP model has been previously proposed[5] which employs Bairstow's method of extracting quadratic factors of a polynomial. Bairstow's method is an iterative method and, as with all iterative methods, suffers from divergence if the *initial guess* is not sufficiently close to the true value. However, properties of the speech signal can be utilised to limit, or at the very least rapidly detect, this effect: it is a reasonable assumption that the LP model will be stable for slowly varying sounds such as vowels, semivowels, and the continuants (this, of course, restricts the application, and hence the Tutor, to slowly varying sounds). The more dynamic plosives cause no problems: the plosives at the commencement of the utterance will be short and hence if the first few analysis frames of the utterance are ignored its effects are minimal. Any plosives at the end of the utterance will also be ignored: the Tutor automatically detects the start and finish of the utterance by monitoring the speech signal energy. The stop before the release of the plosive will be taken as the end of the utterance and the subsequent speech ignored.

It has also been shown that tracking the poles on the z-plane has inherent problems. Tracking requires that each pole from the current analysis frame be related to all poles of the previous frame. For the 8th order model chosen[6] this would require determining the optimal correlation out of $8!=40320$ correlations. There are also problems of discontinuities when complex-conjugate poles close to the real axis become real poles within consecutive frames.

These problems are overcome by tracking the pole movements on the two dimensional real rs-plane as opposed to the complex z-plane[5]. Pole-pairs on the z-plane are represented as a single point on the rs-plane. This non-linear mapping is exploited to reduce the number of permutations considered during the tracking process from 8! to 4!. This reduces the computational load by a factor of 1680. The mapping from the z- to rs-plane is also not a conformal mapping, and this property is exploited to remove the discontinuities associated with complex pole-pairs turning real and vice-versa.

These pole-pairs on the rs-plane are also used to construct templates of the speech segment. These templates can be matched using the exact same process used to track the pole-pair movements. There is, however, a problem associated with the rs-plane. The frequency distortion of the rs-plane gives very poor spectral resolution in the perceptually important low frequency region. If the template matching were restricted to using distances measured on this plane then the distance between low frequency pole-pairs on the target and the utterance template may incorrectly be taken to be insignificant.

To overcome this problem the z- to rs-plane mapping can be warped such that the real-frequency line on the resultant rs-plane is proportional to a tonotopical scale such as the Bark scale[6].

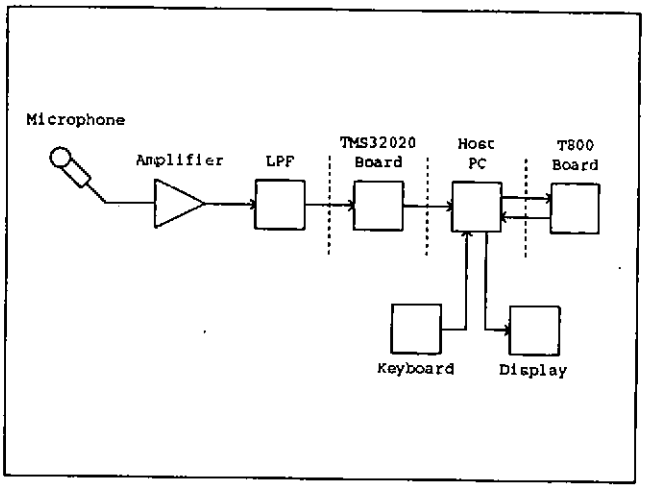

Figure 3: the system hardware.

For the deferred feedback mode, the minimum cumulative inter-template distance between the four pole-pairs of the prototype and current template is used as a distance measure. This distance is visually represented on the dial described in the previous section.

For the immediate feedback mode, the inter-template distance of *each* of the four pole-pairs is represented visually by corresponding to the height of the flags on the flag poles.

4 The System Hardware

The system hardware is based on a portable personal computer with additional computing power provided by a digital signal processor (TMS32020) and a single floating-point transputer (T800).

In addition to the described analysis tasks, the system must amplify the output of a microphone, filter the subsequent speech signal to reduce the effects of aliasing, and digitally sample the speech waveform. The relationship between, the hardware components is illustrated in Figure3.

The host computer for the multi-processor arrangement is a portable PC AT Compatible. It was desired that the system be portable for two reasons. Firstly, Tayside's Speech Therapy Department comprises many departments situated in different hospitals and training centres: the application of the Tutor need not always be in the same building. Secondly, it is standard practise, within the Speech Therapy Department, to lock computer equipment away after use for security reasons. A portable system is therefore far more convenient and practical.

The portable PC has an integral monochrome liquid crystal display (LCD) which can support the screen modes available with IBM's colour graphic adapter. The four colours available on the screen at any one time are represented by varying shades of gray. A sliding contrast control is supplied which caters for varying lighting conditions. On the minus side though, the contrast between on and off pixels varies not only with lighting conditions, but with viewing angle: the therapist will have to be careful that the contrast is high from the patient's view point and may have to put up with poorer contrast his or her self. A back-lit LCD may have overcome this problem.

Secondary storage is in the form of a 5.25" floppy disk, and a 32 Mbyte hard disk. With a hard disk in

the system, it is essential that the hard disk is *parked* prior to transporting the system. In a speech therapy application, it is likely that the system will be moved after each session with the system, and hence a disk parking procedure is included as part of the therapy software.

To accommodate the additional hardware required, the portable PC has six expansion slots, two of which are restricted to half length by the power supply. Two of the full length slots are occupied by a disk controller and the display adapter card. This leaves two half length and two full length slots available for expansion.

The central processing unit of the system is an Intel 80286 processor which can be switched between 6, 8, and 10 MHz. The mother board is populated with 640 k bytes of RAM.

The only equipment external to the PC is a microphone into which the patent will speak. The electrical signal from the microphone (in the order of micro-volts) must be amplified to match the range of the analogue to digital converter. It may also be required by the therapist that samples taken from a recorded session with the patient. The output line voltage from a recorder is 1V pk-pk. The amplification will be switched to cater for both input levels.

To reduce the effects of aliasing, the signal must be passed through a low pass filter (LPF). The speech signal will be digitally sampled at 10 kHz and therefore frequencies above 5 kHz must be reduced as much as possible. This is realised using an eighth order (48 dB/octave fall off) LPF with a cut off frequency of 4 kHz.

The preferred position of the microphone, and the loudness of the speech, will vary from patient to patient. This could cause problems if a patient with a loud voice holds the microphone close. The combined amplification of the amplifier and filter is therefore set to cater for the quietest situation, and a digitally controlled attenuator is added. The Tutor sets the attenuation automatically to match the situation. The amplification, anti-aliasing filter, and the gain control are all fitted on a half length board with BNC connectors for the input and output.

The TMS32020 DSP board for the PC, supplied by Loughborough Sound Images Ltd, integrates Texas Instruments' TMS32020 digital signal processor with fast memory devices and data acquisition hardware. The board is full length, but due to its physical construction, requires an XT (8-bit) slot. The DSP board is mapped into eight locations of the PC's I/O space. The starting I/O address of the block can be set to one of eight locations. The PC has access to the memory external to the processor via two 16-bit ports.

The TMB08 board from Transtech is a motherboard which can accommodate 10 T800 transputer modules (trams). Each tram has a T800, which is an INMOS transputer with a built in hardware floating point arithmetic unit, and can have up to 1 Mbyte of memory.

The board is full length and will operate in either an XT (8-bit) or AT (16-bit) expansion slot. The interface to the PC is compatible with the B004 module specifications from INMOS. It includes a BIOS EPROM which is 32 k-bytes long and is mapped to location 0D000H. Communications between the PC and the TMB08 can either be direct memory access or through eight 8-bit ports mapped onto the PC's I/O address. The base location for these ports is switchable between two locations. For this application requires the motherboard requires to be populated with only one T800 transputer module.

5 Process Distribution

For real time operation, the analysis on the current frame must be complete before the next frame has been sampled. The desired sampling rate is 10 kHz and the frame size is 100 samples, giving a frame duration of 10 ms.

Prior to analysis, the amplified speech signal must be digitally sampled. This task is carried out by the ADC on the TMS32020 board.

A simple distribution approach was used to assign the various analysis tasks: the tasks were implemented one by one on the DSP board until it did not have the processing time left to accommodate another. The remaining were then implemented on the transputer board (which could be expanded if necessary by adding further transputer modules to run in parallel). The PC, which is the slowest processor, was used to facilitate data transfer between the DSP and the transputer.

The DSP had the computational power to take on the task of applying the LP analysis to the sampled speech. This analysis procedure was coded in assembly language for maximum speed. The TMS32020 processor is the second member of the TM320 family of digital signal processors. It is capable of performing multi-function instructions, such as a 16-bit multiply- accumulate with data move, in a single instruction cycle of 200 ns. However, for single cycle operation, it is a requirement that the data to be manipulated is located in the rather limited 544 16-bit memory locations internal to the processor.

Therefore, to minimise the execution time of the LP analysis routines, great care was taken to keep all frequently accessed storage locations within these fast on-chip memory blocks.

The root-finding and pole-pair tracking were assigned to the transputer. The transputer was fast enough for these routines to be coded in C. There was no need to expand the system past the single transputer module.

6 Inter-processor Communications

The communication channel between the DSP board is bi-directional but is essentially under the control of the host PC: the PC can read or write to the DSP's memory but no facilities are supported which allow the DSP access to the host's memory. The PC can access all program and data memory, that does not constitute part of the DSP's 544 words of internal memory, through a pair of 16-bit registers which are mapped into the PC's I/O space.

Once the LP coefficients for a particular frame have been determined, they must be transferred to the transputer board for further processing. Included in this subsequent analysis is the root-finding process which is an iterative in nature, therefore there can be no guarantee that the process will be completed within the duration of the analysis frame.

For this reason it is necessary to insure that any delay in the transputer processing in no way affects the LP analysis. To cater for such a situation the results of the LP analysis are stored in contiguous blocks of memory on the DSP board. This means that the DSP board is not dependent on the state of either the PC or the transputer. The PC on the other hand is wholly dependent on the state of the DSP. If the PC has completed dealing with the previous frame then it must wait until the DSP board has completed analysing the current frame before proceeding. Therefore, it is a requirement that each block of data contains a flag which indicates whether or not the block is valid data: once the PC has dealt with the previous frame, it continually reads this flag until it indicates that the data is valid.

The version of C for the transputer was designed to communicate with the host PC through the INMOS program AFSERVER. The facilities provided by AFSERVER are quite numerous, including file handling, screen and keyboard I/O, and host port access. However, the only facility required for this application was the host port access routines and the complicated protocol within AFSERVER coupled with the fact that byte-access only was supported proved to be a bottle neck for transfer of the LP coefficients from the DSP to the transputer application. To overcome this problem, dedicated communications routines were written in assembly language for the PC which communicated directly with the transputer motherboard's link adapter. The PC was then assigned the task of monitoring the flags on the DSP and, when required, to transfer the desired information to the transputer.

The same communications channel was used to read back the analysis results from the transputer, and the PC could then perform its task of displaying the results graphically.

7 Conclusions

A computer based speech therapy tutor for vowel production has been presented, offering two modes of feedback which has been subjected to some early clinical trials. The therapists involved in field testing the Tutor indicated that the client group requiring therapeutic vowel modification who would respond to a biofeedback method is very small.

To date, only one client trial has been possible. Speech therapists working with cleft palate, hearing

impairment, dysphonia and dysarthia could all identify rare occasions when the tutor would have application, but these were not represented in their ongoing caseload. The speech therapists also considered the specific use of the Tutor and its appeal to their clients. It was felt that the current format would be more suitable for adults, and that colourful and animated displays would have greater appeal to children.

In the trial, a post-pharyngoplasty cleft palate patient, a sixteen year old girl, used the Tutor weekly for a month. A database consisting of her vowels and consonants was created and used to encourage consistent good production, matched to a previously agreed prototype. The therapists were impressed with the ability the Tutor afforded to create new databases, e.g. for client's range of vowels; for any vowel with a range of speakers; for an individual production of repeated vowel; and so on.

To summarise, from the trial it has been seen that the Tutor has several positive features:

1. accurate retention of vowel patterns
2. detailed information is available on the display
3. it has a flexible database structure

The Tutor also has some negative sides that have been identified, and that we will endeavor to eliminate in future work:

1. it addresses the needs of a very small client group
2. the display and graphics require further simplification and increased appeal, especially for children
3. there is a need for auditory feedback to remind users what sound they are aiming for

8 References

[1] Risberg, A. (1968) "Visual aids for speech correction", *American Annals of the Deaf*, **113**, 178–194

[2] Fallside, F. and Brooks, S. (1976) "Real-time areagraph of continuous speech for analysis and speech training" *Electronics Letters*, **12**, 515–516

[3] Bristow, G.S. (1980) "Speech Training with Colour Graphics", *PhD Thesis, University of Cambridge.*

[4] Markel, J.D. and Gray, A.H. Jr. (1976) *Linear Prediction of speech*, Springer-Verlag, Berlin.

[5] Turnbull, J.M. , Sapeluk, A.T. , and Damper, R.I. (1989) "A new method of pole-tracking with application to semivowel recognition", *Proceedings ICASSP '89, IEEE Conference on Acoustics, Speech and Signal Processing* **Vol 1**, 568–571, Glasgow, Scotland.

[6] Turnbull, J.M. , Sapeluk, A.T. , and Damper, R.I. (1990) "Pole Tracking in a Vowel Trainer for Speech Therapy", To be published in *Proceedings of the IASTED International Symposium Signal Processing and Digital Filtering*, 1990, Lugano, Switzerland.