# TOWARDS A COMPUTER BASED SPEECH THERAPY AID

Catherine I. Watson, W K Kennedy, R H T Bates
Department of Electrical and Electronic Engineering,
University of Canterbury,

## Abstract

A computer-based speech therapy is currently being developed at the University of Canterbury Electrical and Electronic Department (UCEEE). At present the aid consists of seven speech analysis modules, which include a vocal tract shape reconstruction from speech signals and a fricative sound indentifier. The real-time hardware for processing and analysing the speech and the software of the aid has all been developed by the UCEEE. The aid has been evaluated by 15 speech therapists, who have all have reacted positively to the aid and its potential.

## INTRODUCTION

In the quest to help their clients improve their speech, speech therapists use many different types of aids. With the advent of computer technology computer based speech therapy aids are now being developed; some aids are already commercially available. The University of Canterbury Electrical and Electronic Engineering Department (UCEEE) has had some 15 years experience in the the field of computer-based instructional aids, firstly with the development of a computer music system (Lamb and Bates, 1978) and later in developing a computer based speech therapy aid (Turner, 1986),(Bates *et al.*, 1987).

The computer based aid, called the CASTT (Computer Aided Speech Therapy Tool) has been developed on an IBM PC. The CASTT currently consists of seven speech analysis modules that all analyse the speech in "real time" and display the information on the computer screen. The seven speech analysis modules are; a fricative identifier, a voice pitch tracker, vocal intensity monitor, a concurrent display of the pitch and loudness variations of the voice, a spectrogram of speech, a display of the consistency and duration of phonated sounds, and a vocal tract shape reconstruction from speech signals.

## THE HARDWARE AND SOFTWARE OF THE CASTT

The CASTT consists of an IBM PC XT, two purpose built speech processing boards, a microphone and speech analysis software.

### Hardware

The special speech processing hardware that resides in the IBMPC XT consists of a custom built analogue to digital (A/D) and digital to analogue (D/A) board plus a Texas Instruments TMS32010 digital signal processing chip with its supporting hardware on a second board. To calculate the speech features with the TMS32010, the appropriate program code is down loaded directly to the TMS32010 program memory by the IBMPC. The IBMPC controls the two special purpose speech boards and can reset the TMS32010 chip if necessary (Bates *et al.*, 1987). Data communication between the PC and the TMS32010 occurs via a mutually accessible 2048 word memory block.

The TMS32010 chip is able to do a 16x16 bit multiply in 200 ns and has some special commands that allow parallel execution of two or three instructions in the same clock cycle. These instructions have been designed specially to handle convolution, windowing and filtering efficiently.

Speech signals are input via a standard microphone to the 12 bit A/D. Speech is sampled at a fixed rate of 10 kHz. The digitized speech is then processed in the TMS32010. The speech features for the current speech modules are all calculated in "real time". It is possible to have audio replay of the stored digitized speech via the 12 bit D/A.

### Software

All seven of the speech analysis modules have two software programs associated with them, one running on the TMS and the other on the PC. The TMS program, written in TMS32010 assembler code, collects the sampled speech from the A/D and calculates the required speech features. The speech features are passed to the IBMPC and the second program governs how these features are displayed on the computer screen. The second program is written either in Microsoft Pascal, Turbo Pascal or Modula 2.

## DESCRIPTION OF THE SPEECH PROCESSING MODULES

In this section we describe the main features of each of the seven speech processing modules.

### Fricative Sound Identifier Module

The estimation of the zero crossing rate of a time domain speech waveform is an accepted method of obtaining a rough but readily calculable approximation its average frequency (Rabiner and Schafer, 1978). Fricative sounds are characterized by noise like waveforms due to the air turbulence employed in their production (Strevens, 1960). Hence the zero crossing rate of the fricatives is greater than for sounds that are not produced by turbulence. The zero crossing rate of sampled speech is calculated in the TMS32010 and passed to the IBMPC. If the rate exceeds a certain preset threshold then the sound is identified as a fricative and there is an appropriate response on the computer screen. The response is a simple longitudinal bar of variable length proportional to the loudness of the fricative.

### Voice Pitch Module

Many different methods for calculating the pitch of a speech signal have been devised (Brieseman, 1984). The technique employed in the Voice Pitch module is a heuristic time domain approach developed by Brieseman (1984) and modified by Turner (1986) . Its principle is based on the estimation of the pitch by visual inspection of the speech waveform.

The pitch estimation algorithm consists of two parts, the peaks of the speech waveform are found and from these the pitch is estimated. In the first part the sampled speech is lowpass filtered digitally at 700 Hz, by a 17 tap FIR filter. The filter, shaped by a Hamming window (Turner, 1986), eliminates all but the major peaks. The main peaks are then detected and stored along with the inter-peak times.

In the second part of the algorithm the first stored peak is compared to the following peaks until another peak with a similar amplitude, within a certain tolerance, is found (call this peak PK). The time between these two peaks is the first estimation of the pitch period. The second and subsequent stored peaks are then systematically checked to see if they have a matching peak at a distance close to the estimate of the pitch. This validation is continued for a length of three times the previous pitch period. If the validation fails anywhere in this range the procedure is immediately begun again, starting with comparing the first stored peak with the next stored peak after PK. If the validation is successful the estimated pitch becomes the new pitch and is passed to the IBMPC.

The pitch data is displayed on the computer screen as illustrated Fig 1(a). The speech therapist's pitch waveform is plotted on the top graph and the client's is on the bottom. For comparison purposes it is possible to superimpose the client's pitch waveform on the speech therapist's either during or after phonation.

### Vocal Intensity Module

The loudness of speech can be obtained from the root mean square (RMS) value of the speech waveform. The RMS value is passed to the IBMPC XT where it is displayed on the computer

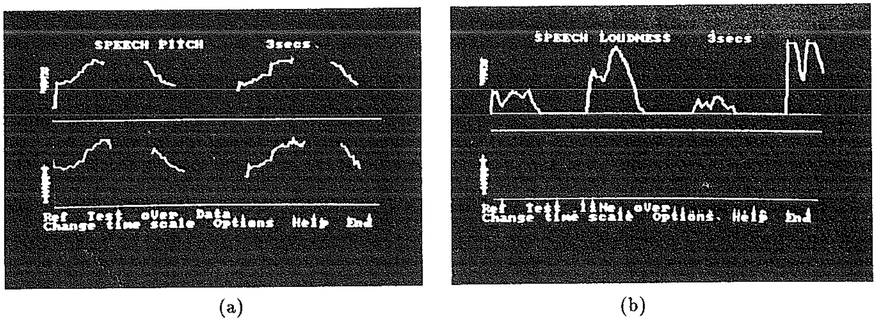

Figure 1: The screen displays of (a) the Voice Pitch module and (b) the Vocal Intensity module for a person saying "hello, hello, hello, hello".

screen (see Fig 1(b)). The screen displays of the voice pitch and vocal intensity modules have a similar layout. For both the modules either three, six or nine seconds of waveforms can be plotted on the graphs; the default is 3 seconds. The pitch and loudness waveforms can be stored on disk for later viewing. Both modules have the facility for limited audio playback.

### Concurrent Pitch and Loudness Module

This module displays the pitch and loudness features of three seconds of speech concurrently. The algorithms that calculate these features are the same as for the voice pitch and vocal intensity modules.

### Spectrogram Module

In the Spectrogram module a new 128 point spectrum is calculated every 20ms. The spectrum is obtained by performing a 256 point fast Fourier Transform (FFT) on the incoming speech samples using the algorithm written by Burrus and Parks (1985). The amplitude of each point in the spectrum is then quantised to 2 bits. The 2 bit number corresponds to the colour the point will be plotted on the screen. ( The graphics card of the CGA in the IBMPC XT can only display up to 4 colours on the screen at any time.)

The pixel addresses for the IBMPC screen are calculated using the TMS32010. If the IBMPC XT assigned the pixel addresses the plotting intensive spectral analysis module would not be able to operate in real time. The spectral display of the incoming speech is plotted in one of two display graphs. The vertical axis is the frequency scale (maximum value 5kHz), and the horizontal axis represents time (maximum 3 seconds).

### Sustained Phonation Module

This program has a game-like format. The clients are required to sustain phonation of a specified sound for a certain length of time. If successful they are rewarded with 1 of 5 pictures chosen randomly by the IBMPC. The speech therapist makes the decision as to whether the client has been successful or not and inputs the information to the computer. The Sustained Phonation Module utilizes the TMS32010 algorithm from the Spectrogram module.

### Vocal Tract Shape Reconstruction

Linear prediction coefficient analysis (LPC) on speech signals (Markel and Gray, Jr., 1976) is used to reconstruct the vocal tract shape approximated by a series of ten interconnecting acoustic tubes. The cross-sectional areas of these tubes are calculated from reflection coefficients, which in turn are obtained from the LPC analysis (Wakita, 1973). The algorithm used in the Vocal

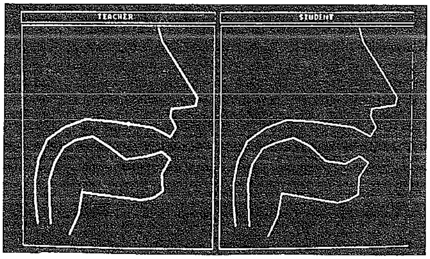

Figure 2: The screen display for the vocal tract reconstruction module showing the reconstucted vocal tract shapes of the sound "ahh" in both the TEACHER and STUDENT windows.

Tract Reconstruction module is based on one given by Markel and Gray (1976).

The IBMPC is passed the ten cross-sectional areas. Using these, the vocal tract shape and the angle the jaw is pivoted are calculated and plotted. The tongue position is not specifically shown but is included in the overall shape of the vocal tract.

The vocal tract shape is displayed within the cross-section of a head, to provide it with a meaningful backdrop, see Fig 2. There are two heads on the screen, one for the speech therapist and the other for the client, allowing for comparisons between vocal tract shapes. For closer comparisons the speech therapist's vocal tract shape can be overlayed on the client's.

## THE EVALUATION OF THE CASTT

The engineering of workable speech analysis modules is only the first step towards an effective speech therapy aid. It is sobering to note that "since the modern electronic era, over 100 different speech training aids have been developed. [All were considered] by their developers to be significant contributions in the area of speech training. However few have been formally evaluated. Very few have had a significant impact on teaching speech to the deaf, and none have come into widespread use." (Braeges and Houde, 1982). To be of any long term benefit a speech therapy aid must be extensively evaluated by the speech therapy community and adapted to meet their requirements.

A series of trials have been conducted by speech therapists on the CASTT, two of these trials have been completed and another is currently in progress. In the first two trials the CASTT was evaluated in various speech therapy clinics for a three week period . Both trials had 6 speech therapists taking part. The therapists filled out a questionnaire on the CASTT compiled by one of the authors,(CIW). The third trial comprises of a series of longterm evaluations, lasting for at least three months. To date, three speech therapists have been involved in this trial.

The first trial established that the speech therapists felt that the aid and the speech analysis modules had much potential. The response of the therapists was overwhelmingly enthusiastic. The first trial also established the areas in which improvements to the speech analysis modules could be made. From the second trial we were able to assess how effective the improved speech analysis modules were and what new changes should be made. The longterm evaluations will give us information on how effective the aid is once the client has got over the novelty value of using a computer.

## THE SPEECH THERAPISTS' RESPONSE TO THE SPEECH ANALYSIS MODULES

The speech therapists did not find all the speech analysis modules equally useful. The voice pitch and vocal intensity modules were universally popular and were usually found to be the most useful modules in the aid. One therapist observed that it took very little training time with the Vocal Intensity module for the children to see when they were speaking too softly or too loudly. The same therapist noted that when using the voice pitch module the children quickly learnt that their vocal behaviour affected the pitch display. They soon discovered how to get some control over the shape of the pitch waveform. For one child, the voice pitch module provided her with the right feedback for her to correct the pitch of her speech, this had previously been a difficult task for her.

The fricative identifier, vocal tract reconstruction, spectral analysis and concurrent pitch and loudness modules had a mixed response from the speech therapists. Several therapists doubted the accuracy of the fricative identifier and vocal tract reconstruction modules. Conversely other therapists found the fricative identifier module very useful. One therapist commented that young children and the intellectually handicapped found the response on the screen when uttering a fricative exciting. Positive response to the vocal tract reconstruction module was less forthcoming. Many therapists found it confusing not having the tongue position separately shown and felt it was not possible to evaluate the module's potential until it was included. However a therapist who worked at a specialist college for the deaf was very enthusiastic about the module's potential and felt it would be a useful speech therapy tool for the disabled.

The crudity of the spectrogram display meant it was not useful as a therapy tool, the spectrogram patterns were not distinct enough for different sounds. The Sustained Phonation module was adapted from the Spectrogram module after a therapist noted, that the "pretty" patterns on the screen from the spectrogram as the sounds were produced, maintained the children's interest. To date the Sustained Phonation module has been evaluated by one speech therapist, in the first long term trial. He felt that the module was very good for teaching breath control and was among the most useful programs in the computer based aid.

The idea of the concurrent pitch and loudness module was suggested in the first evaluation period, by a therapist who had mainly adult clients. The therapists in the second evaluation period, who all had child clients, felt that the program may be useful to adult clients but it was too "advanced" and "complicated" for their clients.

## CURRENT RESEARCH DEVELOPMENTS

Current research on the aid is in three different areas. Firstly there is the on-going assessment of the aid by the speech therapists and consequential improvements.

Secondly a new selective fricative monitor able to distinguish between different fricatives is being developed. The new sorting algorithm will be based on comparing different energy bands of the spectrums of the fricatives (Strevens, 1960), (Hughes and Halle, 1956). Preliminary investigations on the eight English fricatives both spoken in isolation and extracted from words, have given promising results.

The final area of research is on the reconstruction of the vocal tract shape. The vocal tract reconstructions has been scrutinised by a linguistic expert from the Speech Therapy Department at the University of Canterbury, who felt that the reconstructions for the front and neutral vowels were very accurate. However the shapes for the back vowels were imprecise. To improve the accuracy of the speech analysis module two approaches are being followed. The real time algorithm is being modified to investigate the effect of averaging the predictor and area coefficients. The other approach, proposed by one of the authors (RHTB) incorporates Sondhi's (1984) premise that to get a unique solution for the vocal tract shape reconstruction from the lips, as opposed to the glottis, is necessary.

## CONCLUSION

From the series of trials performed on the CASTT it is clear that it has potential to be a useful speech therapy aid. Not all the speech analysis modules were equally as popular, the most "useful" modules depended to some degree on the needs of the clients. More evaluations of the CASTT in clinical environments is necessary and specifically in clinics that deal with deaf young adult clients. The age of the client is a factor in the effectiveness of some of the modules.

Currently the TMS32010 chip has adequate processing power to calculate the speech features in real time. However it may be necessary to update the TMS32010 chip with the much faster TMS32030 chip to accommodate the new computationally intensive fricative sorting algorithm. Also for future expansions to the aid the more powerful TMS32030 chip will allow flexibility in researching and developing more effective speech analysis modules.

ACKNOWLEDGEMENTS

The development of this aid would not have possible with out the support of the following speech therapists: H. Herbert, P. Cosgriff, T. McLeod, P. Macer, S. Claridge, P. Wignall, P. Cosson, E. Terrice, G. Painley, D. Walker, F. Lynn, A. Buckley, C. Musgrave, J. de Serier and Dr M. McLagan. One of us, CIW, acknowledges the financial support of: the Ministry of Research, Science and Technology postgraduate scholarship for Maori, Pacific Islanders and Women; the Federation of New Zealand University Women postgraduate fellowship; EDA Costa Mesa, U.S.A.; and the Royal Society-Canterbury Branch.

## References

Bates, R.H.T., Brieseman, N.P., Clark, T.M., Elder, A.G., Fright, W.R., Garden, K.L., Kennedy, W.K., Squires, P.L., Thorpe, C.W., Turner, S.G. and Jelinek, H.J. (1987), 'Interactive speech-defect diagnostic/therapeutic /prosthetic aid', In Letellier, J.P. (Ed.), *Real Time Signal Processing X*, Proceedings of SPIE - The International Society for Optical Engineering, 20-21 August, 131–139.

Braeges, J.L. and Houde, R.A. (1982), 'Use of speech training aids', In Sims, D., Walter, G. and Whitehead, R.L. (Eds.), *Deafness and Communication: Assessment and Training*, Williams and Wilkins, Baltimore, 222.

Brieseman, N.P. (1984), *A new algorithm for musical pitch estimation*, Master's thesis, University of Canterbury, New Zealand.

Burrus, C.S. and Parks, T.W. (1975), *DFT/FFT and Convolution Algorithms*, John Wiley & Sons Inc, New York.

Hughes, G.W. and Halle, M. (1956), 'Spectral properties of fricative consonants', *Journal of the Acoustical Society of America*, Vol. 28, No. 2, March, 303–310.

Lamb, M.R. and Bates, R.H.T. (1978), 'Computerized aural training: an interactive system designed to help both student and teachers', *Journal of Computer-Based Instruction*, Vol. 5, No. 1 and 2, Aug and Nov, 30–37.

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

Rabiner, L.R. and Schafer, R.W. (1978), *Digital signal processing of speech signals*, Prentice-Hall, 120–130.

Sondhi, M.M. (1984), 'A survey of the vocal tract inverse problem: theory, computations and experiments', In Santosa, F., Pao, Y., Symes, W.S. and Holland, C. (Eds.), *Inverse Problems of Acoustic and Elastic Waves*, Society for Industrial and Applied Mathematics, Philadelphia, 1–19.

Strevens, P. (1960), 'Spectra of fricative noise in human speech', *Language and Speech*, Vol. 3, 32–49.

Turner, S.G. (1986), *Real-time speech analysis for use with impaired speech aids*, Master's thesis, Electrical and Electronic Engineering, University of Canterbury, NZ, March.

Wakita, H. (1973), 'Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms', *IEEE Transactions on Audio and Electroacoustics*, Vol. 21, No. 5, October, 417–427.