A TEST TO ASSESS THE REMEDIAL WORTH OF A COMPUTER-BASED SPEECH THERAPY AID

Catherine I. Watson and John H. Andreae
Department of Electrical and Electronic Engineering,
University of Canterbury,
Christchurch, New Zealand .

Abstract

For a visual speech therapy aid to be useful, its displays must distinguish unacceptable from acceptable speech utterances. Whilst a large number of visual speech therapy aids have been developed, very few have been tested adequately. A Visual Display Test is being developed to assess the visual displays of the CASTT, a computer-based aid developed at the University of Canterbury. The evolution of this test is reported in this paper.

INTRODUCTION

A computer-based real-time speech therapy aid has been developed at the University of Canterbury. The aid, the CASTT, consists of an IBM-PC XT, a special purpose speech board, a microphone and seven speech analysis modules (Watson *et al.*, 1990). An eighth module, the Phoneme Plotter, has recently been developed.

Throughout its development the CASTT has been extensively trialed by fifteen speech therapists in the Christchurch region. The feedback from the therapists strongly suggests that the CASTT is instrumental in the improvement of some of their clients' speech. In the past it has been usual practice at this stage in the development of a computer-based speech therapy aid to perform an evaluation of the aid obtaining quantitative data on the aid's effectiveness in improving clients' speech.

The quantitative assessments of other computer-based speech therapy aids in the past have invariably established that clients who use a computer based aid in speech therapy correct their speech impediments more quickly than those clients who use conventional methods (for example Stark (1972), Bate *et al*(1982) and Arends *et al*(1991)). However never were any reasons given as to why this should be the case. The noted improvements could have been due to the therapists and the clients correctly interpreting the visual displays, or merely due to the clients being motivated by the special treatment they were getting in using a novelty aid. In addition it was never established from these assessments, whether therapists use the visual information on the screen to assess the client's speech or whether they use the information to merely reinforce what they have already heard.

To date there has been very little research on whether the users actually understand the displays of visual computer-based speech therapy aids or, how to test that the information obtained from the displays will be useful in speech therapy. Before quantitative data is collected to establish whether the use of the CASTT in speech therapy enables the speech impaired to correct their speech errors, it is important to establish that correct and incorrect pronunciation can be identified from the visual displays of the CASTT. The purpose of this paper is to propose a visual display test that could be used to assess the potential of a computer-based speech therapy aid.

ASSESSMENT OF VISUAL DISPLAYS OF COMPUTER-BASED SPEECH THERAPY AIDS

A few researchers, such as Bargstadt *et al* (1978) and Maki *et al* (1981), have made some attempt at assessing the displays of specific visual speech aids. Bargstadt *et al* (1978) investigated the success rate of five trained subjects in identifying the eight English fricatives from visual patterns

of the Video Articulator (the patterns were similar to Lissajous figures). Maki *et al* (1981) were interested in the visual displays of the Speech Spectrographic Display. They investigated whether hearing-impaired subjects could identify which of two utterances was the better pronounced from the visual displays and whether the subjects could correctly relate the visual patterns to selected articulatory events (Maki *et al.*, 1981).

Whilst Bargstadt *et al's* assessment checked the consistency of the visual patterns for specific sounds it did not establish that the aid was able to detect errors in speech. The assessment of Maki *et al* did establish this and what is more it sought to establish that the subjects could relate the visual patterns to speech characteristics. An evaluation of this sort could be carried out on the visual displays of the CASTT. One complication would be the number of speech errors that the CASTT would have to be tested for.

The errors in speech fall into three categories, articulation, phonation and fluency(Skinner and Shelton, 1978). Research cited by Braeges and Houde (1982) has demonstrated that there are around 600 types of common speech errors. The CASTT should be able to detect all of these errors because it has modules that visually display prosodic and paralinguistic features of speech (the Voice Pitch Tracker, the Vocal Intensity Module and the Concurrent Pitch and Loudness Module and the Sustained Phonation Module) and modules that visually display the phonemic features of speech (the Spectrogram Module, Fricative Monitor, Vocal Tract Reconstruction and Phoneme Plotter Modules). To test the visual displays of the CASTT for the common 600 speech errors would take a considerable amount of time

The Short Test Of Elementary Error Discrimination

There is one possible visual display test proposed in the literature that could be carried out on the CASTT. Braeges and Houde (1982) basically extended Maki *et al*'s ideas and have developed a test called the Short Test of Elementary Error Discrimination (STEED). The STEED utilises a list of 29 elementary speech errors, which are representative of the 600 common speech errors. Each of these 29 errors is highlighted by a target utterance and an error utterance. The pairs of target and error utterances differ in only one aspect of speech (e.g. articulation intensity, speech quality, etc).

The Elementary Error list can be subdivided into six sets. Each set contains a certain type of speech error. The sets are : the Articulatory Intensity Set which represents errors resulting in intensity change due to misarticulation (as opposed to suprasegmental aspects); the Voiced/Unvoiced Set which represents errors in the voice/unvoiced distinction; the Nasality Set which represents errors in nasality; the Articulation Substitution Set which represents errors in phoneme substitution; the Suprasegmental Set represents errors in suprasegmental aspects of speech; and finally the Speech Quality Set which represents errors in speech quality. Table 1 contains some examples of the twenty nine speech errors and the speech pairs (made up of a target and error utterance) that exemplify the errors.

Braeges and Houde believed that, for each elementary error, if a user could differentiate between the target and error utterance pairs which exemplified it, using a visual speech aid display, then the aid could effectively be used in therapy for all the common speech errors (Braeges and Houde, 1982).

However, Braeges and Houde's STEED neglected to test for the consistency and repeatability of an aid's visual pattern when two utterances are the same. Certainly it is imperative that the difference between two utterances which differ in one aspect of speech, such as a phoneme or pitch, (henceforth referred to as a "different-speech pair") should be immediately noticeable from the displays of a visual speech aid. It is also desirable that two utterances with the same phonetic transcription and prosodic features (henceforth referred to as a "same-speech pair") be seen as more similar than any different-speech pair. Of course, this criterion of "sameness" cannot be satisfied by all same-speech pairs, because there will always be some different-speech

Excerpts From The Elementary Error Speech List	
Articulation Intensity Set	
Detects the release of a complete articulatory closure	boo/boot
Distinguishes sounds whose speech envelopes differ in initial or final rate of change	ban/ran
Voiced/Unvoiced Set	
Distinguishes sustained voice fricative sounds from unvoiced fricative sounds	Sue/zoo
Nasal Quality Set	
Detects errors in the timing of the velar closure or release in nasal sounds.	me/mbee
Articulation Substitution Set	
Distinguishes s/$\int$ substitutions	see/she
Distinguishes far-neighbour back vowel substitutions	do/door
Suprasegmental Set	
Distinguishes differences in terminal pitch contour	now ?/now ! (emphasized)
Speech Quality Set- use any short phrase, vary trials by test measure only	
Distinguishes error of excessive loudness	Normal loudness/ Too loud

Table 1: Examples of some of the speech pairs in the speech list proposed by Braeges and Houde which are representative of the common speech errors

pairs straddling the boundary between acceptable and unacceptable utterances which are more similar than the same-speech pairs spanning the region of acceptable utterances.

THE VISUAL DISPLAY TEST

A new test, the Visual Display Test (VDT) is being developed to investigate the displays of the CASTT. It has been adapted from the STEED. In a preliminary test, the visual displays resulting from both same- and different-speech pairs were shown to participants. From the displays the participants were required to identify the speech pair type. The same- and different-speech pairs were obtained from the elementary error lists.

The Modification Of The Speech List Which Represents The Elementary Errors

A few modifications had to be made to the target and error utterance pairs in the Elementary Error list, for it to be used with the CASTT. The original target and error utterance pair examples had been designed for American English. Only one of the utterance pairs in the list had to be changed for it to be applicable for New Zealand English. This speech list will henceforth be referred to as NZ-SL1.

The Elementary Error list, and hence NZ-SL1, was designed to assess the visual aids which displayed speech characteristics as they varied with time. Hence it could be used to test the Voice Pitch Tracker, the Vocal Intensity module, the Concurrent Pitch and Loudness module, the Spectrogram module and the Sustained Phonation module of the CASTT. The other three modules - the Fricative Monitor, the Vocal Tract Shape Reconstruction Module and the Phoneme Plotter - display speech characteristics for an instant of speech only. Thus a second elementary error list was compiled to test these latter modules, henceforth called NZ-SL2. This list only contained twelve elementary errors. All the speech errors that involved aspects of time were removed. For example from table 1 "me/mbee" highlights errors in timing of velar closure or release in nasals(Braeges and Houde, 1982). Similarly the elementary errors of the suprasegmental aspects of speech and in speech quality were removed, since the three modules were only intended to detect phonemic features of speech.

Since the second group of modules only displayed brief time-invariant speech characteristics, the speech pair examples which represented the errors had to be single phonemes. These were derived from the words and sounds in NZ-SL1, taking into consideration the elementary error highlighted by the original speech pair. For example, from table 1 "Sue/zoo" exemplified the speech error " substituting unvoiced fricatives for sustained voiced fricatives", so the target/error utterance pair in NZ-SL2 became "[s]/[z]". NZ-SL2 can be thought of as a subset of NZ-SL1.

Obtaining The Pre-Recorded Speech For The VDT

For each target/error combination (denoted by X/Y say) from the elementary error lists four utterances were pre-recorded, two utterances of X and two of Y. From these utterances two same-speech pairs (X_1X_2 , Y_1Y_2) and two different-speech pairs (X_1Y_1,X_2Y_2) can be made (a further two different-speech pairs are possible but were not used in the preliminary VDT). From each of these utterances displays are obtained for each visual display type.

Pre-recorded speech was used to ensure that the variations in the visual patterns were due to the utterances themselves rather than because the speech was obtained from many different speakers or one speaker on many different occasions. Also it ensured the varying background conditions were limited, and it enabled all the speech used in the test to be checked for correct pronunciation. Finally, because the VDT was a visual test, pre-recording ensured that the participants could not overhear the utterances as they were being spoken.

The words and sounds of the two speech lists were recorded by a male speaker. The speech was amplified and passed through a 4.5kHz 8 pole elliptic filter into a 16 bit A/D. The filter and A/D were part of an Antex Electronic SX-8 Digital Audio Processor (the SX-8 board resided in an IBM-PC AT). The digitized speech was stored on the hard disk of the IBM-PC after the silences had been edited out. The speech was sampled at 10 kHz. All recording was done in a quiet room.

Preparing The Visual Displays For The Preliminary VDT

Though there are eight speech analysis modules in the CASTT, there are only six distinct visual display types. These are pitch contours, loudness contours, spectral content, frication content, vocal tract shape estimation and Lissajous figures.

All but one of the modules of the CASTT have two display plots. The exception is the Fricative Monitor. In the normal operation of the CASTT one display plot is for the results of the client's speech and the other for a reference display provided by the therapist. The display plots for each visual display type in the preliminary VDT were kept in the same format as for the equivalent speech analysis module of the CASTT, e.g. keeping the scale and posistioning of the display plots the same. In the case of the CASTT's Fricative Monitor which has only one display plot a second display plot had to be added for the preliminary VDT.

All the displays in the VDT were pre-calculated from pre-recorded speech.

The Participants Of The Preliminary VDT

In this preliminary investigation there were nine participants, four women and five men. All were staff or postgraduate students of the Electrical and Electronic Engineering Department at the University of Canterbury except one, who was a school leaver working in the aforementioned department over the summer vacation.

THE RESULTS OF THE PRELIMINARY VDT

The intention of the preliminary VDT was to establish that, for each elementary error exemplified by a target and error utterance pair from either NZ-SL1 or NZ-SL2, there is at least one visual display type in the CASTT from which same- and different-speech pairs can be distinguished. Since the displays allowed only one pair to be displayed at a time, the test relied on the participants to provide their own threshold of difference between "same" and "different". The participants were *not* told whether their responses were correct or incorrect. However they *were* told which target/error combination the pair was obtained from. They had to say whether the displays showed a same-speech pair or a different-speech pair.

The results indicate that even under these conditions the participants did considerably better than they would have by pure guessing. In table 2, the column labelled "8up level" gives the numbers

Successful Identification	NZ-SL1		NZ-SL2	
	8up level	7up level	8up level	7up level
all four speech pairs	*3 out of 29*	*9 out of 29*	*3 out of 12*	*4 out of 12*
both different-speech pairs	*18 out of 29*	*25 out of 29*	*9 out of 12*	*9 out of 12*
both same-speech pairs	*5 out of 29*	*13 out of 29*	*6 out of 12*	*6 out of 12*

Table 2: The results of the preliminary VDT for the CASTT

of correct responses by eight or more of the nine participants. Similarly, the "7up level" gives the numbers of correct responses by seven or more participants. The first row of the table gives the correct responses for identifying all four speech pairs resulting from a target/error combination, the second row just for different-speech pairs and the third just the same-speech pairs. Thus, the top left entry "*3 out of 29*" means that for three out of the twenty nine speech errors, each of the four speech pairs was identified correctly by eight or more of the participants. If they had been using random guessing, even one out of twenty nine would have had a probability of occurrence of less than 0.00001 at the 8up level.

It is not surprising that the participants in this preliminary test were biased in their responses, since they themselves had to provide the threshold between "same" and "different". For the displays resulting from the NZ-SL1 list, 62% of the responses were "different-speech pair" while for the displays resulting from the NZ-SL2 list the percentage was 56.3%. The bias also shows in table 2 where there is considerably greater success in correctly identifying the different-speech pairs. If the responses of correctly identifying different-speech pairs are analysed then we find that 58 of the possible 82 (=2×(29+12)) were unanimously (9 out of 9) selected correctly with at least one visual display type, and only one of the remaining 24 was identified correctly by less than half of the participants on all of the modules.

Discussion Of The Preliminary VDT Results

The preliminary VDT was carried out with a minimal disturbance to the CASTT so the displays would be seen by the participants in the same format as those seen by therapists and clients. Only in the case of the fricative Monitor module did an additional display have to be provided because just one display is normally used with this module. Taken at face value, the results shown in table 2 portray the CASTT as a poor aid which could not be used to distinguish more than a few elementary speech errors. This is at odds with the reports of the speech therapists who have used the aid. The reason for this apparent anomaly becomes clear as soon as we consider the nature of the preliminary VDT.

When a participant in the test judges a speech pair to be "different", there is no doubt that the CASTT displays are showing a difference between the two utterances. For a "same" response we do not know whether the response is due to no difference being observed or because the participant has deemed the difference small enough to count as "same". From this it can be seen that the "different" responses to a different-speech pair are testing the CASTT modules for their ability to distinguish speech errors but the "same" responses are testing the participants' ability to set a threshold on difference for sameness. The preliminary VDT conflates the CASTT errors with participant errors.

The success of the CASTT in displaying differences between the displays for utterances in different-speech pairs is indicated by the numbers given above for the majority judgements of "different" for different speech pairs. Only one of the two different-speech pairs from the elementary errors failed to be judged as different by the majority of the participants using at least one of the modules.

With this support for the CASTT, we must put the blame for the poor results given in table 2 on the preliminary VDT itself. Fortunately, it easy to see how the VDT should be redesigned. Its implementation will involve some temporary extensions to the CASTT.

AN EFFECTIVE VISUAL DISPLAY TEST

The new VDT will test the CASTT both for its ability to separate different-speech pairs and for its ability to display same-speech pairs as less different than different-speech pairs. Instead of asking the test participants whether a pair is "same" or "different", the new test will present displays of three utterances. Two will be a same-speech pair and the third an error utterance. The utterences will once again be drawn from the target/error combinations of NZ-SL1 and NZ-SL2. The participant will only need to judge the relative differences between the pairs of displays. The participant will not need to use any threshold of similarities as was required by the preliminary VDT.

From the results of the new VDT it will be possible to rank the differences obtained for same-speech and different-speech pairs. If the CASTT display modules can be shown to cluster the acceptable utterances and separate the cluster from the error utterances, then its efficiency will have been demonstrated. The new VDT will also allow a detailed study of the features which give rise to the differences in the displays of different utterances.

It is planned to extend the CASTT to provide three displays from each module and to try the new VDT in the near future. The improved test will open up possibilities for examining in detail the effectiveness of the displays and the nature of speech errors. The results can be expected to lead to a better understanding of the aid and to its improvement.

ACKNOWLEDGEMENTS

The authors are grateful for the assistance provided by Dr. M. Mclagan of the University of Canterbury and by Prof. I. Watson of Massey University and to the late Prof. R.H.T Bates who originally conceived the CASTT and who supervised and encouraged CIW in her first three years in the CASTT project.

REFERENCES

AREND S, N., POVEL, D., VAN OS, E., MICHIELSEN, S., CLAASSEN, J. and FEITER, I. (1991), 'An evalution of the visual speech apparatus', *Speech Communication*, Vol. 10, No. 4, Nov., pp. 405–414.

BATE, E.M., FALLSIDE, F., GULIAN, E., HINDS, P. and KEILLER, C. (1982), 'A speech training aid for the deaf with display of voicing, frication, and silence', In *Int. Conf. on acoustics, speech and signal processing*, IEEE, pp. 743–746.

BRAEGES, J.L. and HOUDE, R.A. (1982), 'Use of speech training aids.', In SIMS, D., WALTER, G.G. and WHITEHEAD, R.L. (Eds.), *Deafness and Communication: Assessment and Training*, Williams and Wilkins, Baltimore,MD.

MAKI, J.E., GUSTAFSON, M.S., CONKLIN, J.M. and HUMPHREY-WHITEHEAD, B.K. (1981), 'The speech spectrographic display: Interpretation of visual patterns by hearing-impaired adults', *J. of Speech and Hearing Disorders*, Vol. 46, Nov, pp. 379–387.

SKINNER, P.H. and SHELTON, R.L. (Eds.) (1978), *Speech, Language, and Hearing: Normal Processes and Disorders*, Addison-Wesley Publishing Company.

STARK, R.E. (1972), 'Teaching features of speech to deaf children by means of real-time visual displays', In *Proc. Int. Symp. Speech Comm. and Profound Deafness*, A. G. Bell Assoc., Washington D.C.

WATSON, C.I., KENNEDY, W.K. and BATES, R.H.T. (1990), 'Towards a computer based speech therapy aid', In *Proc. 3rd Int. Australian Conf. on Speech, Science and Technology*, Nov., pp. 234–239.