University of Otago
Te Whare Wananga O Otago
Dunedin, New Zealand

The Visual Display Test: A Test to Assess the Usefulness of a Visual Speech Aid

Catherine I. Watson

The Information Science Discussion Paper Series

Number 94 / 19
November 1994
ISSN 1172-6024

The Visual Display Test: A Test to Assess the Usefulness of a Visual Speech Aid

Dr Catherine I. Watson[1]
Computer and Information Science
University of Otago

November 1994

Abstract

The facility to be able to display features of speech in a visual speech aid does not by itself guarantee that the aid will be effective in speech therapy. An effective visual speech aid must provide a visual representation of an utterance from which a judgement on the ``goodness'' of the utterance can be made. Two things are required for an aid to be effective. Firstly, the clusters of acceptable utterances must be separate from the unacceptable utterances in display space. Secondly, the acoustic features which distinguish acceptable utterances from unacceptable utterances must be evident in the displays of the speech aid. A two part test, called the Visual Display Test (VDT), has been developed to assess a visual speech aid's capacity to fulfil these requirements.

[1] Address correspondence to: Dr C.I. Watson, Research Assistant, Department of Information Science, University of Otago, P.O. Box 56, Dunedin, New Zealand. Fax: +64 3 479 8311 Email: cwatson@commerce.otago.ac.nz

1 Introduction

The ability to speak to be understood is something most of us take for granted. Yet for those people with a speech disability, it can be a very frustrating task. Fortunately through speech therapy it is possible to, if not cure, at least alleviate the speech impairment. The use of a visual feedback system, that can provide information on different aspects of speech, can be beneficial to the speech impaired, in speech therapy.

The three factors which provide the easiest method of differentiating speech sounds are the perceptual features of loudness, pitch, and quality (Ladefoged, 1972). In visual speech aids, it is the acoustic measurements of these features which determine the displays. Over the last 120 years many visual speech aids have been developed. Up to thirty years ago the information provided about speech features was in some form of **current-value-plot**. These are displays which display characteristics of speech for an instant only, for example an s-meter. The 1960's brought the advent of long persistence screens, these enabled the display of **time-plots**. Time-plots are displays which show a record of speech characteristics as they vary over time, for example loudness contours.

The coming of computer and microprocessor technology had a big impact on visual speech aids. Speech features could be calculated via software algorithms rather than via hardware circuitry. Prior to this the speech aids were purpose built, each speech feature was calculated using specialized hardware circuitry. Using digital signal processing techniques, one computer-based speech aid can display many different features of speech. Over the last 20 years many multi-featured speech aids have been developed (see for example the reviews by Braeges and Houde (1982), Bernstein *et al* (1988) and Watson (1994)) One such aid was the Computer Aided Speech Therapy Tool (CASTT), developed at the University of Canterbury.

The CASTT is a real-time speech aid, based on an IBM-PC and a special purpose speech board. Its sole acoustic sensor is a microphone. The CASTT has eight speech analysis modules. These are the Loudness Intensity Monitor, the Voice Pitch Tracker, the Concurrent Pitch and Loudness Module, the Spectrogram, the Sustained Phonation Monitor, the Fricative Monitor, the Vocal Tract Shape module, and the Lissajous figure module (Watson, 1990; Watson, 1992).

The ability to visually display features of speech does not guarantee a speech aid's effectiveness. A visual speech aid must display features of speech in real-time. It must be easy for the therapist and client to use. It must impart information which is of use in speech therapy, and finally it must fit the requirements of speech therapy. These four points can be established, for the most part, through developing the aid interactively with speech therapists.

Throughout the development of the CASTT it has been extensively evaluated by speech therapists, fifteen in total. The evaluations of the CASTT were qualitative. They led to a number of significant improvements in the modules of the aid. In addition, the feedback from the therapists strongly suggested the CASTT was instrumental in the improvements of some of their clients' speech.

There was, however, some contradictory evidence given by one therapist on the usefulness of some of the CASTT's modules. She was using the Fricative Monitor and Vocal Tract Shape module for therapy of sounds for which the modules could not possibly provide any useful information, but she was observing improvements in her client's speech. It was tempting to dismiss all the evidence provided by this therapist, but on careful consideration it became apparent that merely assessing the improvements of client's speech was not enough to establish the worth of the CASTT as a speech aid. Before any definitive comment can be made on the worth of the CASTT as a speech aid, or indeed any speech aid, it is necessary to establish exactly what information one can expect to obtain from the displays. An effective visual speech aid must provide a visual response from which a judgement on the "goodness" of an utterance can be made. At the time there was no adequate test available to test any visual speech aid's capacity to do this, as a consequence, the Visual Display Test (VDT) was developed. The rest of this paper will describe the VDT, and present how well the CASTT performed in the test.

2 The VDT

The VDT has been designed to assess a speech aid's ability to separate acceptable speech from unacceptable speech, and its ability to display crucial speech errors. Two separate tests are required to do this, the VDT part I and VDT part II. VDT part I establishes whether the clusters of acceptable utterances are separate from the clusters of unacceptable utterances in display space. The VDT part II establishes whether the acoustic features which distinguish acceptable utterances from unacceptable utterances, are evident in the displays of the visual speech aid.

To test any visual speech aid, it is necessary to know what speech errors there are. Research by Braeges and Houde (1982) cited 600 common speech errors for the hearing impaired. No known set of common speech errors for people with speech disorders, but who are not hearing impaired, has been compiled. It will be assumed that this set can be represented by the common speech error set for the hearing impaired. To test a visual speech aid for all these errors would take a considerable amount of time. Fortunately Braeges and Houde (1982) compiled a shorter list of 29 elementary errors, which they claim are representative of the 600 common errors. Each of the errors is highlighted by a target/error pair which differ in only one aspect of speech.

Error No.	Description of Elementary Error	Target/Error Combination
1	Detects the release of a complete articulatory closure	boo/boot
6	Distinguishes sustained voiced fricatives from sustained unvoiced fricatives	sue/zoo
21	Distinguishes difference in terminal pitch contours	now?/now!

Table 1: Examples of three of the 29 elementary errors, and the target/error combinations from NZ-SL1 which exemplify these errors (from Braeges and Houde (1982))

The speech list of target and error utterances was actually developed as part of a test, proposed by Braeges and Houde, to asses the potential of a visual speech aid. The test was flawed. It only tested a speech aid's capacity to display difference, when two utterances are judged to be different. It neglected to test for consistency and repeatability of an aid's visual pattern when two utterances are judged to be the same (Watson, 1992). Whilst the test is not useful, the speech list is.

Two modifications were required to the speech list before it could be used in the VDT. First it had to be made applicable to New Zealand English; only one change was necessary (Watson, 1992). The speech list, called NZ-SL1, comprised target utterances and error utterances which were either words or phrases. Table 1 gives several examples of target/error pairs in the list and the speech errors they exemplify, from NZ-SL1.

Braeges and Houde's original speech list, and therefore NZ-SL1, was designed to assess visual speech aids which only have time-plot modules. Five of the CASTT's modules have time-plot displays and, therefore, can be tested with NZ-SL1. The other three modules of the CASTT have current-value-plot displays. Since these modules have no reference to time they cannot be tested for any speech errors that involved aspects of time, for example elementary error 21 (see Table 1). Consequently a second speech list was compiled, called NZ-SL2, in which all the elementary errors which involved a component of time were removed. In addition the target and error utterances which exemplified the elementary errors were single phones, rather than words or phrases as in NZ-SL1. For example in NZ-SL2 the target/error pair that distinguished elementary error 6 (see Table 1) became "[s]/[z]".

The VDT part I

In the VDT part I participants are presented with the plots of three utterances. Two of the plots are from two different spoken versions of a particular sound/word/phrase. The third plot is from an utterance of a different sound/word/phrase. The sets of three plots are called **plot-sets**. The

plots are all obtained from pre-recorded speech.

The participants are required to select which two plots in the plot-sets are the most similar. The participants have no knowledge of what sounds/words/phrases the plots are of. The utterances used in the VDT were drawn from the speech lists. Four plot-sets arise from each of the target/error combinations in the speech list. Let X and Y denote the target and error utterances respectively, and let the subscripts 1 and 2 denote the version of the spoken utterance. Using this notation, the four plot-sets that arise from each elementary error are $X_1X_2Y_2$, $X_1X_2Y_2$, $X_1Y_1Y_2$, and $X_2Y_1Y_2$. The order in which the plots of utterances appear within the plot-sets is random.

Ideally the displays that look the most similar will be the displays of the different spoken versions of a particular sound/word/phrase (i.e X_1 and X_2 or Y_1 and Y_2). For each elementary error, we seek to establish that there is at least one display type in a visual speech aid in which the plots of either X_1 and X_2 or Y_1 and Y_2 can be identified from each of the four plot-sets.

The VDT part II

In the VDT part II evidence of the intended acoustic difference between the target and error utterances is looked for in the displays of the speech aid. For each elementary error, the plots of the two different versions of the target utterances and the error utterances (X_1,X_2,Y_1, and Y_2) are scrutinized. These plots are the same as those from which the plot-sets in VDT part I were formed. The displays are judged according to four criteria:

1. it is required that the plots of the target utterances (X_1 and X_2) be more similar than the plots of target/error pairs (X_1Y_1, X_1Y_2, X_2Y_1, or X_2Y_2);
2. it is required that the plots of the error utterances (Y_1 and Y_2) be more similar than the plots of target/error pairs (X_1Y_1, X_1Y_2, X_2Y_1, or X_2Y_2);
3. the target and error utterances differ in one aspect of speech, the presence or absence of this feature must be obvious in the visual displays;
4. it is required that the visual difference between the target and error plots be related to the intended acoustic difference between the target and error utterances.

If all four conditions are met, then it was said that the display type has remedial potential for the elementary error the target and error utterances exemplify.

The need for the two parts of the VDT

The VDT part I gives the elementary errors in which the difference between the displays of the target and error utterances is obvious. No expert knowledge is needed to distinguish between the target and error displays. The VDT part II shows the elementary errors for which a visual aid has remedial potential if one knows what features should be evident in the target and error displays. The differences between the VDT part I and VDT part II results gives an indication of

the displays of the visual speech aid which could be improved. For example with the CASTT, it was found through comparing the results of the VDT part I and VDT part II that the heights of the Loudness contours had to be made a more obvious feature in the displays.

3 Assessing the CASTT with the VDT

Whilst there are eight modules in the CASTT, there are only six distinct display types: the loudness contours, the pitch contours, the spectral content display, the fricative response, the vocal tract shape reconstruction, and Lissajous figures. The first three in the list being time-plots, the second three being current-value-plots. The VDT was used to assess these six display types.

Test	Speaker	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
VDT I	FEMALE	●			●	●		●			●	●	●	●		●	●		●			●		●			●		●	
VDT I	MALE	●			●						●	●		●		●						●								
VDT II	FEMALE	●		●	●	●	●	●			●	●	●	●	●	●	●	●	●		●	●	●	●	●	●	●	●	●	
VDT II	MALE	●	●	●	●		●	●			●	●		●		●	●	●		●		●	●	●	●	●	●	●	●	●

Table 2: The elementary errors exemplified by female utterances and male by male utterance, for which at least one of the time-plots of the CASTT had remedial potential in VDT part I and part II.

When the CASTT was assessed by the VDT, 31 participants were involved in VDT part I and one was involved in VDT part II. The plots were obtained from the utterances of one man and two women. For each elementary error, two spoken versions of the target and error utterances were recorded by the male speaker and one of the two female speakers.

Table 2 gives the elementary errors exemplified by female utterances and by male utterances for which at least one of the time-plots of the CASTT had remedial potential, in both VDT part I and VDT part II. Through judging the displays, purely on the basis of " which displays are the most similar" it is possible to use the displays of the CASTT to distinguish between the target and error utterances which exemplify a sizeable number of elementary errors. There is another, and much larger set of elementary errors for which the CASTT has remedial potential if one has knowledge of what features to look for in the displays. The VDT indicates that the time-plots of the CASTT have the potential for the remediation of many speech errors. This finding concurs with the evidence provided by the speech therapists.

Another finding from the VDT about the time-plots was the areas in which the time-plot displays could be improved. The VDT revealed that attention had to be drawn to the height and

width of the loudness contours and to the height of the pitch contours. To many of the 31 participants in VDT part I it was not obvious these were important features of which to take note. An anomaly in the pitch tracking algorithm often caused a glitch in the beginning of the pitch contours. This glitch distracted many of the 31 participants in VDT part I, resulting in many of the plots of the X_1 and X_2 utterances, or Y_1 and Y_2 utterances, not being judged as being most similar. Thus improvement to the pitch tracking algorithm is required. Finally the VDT revealed the sensitivity of the spectral content plots to the loudness of an utterance, was a distracting feature. Some way of reducing this sensitivity must be found.

In contrast to the time-plots, the VDT revealed that the current-value-plots of the CASTT yielded very little useful information about speech and were of little use as speech aids. As a consequence new analysis algorithms are being investigated for both the Fricative Monitor and the Vocal Tract Shape module. The Lissajous figure module also performed very badly in the VDT, in the VDT part II it had no remedial potential for any of the elementary errors. However it will be retained in the CASTT as an aid for phonation. The 29 elementary errors did not represent anything as fundamental as the ability to phonate.

The VDT and the assessment of the CASTT has been described in greater detail in Watson (1994).

4 Conclusions

The Visual Display Test has been presented. It is a two part test which assesses the ability of a speech aid to provide a visual response from which a judgement on the "goodness" of an utterance can be made. The VDT can be used by developers of visual speech aids to ensure that useful information about speech can be obtained from the visual displays. They can also use the VDT to find out what aspects of the speech displays could be improved. A speech therapist, or any potential visual speech user, could perform the VDT part II on a visual speech aid, providing it was possible to get two plots of the target utterances and two plots of the error utterances on the screen at one time. In doing this they would be able to assess the remedial worth of a visual speech aid and to assess whether the aid would be of any use to them.

Acknowledgements

The author would like to thank Dr J. Andreae for the many helpful discussions she has had with him in order to formulate this work.

References

BERNSTEIN, L.E., GOLDSTEIN, M. H. and MASHIE, J. J. (1988)," Speech training aids for hearing-impaired individuals: I. overview and aims", *Journal of Rehabilitation and Development*, Vol 25, No. 4, p53-62.

BRAEGES, J. L. and Houde, R. A. (1989), " 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, p222-244

LADEFOGED, P. (1972), *Elements of Acoustic Phonetics,* The University of Chicago Press.

WATSON, C.I., KENNEDY, W.K., and BATES, R.H.T. (1990), "Towards A Computer-Based Speech Aid", *In Proc. 3rd Australian International Conference on Speech Science and Technology",* Melbourne, November, p234-239.

WATSON, C. I. and ANDREAE, J. H. (1992), `A test to assess the remedial worth of a computer-based speech therapy aid', *In the Proceedings of the 4th Australian International Conference on Speech Science and Technology*, Brisbane, p279-284.

WATSON, C. I. (1994), *Investigation and Implementation of Computer-Based Aids for the Speech Impaired,* PhD thesis, University of Canterbury.

University of Otago

Department of Information Science

The Department of Information Science is one of six departments that make up the Division of Commerce at the University of Otago. The department offers courses of study leading to a major in Information Science within the BCom, BA and BSc degrees. In addition to undergraduate teaching, the department is also strongly involved in postgraduate programmes leading to the MBA, MCom and PhD degrees. Research projects in software engineering and software development, information engineering and database, artificial intelligence/expert systems, geographic information systems, advanced information systems management and data communications are particularly well supported at present.

Correspondence

This paper represents work to date and may not necessarily form the basis for the authors' final conclusions relating to this topic. It is likely, however, that the paper will appear in some form in a journal or in conference proceedings in the near future. The authors would be pleased to receive correspondence in connection with any of the issues raised in this paper. Please write to the authors at the address provided at the foot of the first page.

Any other correspondence concerning the Series should be sent to:

DPS Co-ordinator
Department of Information Science
University of Otago
P O Box 56
Dunedin
NEW ZEALAND
Fax: +64 3 479 8311
email: workpapers@commerce.otago.ac.nz