Proceedings of The Institute of Acoustics

VISUAL FEEDBACK FOR THE DEAF

N.D. Black (1) and E.L. Gailey (2)

(1) Dept. of Elect. Eng. University of Ulster
(2) Dept. of Comm. University of Ulster

INTRODUCTION

A large number of people suffer from a significant degree of hearing loss. Exact numbers are difficult to determine since there is no statutory register for hearing impairment, but it has been estimated that about 45,000 are affected in Northern Ireland alone. Of these, around 40,000 have what is termed postlingual deafness, meaning that they acquired their hearing loss after learning to speak. The most obvious affect on communication is that the individual can no longer hear what is said, but severe or profound deafness can also adversely effect the person's previously normal speech. Such speech deterioration, along with evidence of the serious effects this has on the attitudes of other people towards the deafened individual, has been documented by Cowie and Douglas-Cowie [1]. One of the areas in which deterioration is noted is vowel quality; although in written English the omission of vowels leaves the discourse fairly intelligible, recent research by Massen and Povel [2] has shown that the correction of vowels in the spoken language of profoundly hearing impaired subjects dramatically improves their intelligibility.

It is widely (though not universally) held to be the case that auditory feedback is necessary to maintain normal speech production (Borden [3], Zimmerman and Rettaliata [4]) and that it is the absence of this feedback which results in the deterioration of speech following postlingual hearing loss. This assumption underlies the many efforts that have been directed towards finding a replacement for feedback which is missing in deafness: typically some correlate of the speech signal which is normally hidden from the subject but which can be made tangible by some means. Recently technological developments have enabled the hearing impaired in general to benefit from several different types of stimulant primarily presented in the form of a visual display. For example, Nickerson et. al. [4], Brooke et. al. [5] and Povel and Wansink [6] have developed devices specifically designed to teach vowel production by visual means. Typically in these systems some feature of the articulatory process is used to control some event on the screen such as a tracking sequence or a game procedure, or is plotted onto some articulatory target such as a vowel space display. While these systems provide different types of unambigious displays, they do not give the subject any articulatory correlate which he may refer to in unaided discourse.

VISUAL FEEDBACK FOR THE DEAF

The device proposed in this paper, intended primarily for use in speech conservation work with the postlingually deaf in the area of articulatory placement of vowels, provides a direct visual display of the tongue position in profile. For each of several vowels, the display shows the resting tongue position, the target position and the actual position achieved by the subject. The subject thus has immediate information on any discrepancies between where his tongue should have been and where it was, alerting him directly to the oral sensations involved and thereby facilitating carry-over when the display is not in use. Flexible pattern recognition is an additional feature, enabling the degree of accuracy of the subject to be manipulated. This ensures that reinforcement and feedback can be given to a subject even in the early stages of training when his productions may bear little resemblance to the target template.

SYSTEM OVERVIEW

The system to be described here is an extension to that described by Boyd et. al.[8], differing only in software and display. Briefly, it conveniently divides into three components: preprocessing hardware, microcomputer and display as shown in fig. 1. The hardware consists of a conventional dynamic microphone coupled to automatic gain control and preamplifier circuits which condition the incoming speech signal. Spectral analysis of this signal is achieved via a monolithic audio spectrum analyser [9]. This chip is designed specifically for speech recognition systems and features 16 channels of bandpass filtering, signal detection, and post filtering covering a frequency range from 260 Hz - 6kHz. The outputs are available separately but for economic reasons Boyd et. al. chose to make use of the on chip multiplexer and an additional fast 8-bit analogue to digital converter (ADC) to select them. The microprocessor performs three tasks: collection of the input data by providing control of the multiplexer and ADC, processing of the data to extract required features and presentation of the display. The micro used in this system is a BBC model B chosen because (i) it is competitively priced, (ii) it is widely used in special educational and clinical establishments and (iii) it has excellent graphics facilities.

Two different types of display are used in the system. A two-dimensional frequency-amplitude display of a sustained sound is maintained from the original system because of its usefulness in differentiating between vowels which look similar from a lip reading perspective. With this type of display the screen is divided into a teacher's display area and a student display area. The teacher produces a reference sound pattern in the upper half of the screen and the student attempts to mimic it on the bottom half. The second, and major, display of this system shows a mid-sagittal view of the vocal apparatus which is constructed from

the first three formant frequencies F1, F2, F3; extracted form the students utterance, using the procedure proposed by Ladefoged et. al [10]. Using this technique, they showed that the shape of the tongue during vowel utterances could be characterised in terms of two factors, a front raising component R_F and a back raising component R_B. They found the optimum proportions of these two parameters to be related to the formant frequencies as follows:-

$$W_f = 2.309(F_2/F_1) + 2.105(F_1/F_3) + 0.117(F_3/F_1) - 2.446 \quad (1)$$

for front raising and

$$W_B = -1.913(F_1/F_2) - 0.245(F_2/F_1) + 0.188(F_3/F_1) + 0.584 \quad (2)$$

for back raising.
Thus given the first three formant frequencies, eq. 1 and eq. 2 can be used to plot the corresponding tongue shape. In this system the display shows two tongue positions, one for a reference vowel and a second, the subject's attempt to reproduce it. Fig. 2 shows the display for attempts at the vowels /ɪ/ as in the word 'bit' and /U/ as in the word 'boot'.

Formant Selection

One disadvantage in generating this type of display is the dependence upon formant frequency extraction. This is problematic at the best of times but the difficulties are compounded when extraction is required from corrupted speech such as that produced by the hearing impaired. However, there are a number of reasons why this technique is suitable in this situation. Firstly, Ladefoged et. al's. model is perceptually quite robust in terms of formant deviations. Secondly, gross tongue shapes suffice for current needs and lastly, because of the system's hardware arrangement it offers the most economical solution to this type of display. Unfortunately, because of the wide spacing between the filters, the audio spectrum analyser does not lend itself conveintly to formant extraction using simple techniques like peak picking. Nevertheless, a peak picking algorithm was chosen, supplemented by a procedure which was proposed by Millar [11] to overcome the influence of voicing on formants within speech spectra. As an example of Millar's technique, consider a portion of a typical output from the audio spectrum analyser as shown in fig. 3. A_n and F_n represent the amplitude and frequency respectively of the n^{th} filter with the actual formant frequency F_F having amplitude of A_F. By assuming the spectral envelope to be triangular, Millar is able to solve for F_F using linear equation theory. He showed that:

$$F_F = \frac{F_{n+1} + F_n}{2} + \frac{A_{n+1} - A_n}{2m} \qquad A_{n-1} < A_{n+1} \quad (3)$$

where $m = (A_F - A_{n+1})/(F_{n+1} - F_F)$
and

$$F_F = \frac{F_{n-1} + F_n}{2} + \frac{A_F - A_{n+1}}{2m} \qquad A_{n-1} > A_{n+1} \qquad (4)$$

where $m = (A_{n+1} - A_n)/(F_{n+1} - F_n)$

Thus the strategy used in formant estimation is to firstly pick the peaks of the spectrum and determine the most likely candidates. F_1, F_2 and F_3 are then chosen by application of eqs. 3 & 4.

Pattern Recognition

Formant frequencies for several 'reference' vowels are stored in memory in the form of templates. The templates can of course be constructed from several sources; teacher, third party or the subjects themselves. For normal hearing subjects, only one utterance is required to do this. With a deafened subject, several utterances are required which are then assessed by a speech therapist to select the best few attempts. Formants are then for these, again using the method outlined. Finally the formants are averaged to give the template data. During the therapy session a simple pattern recognition technique is performed to compare the subject's utterance with his own templates. If the utterance and template are not sufficiently close to achieve a perfect match, the therapist is given the opportunity to relax the degree of match. This can best be explained with reference to fig. 4. The subject's utterance has been reduced to the first three formants. The template can be considered as a grid with windows which correspond to reference formant frequencies. In a perfect match the formants line up exactly with the windows on the grid. When the formants do not line up exactly the width of the windows can be altered to allow a match to take place.

FIELD TRIALS

The device has been used in field trials by a speech therapist working with deafened individuals whose pattern of speech deterioration indicated that vowel placement correction would be of value. Because of the necessity to ensure comparability between subjects for the purposes of these empirical trials, and the difficulty of matching deafened people on such factors as type and severity of deafness, age of onset, pattern of speech deterioration, previous exposure to speech therapy, and other similarly important variables, only three subjects were selected for intensive therapy. They were two males and one female, ranging in age from their early twenties to mid forties, who had all suffered profound hearing loss in their early teens, whose speech contained vowel distortions, and who had had no previous speech therapy. These were matched with three normal hearing controls

who were given the same degree of exposure to the device; in order to compare the value of the visual information in the presence and absence of auditory feedback.

Each subject participated in ten sessions where a therapy session consisted of work on the same four English vowels. A minimum of twenty attempts were made at each vowel: the first five were used to establish the window width which would be accepted for that vowel during the session, then the subject continued until he had reached fifteen correct productions. There are therefore two measures which can be examined to gauge improvement over time: the window width allowed, a smaller window meaning greater accuracy in the subject's attempts, and the number of attempts needed to reach the criterion level of fifteen productions within the window.

TABLE 1 Mean percentage window widths allowed over ten sessions

Subjects	Sessions										
	1	2	3	4	5	6	7	8	9	10	Overall
Deaf	23	20	18	16	15	18	15	13	11	11	16
Hearing	21	15	11	8	5	6	6	5	5	6	8

TABLE 2 Mean number of attempts to criterion over ten sessions

Subjects	Sessions										
	1	2	3	4	5	6	7	8	9	10	Overall
Deaf	28	24	23	23	21	22	20	18	18	18	22
Hearing	23	20	18	16	15	15	15	15	15	15	17

Table 1 displays the window widths averaged over the three deaf and three hearing subjects, across ten sessions. It can be seen that both groups became more accurate over time, requiring a smaller window, but the deaf subjects required windows still about twice as large as the hearing group at the end of the sessions. Further trials would be necessary to discover whether the deaf subjects could eventually achieve the same widths as the hearing; they do not appear to have reached a plateau in their scores yet, while the hearing subjects' windows settled down around 5% after the first five trials.

Table 2 shows the mean number of attempts made by each group per session to achieve fifteen productions within the set window. As in the pattern of results for the window widths, both groups show marked improvement over time, but the rate of increase is much slower for the deaf subjects and they have not reached the optimum performance after ten sessions.

DISCUSSION

The results of these trials clearly indicate that deaf subjects can increase the accuracy and consistency of their vowel productions with the help of the visual display device. One unanticipated outcome of the trials was that all subjects showed an interest in and an increasing ability to interpret directly the spectral display which preceded the facial display. This raises the possibility of working with more complex displays at a later stage of therapy. It also reinforces the current intentions to develop a more comprehensive system of augmentative feedback for the hearing impaired, and to incorporate the present display into such a system.

Hearing subjects also benefit from the visual feedback, but are able to make use of their auditory mechanism to accelerate the process. This renders the device of interest to phonetics training courses.

A full evaluation of the device would involve training subjects (both hearing and deaf) on some vowels using this visual feedback and other vowels using standard verbal feedback by the therapist as to the accuracy of each attempt, then comparing the rate of progress on each. Such an evaluation is currently underway; preliminary results indicate that without the direct articulatory information provided by the device hearing subjects show only a slightly slower rate of progress, but the effect on the deaf subjects is dramatic - they take very much longer to achieve consistency within a session, and show almost no carry-over to subsequent sessions. The implication is that the articulatory information displayed by the device was the key factor in the improvements noted in this evaluation.

One issue which should always be addressed in a therapeutic environment is the degree to which any improvement noted in therapy carries over into the client's everyday life. No formal evaluation was attempted of the extent of carry-over from work on isolated vowels to the deaf subjects' connected discourse, but informal observation indicates that this was minimal after this relatively short exposure to the device. Any future work should be designed to incorporate this essential element.

ACKNOWLEDGEMENTS

The authors would like to thank Ivan Boyd and William Millar for their contributions to this paper.

References

[1] R.I.D. Cowie and E. Douglas-Cowie, 'Use of feedback in established and developed speech', In M.E. Lutman and M.P. Haggard (Eds), Hearing science and hearing disorders, 183-

203. New York Acd. Press (1983).

[2] B. Massen and D.J. Povel, 'The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech', J.A.S.A., Vol. 78, 877-886, (1985).

[3] G. J. Borden, 'Use of feedback in established and developing speech.'in N.J. Lass (Ed), Speech and Languages: Advances in basic research and practice., Vol. 3, 223-242. New York: Acd. Press, (1980).

[4] G. Zimmerman and P. Rettaliata, 'Articulatory patterns of adventitiously deaf speaker: Implications for the role of auditory information in speech production.' J. Speech and Hear. Res., Vol. 24,169-178, (1981).

[5] R.S. Nickerson, D.N. Kalikow and K.N. Stevens, 'Computer-aided speech training for the deaf.' J. Speech and Hearing Disorders, Vol. 41, 120-132, (1976).

[6] S. Brooks, F. Fallside, E. Gulian and P. Hinds, 'Teaching Vowel Articulation with the Computer Vowel Trainer,' Brit. J. Audiology, Vol. 15, 151-163 (1981).

[7] D.J. Povel and N. Wansink, 'A Computer-Controlled Vowel Corrector for the Hearing Impaired', J. Speech and Hearing Res., Vol. 29, 99-105, (1986).

[8] I. Boyd, W. Millar, L.M. Boyd and E.L. Gailey, 'A Spectral Visual Feedback System for the Hearing Impaired', I.E.E. Int. Conf. Speech Input/Output; Techniques and Applications, Conf. Pub. no. 258, 257-262, (1986).

[9] L.T. Lin,H.F. Tseng, D.B. Cox, S.S. Viglione, D.P. Conrad and R.G. Runge, 'A Monolithic Audio Spectrum Analyzer', I.E.E.E. J. Solid-State Cots., Vol. SC-18, No.1, (1983).

[10] P. Ladefoged, R. Harshman, L. Goldstein and L. Rice, 'Generating vocal tract shapes from formant frequencies', J.A.S.A., Vol. 64, No. 4, (1978).

[11] W. Millar Unpublished Ph. D. dissertation, Q.U.B. (1981).

VISUAL FEEDBACK FOR THE DEAF

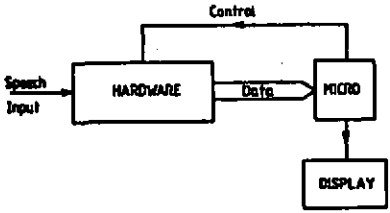

Fig 1 SYSTEM COMPONENTS

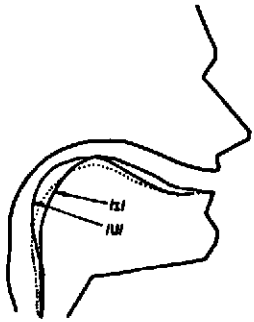

Fig 2 DISPLAY SHOWING TONGUE SHAPES FOR THE VOWELS /I/ AND /U/ WITH RESPECT TO THE NEUTRAL POSITION

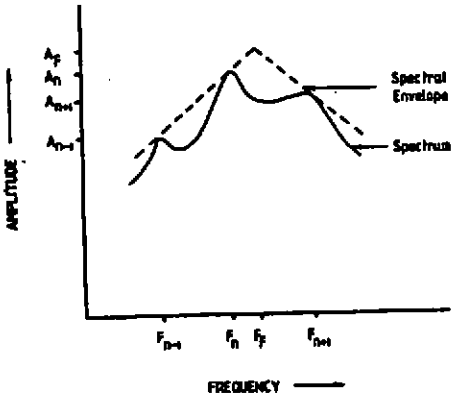

Fig 3 FORMANT ESTIMATION FROM SPECIAL OUTPUT

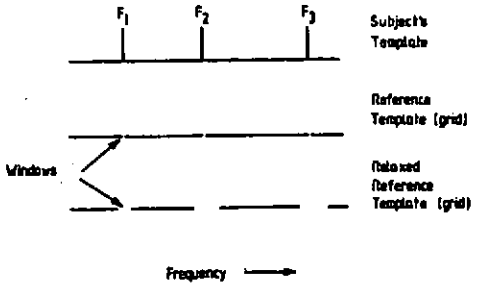

Fig 4 ILLUSTRATION OF FLEXIBLE PATTERN RECOGNITION