VISUAL FEEDBACK OF SPEECH: PHONETIC ASPECTS OF THE IBM SPEECH TRAINER

Florien J. Koopmans-van Beinum

1.0 INTRODUCTION

In 1983 VISI-C, the Dutch Foundation of Visual Speech Information, has been established to promote the use of Cued Speech in The Netherlands. Cued Speech was invented in 1967 by Dr. R. Orin Cornett of the Gallaudet College, Washington DC, in order to reduce or eliminate the ambiguities in lip-reading. In his system the speaker can code additional phonological or phonetic speech information by combining positions and shapes of his/her hand (e.g. Cornett, 1972).
The Dutch Foundation VISI-C has recently extended its activities since IBM-Holland provided it in January 1985 with three so-called 'speech trainers'. These are IBM Personal Computers equipped with a special prototype board for speech input and processing, developed by the IBM France Scientific Center within their Deaf Children Project (Denoix, 1984).
The three speech trainers have been provided to VISI-C in order to get information concerning their use in practice and to get them evaluated in various Dutch institutes and schools for deaf and hearing impaired children. Moreover, a scientific committee, consisting of researchers of three Dutch Universities (of Utrecht, Nijmegen, and Amsterdam) was formed to supervise this Dutch speech trainer project.
Although a trial period of only six months is far too short to present a systematic evaluation of the speech trainers as applied in the institutes and schools for deaf and hearing impaired children, I nevertheless want to describe here the first impressions from teaching practice and from our own laboratory trials. This paper therefore presents only the personal views and experiences of the author.

The main philosophy of the designers of the speech trainer seems to arise from the rage concerning video or computer games. If guiding an animal or an object can be achieved by voice input, this rage as it were might be employed just as well in speech training projects for deaf and hearing impaired children, with the aim to teach the deaf child a better control of its voice. By visualizing parameters of the voice on a screen and by comparing these parameters with similar parameter representations produced by the teacher's voice, it is believed the child can learn how to improve his speech in a gradual way.

2.0 POSSIBILITIES OF THE SPEECH TRAINER

First of all the present capabilities of the IBM speech trainer will be described here, followed by some comments.
All programs run under the DOS 2 operating system. Normally, as said in the short subjoined manual, drive A would hold a DOS diskette, and application

programs would reside on a diskette in drive B. In our case, however, use was made of a single diskette holding speech training programs as well as DOS commands. Since no PCSE.BAS APPLICATION PACKAGE was provided, we got no possibilities to acquire acoustic parameters and to perform parameter computations. So the present report will concern only the speech training programs as we used them and as we saw them in various versions to be used in speech training practice in institutes and schools.

2.1 Program selection

A MENU program displays the list of available speech training programs from which one can select one program at a time by pressing the corresponding function key or, in a later version, by stepwise moving a cursor to the name of the wanted program. Different languages can be selected, meaning that the program names will be displayed in the selected language and that this language will be used thereafter in the programs too. When we started the project only English, French and German were available, in a later version Dutch was fed in as well.

2.2 Amplitude adjustment

The first thing to do when the computer is powered on is to adjust the potentiometer for a convenient signal amplitude. While pronouncing a sustained vowel sound the screen displays intensity on a dB-scale together with the amplitude of the signal (on a scale going from 1 to 11). If necessary a potentiometer in the microphone amplifier can easily be adjusted to the desired level. As in the other programs the Escape key is used to exit this program and to return to the MENU.

2.3 Project presentation

This option of the MENU offers a general presentation of the IBM research project, displaying information on deafness, visual feedback in speech training, Cued Speech and lip-reading. As said in the manual this program is not really self sufficient, but it should be accompanied by an oral presentation. Within this neat demonstration program one can pronounce a sentence and in real time produce pitch graphs in three possible colors. By pressing the space bar one can proceed from one page of the presentation to the next.

2.4 Acoustical parameters

The program DISPSENT displays acoustical parameters for short sentences, processing directly the digitized speech signal. The screen is divided into two parts: the upper part displaying in real time the zero crossing rate of the signal in green, and the intensity (the energy level in decibels) in the

same graph in red. The maximum duration of a sentence to be processed is 3.28 sec. A triangular cursor can be moved across the sentence display and pinpoint a given frame to be displayed in more detail in the lower part of the screen.
For the display in the lower part one can choose from two options: in the first place a fragment of the signal waveform, sampled at 10 kHz, 12 bits per sample. Part of this waveform, corresponding to the indicated frame of 128 samples, i.e. 12.8 ms, is represented in red, the remaining part in yellow. A scale factor can be used to achieve a better representation in case of low amplitude waveforms.
The second option for the lower part represents the autocorrelation coefficients, linear predictive coding coefficients, and the frequency spectrum from 0 to 5 kHz. Here an option provides computation of either normal correlation coefficients from the derived signal multiplied by a Hamming window, or pitch synchronous autocorrelation coefficients.
Each time the cursor is moved one frame to the right or to the left by using the corresponding arrow key, the display is updated. If necessary the cursor can be moved at a greater speed across the screen to reach the selected speech instant.

2.5 Pitch and intensity graphs

The program PITCHINT displays graphs of pitch and/or intensity and is intended to be used interactively by the teacher and the pupil in speech training sessions. The screen can be used in two modes. One option provides a split screen displaying the teacher's screen in the upper half and the pupil's screen in the lower half. The other option is a mixed screen mode representing the pupil's graph overlaid over the teacher's, using different colors. A flashing symbol indicates which screen is active, since only one can be active at a time. An option provides either pitch display or intensity, or both together. A musical note symbol indicates the pitch mode, a sparkling sun symbol indicates the intensity mode. Both symbols together are displayed to indicate the combination mode.
Pitch and intensity graphs can be produced from left to right or from right to left. In the option right to left no combination of pitch and intensity is possible, but the display is continuous until the user stops it. In the option left to right the display is either one-off or continuous.
The display area corresponds to a time span of 3.17 secs, while every 12.8 ms a new point in a graph is displayed. Frequency and intensity scales can be set separately for teacher and pupil and are represented in the right of the screen. For the active part of the screen the display parameters can always be modified. Standard intensity scale values can be selected: 15-60 dB for male voice, 20-70 dB for female voice, and 25-60 dB for child voice. Frequency scales can be selected: 80-300 Hz for male voice, 200-400 Hz for female voice, and 250-430 Hz for child voice. Optionally the preselected frequency scales can be changed, thus vertically changing the graph accordingly.
Switching from split screen to mixed screen overlays both graphs currently displayed on split screen. Optionally an optimal horizontal matching of both

graphs can be computed and, if selected, the displayed graph is centered correctly according to the average and scaled according to the standard deviation of the displayed pitch or intensity values.

2.6 Pitch controlled games

The program PITGAME provides the user with a number of games very much alike the well-known video or computer games in which the user has to guide a mobile object or a being across the screen to reach targets while avoiding obstacles. However, PITGAME does not ask for a joystick or some other handle, but for voice input: the object or animal is pitch controlled. As soon as a voiced sound is produced the mobile starts moving to the right of the screen and keeps moving as long as the sound is sustained. To guide the mobile along the obstacles the user has to higher or lower the pitch while voicing. Three difficulty levels can be selected. At first four versions were available: a camel drinking lakes and avoiding palm trees, a duck eating worms and running away from wolves, a dolphin catching fishes and keeping clear of squids and octopuses, and a car trying to reach the gasoline stations and avoiding pedestrians. Since it seems to be very easy to create new games a later version contained also a mouse eating cheese and dodging cats, and a more or less winding path that had to be followed without going beyond the bounds.

2.7 Articulation game

The articulation game program provides a set of labyrinth games based on automatic vowel recognition. To get around part of the problems of speaker normalization the user has to select beforehand whether a male, a female or a child is speaking. Next one has to select one set out of several series of vowels, and choose the type of labyrinth (e.g. rectangles, a spiral, blocks of houses in a city) from which one has to find his/her way out. When the intended vowel is pronounced a mobile moves into the direction indicated by the vowel and keeps moving into that direction as long as the vowel is sustained until the mobile bumps into the walls. Then one of the vowels indicating another direction has to be pronounced. In this way the game goes on until the exit has been reached.

2.8 Word recognition

The word recognition program comprises two possibilities. In the first optional program the user types a maximum of ten words. By pronouncing the typed words twice in succession, one trains the recognizer. Again, to overcome part of the speaker normalization problems the user has to select beforehand male, female or child speech mode. Next, when pronouncing one of the trained words, the program tries to recognize it and displays one of the ten words, the result being correct or wrong.
The second option provides a word recognition game very much like the

'mouse-game' of Bell Labs (e.g. Sorace et al., 1983), as demonstrated at Epcott Center in Orlando, FL. In the word recognition game of the speech trainer the user has to train the computer by pronouncing five words, indicating left, right, up, down, and stop. Every word the user wants to pronounce for this purpose in whatever language, may be selected. After the training phase a mobile, in this case a large square block, has to 'eat' all small blocks from the screen, while it is guided by the user by means of the five spoken commands.

2.9 Lately added games: CAR DRIVER and BALLOON

In a later version of the speech trainer two more games have been added: CAR DRIVER and BALLOON. The first one cannot be described here since I never saw it functioning. The other one, BALLOON, is a Japanese addition, as can be seen moreover from the Japanese characters in the display. The game is intended for the training of respiration combined with phonation. First the teacher or the pupil him/herself creates mountains or plateaus in a flat landscape by producing a voiced sound and sustaining it as long as the plateau has to be extended. The landscape is created in the upper half of the screen and can be continued in the lower half. Next a mobile balloon has to be lifted over the mountains by means of phonation. Since there are several mountains in the landscape on irregular distances of each other, the game calls for exact timing of respiration combined with phonation, otherwise the balloon will be smashed against the mountain-side.

3.0 COMMENTS ON THE USEFULNESS OF THE OFFERED PROGRAMS

As said already in the introduction a trial period of six months is far too short for a systematic evaluation of all aspects of the Dutch speech trainer project. In addition, rather much time was lost in the first months because of some technical problems concerning signal amplitude adjustment. Nevertheless we consider it wise to discuss in an early stage the positive and negative sides of the speech trainer as it is at this moment.
One of the main problems while working with children in training practice is the time consuming starting procedure of the speech trainer and the relatively long duration for going from one used program to another wanted program. Since training sessions often take only fifteen minutes, it is necessary to avoid idle waiting time. If accidentally the teacher makes a wrong program selection the whole time consuming procedure has to be restarted. Moreover, in several programs it is impossible to restart rapidly in case of an incorrect speech production.
Another type of problem is the fact that the speech trainer, as it is offered to us, is still in a developmental stage. This means that a number of the programs described above did not work at all or fell short, which is demotivating for the teachers. Also the lack of a clear and complete manual was cause of unnecessary problems.
In the next paragraphs we shall describe merits and demerits of the successive programs, on the basis of our own experiences or as reported by

the teachers or the other committee members.

DISPSENT: acoustical parameters.
Although this program explicitly is not intended for the children, but for the teachers, it is of high interest for every phonetician working in the field of acoustics. In its present form however, it is not useful for several reasons: e.g. the user has no access to parameter values; the displays are not very detailed; the time-axis is invariable; the speech fragment cannot be re-heard; the useful spectrum is marginally displayed as compared to the marked representation of the LPC-parameters. In addition it might be more instructive to display the waveform in the upper half of the screen and optionally data of analysis (zero-crossings, energy, LPC-parameters, etc.) in the lower half.
Nevertheless it is clear that this program contains many of the elements necessary for speech training. The main question, however, is which are the features to be displayed and in what form.

PITCHINT: pitch and intensity graphs.
This program can be seen as the main corpus of the programs available in the speech trainer and seems to be the most elaborate one. Although the operation procedure is rather time consuming, the program turned out to be very useful in speech training practice. Especially the intensity display makes children aware of e.g. incorrect durations of speech sounds or stops, of unwanted syllabification or vowel addition, of leaving out consonants in clusters, of low intensity compared to the pattern as produced by the teacher. The children seem to be highly motivated by the 'objective' visual feedback of their speech productions, rather than the 'subjective' approval or disapproval of the (hearing) teacher.
Young children can be trained here very well in diadochokinetic exercises: trying to increase the number of repetitive speech movements (babbling) becomes a self controllable game !
Another additional positive aspect of the program PITCHINT is the required interaction between teacher and child: turn taking is an important element in the procedure of this training program and can be trained in passing.
The pitch graph requires much more scaling activities in order to achieve an acceptable display and is therefore less attractive in practice, whereas the combination graph of pitch and intensity mainly functions as a thick intonation contour without information on intensity.

PITGAME: pitch controlled games.
Although these type of games turn out to be very attractive for children, one may ask whether they will not be horribly boring after some weeks. However, speech training practice in schools and institutes never offers the possibility of playing these games night and day.
As for the usefulness of the pitch controlled games we consider them inappropriate, at least in their present form. The 'path' game is the only significant one since the children have to follow here an imposed pitch contour. All other games invite incorrect and uncontrolled phonation since large intonation jumps guiding the mobile right across the obstacles, yield the best results! Only if the child is obliged to use murmurs, the game

answers it purpose of controlling pitch movements more or less. Making the intonation track as produced by the child visible to avoid extreme pitch movements, might be in my opinion of great profit.

ARTICULATION GAME: labyrinths.
Although the ideas behind these labyrinth games are very useful in speech training practice, the games in their present form are absolutely inappropriate. Apart from the fact that technically the game doesn't function very well and only part of the offered choices actually can be selected, the game is inappropriate still because of the French character of the vowels to be used. Trying out this game by using synthetic vowels revealed that the more unnatural a vowel sounded, the greater success was scored. Moreover it is very annoying that after making a new labyrinth selection, the display does not change unless some noise has been produced. This holds even for the option ESCAPE!
Nevertheless this type of games might be a valuable tool in speech training if these problems could be overcome.

WORD RECOGNITION.
It is not clear why the word recognition program is offered in the speech trainer packet. First of all the word recognition is of very poor quality, duration seeming to be the only cue. Even the word recognition game with only five words to train, performs badly. In the second place, however, one may wonder whether it makes sense for a deaf or hearing impaired child to train a word recognizer by means of incorrectly articulated words. Ideally the teacher should train the recognizer, but as long as problems of speaker normalization have not been solved adequately, word recognition cannot be considered to be a meaningful tool in speech training.

CAR DRIVER did not function in the versions I worked with.

BALLOON: respiration and phonation training.
This speech training game turned out to be a very basic training tool since timing of respiration and phonation are essential in this game.

4.0 CONCLUSION

The idea of giving deaf and hearing impaired people a visual feedback of their speech, is not new at all. However, the coming of personal computers brings this kind of feedback within the scope of speech training in schools and institutes. But too often incomplete and even inadequate products with a lot of growing pains are sent into the world, promising mountains of gold. However, if one opts for this kind of product promotion, it must imply high requirements as for usefulness, and a carefully attuning of needs and technical possibilities. The main question, as said before, is the need to explore which features are necessary to display and in which form this will be useful and teachable in speech training. These essential questions have not been solved yet in the IBM speech trainer, but the system itself provides many possibilities.