Institute of Phonetic Sciences,
University of Amsterdam
Proceedings 15(1991), 121-129.

THE IBM SPEECHVIEWER

Ton G. Wempe and Marlien van Lunen[*]

Abstract

The 'SpeechViewer' system, developed by IBM for using with their PS/2 microcomputers or compatibles, is meant as an aid for speech therapists while working with their clients. This paper reports about its speech training and analyzing facilities that are tested in practice to some extent. Using a hardware vowel generator, the accuracy of the detection of the speech features is tested as well. Finally, some evaluation remarks are given.

1 Introduction

The IBM 'SpeechViewer' is developed in France by IBM to support the treatment of speech and hearing impairments and can be seen as a training instrument and a diagnostic aid for clients under guidance of a speech therapist. The system is meant to enhance existing therapies and to improve diagnostic tests. Speech therapists could teach their clients how to get control of such speech features as pitch, loudness, pronunciation and rhythm.

The 'SpeechViewer' consists of an (8 bit slot) extension board which fits into an IBM pc or compatible, a microphone, a loudspeaker and a software package. When applied in a pc (preferably with colour display) the system is able to digitize speech signals from the microphone (or other audio source), process them in real time for interactive training and analysis, and display their waveforms or other data. The digitized speech can be made audible again.

While processing speech the various sections of the program give immediate feedback in terms of loudness, pitch, temporal aspects, voicing, spectral data, etc.. In addition, it is possible to recognise clients' vowel sounds spoken into the microphone by comparing them with model vowels from a built-in standard Dutch vowel model. New vowel models can be brought in via the microphone as well. From the speech sounds some statistical data can be extracted (mean and standard deviation of pitch, amount of voicing, etc.).

The 'Centrum Informatica voor Gehandicapten' (Dutch Informatics Center for the Handicapped) provided us, for the duration of several months, with such a system for evaluation purposes so that we could test its functions and games.

Rather than an exhaustive evaluation of all possible functions, as is done with an earlier version of the system (de Leeuw, 1987), our activities with the system were limited to exploration and judging of its functions by a speech therapist and to checking

[*] Speech therapist

of response and accuracy with the aid of an artificial vowel generator which produced stable vowel signals with well known parameters.

Some experiences and comments are presented in this paper.

2 Summary of the systems' features

In this chapter a global description of the SpeechViewer's programmes is given. We will use the classification and names of program sections according to the main menu of the system itself.

2.1 Section I 'Kennismaking' (introduction)

- Program: 'Stemgeving' (voicing)

When the sound picked up by the microphone is voiced and exceeds a certain level, a kaleidoscope is displayed which continuously varies in form and colour during the time the signal meets these conditions. As soon as the sound level drops below the (adjustable) threshold or the sound becomes voiceless, the duration of the latest uninterrupted sound is displayed in digits.

- Program: 'Luidheid' (loudness)

On the screen a balloon is displayed which varies in size according to the loudness of the sounds uttered.

- Program: 'Toonhoogte' (pitch)

On the screen a thermometer is diplayed which shows the current pitch as long as the incoming sound is voiced. The scale has 100 Hz steps. In addition, the highest and lowest pitch that occurred are displayed in digits.

- Program: 'Steminzet' (voice onset)

The screen shows a train on a track. Every time the voiced part of the incoming signal starts, the train moves a little forward. The minimum duration of the voiced part must be 0.1 s for activating the train.

- Program: 'Luidheid en stem' (loudness and voicing)

A clown with a green bow tie is displayed. When the produced signal is voiced, red spots appear on the tie. When the volume of the signal is raised, the mouth of the clown becomes bigger.

2.2 Section II 'Vaardigheidstraining' (speech training)

2.2.1 Program: 'Toonhoogte' (pitch)

At the left side of the display a movable figure appears. Scattered all over the screen, a number of targets and obstacles are displayed. The purpose of the game is to move the figure to the targets, avoiding the obstacles, by raising or lowering the pitch.

Several different figures can be chosen. Varying levels of difficulty can be selected. The figure has to be moved by producing gliding tones. When all targets have been hit and all obstacles are passed without errors the client is rewarded with pictures resembling fireworks on the screen and noises through the loudspeaker.

Further options which can be selected are:

- alteration of the maximum speed of pitch change (when this maximum speed is exceeded the screen marks the place and direction of that change)
- alteration of pitch range (can be stored on disk for each client)
- alteration of horizontal speed
- manual positioning of any number of targets and obstacles

2.2.2 Program: 'Stemgeving' (voicing)

First, a target sentence can be produced during which a balloon travels slowly in horizontal direction over the screen and draws a line at the same time. When the incoming signal is voiced, the balloon rises so that 'mountains' are formed.

Next, the client must lead the balloon over these 'mountains' by producing voiced speech at the right moments. When the pupil succeeds, the program gives a reward via the screen and the loudspeaker.

The mountain-pictures can be saved to and loaded from disk.

Three different horizontal velocities can be selected.

2.2.3 Program: 'Klinkercorrectie' (vowel correction)

A vowel must be imitated. The more the sound resembles the target vowel, the higher a monkey climbs a tree, pushing a little monkey up the tree. Each time a new sound is started, the first monkey starts from the ground. The second monkey, however, stays at the highest point that was reached in the tree, thus showing how close the sound matched the target. When the top is reached a coconut falls from the tree and an accompanying sound is heard. Now the client can start again from the beginning. A target vowel can be chosen in advance.

To let the monkey start, the deviation from the target vowel has to be less than the lower threshold; to reach the top, the deviation must be less than the upper threshold. Both thresholds can be adjusted. The various levels of matching are also displayed in digits. The target vowels can either be selected from a built-in model or can be newly defined in the program 'Forming vowel models' (see below).

2.2.4 Program: 'Klinkerdiscriminatie' (vowel discrimination)

A small square must be moved through a maze by producing vowel sounds.

To move the square in all four different directions, any four vowels can be selected from a vowel model. In the maze, the pupil sees which vowel must be produced for each direction. The maximum difference between the sound produced and the target sound, that is allowed for determining a vowel, can be adjusted ('tolerance threshold'). The form of the maze can be regenerated at random any time. The complexity of the maze can be altered as well. Also, the travelling speed can be changed.

2.2.5 Program: 'Klinkermodelformatie' (forming vowel models)

Four different selections are presented:

1. A new vowel model can be brought in by producing sustained vowels into the microphone. Every 53.2 ms a spectrum is extracted from the sampled vowel signal. A maximum of 25 spectra per vowel are extracted for collecting spectral variations.
2. A vowel-recognising test can be run to check vowel discrimination. Any model present on disk can be chosen as the reference vowel model (a standard Dutch vowel set ('ABN') is included in the software). A 'tolerance threshold' can be adjusted for accepting different amounts of spectral variations.
3. A matrix of relative spectral differences between all vowels can be displayed ('statistical data').
4. Vowel models can be 'compressed' (the number of spectra with differences less than an adjustable threshold is reduced), or merged with other models.

2.3 Section III 'Imitatie' (imitation)

2.3.1 Program: 'Intonatie en uitspraak' (Intonation and pronounciation)

On the screen a real-time plot of the pitch and/or loudness of the incoming sound is displayed. Voiced and voiceless sections of the sound are represented by different colours. For imitation purposes, it is possible to display the pitch/loudness of the current microphone signal and of a target utterance simultaneously by using a split screen or a 'front/back' display. The range of the time axis can be set to 4, 8 or 12 s. Different pitch and loudness ranges can be selected.

In the 'Statistical data' mode two histograms are displayed: one for the percentage of occurrences of each pitch value and another for the percentage of occurences of each relative loudness value. For each graph the median, mean value and standard deviation are presented in digits (the pitch data are presented only for the voiced parts of the signal). In addition, the total duration of the time the signal was voiced is given. The recorded speech samples can be saved on disk (in 8 bit format) for later use. Speech samples from this program can be used in the program 'Speech analysis' and vice versa.

2.3.2 Program: 'Spraakanalyse' (speech analysis)

The upper section of the screen displays a graph of the pitch and loudness of the incoming signal as functions of time (the total time sweep on the screen is slightly more than 4 s). Using cursors, a time segment of the graph can be selected with an accuracy of 13.3 ms steps. The lower part of the display shows a waveform representation of that section of the signal which is demarcated with the cursors in the upper display area. The vertical range of the waveform graph can be altered by using a horizontal cursor in the upper display area. The selected segment can also be made audible.

From the selected segment statistical data can be presented in the same way as described above.

It is possible to save the recorded speech samples on disk for later use (samples are stored in 8 bit format). Speech samples from this program can be used in the program 'Intonation and pronunciation' and vice versa.

2.3.3 Program: 'Spectra'

The screen shows a continuously updated spectrum of the speech sounds that are currently produced. The graph displays the spectral amplitude (logarithmic scale, 80 dB total range) versus the frequency (linear scale, total range 3.3 kHz). The applied time window is 52 ms. Spectra can be 'frozen' for comparison and imitation. Spectrum extraction only occurs when the input signal exceeds a specific amplitude threshold which can be adjusted in 5% steps.

3 Experiences with the system

3.1 Section I (introduction)

The programs in this section, which are very simple, can be used to let the clients become aware of the way they produce sound and voicing. Instead of displaying values of loudness or pitch the screen merely reacts on the existence of sound or voicing in order to familiarise the client with the control of these variables. The program 'Pitch' forms an exception: there, the thermometer shows the current pitch value. One may expect that the program 'Loudness' forms an exception too: the size of the balloon should indicate the amount of loudness. However, it turns out to be difficult to estimate the different sizes which the balloon assumes, as the loudness area between minimum size and half size is very small.

In the 'Voice onset' program the decisions made between voiced/unvoiced sounds are not always satisfactory. The train should move up a bit only on the onset of voice; some voiced consonants, however, are apparently treated as voiceless sounds (i.e. voiced [d], voiced [v], voiced [r] and voiced [g]). It seems that the microphone input sensitivity plays a part too.

The scale on the thermometer in the program 'Pitch' is heavily compressed at the higher end, probably due to the growing inaccuracy at increasing frequencies, caused by the limited sample frequency. The distance between 100 Hz and 200 Hz is about twice the distance between 200 Hz and 400 Hz. A logarithmic scale would be more convenient, however.

The pitch range, displayed in digits, cannot be 'transported' to other programs, for purposes of pitch scaling (see the remark about this subject in chapter 4).

3.2 Section II (training)

- Pitch program: (screen with targets and obstacles)

Adaptation of the pitch range to that of the client is necessary for achieving a fair chance of success in this game. When using the built-in screen layouts, even in the case of 'normal' speakers, it appears to be rather difficult to hit all targets and to avoid all obstacles.

The 'navigation' between editing screens (for manual positioning of targets and obstacles) and playing screens does not always seem straightforward (see also chapter 4 about user-friendliness).

- Voicing program: (travelling balloon)

A simple game for training the timing of voiced segments. With the target sentence the degree of difficulty can be varied widely.

- Vowel correction program: (monkey in tree)

The reliability of vowel recognition within this program depends much on the use of an adequate vowel model. A minimum requirement would be that, with the program 'Forming vowel models', separate vowel models are made of male, female and children's voices, and in this program (and all other vowel recognising programs) used for clients correspondingly. The built-in Standard vowel model for Dutch has a quite limited value in this respect. Furthermore, the therapist should adapt the upper and lower 'recognition thresholds' to the client's skills.

- Vowel discrimination program: (maze)

The same comments as were made above on the vowel models apply here as well. The adjustable 'tolerance threshold' should be set for each client individually.

It is obvious that only combinations of vowels should be selected from which the program can recognise all four vowels without confusion.

Both vowel recognising programs ('Vowel correction' and 'Vowel discrimination') have the disadvantage that the clients, while attempting to produce vowels correctly, get no information on the ways to improve their vowel sounds. They simply have to explore the right pronunciation by trial and error. A method as applied in the 'Vowel Corrector' from the 'Visual Speech Apparatus' (Povel et. al., 1991) seems preferable because its screen displays vowels as points in a plane (formant-like representation), so that the client should be able te explore the relation between articulatory movements and vowel position.

- Forming vowel models program:

As can be expected from this vowel-recognising system based on spectral differences, it is not possible to assemble a vowel model which can be used to recognise all vowels without confusion. Because the discrimination occurs statically, i.e. on spectral points instead of spectral traces, diphthongs must be left out of consideration.

The 'user-friendliness' of the programmes was found satisfactory, although the explanation of the 'tolerance threshold' and 'compression threshold' in the user's manual (IBM, 1988) was not immediately clear.

3.3 Section III (imitation)

- Program: 'Intonation and pronunciation'

This title promises more than the program offers. Only the variables pitch, loudness and voicing are displayed as a function of time, and no information on pronunciation of vowels or, let alone on consonants is presented.

The voiced/unvoiced detection is not always entirely correct. Signal parts of lower levels often are displayed as being voiceless.

In 'split screen' mode it could be rather difficult to imitate the target sentence because the recording time could contain many syllables, even when set to its minimum (4 seconds). Besides, in this program it is not possible to use cursors for demarcation and replaying of fragments. Such a feature probably would add extra training facilities. (The following program does have this feature, but its display was not designed for imitation purposes.)

- Program: 'Speech analysis'

From a 'phonetic' point of view this program (and the program 'Spectra', see below) seemed the most interesting part of the system because of its facilities for segmentation and analysis. Unfortunately, although a selected fragment can be replayed or its statistical data displayed, it cannot be imported into the program 'Spectra' for analysis.

Speech samples obtained in this program can be imported into the 'Intonation and pronunciation' program. However, this can only be done with the recording as a whole, not with a selected segment.

Of all programs, this program showed the most serious problems in the voiced/unvoiced detection: even in cases of skilled speakers substantial parts of their utterances were displayed as voiceless, while they were voiced in reality. Also, when we used recorded speech material of proven quality from male speakers, the amount of speech that was inadvertedly displayed as voiceless was even greater! As we made sure that the signal level of the tape recorder was adjusted properly and that no frequencies were produced outside the F_0 range, we had to conclude that, at least for F_0 contours, this part of the program cannot be used for a large number of speakers.

- Program: 'Spectra'

It turned out to be impossible to obtain steady graphs of spectra while producing sustained vowel sounds with constant parameters: the height of the formant peaks varied heavily. In addition, the positions of formant peaks on the frequency axis often seemed incorrect. (See chapter 4 for testing the program's accuracy).

The only way to provide this program with speech is via the microphone. If it were possible to make spectra of recorded speech files from other programs, for example from selected segments in the 'Speech analysis' program, the usefulness of this program would have been greater.

4 Accuracy

Some basic limitations of the accuracy of the Speech Viewer can be derived from the (very sparse) 'technical data' that can be found in the IBM Users Manual:

- AD-conversion is 12 bit (8 bit when saved on disk).
- Number of samples per second: 9600.
- Anti-alias filter: 3.8 kHz cut-off with 28 dB/octave.
- Time window: 13.3 ms (128 samples).
- Fundamental frequency range: 75 - 960 Hz.
- The loudness is determined in every window as the maximum amplitude within the window in percentages of the system's total amplitude range.

The sample rate used has a time grid of 104.2 µs, which means that the inaccuracy of the fundamental frequency measurements reaches from 0.6 Hz at the low end of the scale (75 Hz) to 87 Hz at the high end (960 Hz), i.e. the inaccuracy due to the sampling process varies from 0.8 % to 10 % approximately.

In order to get some information about the accuracy of the display of speech variables like formants and fundamental frequencies we used a (hardware) formant generator, built by Wempe and Van Maanen (1981), which can produce vowel-like sounds with accurate adjustable parameters.

During the feeding-in of vowel sounds with natural parameter values in all programs where F_0 is concerned, it appeared that most(!) parts of the signals were inadvertedly treated as voiceless. The voiced/unvoiced detection seemed to appear at random, while the produced signal was continuous and steady. After manipulating a great deal with the speech variables it became clear that the F_0 extraction only works satisfactorily when the signals contain relatively high energy in the lower frequency range (the first few hundred Hz). This could be an explanation for the disappointing response of the system to trained voices as these tend to have steeper glottal pulses and therefore relatively weaker low frequencies.

The accuracy of the fundamental frequency display was found according to the inherent limited accuracy caused by the sampling process (see above).

When we applied the formant generator for testing the 'Spectra' program the inaccuracy and instability of the heights of the formant peaks turned out to be 10 dB approximately.

The generated values of the formant frequencies were reasonably well represented by the frequency position of formant peaks on the display. However, most settings of speech variables resulted in the display of some additional peaks that impossibly could have been caused by the generator as the frequency components due to distortion could not produce frequency components in those regions. Although the heights of these 'alias' peaks were roughly 40 dB lower than the 'real' peaks, their mere presence and unstableness was quite disturbing.

5 Overall comments and conclusion

One of our first overall impressions of the system was its playful use of colour graphics. We assume that it will highly motivate clients, especially children. Although the computer system had VGA quality, the definition of the graphics seemed to be according to the IBM 'Colour Graphics Adapter' (one of the older graphics display boards).

The software package consists of many stand-alone programs that cannot interact with each other (although there are a few exceptions). Therefore it is generally not possible to supply a specific program with input data that are produced by an other program. For example, in the program 'Speech analysis', a selected part of a speech waveform which, using the cursors, has been selected from a complete sentence, could possibly be fed into the 'Spectra' program (both from Section III). A possibility like that would offer many extra features for analysis of the spoken utterances. Another example would be the possibility to use the measured pitch range in the program 'Pitch' from Section I (introduction) as input information in the program 'Pitch' from Section II (training). This would avoid frustration for the client when training pitch with a wrong pitch range of the pitch detection program.

In almost all programs (except those from Section III: imitation) the sensitivity of the microphone input cannot be adjusted. Because the loudness of the voices of the patients can vary heavily, the visual feedback in some programs depending on the intensity of the speech produced can be often out of range or not visible at all. Besides, the signal fed into the (12 bit) AD convertor being either too high or too low could of course result in distortion or a poor signal to noise ratio respectively. The programs from Section III (Intonation and pronunciation, Speech analysis, Spectra) have the ability to adjust amplitude range. However, after the user has left a program, the adjusted range is lost. Preferably it should be possible to set the input sensitivity in Section I (familiarizing) when a client starts, and to use that setting for all other programs during the complete session of the client.

The maximum pitch range for all programs where pitch is involved is 75 - 960 Hz. However, a considerable number of male voices produce pitches below 75 Hz. The parts of the signal where this is the case are always considered to be voiceless which causes the display to behave in an unwanted or even frustrating way.

The fundamental frequency steps at the higher range of the scale are so large that the graphs showing the fundamental frequency become quite rough staircase-shaped curves. As the display must react in real time, post-smoothing techniques would offer no real solution; obviously the best cure is using a higher sample frequency.

The limitation of the signal processing bandwidth is obvious: the sample rate used offers not much more than 'telephone quality' when utterances are replayed. However, for most of the interactive games and training this limitation is not important, as the system mainly deals with voiced signals, having their main energy at the lower part of the frequency range.

The loudness measuring method applied in the 'SpeechViewer' is not correct, strictly speaking. Although computation of the R.M.S. values would be a better estimation of loudness, the method used here turns out to be sufficient for the purpose.

The errors in the voiced/unvoiced detection of the system seem to depend very much on the spectral distribution of the voices applied. For some voices the programmes behaved in a quite unacceptable way.

The continuously updated graphs in the program 'Spectra' appear in such an unstable way that the usefulness of the program seems very limited.

The 'user-friendliness' of the software was found quite reasonable. Still it could be improved somewhat on the following items:

- In all screens the way to return to an earlier selection level should be displayed. A natural tendency exists to select ESC, which makes it necessary to start the program all over again; in other cases it is only possible to go one level back with the ESC key and sometimes the PgUp key must be used.
- The way to select specific options or settings should be displayed as much as possible on the current display screen. The way it works now means that a selection must be chosen before one sees how a setting can be altered, even when using the F1 (help) function.

References

IBM, (1988). *IBM PS/2 SpeechViewer Programs, User's manual*, IBM International Products Limited U.K. 129pp. (in Dutch)

Wempe, A.G. & Maanen, A.W. van (1981): "Speech Formant Generator", *Proceedings of the Institute of Phonetic Sciences Amsterdam* **6**: 47-55.

Povel, D.-J. & Arends, N. (1991): "The Visual Speech Apparatus: Theoretical and practical aspects", *Speech Communication* **10(1)**: 59-80.

Leeuw, I. de (1987): "The IBM Speechtrainer: An inventory of applications.", *Report of the Institute of Phonetic Sciences Amsterdam* **92**, 57pp. (in Dutch)