THE USE OF THE VOICE VISUALIZER IN
LANGUAGE TRAINING AND LINGUISTICS

By Thomas Guerin

The electronic Voice Visualizer is a machine originally developed by Nihon Kohden for the purpose of visualizing utterances in the form of electronic waves on a cathode ray tube for people who have speech defects or are unable to hear their own voice through hearing defects. In the original concept, the machine was intended to be used by such a person while comparing his own pronunciation of a particular word or phrase with that of a correct model through the use of a photograph of the model or with an instructor producing the correct model on the cathode screen alternately with the student.

Description of the Visualizer

The cathode tube’s electron sweep is either 2.5 seconds or twice that speed, i.e. 1.25 seconds. The faster sweep serves, in effect, to spread out an utterance so that it can be examined in more detail. There is a headphone set attached which has a microphone connected for student input. There is also a separate microphone input which can be used simultaneously by an instructor. The volume of the input can be adjusted and the wave pattern can be filtered to give contour rather than sharp wave patterns. The visualizer is equipped with an orange filter to sharpen contrast for viewing. The electron wave can also be made brighter or weaker in intensity.

There is a tape recorder attached with a 2.5 second endless tape which exactly matches the slower electron sweep of the cathode tube. The tape itself triggers the electron sweep so that the pattern always occurs in the same part of the screen when the tape recorder is used for input. The sweep is 100mm in length so that time and length are convertible at a rate of 1mm=0.025 sec.

The endless tape was intended to make it possible for the student or trainee to see his own utterance repeated and have the instructor be able to point out the differences between them and the model. But the tape recorder is also useful when doing linguistic analysis because it can be used to center a particular subject of study on the screen for photographing.

Use of the Visualizer in Instruction

The machine was originally intended for remedial practice in pronunciation. In Japanese it is called 発音直視装置, and was intended for those who had hearing defects rendering them incapable of hearing their own voice. The subject is asked to pronounce a certain sound or word into the microphone of the visualizer and while the subject’s voice is being repeated on the screen from the built-in endless tape, the instructor points out the differences in the subject’s pronunciation with his or other
recorded pronunciation produced alternately on the screen or with a photograph. The student then tries to approximate the wave pattern of the model pronunciation with his own pronunciation.

The machine, as it stands, can be used for language instruction in pronunciation and phonetics and remedial pronunciation instruction, especially in an oral production study situation. One of the difficulties with its use in this regard, however, is the necessity of using it with only one student at a time. It may be very useful in classes with a small number of students, such as seminars. But it may also be attached to the control console of the language laboratory so that the individual student could, at the direction of the instructor, put his voice into the visualizer from his place in the LL classroom. The wave pattern can be shown through the TV monitors by using the video camera in the control booth. Thus, there is a possibility of using the visualizer in a large classroom situation to some extent.

It can be foreseen that one difficulty with use in the classroom would be the pronunciation used as a model. Ordinarily, this requires someone with near native speaking ability in the language being studied. Since most instructors are hesitant to use their own pronunciation as models, the visualizer which Sapporo University has acquired as an adjunct to the Language Laboratory was modified so that an ordinary tape recorder can be played into it. Thus, the model voice can be taken from recorded material, or pre-recorded by a native speaker of the desired language.

*Recording the Wave Pattern Permanently*

There are two possible methods of permanently recording wave patterns visually. One is by roll stylus, in which the wave pattern is copied on a sheet of paper as it moves under a stylus which moves vertically to record amplitude in the same way that earthquakes are recorded in seismology. The roll stylus machinery is very expensive in initial outlay, but succeeding costs are rather low. However, the stylus machinery requires rather elaborate preparation and can be considered to lack mobility and ease of handling, and is not as detailed as the actual wave pattern produced on the cathode tube screen.

The other method is the Polaroid camera which is produced specifically to record cathode ray tracings. The camera has a hood which fits against the screen of the cathode tube to shut out other light, and it has a fixed focus which is exactly the distance from the lens to the outer edge of the hood. It has a trigger grip for tripping the shutter. The camera is relatively inexpensive, i.e. about ¥70,000, but succeeding costs for film are rather high, about ¥200 per picture. However, photography is by far the easiest and simplest way of recording wave patterns, and most accurately shows the patterns for imitation by the student. Also, the Polaroid camera, being almost instantaneous in developing the pictures, is the most flexible for use in the classroom situation and for quick analysis in the linguistic studies for which it may be employed.

Other types of cameras may, of course be used, but there are problems of close
up focusing, prevention of light leakage, and most of all delayed development so that each sound or word or phrase can only be analyzed hours or days later.

There are more elaborate cameras and attachment equipment for taking wave patterns of cathode ray tubes or oscilloscope tracings. These, however, are usually intended for engineering and electronic laboratories and offer little advantage over the above mentioned camera for linguistic purposes.

In taking pictures of the cathode ray tracings it is best to remove the orange filter since it tends to blur the image with reflected light.

*Linguistic Analysis Using the Visualizer*

The Visualizer basically records sound amplitude, but in very small segments, so that instantaneous variances are produced on the screen. Thus, stops are depicted as dampening the wave amplitude while vowels are shown as increase in wave height in proportion to their openness.

![Image of some English vowels](image)

"ee" "i" "ei" "æ" "ah" "u" "uw"

Fig. 1. Some English vowels.

Three other factors influence the wave pattern besides openness. One is stress accent, in other words volume increase. This is very necessary in analyzing English accent, and to a great extent, Japanese accent also, since the Visualizer indicates that the tonal accent of Japanese is not entirely divorced from stress.

A second factor is breath sound which produces rather sharp peaks, especially in the case of plosives such as the English 'p' though the effect can be somewhat limited by switching in the wave filter to give more contour than individual wave peaks if so desired. Other unvoiced plosives and fricatives have distinctive patterns, though not usually as peaked as the 'p' sound.

A third factor is tone. The visualizer shows tonal height by the number of sub-waves and their amplitude. The higher the tone, the fewer sub-waves are visible. The sub-waves become commensurate with the main contour wave making it brighter and smoother in texture. Even relatively small changes in tone can be distinguished after a certain amount of study.
Fig. 2. Plosive 'p', without wave filter.

In this phrase there are eight 'p' sounds, two 't' sounds and three 'k' sounds. (One 't' is combined with 'k' in 'picked'). The 'k's are readily visible as the lower peaks. In figure 3 the wave filter was used to diminish the sharp peaks and show contour. Especially noticeable is Fig. 3 are the 'er' sounds in 'Peter' 'Piper' and 'peppers'.

Fig. 3. With wave filter.

Fig. 4. English accent. In this figure, the first phrase has the stress on the 'Eng' of 'English' and in the second on the 'tea' of 'teacher'. If the 'Eng' of the first phrase is compared with that of the second, the first one is obviously quite a bit higher, while the reverse is true for the 'tea' of 'teacher'.

Also, the 'ee' sound in 'English' in the first phrase is broader and somewhat more rounded at the edge, indicating that a higher pitch was used than in the second phrase, while, again, the opposite may be said concerning the 'ee' sound in 'teacher' in the first and second phrases.
Otherwise, the first phrase may be broken down into its components as follows: a small preparatory sound as the voice begins to form the vowel ‘ee’; the sharp distinction between ‘ee’ and ‘η’ and the sharp distinction again between ‘η’ and ‘li’; the rough breath sound produced by ‘sh’; the unvoiced ‘t’ totally separate from the following ‘ee’; the rather large gap between ‘tea’ and ‘cher’; the sharp ‘ch’ sound followed by ‘er’. This final ‘er’ also has a rather serrate edge in the wave pattern indicating a falling or low pitch, and is quite typical of all final ‘er’ syllables.

Fig. 5. Japanese accent. In this figure the difference in accent is quite visible from the amplitude of the waves. However, the slightly darker spot toward the center of the ‘sh’ of 橋 indicates that, in comparison, the pitch of ‘shi’ is higher than that of ‘ha’. Although the main differences between the first an second words can probably be attributed to volume rather than pitch, the Japanese girl who recorded these words was aware that I was looking for the difference in accent between the two words, and probably over emphasized the accent, causing an unnaturally strong stress.

However, the Visualizer does also record differences in tone or pitch.

Fig. 6. Indication of Pitch. This figure shows two sounds of about a fifth apart in the musical scale given by a Japanese girl. The lower pitch sound is on the left.
Fig. 7. Organ Notes. These are the configurations of notes on a small electric organ. The notes are in ascending order from the left. The first is the lowest 'C' on the organ; the next is the 'A' an octave and a half above; next is 'E' above middle 'C' and the final figure is the 'A' above it. The sub-wave patterns progressively disappear as the pitch rises and the smaller waves fill out the full configuration.

The actual frequency in cycles of the tone can be computed by counting the number of peaks in a wave. In order to do this it is better to use the fast sweep of the Visualizer in order to spread out the waves, and then calculate by means of counting the wave peaks. The formula then is:

\[ @ \times 10 + 1.25 = \text{c.p.s.} \]

@ indicates the number of wave peaks over 8mm; the reason for multiplying by 8mm instead of 10mm is that in the photographs the size of the waves are reduced by a factor of 20%. 1.25 is the speed of the electron sweep at fast speed.

For fig. 8, which is the first note of fig. 7 at fast sweep, the result is 72 c.p.s. since there are 9 peaks on the edge over 8mm, in the photograph.\textsuperscript{1)}

Comparative Analysis Using the Visualizer

In comparing pronunciation, accent and intonation, the differences between proper and poor production influenced by another language becomes very obvious.

Fig. 9. In this figure the first was uttered by a Japanese female, the second by an American male. The differences in pitch of voice are immediately discernible. In the
first, the fricative 'k' is followed by a very high peak representing 'ou' and then diminishes to 'm' and opens again to a lesser extent for 'o' and then diminishes to the alveolar stop for the Japanese rolled 'r' and finishes with the closed vowel 'ee'. The second large peak for 'mo' is noticeably lower in pitch than the 'ou' of 'kou'.

The second 'kounori' was spoken by an American male unfamiliar with the word and reading it from Hepburn style Romaji. The fricative 'k' is somewhat larger than in the first word uttered by the Japanese, but the 'ou' is very small by comparison while the 'mo' is very large and perceptibly higher in pitch. There is also a lack of dampening between 'mo' and 'ri' indicating that the open American 'r' was used. Thus the displacement of accent, perhaps due to the tendency to place the accent on the anti-penult in words such as this, is quite evident, in addition to other pronunciation errors.

Besides pronunciation, accent and intonation; rhythm, blendings and elisions can also be analyzed for phrases using the visualizer.

Fig. 10. The first phrase was spoken by an American male and the second by a Japanese female who has spent about a year in the United States and has a rather good command of conversational English.

The first thing that is strikingly different between these two sentences is that they are of different length. Sentence two is 32% longer than sentence one. In actual time they are 2.03 seconds for sentence one and 2.96 seconds for sentence two. (length +10 x 2.5 sec.) Next, it is easily noted that the space between the words for the sentence of the Japanese speaker is notably more than that of the American. The pauses constitute about 0.25 seconds for the Japanese but only 0.06 seconds for the American. In blendings, the 'my' and the 'Eng' of 'English' have been joined, and the 'is' has been contracted to 's' after 'teacher'.

The length of syllables is also quite different in the two sentences. The American speaker did not lengthen every syllable proportionately, but lengthened certain ones in accordance with a certain rhythm. For example, the comparison of the word 'Colorado' as spoken by each speaker shows that, though the total word was given about the same amount of time, the syllables were somewhat different in length.
<table>
<thead>
<tr>
<th></th>
<th>Co</th>
<th>lo</th>
<th>ra</th>
<th>do</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese</td>
<td>0.156</td>
<td>0.125</td>
<td>0.187</td>
<td>0.015</td>
<td>0.171 = 0.654</td>
</tr>
<tr>
<td>American</td>
<td>0.171</td>
<td>0.109</td>
<td>0.203</td>
<td>0.031</td>
<td>0.156 = 0.670</td>
</tr>
</tbody>
</table>

The result of this comparison is that the difference between the 'ra' and 'lo' is much more in the American's utterance than the Japanese' utterance and less in regard to 'Co'. This would indicate that the American's tendency is to lengthen and strengthen accents at the expense of the preceding and proceeding syllables, while this is present to a much lesser extent in the Japanese' utterance. If such studies were carried out over a broad range of utterances in regard to both Japanese and English, the basic differences in accentuation between the two languages could be made much clearer.

There are several other significant factors in these photographs of these utterances. Among them the considerably higher tone given to 'Eng' in the American's utterance, which is recognizable because of the opaqueness of the wave rather than the amplitude, which is small since it has a closed vowel. However, at this point the number possibilities for linguistic studies for the Voice Visualizer must certainly be clear. Rather than give an exhaustive description of them here, which would probably be impossible, I leave it to my colleagues in their various studies to recognize the possible uses. I have only touched slightly on phonetics, almost not at all on prosody in which our colleague, Professor Kumagai has made use of the Visualizer. There are others, I am sure who have some field in which they could make use of the machine.

So, at this point I commend it to your talents, and will be more than happy to demonstrate its use to anyone so desiring.

1) The photographs have been further reduced in printig by about 50% so that this formula would have to be changed to fit the new dimension.