Pitch information displayed on a vibrator matrix as a speech reading aid. Some preliminary results

Spens, K-E.

journal: STL-QPSR
volume: 16
number: 2-3
year: 1975
pages: 034-039

http://www.speech.kth.se/qpsr
II. SPEECH AND HEARING DEFECTS AND AIDS

A. PITCH INFORMATION DISPLAYED ON A VIBRATOR MATRIX AS A SPEECH READING AID. SOME PRELIMINARY RESULTS *

K-E. Spens

Abstract

An experiment with the tactile conveying of intonation as a support to speech-reading is described. In earlier versions of tactile pitch displays, frequency information has been transformed to that of locus by the use of vibrators placed in a row on the arm or the finger (Engelmann & Rosov in Exceptional Children, 1975; Stratton in Volta Review, 1974). The question was if the information could be better decoded by the skin (or actually the brain), if the display was expanded to include a time dimension also. During the experiment a certain time length of the intonation contour was continuously displayed to the subject on a vibrator matrix. Results suggest that the intonation contour has some informative cues which are best conveyed to the sense of touch when the displayed length of the contour is around 90 ms.

* This is a revised version of a paper with the same title presented at The International Congress on Education of the Deaf in Tokyo, August 1975.
Introduction

Experiments on conveying speech to the deaf and deaf-blind persons by transforming the acoustic speech signal into some form of vibratory display have been made during the last 50 years. In 1973 Kirman made an extensive review of the most important work in the field of tactile speech perception.

When conveying information to the sense of touch a very crucial point is, as Kirman (1973) points out, the organization of the stimuli and he concludes that a properly designed tactile display of speech should give the stimuli an organization which suits the ability of the skin to recognize spatio-temporal patterns. This means that the character of the stimuli should resemble those stimuli generated when, for instance, the fingers are exploring an object by touching its surface, edges, and contours.

In our work with tactile aids for the deaf at the Department of Speech Communication in Stockholm the complex problem of a complete tactile speech display is somewhat simplified by our aim of creating a tactile aid which will support the speech-reader. The important question will then be which parameter or parameters extracted from the speech signal should be the best complement to the visual signal.

Many investigators consider the prosodic information of pitch or intonation to be a very important parameter (Blesser 1969; Risberg 1974; Stratton 1974; Svensson 1974) in speech perception. The fact that pitch is a parameter, which is very difficult to speech-read, is reflected in the very bad intonation of most congenitally deaf talkers. Efforts have been made to display prosodic information on single vibrators (Gault 1926 and others), but the results have been poor. Fitch displays of the frequency-to-locus type have given more promising results (Engelmann and Rosov 1975; Stratton 1974). This indicates that the organization of stimuli in a more spatio-temporal manner can, indeed, increase the ability of the skin to extract relevant information. There was a question as to the intonation decoding ability of the skin could be further increased by adding another spatial dimension, i.e. to use a surface of vibrators instead of a single row. There are indications that this is the case. Taenzer (1971) has shown that an increase in the number of columns from one to six on the Optacon (a reading device for the blind displaying the letters on a vibrator matrix) does almost linearly increase the tactual reading rate of
printed letters with the fingertip. The most straightforward spatio-temporal organization of the pitch stimuli is shown in Fig. II-A-1. This is an expansion of the single column frequency-to-locus display with similar columns distributed along a negative time scale. The frequency information on each column is shifted to its left neighbor at certain time intervals, and at the same time new data are entered in the rightmost column. The display will therefore cover a time period which will be referred to as the time window. The entering of new data at too slow a rate, i.e. too long a shift interval, will, of course, reduce the information and a very short shift interval will make the display similar to a single column display as almost the same information is displayed on all the columns.

The aim of the described experiment was to find out if there is an optimal shift interval (time window) which then should suggest that it is worthwhile to further evaluate a pitch display with this type of organization.

**Apparatus**

The vibrator matrix of an Optacon was used as tactile display. The matrix consists of 6x24 vibrators covering an area of 10x20 mm suiting the tip of the index finger. As the apparatus originally was a reading device for blind persons and therefore interfaced to a camera, a special interface was designed to make it possible to display a time-dependent parameter like the pitch. The interface has been reported on in a recent article (Spens 1974). The frequency scale was quantized into the following intervals: 80-100, 100-120, 120-140, 140-160 and over 160 Hz. Each interval was represented by two vibrators making the frequency scale cover a vertical range of 10 mm. Only one interval per column was activated at the same time (see Fig. II-A-1). Each new sample of pitch frequency was introduced on the rightmost column of the display.

Data were displayed at a rate of 250 frames/sec which is equivalent to an interval of 4 ms between successive frames. The arrangement was analogous to a tactile motion picture with 250 frames/sec. The stimuli of one column could be shifted to its left neighbor at any multiple of the frame interval (4 ms). For reasons of technical ease the time window is defined as six times the shift interval, which means that the length of the time window could be any multiple of 6x4 ms = 24 ms.
An intonation contour sampled at different intervals

Vibrator matrix

Time window

short

optimal

too long

neg. time

0

indicate active vibrator

Fig. II-A-1. This figure illustrates how the same intonation contour can be differently displayed on the vibrator matrix by the use of different sampling intervals. The displayed time window = six sample intervals. A short-time window generates an almost horizontal line moving up and down with the frequency of intonation, while a longer time window generates more spatially complex patterns. A too long time window will result in loss of information. This is demonstrated by the poor indication of the peak of the intonation contour.
In order to keep the attention level as high as possible, the subjects got immediate feedback on the accuracy of the response. This introduced some learning into the experiment which was taken into account by the mentioned balancing of order. To prevent auditory cues from the vibrators hearing protectors and masking noise were used.

Results

Seven of the nine subjects show a positive peak at a time window length of 96 ms. A likely explanation of this is that some informative cues about intonation is best conveyed to the sense of touch if the length of the displayed part of the contour is around 90 ms. Many factors may of course affect this optimal length. To mention which and to what extent could only be speculative. The perceptual tactile short-term memory, the average length of a "prosodem" (the minimum information carrying length of the intonation contour), the intonation range and intonation speed of the talker are some of the factors which should be further investigated.

Subject 1 spontaneously expressed that she experienced absolutely no information from the tactile display and that she had concentrated only on speech-reading. Her experience is well supported by the data which do not show any influence from the change of time window length. Also subject 3 seemed to be influenced very slightly by the tactile information.

Conceivably, there might be two different ways to benefit from the tactual display of the intonation contour. One is to use the ability of the skin to perceive rhythms when decoding speech. With a short, displayed portion of the pitch contour, i.e. a short-time window, the rhythm of voiced and voiceless segments will best support the speech decoding. A longer time window will cause integration and masking and the ability to decode rhythms will decrease.

The other is to use the pattern recognition ability of the skin, in this case the ability to recognize and decode the rises, falls and peaks of the intonation, which should be a support to speech-reading. This decoding strategy probably needs a longer displayed portion of the intonation contour to work well and the data could be interpreted as indicating that this strategy works best if the displayed length of the intonation contour is about 90 ms. A further expansion of the time window will reduce the
The pitch extractor consisted of one low-pass and one high-pass filter to condition the speech signal, a frequency-to-voltage converter working on the peak-picking principle, and a smoothing low-pass filter. A special electronic circuit was used to activate one of five inputs of the Optacon interface. These inputs represented the five frequency intervals. The time interval between successive samples, i.e. the time window, could be varied with the interface.

**Stimuli**

The stimuli were four sets of sentences recorded on video-tape. Each set consisted of 15 different sentences presented three times each in randomized order. All sentences were statements of about the same length, on the average five words. The sound signal was recorded on the video-tape with a contact microphone that was used to permit a more correct extraction of the pitch signal. The reader was a male colleague with a typical pitch of 130 Hz and with normal intonation. To decrease the visual information the sentences contained no labial and labio-dental sounds.

The subjects were divided into two groups of four each. Each subject in a group speech-read all the four sets of sentences in different orders. The speech-reading was supported by the tactile pitch display on which the subject held his left hand index finger. The four time windows 24, 48, 96, and 144 ms were used. No combination of time window, sentence set, and order was repeated.

**The subjects**

The subjects were 8+1 colleagues with no or very little experience of speech-reading and with no experience at all of the tactile display.

**Procedure**

The subject was placed 1.5 m from the video-monitor which displayed the head of the reader in about natural size. The left hand index finger was held on the vibrator matrix. The responses were given on a forced choice basis. The subject could choose from a list containing the 15 sentences of the set. Before each session the sentence list was studied by the subject for about five minutes.
Fig. II-A-2. The percent correct responses as a function of the length of the time window. The peaks in the performance at 96 msec indicate that there are some informative cues in the intonation which are at best conveyed to the skin at that length of the time window. The result of group one indicates that there may also be some cues which are at best conveyed at very short time windows. (see the text).
Fig. II-A-3. Individual performance as a function of the length of the time window. All subjects but two show a peak at 96 ms (see the text). The individual data are normalized according to learning and differences of the sentence sets. A ninth subject who by mistake was presented the wrong order of time windows and hence excluded from the balanced group data is also plotted in this figure.
possibility to resolve these structures also (peaks and valleys of the intonation contour) and the data suggest that this happens for time window lengths above 96 ms. Subjects may utilize these two decoding strategies differently which is indicated by the different results of groups 1 and 2 (see Fig. II-A-2).

The results obtained from group 1 may be interpreted as an indication that it relies more on the rhythm extracting strategy although it also uses the pattern recognition strategy.

Data in Fig. II-A-3 show that, except from subjects 1 and 3 who make little or no use of the vibrator, five subjects have their best performance at a time window length of 96 ms and two subjects (no. 2 and 4) do better at 24 ms although both have a peak at 96 ms. The fact that seven subjects show a peak in their performance around 96 ms without any previous training may indicate that this spatio-temporal organization of the tactile intonation stimuli suits the pattern recognition ability of the skin better than does a single row display, which in turn could simplify and speed up the training to make it a support to deaf speech-readers.

References: