Voice analysis in depressed patients: Rate of change of fundamental frequency related to mental state

Askenfelt, A. and Sjölin-Nilsonne, Å.

journal: STL-QPSR
volume: 21
number: 2-3
year: 1980
pages: 071-084

http://www.speech.kth.se/qpsr
B. VOICE ANALYSIS IN DEPRESSED PATIENTS: RATE OF CHANGE OF FUNDAMENTAL FREQUENCY RELATED TO MENTAL STATE

A. Askenfelt and A. Sjölin*

Abstract

Two pilot studies are described. An attempt has been made to find a voice parameter that reflects the voice change in depression. A first pilot study using fundamental frequency histograms showed no difference in the mean pitch of the voices of depressed psychiatric patients as compared to non-depressed controls. There was, however, a tendency towards a greater pitch variation after sleep deprivation in the depressive patients. In the second pilot study the average rate of change of the fundamental frequency of the voice when reading a standard text was measured in three depressed patients during their illness and after recovery. The results suggest that the rate of change of fundamental frequency may be a parameter sensitive to voice change in depression.

Introduction

The non-verbal communication of affective state is a phylogenetically old function upon which faculties of speech and language have been superimposed in man. Non-linguistic auditory cues transmit a great deal of information about emotional state as can be seen by the accuracy with which listeners perceive the emotional tone of speech (Kramer, 1963).

Roessler & Lester (1976) used spectral analysis to study rapid voice changes in a patient during psychotherapy sessions. They found that depressed mood correlated with lowered fundamental frequency ($F_0$) and decreased acoustic power. The authors are careful, however, to point out that results obtained from a single patient do not permit general conclusions. Scherer et al. (1972) tested listener/judges' ability to recognize specific emotions of actors in recordings. All sequential speech variables were eliminated by the use of randomized splicing, low-pass filtering and a combination of both techniques. This procedure leaves short-term amplitude variation, mean amplitude, mean pitch and rate of speech as a basis for

* Karolinska Institute, Department of Psychiatry, St. Göran's Hospital, Stockholm, Sweden
judgement of affect. The judges could identify the emotions expressed with better than chance probability. This suggests that many vocal cues are redundant and that a small set of parameters might suffice to transmit emotional content.

Kotlyas & Morozov (1976) used singers in a study with similar design. Oscillograms were made of the recordings. Sorrow turned out to be the easiest emotion to identify, it was associated with a long syllable duration and slow rise and decay in the syllables.

Williams & Stevens (1972) prepared narrow-band spectrograms of recordings of actors. $F_o$ was measured every 0.15 sec, distribution curves were drawn to determine the median $F_o$ and the range of $F_o$ (10th-90th percentile). The $F_o$ was lower for actors speaking in sorrow situations, the range of $F_o$ was then narrow and there were more pauses and voicing irregularity.

A second experimental approach is used by Scherer (1974). In an attempt to assess the way in which inferences of emotional content are based on specific acoustic cues he manipulated pitch level, pitch variation and tempo in synthesized tone sequences. He found that the combination of slow tempo, low pitch level, descending pitch contours and lack of overtones were reliably perceived as sadness by listeners.

Scherer (1979) has reviewed studies of voice changes in mental depression. He quotes Zuberbier who found similar mean $F_o$ values in depressed patients as in normals but lower intensity, reduced dynamic range and lack of emphatic accents.

It is necessary to define reliable objective parameters correlated with the depressed state in patients, especially for the evaluation of longitudinal therapeutic trials.

In the present study the voice of mentally depressed patients has been analyzed. An attempt has been made to find more revealing voice parameters than those previously used.
Subjects

All subjects in this report (pilot study 1 and pilot study 2) were examined and recorded as part of depression research projects at the Department of Psychiatry, St. Göran's Hospital, Stockholm.

The voice recordings in the pilot study 1 were performed in collaboration with the Department of Phoniatrics, Huddinge Hospital (head: Björn Fritzell). In the first study four patients with mental depression (D1, D2, D3, D4) and four non-depressed controls (C1, C2, C3, C4) were recorded at 2 hr intervals over 24 hrs during sleep deprivation therapy. In the second study three depressed patients were recorded in depressed state as well as after recovery from depression.

The clinical status of the patients was rated according to 22 items from the Comprehensive Psychopathological Rating Scale (CPRS) (Asberg et al, 1978). Two examiners rate 0-3 points for each item, 3 representing the most extreme manifestation. The items are divided into "observed", such as restlessness, slowing of speech - and "reported", such as thoughts of suicide or delusions of guilt. A global assessment is also made - the rater summarizes his evaluation of the degree of illness. 0 represents no signs of depression - 3 is a severely depressed state.

Method and analysis

The same recording and F₀ tracking technique was used in both studies: The recordings were made on a 2-track audio tape, see Fig. IV-B-1. On one track the signal from a small air microphone was recorded. This microphone was attached to the rim of a pair of glasses in order to ensure a constant distance from the mouth. The other track contained the signal from a contact microphone (i.e., accelerometer) fastened to the neck below the larynx.

The first step in the analysis is to detect the F₀ in speech, Fig. IV-B-1. We have used rather simple existing hardware equipment developed for a different voice quality project. When combined with the special recording technique just described, this
equipment, which operates in real time, gives a satisfactory result, as shown by Askenfelt et al (1980). The fundamental period time ($T_0$) is sampled at the end of every pitch period, converted to frequency, and stored in the computer disc memory.

The intensity level of the speech is sampled simultaneously. The signal recorded from the air microphone is used for monitoring purposes only.

![Diagram](image)

**Fig. IV-B-1.** Schematic view of the equipment hardware used for recording and analysis. The $F_0$-detection branch comprises an LP-filter, $F_0$-detector, and a period time counter. The contact microphone facilitates the $F_0$-detection. Fundamental period time ($T_0$), fundamental frequency ($F_0$), and intensity ($I$) are sampled at the end of each pitch period and stored in the computer disc memory.

**Pilot study 1**

This study was undertaken in order to test the usefulness of an $F_0$-histogram (see Fig. IV-B-2), which shows a typical distribution of $F_0$ during speech. The computer calculates the arithmetic mean of the $F_0$-distribution ($MF_0$) and the range of $F_0$-variation in terms...
of the base width of a triangle approximation of the histogram (Fig. IV-B-2). The MFo will correspond to the general pitch level; the base of the triangle will reflect the amount of variation of pitch within the specific speech sample. The voice samples in this study were obtained by letting patients and controls read different stories from a book of fables; as a rule, different fables were read in each recording session.

\[ F_{max} = 191 \text{ Hz} \]

\[ MFo = 183 \text{ Hz} \]

Fig. IV-B-2. \( F_o \)-histogram showing the distribution of \( F_o \) in a standard text. The base width of the triangle approximation of the distribution is used as a measure of the \( F_o \)-range, see text. MFo is the mean of the distribution.

The patients were treated with sleep deprivation therapy and their mental state was rated the day before and after the vigil using the Comprehensive Psychological Rating Scale (CPRS).
Subject D 1 showed marked improvement, D 2 showed higher points in the second rating. This was probably due to anxiety associated with alcohol withdrawal symptoms. D 3 and D 4 showed some improvement with respect to objective and subjective ratings but not in the global rating (Table IV-B-I).

An examination of the histogram parameters revealed no differences between the controls and the depressed patients as regards pitch level or range over 24 hours. However, a clear increase of the pitch variability measure could be seen in the later recordings in one member of the depressed group, in another there was a trend in the same direction. This led us to try subdividing the recordings into two groups, one comprising recordings 1-7 and the other nos. 8-13. The mean values of the pitch variability as measured by the base of the triangle were compared (see Table IV-B-I). Δ represents the difference in mean value between the first and second groups of recordings.

Table IV-B-I. Average pitch variability (MV) as estimated from the base width of a triangle approximation of F0-histograms ± SD for the first seven recordings as compared to recordings nos. eight to thirteen. Δ represents the difference in mean value between the first and the second groups of recordings. CPRS = the Comprehensive Psychopathological Rating Scale, Obj = objective items, Subj = subjective items. The scores represent the mean value of the points given by two raters.

<table>
<thead>
<tr>
<th></th>
<th>MV 1-7 (cent) ± SD</th>
<th>MV 8-13 (cent) ± SD</th>
<th>Δ</th>
<th>CPRS Score Before SDT subj obj global</th>
<th>After SDT subj obj global</th>
</tr>
</thead>
<tbody>
<tr>
<td>D 1</td>
<td>979 ± 103</td>
<td>998 ± 57 19</td>
<td></td>
<td>1.5 0.75 1.5</td>
<td>0 0 0</td>
</tr>
<tr>
<td>D 2</td>
<td>983 ± 62</td>
<td>1271 ± 177 288</td>
<td></td>
<td>1 1 2 1.75</td>
<td>2</td>
</tr>
<tr>
<td>D 3</td>
<td>111 ± 173</td>
<td>1386 ± 208 275</td>
<td></td>
<td>1 1 1 0.25 0.25</td>
<td>1</td>
</tr>
<tr>
<td>D 4</td>
<td>692 ± 26</td>
<td>714 ± 120 22</td>
<td></td>
<td>1.5 1.25 2</td>
<td>0.75 0.5 2</td>
</tr>
<tr>
<td>C 1</td>
<td>1113 ± 178</td>
<td>1245 ± 83 132</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 2</td>
<td>1595 ± 39</td>
<td>1347 ± 234 -208</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 3</td>
<td>765 ± 47</td>
<td>755 ± 32 -10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 4</td>
<td>1259 ± 156</td>
<td>1327 ± 224 68</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pilot study 2

In the second pilot study we investigated the average rate of change of \( F_0 \) during reading to test if this would be a more revealing measure of \( F_0 \)-variability. We also decided to use patients as their own controls by comparing recordings made during depression and after recovery.

Each recording contained an initial period of free speech and a reading of a standard text.

Analysis

Before data can be computed from the stored \( F_0 \)-samples, special precautions must be taken. This is due to the fact that the quantity we wanted to compute is the rate of change in \( F_0 \) which is extremely sensitive to errors in the measurements underlying the \( F_0 \)-curve. For this reason all discontinuities in the \( F_0 \)-curves must be eliminated by a complex filtering procedure. An example of this cleaning process can be seen in Fig. IV-B-3.

The curve shown in Fig. IV-B-3a represents the original \( F_0 \)-recording vs time. The curve includes all parts of the speech - voiced as well as voiceless and also pauses. The curve in Fig. IV-B-3e represents the intensity of the speech in dB.

The first step in the "cleaning" procedure is to eliminate the measuring errors in the voiced parts of the speech sample. This is done by applying two different digital filters: first, a running median filter which eliminates sharp discontinuities shorter than a given duration, and second, a low pass filter which rounds off remaining sharp corners. The result of these filtering procedures can be seen in Fig. IV-B-3b and IV-B-3c, respectively.

Finally, the voiceless parts of the recording are eliminated by introducing a threshold criterion in the intensity curve, as shown in Fig. IV-B-3e. The parts of the recording with an intensity below the threshold are omitted in the following computations. This is represented graphically by setting the \( F_0 \) of these parts to a very low value (50 Hz). The final result of the cleaning process can be studied in Fig. IV-B-3d.
Fig. IV-B-3. Effects of different types of filtering.
(a) Original $F_0$-recording.
(b) The signal in (a) after a 9-point median filter.
(c) The signal in (b) after a 9-point low pass filter.
(d) The signal in (c) after elimination of the parts in the speech with an intensity below a threshold, see text. The $F_0$-value of the eliminated parts is set to 50 Hz.
(e) Intensity of the speech. The threshold used in (d) is indicated with a dotted line.

The recorded data have now been reduced to piece-wise continuous smooth parts of the $F_0$-curve, as schematically indicated in Fig. IV-B-4a. Using the symbols defined in Fig. IV-B-4b the Rate of Change of Fundamental Frequency (RCF) at a certain point on the curve is now computed as:
\[ \text{RCFF} = \frac{\frac{1}{w} |\Delta F_0|}{\bar{F}_0} \cdot 100 \]

where \( w \) is an appropriate time window in ms and \( \bar{F}_0 \) is the average \( F_0 \) within this window. The RCFF-formula is very close to an approximation of the time derivative at a point on the \( F_0 \)-curve:

\[ \frac{d F_0}{dt} \approx \frac{\Delta F_0}{w} \]

The change in \( F_0 (\Delta F_0) \) during the time interval \( w \) is normalized by dividing by \( \bar{F}_0 \). The "unit" of the RCFF is consequently percent change in \( F_0 \) per ms. This makes the RCFF independent of the speaker's average pitch and this serves equally well for both males and females. No attention is paid to the sign of \( \Delta F_0 \). RCFF indicates the rate of change of \( F_0 \) only, not the direction.

**Fig. IV-B-4.** (a) Schematic figure of a filtered \( F_0 \)-curve. The encircled part is shown enlarged in (b). (b) Explanation of symbols used in the RCFF-computation.
The $F_o$-contour and its associated RCFF curve of a sentence are shown in Fig. IV-B-5. The principal property of the conception of the derivative is easily seen: a sequence with a high rate of change in $F_o$ shows a high value in the RCFF-curve. When RCFF has been determined, $w$ is moved in time by one $T_o$ and a new RCFF is computed. In order to obtain one single measure representative of the average rate of pitch change, the RCFF for all time windows are averaged. This cannot be done without taking special precautions. First, in spite of the preceding filtering of the $F_o$-curve there might remain some errors in the $F_o$-data which will manifest themselves as abnormally high values in RCFF. Second, even a very steadily phonated vowel will show a minute fluctuation in fundamental frequency and thus produce a small RCFF-value. Both these phenomena can mask important differences in pitch behavior between patients.
Consequently, RCFF-values over a maxlevel (DERMAX) and below a minlevel (DERMIN) are omitted before computing the RCFF-value for the entire speech samples.

The summarized measure of the pitch variations in a recorded speech sample - henceforth Average Rate of Change of Fundamental Frequency (ARCFF) is thus computed as:

$$\text{ARCFF} = \frac{1}{M} \sum_{i} \text{RCFF}(M)$$

for all DERMIN < RCFF(M) < DERMAX. M is the number of time windows in the recorded text. The numerical values used in this investigation were:

$$M \approx 3000$$
$$w = 50 \text{ msec}$$

The highest RCFF a voice can normally reach is about 1%/ms (Sundberg, 1979). It should be noted that the ARCFF is very sensitive to the values of DERMAX and DERMIN. In addition to the ARCFF an $F_0$-histogram was computed, as described earlier. As a measure of $F_0$-variation the Standard Deviation (SD) was now used instead of the base width used in the first pilot study. In order to obtain a measure of the rate of speech, the total time ($T_{tot}$) used for reading the standard text was measured by the computer.

Our hypothesis is that ARCFF, computed as described, would be one of the voice parameters that varies with the emotional state of the speaker.

**Results**

<table>
<thead>
<tr>
<th>PATIENTS</th>
<th>RECORDING</th>
<th>CPRS global score</th>
<th>HISTOGRAM</th>
<th>$T_{tot}$</th>
<th>PAUSES</th>
<th>ARCFF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$F_0$</td>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Hz</td>
<td>Hz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>183</td>
<td>17</td>
<td>36</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
<td>174</td>
<td>32</td>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>117</td>
<td>16</td>
<td>40</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>114</td>
<td>17</td>
<td>40</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>140</td>
<td>13</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0-1</td>
<td>135</td>
<td>17</td>
<td>25</td>
<td>22</td>
</tr>
</tbody>
</table>
Of the three patients as yet studied, patient no. 1 - a 73 year old woman with bipolar affective disease, was the only one to show greater variation of $F_o$, as expressed by the histogram SD, faster reading time, and fewer long pauses after recovery (Table IV-B-II). No. 2 - a 42 year old man with an acute depression showed complete clinical recovery but none of the above mentioned voice changes. Essentially, the same is true for patient no. 3 - a 44 year old woman, however, her rapid reading accelerated after recovery. With respect to the ARCFF, however, we found that all these three patients showed an increase as they improved clinically. Thus, it is possible that ARCFF may be a sensitive objective parameter of voice change in depression, uninfluenced by the rate of speech.

Discussion and conclusions

1) Choice of speech sample

In order to be able to compare reading time and to eliminate differences due to intonation patterns associated with the linguistic content of the text, we chose to use a standard text for the tests in Pilot Study II. The emotional content of the text could influence the manner in which it is read (Kitzing, 1979).

The reactions of the patients to reading the same text on many occasions differed. The severely depressed were unable to concentrate on the meaning of the text. There was no facilitation or other learning effect until they had recovered sufficiently to make sense of the story.

Further information may be collected by comparing the ARCFF of the free speech with that of the standard text.

2) Choice of parameters

The depressed patient often shows psychomotor inhibition which also affects breathing and articulation. This is thought to be one of the factors leading to a slowing of the speech rate. The connection between the rate of speech and the general degree of motor deceleration is not known. There could be additional factors that would influence the speech rate.
In order to obtain a measure of the F0-variation we have tried using the ARCFIF which allows us to register the mean speed with which a person varies his F0. In all three patients studied a covariance between ARCFIF and the degree of depression was observed. From the point of view of correlation to clinical state, the ARCFIF seems to be a promising voice parameter.

In the present study pause measurements were also included. Greden et al (1980) suggest that pause distribution may be of diagnostic value in different types of depression. In addition, there are other voice parameters that might prove rewarding to examine, for example, a measure of the distribution of the RCFF values, intonation and intensity patterns.

Acknowledgments

The authors are indebted to Johan Sundberg and Lennart Wetterberg for scientific advice and encouragement. The work was supported by grants from HSFR, NFR, FRN to Johan Sundberg, and the Swedish Medical Research Council, grant no. 3371, to Lennart Wetterberg.

References


