Vocal indicators of psychiatric treatment effects in depressives and schizophrenics

TOLKMITT, F., et al.

Abstract

Voice and speech changes as a result of clinical treatment for 17 depressive and 15 schizophrenic patients were investigated. Speech samples taken at interviews before and after treatment were analyzed with regard to f0, spectral energy distribution, and formant frequencies of vowels occurring in identical phonetical context. Both groups of patients showed a decrease in f0 after therapy, which was interpreted as a reduction of general arousal. Differential results with regard to spectral energy distribution suggested that the voice of depressives became more relaxed after therapy, but the opposite seemed to be true for schizophrenics. Significant formant changes obtained for the first formants of two vowels were interpreted as differences in the precision of articulation before and after therapy.

Reference


DOI: 10.1016/0021-9924(82)90034-X
VOCAL INDICATORS OF PSYCHIATRIC TREATMENT EFFECTS IN DEPRESSIVES AND SCHIZOPHRENICS

F. TOLKMITT, H. HELFRICH, R. STANDKE, and K.R. SCHERER
University of Giessen, W. Germany

Voice and speech changes as a result of clinical treatment for 17 depressive and 15 schizophrenic patients were investigated. Speech samples taken at interviews before and after treatment were analyzed with regard to $f_0$, spectral energy distribution, and formant frequencies of vowels occurring in identical phonetical context. Both groups of patients showed a decrease in $f_0$ after therapy, which was interpreted as a reduction of general arousal. Differential results with regard to spectral energy distribution suggested that the voice of depressives became more relaxed after therapy, but the opposite seemed to be true for schizophrenics. Significant formant changes obtained for the first formants of two vowels were interpreted as differences in the precision of articulation before and after therapy.

Introduction

It has often been suggested that the voice of mentally and emotionally disturbed persons differs from that of normals (Moses, 1954; Ostwald, 1963). Investigations of the differences in nonlinguistic features of speech and voice between disturbed and normal persons are hampered by the relative scarcity of well-defined, objectively measurable acoustical features that can be expected to covary with changes in affective states. The few relevant studies widely scattered across different disciplines have recently been summarized by Darby and Sherk (1979) and Scherer (1979). Those reviews show that voice and speech patterns seem to reliably accompany psychopathological symptoms although the evidence is still weak and the findings are rarely replicated. The authors suggest that inconsistencies among the results in some of the studies may be due to differences in research methodology as well as the lack of clearly defined symptoms and syndromes for many clinical categories. Among the possible measures of paralinguistic speech characteristics, fundamental frequency ($f_0$) and energy distribution in the spectrum of the voice seem to be most promising. Many studies report significant differences in level and variation of $f_0$ between disturbed patients and normals although the direction of the differences is not always consistent and may depend on such factors as symptom, sex, type of voice sample (free speech vs reading), and treatment.
Scherer (1979) pointed out that there are only very few studies on voice and speech changes as a result of treatment even though the continuous monitoring of such changes could provide important cues for patients' improvement during the course of therapy. Except for a few studies in which voice parameters before and after electroconvulsive shock were compared (Ostwald, 1963; Darby and Hollien, 1977), speech behavior of disturbed patients is mostly compared with that of normals. Given the many problems concerning sampling and performance comparability of clinical and control groups in such studies, the diagnostic validity of various nonlinguistic features might be more appropriately studied in patients during treatment from "abnormal" to "normal" behavior, using each patient with his or her idiosyncratic speech patterns as his or her own control.

This paper reports such an investigation of voice changes as a result of treatment in two clinical groups. Speech samples of hospitalized female patients, schizophrenics and depressives, were obtained before and after therapy. These speech samples were subjected to acoustic analyses. Dependent measures were $f_0$ and energy values for one-third-octave bands in the long-term voice spectrum as well as formant frequencies of selected vowels. Formant frequencies, though not yet investigated with regard to possible differences in the speech of disturbed and normal persons thus far, were included in the present analysis since they are known to be useful parameters for speaker identification (Goldstein, 1976). Since the extraction of $f_0$ and spectral data, on the one hand, and formant frequencies, on the other, requires different procedural and analytical treatment, the respective method and result sections are presented separately.

Analysis of $f_0$ and Long-Term Spectra

Method

Patients and speech samples. Ekman and Friesen provided the materials and the overall design of the study. In order to study body movement and facial expression in relation to emotion, mood, and psychopathology, they obtained sound motion picture film records of a standardized interview conducted at the time of admission and near the time of discharge from a mental hospital with patients diagnosed as depressive and schizophrenic (cf. Ekman and Friesen, 1974). Further details on the interview schedule, diagnoses, types of therapy, self-rating data, and findings on body movement can be found in Ekman and Friesen, 1968, 1974, and Kiritz, 1971.

Design. The independent variables consisted of the between-subject factor "Syndrome" (S), with the two levels Schizophrenic Syndrome (SS) versus Depressive Syndrome (DS), and the within-subject factor "Interview" (I), with
the two levels Admission Interview (AI) versus Discharge Interview (DI). The dependent variables were fundamental frequency ($f_0$) and half-octave band energy values from long-term spectra.

In order to assess whether $f_0$ or long-term spectra, or both, are indicative of a patient's affect state, it is necessary to obtain a sample of a patient's utterances during a speech situation which is representative of his or her characteristic enduring affect or mood state. It was assumed that the first few minutes of each of the recorded interviews came closest to fulfilling this requirement since later parts might have been affected by the interaction with the clinician and the topics discussed. It was therefore decided to use only those of the patients' vocalizations which occurred during the first 4 min of the interview. After excluding patients whose tape-recorded interviews contained background noise or other quality defects hampering acoustic analysis during the first 4 min, speech samples for 11 schizophrenic and 9 depressive patients were analyzed. Their utterances varied between 10 sec and 3 min, with an average duration of 1 min.

**Acoustic analysis.** The acoustic analyses were performed on the Giessen Speech Analysis System, a PDP 11/35 based system for digital speech analysis. All of the patients' vocalizations during the first 4 min of an interview were digitized with a sampling rate of 16,384 Hz, using an anti-aliasing filter cut-off frequency of 6 KHz and stored on digital tape as data blocks of 62.5 msec each. Silent pauses were eliminated using a threshold detection program (Helfrich, 1980). By means of the autocorrelation method, $f_0$ was extracted for each data block and then averaged. Power spectra were calculated for successive data blocks using hardware fast Fourier transform (FFT) processing. The mean of these spectra yielded the long-term spectrum that was partitioned into 13 one-third octave bands ranging from 260 Hz (to exclude the region of $f_0$ energy variation) to 5760 Hz. Since the gain setting during recording was not controlled, the energy values of all 15 bands were normalized to allow for comparison between patients and interviews.

To check on the accuracy of $f_0$ extraction using fixed-length windows of 62.5 msec, the start of the respective data blocks was shifted by a constant factor in a reliability analysis, and $f_0$ values for the original and the shifted blocks were correlated over the whole speech sample. Correlation coefficients ranged from 0.96 to 0.98, and it seems safe to assume that the $f_0$ of the speech waves was extracted with reasonable accuracy.

The reliability of the long-term spectra was tested by means of an odd–even method. For each patient and interview two separate long-term spectra were computed; one based only on the odd data blocks, the other only on the even. The energy difference between corresponding one-third octave bands proved to be very small (less than 7% deviation).
Results

Fundamental frequency

Mean $f_0$ values as a function of the two independent factors, syndrome and interview, are shown in Figure 1. A two-way analysis of variance with repeated measures for the interview factor yielded a main effect for interview at the $p = 0.06$ confidence level (two-tailed, $F = 4.01$). Apparently, therapy or the fact that the patients were hospitalized caused $f_0$ to decrease for both clinical groups, with a larger decrement for the group of depressives.

Third-octave bands in the long-term spectrum

The relative spectral energy of the 13 one-third octave bands for all four experimental conditions is given in Figure 2. The proportion of spectral energy below 500 Hz (bands 1 and 2) decreased for schizophrenic patients from AI to DI, but increased for depressive patients. The opposite seems true for both groups in the frequency range between 500 and 1000 Hz (3rd, 4th, and 5th

![Figure 1](image-url)  
Fig. 1. Mean $f_0$ values as a function of the interview factor with syndrome as a parameter (△ schizophrenics, • depressives).
Fig. 2. The relative spectral energy of the 13 one-third octave bands for the four experimental conditions (schizophrenic: Δ before therapy, ▲ after therapy; depressive: ○ before therapy, ● after therapy).
This would be expected since most of the energy is located in the region below 1000 Hz and since spectra had been normalized. In order to test for significance, separate analyses of variance were performed on the LFR (low frequency ratio, i.e., relative energy below 500 Hz) and MFR (medium frequency ratio, between 500 and 1000 Hz). For both analyses (which should yield basically similar results since LFR and MFR are not entirely independent) only the interactions between syndrome and interview turned out to be significant at $p < 0.05$ (Fig. 3). For LFR the interaction was caused by a significant energy difference between schizophrenics and depressives at DI (Fisher's LSD [least significant difference]—test; Fryer, 1966) and the energy increment for depressives from AI to DI came close to significance. In the case of MFR, the interaction was caused by the energy decrement from AI to DI for the depressives (Fisher's LSD-test); the energy difference between both groups at AI fell slightly short of significance.

Analysis of Formant Frequencies

**Method**

**Material.** At the end of each interview the patients had to participate in an affect posing session (also recorded) that required them to repeat the following sentences three times with as much appropriate feeling as possible:

- I feel angry
- I feel ashamed
- I feel afraid
- I feel disgusted
- I feel sad
- I feel happy.

The vowels of the adjectives were well suited for formant analysis for they occurred within the same linguistic context three times before and after therapy. Since vowels at the beginning or end of a word are often very short, only the embedded vowels were selected for the formant analyses. The exception was the vowel “y” in “happy” which seemed to be long enough to guarantee adequate analyses. The selected vowels were therefore /i/ in “happy,” /æ/ in “ashamed,” /æ/ in “sad” and “happy,” /æ/ in “disgusted,” and /ei/ in “afraid.” The vowel segments from those patients who had repeated the stimulus sentences three times during AI and DI were analyzed.

**Design.** As the number of schizophrenic and depressive patients available for
The relative spectral energies below 500 Hz (LFR, left) and between 500 and 1000 Hz (MFR, right) for the four experimental conditions (LFR, low frequency ratio [240–440 Hz]; MFR, medium frequency ratio [440–920 Hz]; AI, admission interview; DI, discharge interview).

Analysis varied considerably from phoneme to phoneme, it was not possible to use syndrome as an independent variable as in the earlier analysis. Instead, for each phoneme all patients were treated as one group and independent factors were the two repeated measurement interviews (AI and DI) and repetition of the utterances (three times). Dependent variables were the formant frequencies which were extracted by means of linear prediction routines (Markel and Gray, 1976).

Procedure. The adjectives were transferred from analog tape onto digital tape with a 16,384 Hz sampling frequency and a 6 Hz anti-aliasing filter cut-off frequency. Using GISYS, the Giessen System developed for speech editing purposes, the time signal of each digitized adjective was displayed on a graphics display screen and the stationary portion of the critical phoneme was segmented using a movable cursor and separately stored for analysis. In some cases the phoneme was too short (< 0.086 sec) to allow the patient's data to be used for
that particular phoneme analysis. The number of patients remaining for the analyses of the five phonemes for AI and DI was as follows:

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Remaining Patients</th>
<th>Total Segments</th>
</tr>
</thead>
<tbody>
<tr>
<td>/i/ (happy) 17 SS (× six repetitions)</td>
<td>= 102 segments</td>
<td></td>
</tr>
<tr>
<td>/e/ (ashamed) 21 SS</td>
<td>= 126 segments</td>
<td></td>
</tr>
<tr>
<td>/æ/ (sad) 20 SS</td>
<td>= 120 segments</td>
<td></td>
</tr>
<tr>
<td>/æ/ (happy) 16 SS</td>
<td>= 96 segments</td>
<td></td>
</tr>
<tr>
<td>/ɑ/ (disgusted) 20 SS</td>
<td>= 120 segments</td>
<td></td>
</tr>
<tr>
<td>/ei/ (afraid) 21 SS</td>
<td>= 126 segments</td>
<td></td>
</tr>
</tbody>
</table>

The formant frequencies were extracted for each of the 696 phoneme segments by means of formant extraction routines based on the linear prediction model of the Interactive Laboratory System (ILS) developed at the Speech Communication Research Laboratory, California. In the course of the formant analysis the first ten frames were analyzed for each vowel, each frame consisting of 256 data points. As adjacent frames always overlapped by 128 data points, a total of 1408 data points resulted for the ten frames, which is equivalent to 0.086 sec. (1408 data points/16,384 Hz). The parameters of the model were set to extract a maximum of eight formants, but only the first three were used for analysis since higher ones could not be extracted reliably. As an example, the sentence "I feel sad" is presented in Figure 4: In (a) the time signal is shown; in (b) a segment of the vowel /æ/ is displayed; and in (c) the tracts for the first formants are plotted.

Results

Formant extraction for each phoneme was performed at the frame level, and ideally the ten frame values for each of the three formants should have been averaged to obtain the mean formant frequencies. However, the linear prediction model did not always succeed in extracting the first three formants, and quite frequently one or two of them were missing. To identify the obtained frequencies, these formants had to be compared with formant criteria. The criterion values for a phoneme were obtained by mapping all extracted frequencies into a distribution with 100 Hz intervals. The peak values that corresponded with spectral peaks of the respective synthetic vowel (Fant, 1968) were taken as the formant criteria for that phoneme (Table 1). A bandwidth of 600 Hz around the formant criteria was chosen and only those frequencies of a given phoneme that fell within these tolerance limits were accepted as actual formants. In the case of the first formant, all values below 240 Hz were eliminated as they could have

1Although the second vowels in "ashamed" and "afraid" sound similar, the formant frequencies obtained justify the different phonetic transcriptions used.
Fig. 4. Various graphical representations of the sentence “I feel sad.” (a) The time signal; (b) a segment of the vowel /æ/ from “sad”; and (c) the tracts of the first four formants of /æ/.
TABLE 1
Formant Criteria for F1, F2, and F3 That Were Used in This Study

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>350</td>
<td>2400</td>
<td>3000</td>
</tr>
<tr>
<td>e</td>
<td>500</td>
<td>2300</td>
<td>2950</td>
</tr>
<tr>
<td>æ1</td>
<td>750</td>
<td>2000</td>
<td>2850</td>
</tr>
<tr>
<td>æ2</td>
<td>850</td>
<td>1900</td>
<td>2750</td>
</tr>
<tr>
<td>A</td>
<td>650</td>
<td>1750</td>
<td>2650</td>
</tr>
<tr>
<td>ei</td>
<td>600</td>
<td>1900</td>
<td>2550</td>
</tr>
</tbody>
</table>

been estimates of the fundamental frequency. To guarantee reliable formant estimates, only those formant averages that were based on at least five frame values were entered into the analyses. These selection procedures further reduced the available data.

Since the variation of formant structure within the three repetitions at AI and DI were not of experimental interest and could be considered to be nonsystematic, the data were pooled over this factor, leaving only syndrome as an independent variable. The means of the first three formants of the five phonemes at AI and DI are listed in Table 2. The number of subjects that each mean is based on varies and is given in column N.

All formant changes of more than 20 Hz (column Δ in Table 2), except for /i/, were of an increasing nature; that is, the formant frequencies increased from AI to DI. In order to test for the significance of these changes, separate analyses of variance were performed for each vowel. Only the first formant change of the phonemes /A/ and /ei/ turned out to be significant at $p < 0.05$. Apparently, intrasubject variability as well as the small number of subjects, especially in the case of F1 for /e/, prevented even more sizeable formant differences from turning out to be significant.

Discussion

This study was aimed at investigating whether or not psychiatric treatment leads to changes in certain nonlinguistic voice parameters. Analyses of $f_0$, LFR, and MFR before and after therapy yielded some interesting leads.

A main effect for the factor "Interview" indicated a decrease of $f_0$ after therapy ($p = 0.06$). Assuming that a high $f_0$ might be indicative of higher tension in the vocal cords which in turn might be due to higher muscle tone, it could be argued that by the end of therapy patients were more relaxed as a result of the effect of treatment. This is consistent with earlier research showing a fairly strong relationship between $f_0$ and general arousal (cf. Scherer, 1979). However,
<table>
<thead>
<tr>
<th>Phoneme</th>
<th>FI</th>
<th>DI</th>
<th>Δ</th>
<th>N</th>
<th>FI</th>
<th>DI</th>
<th>Δ</th>
<th>N</th>
<th>FI</th>
<th>DI</th>
<th>Δ</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>happy /i/</td>
<td>400.2</td>
<td>370.3</td>
<td>-29.9</td>
<td>5</td>
<td>2436.4</td>
<td>2444.6</td>
<td>8.2</td>
<td>9</td>
<td>2895.9</td>
<td>2838.6</td>
<td>-57.3</td>
<td>10</td>
</tr>
<tr>
<td>ashamed /e/</td>
<td>499.6</td>
<td>556.7</td>
<td>57.1</td>
<td>6</td>
<td>2248.7</td>
<td>2266.8</td>
<td>18.1</td>
<td>15</td>
<td>2896.3</td>
<td>2938.6</td>
<td>42.3</td>
<td>14</td>
</tr>
<tr>
<td>sad /æ,ı/</td>
<td>702.1</td>
<td>690.7</td>
<td>-11.4</td>
<td>10</td>
<td>1959.6</td>
<td>1981.6</td>
<td>22.0</td>
<td>13</td>
<td>2812.8</td>
<td>2810.9</td>
<td>-1.9</td>
<td>14</td>
</tr>
<tr>
<td>happy /e,ı/</td>
<td>867.8</td>
<td>857.6</td>
<td>-10.2</td>
<td>9</td>
<td>1848.6</td>
<td>1887.2</td>
<td>38.6</td>
<td>15</td>
<td>2739.8</td>
<td>2729.2</td>
<td>-10.6</td>
<td>10</td>
</tr>
<tr>
<td>disgusted /ʌ/</td>
<td>601.1</td>
<td>634.0</td>
<td>32.9</td>
<td>16</td>
<td>1744.8</td>
<td>1744.0</td>
<td>-0.8</td>
<td>18</td>
<td>2650.1</td>
<td>2642.6</td>
<td>-7.5</td>
<td>13</td>
</tr>
<tr>
<td>afraid /ei/</td>
<td>540.0</td>
<td>573.4</td>
<td>33.4</td>
<td>18</td>
<td>1875.9</td>
<td>1882.7</td>
<td>6.8</td>
<td>16</td>
<td>2592.6</td>
<td>2576.7</td>
<td>-15.9</td>
<td>2</td>
</tr>
</tbody>
</table>

AI, admission interview; DI, discharge interview; Δ, difference between AI + DI; N, number of subjects.
alternative explanations are possible. The initial admission interview could have caused higher anxiety than the discharge interview, at which time all patients were well acquainted with the clinical environment and the interview procedures.

Differential results with regard to the two groups were obtained for energy distribution. The proportion of spectral energy below 500 Hz, the low frequency ratio (LFR), decreased after therapy for schizophrenics, but increased for depressives. The reverse was true for the proportion of energy between 500 and 1000 Hz, the medium frequency ratio (MFR). This is consistent with Laver's (1975) distinction between tense and lax voices (the former sounding metallic or sharp; the latter, mellow or muffled). He argues that tense voices have an energy concentration between 500 and 1000 Hz and lax voices have a relatively greater proportion of total energy below that frequency region. Evidence for such an energy shift during states of high tension or affect has been reported by various researchers. Roessler and Lester (1979) demonstrated such an energy shift when they compared the spectral energy distribution of a female voice in a low and high affective state. Hecker et al. (1968) noted an increase in the amount of high-frequency energy in the glottal pulses when under stress. Support comes also from aviation research: Popov et al. (1971) related increasing emotional stress as measured by heart rate to augmentation of the centroid of spectral energy between 300 and 1200 Hz. Furthermore, Frøkjær-Jensen and Prytz (1976) have found a comparison between lower and higher parts of the spectral energy distribution useful in distinguishing pathological qualities. Presumably this energy shift is of laryngeal origin and is supposed to be caused by the damping characteristics of the ventricular folds (Pepinsky, 1942; Van den Berg, 1955). During normal (lax) voice the ventricle acts as a low-pass filter suppressing energies above 500 Hz. When under muscular tension the ventricular folds disappear and the higher harmonics are not filtered, giving the voice a harsh and tense quality. For the present results this may imply that the voices of depressive patients became more lax and possibly more "normal" after therapy. Since states of high anxiety, which occur frequently during acute phases of depression, are characterized by abnormally high muscular tension (Goldstein, 1965; Malmo, 1975), anxiety reduction because of therapeutic improvement may have led to a general decrease in muscle tone with corresponding changes in voice production.

Analyses of vowel formant structures resulted in a significant frequency increment from AI to DI for the first formant of the vowels /A/ and /ei/. A comparison of the formant frequencies of these two vowels with those of the remaining vowels shows that the former correspond more closely to the neutral formant frequencies 500, 1500, and 2500 Hz. These formants are generated when the vocal tract is in the resting position. In this position the shape of the vocal tract is similar to a cylinder of about 17 cm in length and 4 cm in diameter (Minifie, 1973). Fant (1968) uses the symbol /E/ for the neutral vowel. It could
be argued that /ʌ/ and /ei/ were pronounced with less articulatory effort before therapy than after because their first formants were closer to 500 Hz at AI than at DI. Inspecting the first formants of /e/ and /i/ one can notice a similar tendency. The respective differences between AI and DI are probably not significant because of the small number of subjects these means are based on.

Although there is no obvious explanation for these results, it is possible to speculate about the underlying processes. Obviously the utterance of a neutral vowel requires the least articulatory effort compared with all other vowels. Hecker et al. (1968) mentioned that under stress, subjects have a tendency to speak with a less constricted vocal tract which basically means that their articulatory effort becomes smaller. There is also evidence in the literature that depressive patients tend to speak with lax articulation (Newman and Mather, 1938). Since negligent articulation seems to coincide with a less constricted vocal tract, this could be the reason why vowel formants before therapy are closer to the formant frequencies of the neutral vowel than after therapy. It could then be argued that as a result of therapy the patients exert greater articulatory efforts such that vocal tract constrictions increase and, consequently, vowel formants reach the expected values.

Two main results were obtained when comparing the voice of clinical patients before and after therapy. A spectral energy shift occurred that was partly dependent on the patients syndrome, and a tendency before therapy was observed to pronounce vowels in such a way that their formants were close to those of the neutral vowel. The first phenomenon seems to be supported by similar results in the literature, however, the second one has not been reported so far. The proposed mechanism for the latter phenomenon is highly speculative but provides interesting leads for further research.

The authors gratefully acknowledge the support of Paul Ekman and Wallace Friesen who supplied the audiotapes of the interviews that were analyzed in this study.

References


