Presented at the *3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications* (MAVEBA) December, 2003, Firenze, Italy. To appear in the conference Proceedings.

AUTOMATIC ANALYSIS OF VOCAL MANIFESTATIONS OF APPARENT MOOD OR AFFECT

E. P. Rosenfeld[1], D. W. Massaro[2], J. Bernstein[1]
[1]Ordinate Corporation, Menlo Park, California, USA
[2]Department of Psychology, University of California at Santa Cruz, USA

Abstract: **Skilled clinicians are able to integrate linguistic, paralinguistic, and non-linguistic cues in the assessment of mood disorders. This project identified duration- and amplitude-based aspects of the speech signal that can be measured automatically by computer and which provide paralinguistic information about the apparent affect of a speech sample. A group of 40 experimental subjects produced 1440 spoken renditions of sentences, in 3 conditions, uninstructed, depressive, or manic. An automatic speech recognition system extracted 10 paralinguistic parameter values from each of these spoken responses. Psychotherapists have a relatively uniform model of depressive and manic speech patterns, which shows up in distinct paralinguistic features of their speech when simulating these states. Several features are significantly different in the three simulated emotional states and these features can be detected automatically.**

Keywords: automatic, speech recognition, mood, affect.

I. INTRODUCTION

Skilled clinicians are able to integrate linguistic and non-linguistic cues in the assessment of mood disorders. This ability is part of what makes a skilled clinical interview the preferred method of assessment for mood disorders. Among all the non-linguistic aspects of a patient's behavior, non-linguistic aspects of speech may be the easiest to record and analyze. These paralinguistic aspects of the manner of speaking can be collected unobtrusively and analyzed objectively. Previous research has identified stable patterns of acoustic indicators of mood and emotion [1-15]. Among many reported patterns, sad or depressive speech tends to be quieter, slower, with longer pauses, lower in pitch and more monotonous than normal speech.

This research project [16] identified certain duration- or amplitude-based aspects of the speech signal that can be measured completely automatically by computer and which provide paralinguistic information about the apparent affect of a speech sample. Specifically, the project identified measurable physical differences in speech signals that can be used to estimate how depressed or elated a person would sound to a panel of experienced clinicians.

The purpose of the project was the development and evaluation of techniques that may contribute to the measurement of affective states like depression. This project is an empirical study preliminary to building an integrated computerized instrument for administering structured interviews to patients, via the telephone, that can provide non-obtrusive, objective data that may improve assessment accuracy and validity. The project created a corpus of elicited speech and developed an automatic analysis of the recordings. The experiment reported here accomplished two preliminary objectives:

A. Replicate the reported relations between timing and amplitude of speech and perceived affect, for example, by [9, 11], but using fully automatic means;
B. Find and verify additional temporal manifestations of affect in speech signals.

The project focused on answering three main questions:

1) Which measurable paralinguistic characteristics of speech (e.g. response latency, speech rate, amplitude) can be reliably related to the simulated mood of a speaker?
2) Which of these characteristics can be derived automatically from the acoustic signals of spoken responses to test questions?
3) Which observed measures of paralinguistic variables show significant differences across speakers, and which show significant differences only within speakers?

II. METHODOLOGY

The data collection procedure followed a single session experimental design, wherein each speaking subject took a seven-minute speaking test by telephone, three times in succession, under three different conditions: once without instruction, once instructed to speak as if severely depressed, and once instructed to speak as if they were extremely manic (the order of the second and third conditions was counterbalanced). The seven-minute speaking test is a "sample" form of the PhonePass SET-10, a language test developed by Ordinate Corporation in California [17] to measure spoken English proficiency.

The experiment compared acoustic variables extracted from the speech samples corresponding to the uninstructed (or "normal") renditions, to the same variables from the speech samples that the speaking

subjects intended to be simulations of depressive and manic speech. Analysis of data from this experiment was intended to determine whether or not there were observable differences in the speech samples according to the speakers' intentions.

Subjects: The Speaking Subjects comprised 40 psychotherapists who were all native speakers of English, between the ages of 30 and 71; mean age was 53 years old. Of the 40 speaking subjects, 23 were women and 17 were men. Each speaking subject spent approximately 35 minutes of time in the experiment.

Instrumentation: The PhonePass SET-10 Sample Form comprises a set of 32 items administered in a 7-minute telephone call. Each item presents a recorded prompt over the telephone that solicits a spoken response from the subject that is recorded via telephone. The items used in this experiment form part of a single test form that prompts a subject to speak 32 times. Items of five different types are presented to the examinee: first, six one-sentence readings, then eight elicited repetitions of sentences, then eight opposite words, then eight short-answer questions, and, finally two open questions – each allowing the subjects thirty seconds to deliver their response. Most items elicit one-sentence responses or one-word responses that are about 0.5 to 5 seconds in duration.

Assuming that the average response length is six words and an average word has four phonemes, with 26 spoken responses measured per subject per condition, the data set potentially contains about 624 dependent measures per subject condition and more than 1800 dependent measures per speaking subject.

Ten dependent variables were measured:

TST: total speaking time (milliseconds)
TPT: total pause time (milliseconds)
TUT: total utterance time (milliseconds)
ROS: rate of speech (phonemes/second)
ART: articulation rate (phonemes/second)
LAT: response latency (milliseconds)
MPD: mean pause duration (milliseconds)
SDP: segment duration probability (log probability)
PDP: pause duration probability (log probability)
MaxSA: maximum speech amplitude (signal value)

III. RESULTS

The results are presented numerically in Tables 1 and 2. Table 1 presents the data grouped across subjects, each cell showing the mean and standard deviation of each sample of 480 responses (12 selected responses x 40 subjects) per condition as measured on each of the 10 paralinguistic acoustic parameters under study. Table 2 presents the data organized by within-subject, within-item differences when the Speaker Subjects responded to the same item with two different intended moods.

The data as presented in Table 1 represent a comparison of groups of Speaking Subjects according to their instructed intentions. Table 1 presents measures that describe the central tendency and dispersion of the paralinguistic parameters of the responses when these Speaking Subjects talked in three different moods, as these parameters were automatically estimated by the speech recognition and signal processing internal to the PhonePass system.

Table1:
Mean, s.d. of Parameters for Intended Mood (N=480)

Param	D (=Depressed)		N (=Normal)		M (=Manic)	
	mean	s.d.	mean	s.d.	mean	s.d.
TST	2891.67	985.83	2556.67	787.10	2234.92	925.30
TPT	178.60	360.61	40.29	192.50	124.67	580.90
TUT	3070.27	1148.26	2596.96	836.97	2359.58	1276.64
ROS	9.72	1.97	11.37	1.68	13.06	2.72
ART	10.15	1.73	11.50	1.62	13.33	2.39
LAT	1360.21	759.0	656.79	287.60	533.58	498.38
MPD	20.12	41.24	4.57	20.43	12.51	59.16
SDP	-5.23	0.39	-4.90	0.29	-5.05	0.27
PDP	-2.63	0.93	-2.32	0.79	-2.20	0.82
MaxSA	6.62	4.24	9.96	4.74	15.35	8.23

The columns in Table 1 are ordered D – N – M (Depressed, Normal, Manic) in the expectation that the parameter values will generally be increasing or decreasing in that order. That is, from the literature, one would expect the Normal value of most of these parameters to be between the Depressed and the Manic value. This presumed ordering was observed for seven of the ten paralinguistic parameters in this study.

Table 2: Within-Subject Within-Item Paired Differences

Param	D-N (N = 375)		M-N (N = 373)		M-D (N = 386)	
	mean	s.d.	mean	s.d.	mean	s.d.
TST	344.80	595.62	332.44	730.24	655.80	838.19
TPT	141.07	399.59	56.94	413.48	76.14	517.90
TUT	485.87	809.41	275.50	989.47	731.94	1163.39
ROS	-1.75	2.14	1.66	2.53	3.32	2.69
ART	1.43	1.85	1.79	2.22	3.15	2.26
LAT	672.96	697.75	136.59	465.24	806.27	739.19
MPD	15.38	44.77	6.34	54.71	8.88	67.12
SDP	-0.34	0.43	-0.15	0.37	0.19	0.46
PDP	-0.33	1.09	0.03	0.94	0.40	1.15
MaxSA	3.41	4.05	5.35	6.72	9.00	7.44

Table 2 presents the data in a way that is more relevant to the ultimate question: how well would one expect an automatic system to detect changes in a known speaker's paralinguistic parameters under the instructions of this experiment. Table 2 presents paired differences. Each normal item response by each subject is a control on the measures for that item in the other two conditions. This way of treating the data should eliminate expected inter-subject and inter-item variance, yielding smaller standard errors of the mean, while the mean differences are approximately equal to the differences in the means for the various moods. This expected reduction in variance

should promote rejection of the null (no-difference) hypotheses.

To test the significance in the differences in the mean parameter values, as shown in Table 1, across the population of speaking subjects and across the various sets of 12 items measured per call, a t-test for two population means with variances unknown and unequal [18] was used. The results indicate that 29 of the 30 observed differences in means are significantly different from zero ($t > 1.96$, $p = 0.05$), and even under the stricter criterion corrected for 10 simultaneous variables ($t > 2.81$), 26 of the 30 t-tests are still significant ($p < 0.0025$). All four of the comparisons that fail the stricter significance test, TPT (M-D), MPD (M-N, M-D), and PDP (M-N), are based in part on the measures of pause time in the manic experimental condition.

A simple and conservative test of the statistical significance of the differences between intended normal, depressive and manic speaking on the 10 paralinguistic acoustic parameters is a sign test [19]. The sign test assumes related samples, considered in pairs where members of the pairs can be ranked. The sign test does not assume that the data under study carry more than ordinal information, and it does not assume a normal distribution. The differences in 28 of the 30 possible comparisons are statistically significant ($z > 1.96$, $p = 0.05$). Only the manic-normal differences for TPT and MPD fail to reject the null hypothesis of no difference. If a 10-variable correction is accepted, and the rejection region is divided by 10 so that $p < 0.0025$ is the criterion for significance, the boundary of significance for the statistic increases from 1.96 to 2.81.Under this stricter criterion and with a test that makes no assumptions about distribution shape, 28 of the 30 tests show the mean difference to be significantly different from zero. Note that differences with values of zero were not counted in the calculation of the sign test.

Table 3: Values of d′ for depressed vs. normal speech within- and across-subject groupings

Parameter	d′ (population)	d′ (person-item)
TST	0.376	0.819
TPT	0.478	0.499
TUT	0.477	0.849
ROS	0.902	1.155
ART	0.805	1.097
LAT	1.226	1.365
MPD	0.478	0.486
SDP	0.962	1.120
PDP	0.365	0.436
MaxSA	0.744	1.192

A convenient measure of discriminability is “d-prime” (written d′). The parameter d′ is a standardized difference between two means [20]. Table 3 displays the value of d′ for depressed speech when this condition is to be discriminated from normal speech. The d′ is a normalized standard score. A d′ value of 0.0 indicates that there is no information useful in discriminating the depressed speech samples from the background expectation of normal speech. Larger d′ values indicate greater discriminability in a parameter and greater usefulness for automatic categorization of signals.

Table 4: Agreement of significant experimental results from literature reviews

Parameter	Significant Order	Scherer (1986) agrees	Murray & Arnott (1993) agree
TST	D > N > M	yes	yes
TPT	D > N > M	no info	no info
TUT	D > N > M	yes	yes
ROS	M > N > D	yes	yes
ART	M > N > D	yes	yes
LAT	D > N > M	no info	no info
MPD	D > N , D > M	no info	no info
SDP	N > M > D	no info	no info
PDP	M > N > D	no info	no info
MaxSA	M > N > D	yes	yes

IV. Discussion

The data are generally consistent with an alternative hypothesis that psychotherapists have a relatively uniform model of depressive and manic speech patterns that do show up in their simulations and agree with the patterns reported in the literature. Of the parameters (often vaguely specified in the literature) that seem to have an analog in the parameters of this experiment, the significant observed orders are uniformly in accordance with published literature reviews, as is shown in Table 4.

Many of the statistical tests show effects that are extremely unlikely under the null hypothesis, yet the single parameter d′ values are not particularly large, which indicates that a device that used any single one of these parameters to classify an unknown person could make a substantial number of errors. The d′ values are generally greater for the within-speaker comparisons, which supports the intuitive and expected result that a device or a person would do better using paralinguistic information to discriminate among the moods of a known person than to identify the moods of an unknown person. From a single voice recording by itself, a listener can presumably recognize a mood shift in a friend more reliably than that same listener could classify the mood of a stranger.

All ten of the paralinguistic acoustic variables that were studied had statistically significant association with one or the other of the two moods (depressed or manic) that were intentionally simulated by the psychotherapists who served as speaking subjects; eight out of ten parameters were significantly different in both moods from the uninstructed (normal) condition. Two variables failed the test of significance for the manic speech only in manic versus normal comparisons.

When analyzed within subject and within item, both

simulated moods are significantly different from the uninstructed (normal) mood in nine of the ten parameters, instead of the eight of ten in the group comparison. The only failure of significance was in one manic versus normal comparison.

Certain conditions of this experiment limit the scope of the conclusions. The foremost limitation concerns the use of psychotherapists as subjects. The variety of initial speaking patterns and courses of change over time that is found in real clinical populations is simply not found in the speech data from people simulating moods. Likewise, there is no possibility to compare the speech data with concurrent scores on cognitive, emotional, physiological, or motor-performance tests. Thus, none of the hypotheses about the cognitive or psychomotoric nature of mood disorders as discussed by [7] or by [14] can be tested with this new data. Finally, an important limitation is that voice fundamental frequency (F0) was not measured and therefore not analyzed.

V. Conclusion

Psychotherapists can imitate (without any instruction or guidance) some of the vocal patterns of depressed and manic people in a way that is relatively consistent over the population of therapists and is also consistent with the paralinguistic changes reported in the literature on speech in emotion and mood disorders. For many traditional paralinguistic parameters, the ordering of {depressed, normal, manic} is monotonic increasing or decreasing. Generally, for the psychotherapists simulating mood or pathology, the depressed direction from normal is more reliably and distinguishably produced.

The differences in paralinguistic parameters between groups of people when speaking normally and when simulating moods are very significant, but these differences may be relatively difficult to use for mood identification from any single one of the duration- or amplitude-based parameters that were studied in this project.

If these results can be replicated with an appropriate clinical population, then this study provides a system and the core of an algorithm for rating the paralinguistic evidence of mood disorders by telephone, automatically, on demand. Note that to be useful or interesting, the system does not have to be highly accurate, it may suffice that the system performs as well as a skilled therapist, and only on that aspect of the therapist's judgment that relates to manner of speaking.

References

[1] M. Alpert, A. Rosen, J. Welkowitz, C. Sobin & J. Borod, "Vocal acoustic correlates of flat affect in Schizophrenia". *British Journal of Psychiatry*. 154, 51-56, 1989.

[2] R. Banse & K. Scherer, "Acoustic profiles in vocal emotion expression". *Journal of Personality and Social Psychology*, 70 (3), 614-636, 1996.

[3] R. Cowie & R.R. Cornelius, "Describing the emotional states that are expressed in speech". *Speech Communication*, 40, (1-2),5-32, 2003

[4] H. Ellgring, & K. Scherer, "Vocal indicators of mood change in depression". *Journal of Nonverbal Behavior*, 20 (2), 1996.

[5] G. Fairbanks & L. Hoagland,. "An experimental study of the durational characteristics of the voice during the expression of emotion". *Speech Monographs*, 8, 85-90. 1941.

[6] T. Goldbeck, F Tolkmitt, & K. Scherer, "Experimental studies on vocal affect communication". In K. Scherer (Ed.), *Facets of Emotion. Recent Research*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988, pp. 119-137.

[7] J. Greden & B. Carroll, "Decrease in speech pause times with treatment of endogenous depression". *Biological Psychiatry*, 15 (4), 575-587, 1980.

[8] S. Kuny & H. Stassen "Speaking behavior and voice sound characteristics in depressive patients during recovery". *Journal of Psychiatric Research*, 27 (3), 289-307. 1993.

[9] I. Murray & J. Arnott, "Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion". *Journal of the Acoustical Society of America,* 93 (2), 1097-1108, 1993.

[10] J. Pittam & K. Scherer, "Vocal expression and communication of emotion". In M. Lewis & J.M. Haviland (Eds.), *Handbook of Emotion,* New York: Guilford Press, 1993, pp. 185-198.

[11] K. Scherer, "Vocal affect expression: A review and a model for future research". *Psychological Bulletin*, 99 (2), 143-165, 1986.

[12] H. Stassen, G. Bomben & E. Gunther, "Speech characteristics in depression". *Psychopathology*, 24, 88-105, 1991.

[13] E. Szabadi, C. Bradshaw & J. Besson, "Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression". *British Journal of Psychiatry*, 129, 592-597, 1976.

[14] D. Widlocher & A. Ghozlan, "The measurement of retardation in depression". In I. Hindmarch & P. Stonier (Eds.), *Human Psychopharmacology: Measures and Methods,* Vol. 2. New York: John Wiley & Sons Ltd., 1989, pp. 1-22.

[15] C. Williams & K. Stevens, "Emotions and speech: some acoustical correlates". *Journal of the Acoustic Society of America,* 52, 1238-1250, 1972.

[16] E. P. Rosenfeld, "Automatic analysis of vocal manifestations of psychological states". Unpublished doctoral dissertation. Western Graduate School of Psychology, Palo Alto, California, 2000.

[17] Ordinate Corporation, *Validation Summary for the SET-10 Test,* Menlo Park, CA: Ordinate Corporation, 2000.

[18] G. Kanji, *100 Statistical Tests*. London: SAGE Publications, 1993.

[19] S. Siegel, *Non-parametric Statistics*. New York: McGraw-Hill, 1956.

[20] C. Coombs, R. Dawes & A. Tversky, *Mathematical Psychology.* Englewood Cliffs, NJ: Prentice-Hall, 1970.