Proceedings of the 25th Annual International Conference of the IEEE EMBS
Cancun, Mexico • September 17-21, 2003

Analysis of Prosodic Variation in Speech for Clinical Depression

Elliot Moore II*, Mark Clements*, John Peifer† and Lydia Weisser‡
*School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
†Interactive Media Technology Center, Georgia Institute of Technology, Atlanta, GA, USA
‡ Department of Psychiatry and Behavioral Health, Medical College of Georgia, Augusta, GA, USA

***Abstract*— Understanding *how* someone is speaking can be equally important to *what* they are saying when evaluating emotional disorders, such as depression. In this study, we use the acoustic speech signal to analyze variations in prosodic feature statistics for subjects suffering from a depressive disorder. A new sample database of subjects with and without a depressive disorder is collected and pitch, energy, and speaking rate feature statistics are generated at a sentence level and grouped into a series of observations (subset of sentences) for analysis. A common technique in quantifying an observation had been to simply use the *average* of the feature statistic for the subset of sentences within an observation. However, we investigate the merit of a series of statistical measures as a means of quantifying a subset of feature statistics to capture emotional variations from sentence to sentence within a single observation. Comparisons with the exclusive use of the *average* show an improvement in overall separation accuracy for other quantifying statistics.**

***Keywords*— Affect, depression, prosodics, speech**

I. Introduction

The human body constantly emits a multitude of signals (heart rate, perspiration, etc.) expressing its current condition. Quality health care depends on the ability to extract and properly interpret these signals. One of the easier body signals to extract is the acoustic speech waveform. Speech signals can carry an extreme amount of information on the condition of the body and mind simply by *how* it is expressed. We refer to the "*how*" of speech perception as vocal affect. Vocal affect refers to the emotional expression of speech and its relation to the overall state of the speaker. Prosodics are the most prominent set of features studied in relation to vocal affect due to their ease of measurement and strong correlation to human perception. The prosodics of speech relate to measurements of pitch (F_0), speaking rate and energy. The use of prosodics and other speech features have led to a number of studies on distinguishing discrete states of emotion, such as anger, happiness, and fear, ([1], [2], [3], [4], [5], [6]) and emotional disorders, such as clinical depression ([7], [8], [9]), which is the focus of our study.

Clinical depression is one of the more common emotional disorders ([7]) making it an important choice of study when analyzing features related to the emotional state of a speaker in a clinical setting. While not all studies on depression have agreed on the role of prosodic features ([4], [8]), many studies have found prosodic features as useful indicators of depression in the voice. An important point is that a speakers overall mental state is evaluated over a period of time. Identifying statistics that are sensitive to variations in vocal affect over successive intervals of time can be crucial in making proper decisions. In this study, we extend the use of prosodics in the analysis of clinical depression by utilizing an analysis and classification method designed to capture variations in affective expression among a series of observations over time.

II. Speech Database

A collaborative effort with the Department of Psychiatry and Behavioral Health at the Medical College of Georgia (MCG) produced a new speech database consisting of subjects with and without a depressive disorder. The depression subjects (experimental group) in the study were taken from the outpatient clinic of the psychiatry department of MCG. While using subjects from the outpatient clinic may have resulted in a wider variety of depressive disorders, it was assumed for the purpose of this study that any diagnosis of depression for which the subject was still receiving treatment was sufficient for creating a study group. The speech corpus consisted of voice samples of both males (6 patients, 9 controls) and females (9 patients, 9 controls). Patients included in the study were required to meet the following criteria:

1) Diagnosis of Major Depression or other depressive illness
2) Able to comprehend and satisfactorily comply with the protocol requirements
3) No clinical picture and/or history that was consistent with delirium, dementia, amnestic or other cognitive disorder
4) No diagnosis of bipolar disorder
5) Had not met DSM-IV criteria for any significant substance use disorder within the past three months
6) Spoke American English fluently (preferably as first language)

The control group was required to meet similar requirements with the obvious exclusion of having any prior diagnosis of a depressive illness. All recordings were made over a period of approximately a year and collected in the same room with sufficient acoustics to provide a relatively clean audio sample. The ages of the patients ranged from 19-57 for the females and 33-50 for the males. While most of the patients were on medication for their illness, none of the medication adversely affected the physical process of speech production. The ages

0-7803-7789-3/03/$17.00 ©2003 IEEE

for the controls ranged from 29-53 for the females and 25-39 for the males. The controls were volunteers taken from within the MCG clinic, which included students and employees. Each subject's voice was recorded while reading a short story that provided at least 3 minutes of speech. The recording samples were stored in 16 kHz PCM wave file format, and broken up into 65 separate sentences covering the entire reading session.

III. Feature Extraction

The features included in this study were the pitch (fundamental frequency, F_0), energy, and speaking rate. Fundamental frequency measures the approximate periodic rate of glottal opening and closing during voiced sections of speech. The F_0 extraction algorithm for this study was similar to an algorithm in [10] and based on a slight modification of the Markel algorithm [11]. Although more recent pitch extraction techniques existed, the algorithm had shown to be efficient for clean speech and easy to implement. The energy contour of the signal measured the approximate intensity of the speech over time. It was measured on a frame by frame basis and computed according to (1)

$$E[k] = \frac{1}{N}\sum_{n=0}^{N} s^k[n]^2 \quad (1)$$

where,

$$s^k[n] = w_r[n] * s[n + kN] \quad (2)$$

represented the k^{th} frame of speech being analyzed and $w_r[n]$ was a rectangular window of length N. With the understanding that recording conditions (i.e., microphone distance, recording level) varied slightly from session to session, our analysis was based on energy perturbation measures among the voiced sections of each contour as opposed to direct energy values. Energy contour statistics were divided into two categories: the energy deviation statistics (EDS) and energy median statistics (EMS). The computation of these statistics required first dividing the contour into its voiced sections and making direct statistical measurements (see section IV) on each voiced section within the sentence. After making these measurements, the EDS and EMS were computed by taking all of the statistics from the voiced sections and computing the standard deviation (EDS) and the median values (EMS) of the i^{th} statistic according to

$$EDS_i = STD(ES_i[E_V]), i = 1, \ldots, N \quad (3)$$

$$EDS_i = MED(ES_i[E_V]), i = 1, \ldots, N \quad (4)$$

where ES_i is the i^{th} statistic computed on the voiced sections, E_V is the energy contour broken up into V voiced sections, and N is the total number of statistics computed on each voiced section. The EDS and EMS measures were experimental measures developed for this study. Speaking rate was measured in terms of words/sec (wds/sec) and syllables/sec (syl/sec) using prior knowledge of the approximate number of words and syllables for each recorded text. There was some variation in the exact number of words/syllables for each speaker since some syllables or words may have been omitted due to reading error or pronunciation variability, but the variation was considered to be negligible and not to have a significant effect on the measurement results.

IV. Analysis Method

Feature extraction was performed at a sentence level. We referred to the direct feature statistical measures within each sentence as *intra-sentence* statistics. The statistical measures taken on the extracted features is shown in tables I and II. For analysis, two different observational groupings were used. The first grouping (*G1*) divided the passage for each subject into 13 observations of 5 sentences each and the second grouping *G2* divided the passage into 5 observations of 13 sentences each. In this way, the entire passage was covered and the strength of the statistical separation could be tested for various observation lengths. The statistics listed in table II were only measured on the pitch and energy contours. Unless otherwise noted, the features in table II were extracted from both the voiced F_0 and energy (EDS, EMS) contours. The 1^{st} order polynomial coefficients (a_1 - contour slope; a_0 - intercept) were measured from the voiced F_0 contour by removing all unvoiced frames and setting all voiced frames adjacent to one another to determine the general flatness and direction of the contour. The coefficients for the energy contour were measured on each voiced section and then the EDS and EMS were computed according to (3) and (4) respectively. Both were measured using a 1^{st} order polynomial fitting function in MATLAB. The JF (Jitter Factor), Avg_1 (1^{st} order perturbation average), Med_1 (1^{st} order perturbation median), STD_1 (1^{st} order perturbation standard deviation), MR (mean rectified), and RAP (relative average perturbation) measures were based on perturbation functions in [12].

A common technique in quantifying an observation had been to simply use the *average* of the feature statistic from the subset of sentences within an observation. However, we did not believe that this could fully represent the affective variation within the observation. Therefore, we investigated the merit of analyzing all of the statistics listed in table I as a means of quantifying the subset of feature statistics for an observation. For example, a common means of quantifying a

TABLE I
BASIC STATISTICAL MEASURES

Statistic	Equation
Average (AVG)	$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
Median (MED)	$50^{th} percentile$
Standard Deviation (STD)	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$
Minimum (MIN)	$5^{th} percentile$
Maximum (MAX)	$95^{th} percentile$
Range (RNG)	$MAX - MIN$
Dynamic Range (DRNG)	$log_{10}(MAX) - log_{10}(MIN)$
Interquartile Range (IQR)	$75^{th} percentile - 25^{th} percentile$

TABLE II

OTHER STATISTICAL MEASURES

Statistic	Description
$10P$-$90P$ (F_0 only)	10^{th}-90^{th} percentiles
MR	$MR^1 = \frac{1}{N-1}\sum \lvert p_i^1 \rvert$
JF	$\frac{MR^1}{\bar{x}}$
AVG_1	$\overline{p^1} = \frac{1}{N}\sum p_i^1$
MED_1 (EDS, EMS only)	Median of 1^{st} order perturbation
STD_1	$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(p_i^1 - \overline{p^1})^2}$
RAP	$\frac{MR^2}{\bar{x}}$
a_0, a_1	polynomial coef.: $a_0 + a_1 x$

subset of *intra-sentence* statistical measures (i.e., the *standard deviation* of F_0) representing a single observation would be to use the *average* over that interval. We simply added other quantification statistics based on table I creating eight quantifiers of each *intra-sentence* statistical measure over the same interval. We referred to these quantifiers as *inter-sentence* statistics. In this way, over 400 statistics per speaker were generated for analysis.

A one-way analysis of variance (ANOVA) was performed for each grouping (*G1*, *G2*) on each *inter-sentence* statistic grouped by gender (male, female) and subject type (patient, control) to investigate potential significance among the features. A Gaussian Mixture Model was used on each *inter-sentence* statistic displaying potential significance to model the two subject types (i.e.,λ_C - controls; λ_E - patients). The classification of **X** (a sequence of feature vectors $x_1, \ldots, x_N$representing independent observations) into the appropriate class was done by finding the model that yielded the greatest log-sum probability based on the entire set of observations. Because of the small data size, the leave-one out (LOO) method was used in training and testing. (i.e., training excluded one subject and tested with that subject). The separation accuracy of each *inter-sentence* feature statistic with an appropriate significance level was evaluated utilizing the entire set of observations for each grouping. This was done so that variations in affective expression could be monitored on each subject for the entire session (~3-5 minutes of speech) before a decision was made.

V. RESULTS

A. Males

Tables III and IV show the feature statistics that yielded the top accuracy levels for grouping *G1* and *G2*, respectively. The **Average** column refers to the features identified using the *average* as the means of quantification and the **Study** column refers to the use of the other statistics from table I. The columns are divided by feature category, *intra-sentence* (***Intra***) statistical measure, *inter-sentence* (***Inter***) statistical quantifier, and accuracy level (**Acc**). Energy and speaking rate features did not play a significant role in separation for the males but several feature statistics related to F_0 showed good separation. Both tables show that the use of other quantifying statistics besides the *average* provides accuracy levels that are, on a whole, better than the exclusive use of the *average* alone. A maximum classification accuracy of 87% (2 misclassified controls) was only seen with quantifying statistics other than the *average* (the *STD* of AVG_1 for F_0). It should be noted that the classification errors are all control subjects. The only apparent difference among the different groupings (*G1*, *G2*) is in the type and number of feature statistics identified and used for classification.

B. Females

Tables V and VI show the feature statistics that yielded the top accuracy levels for grouping *G1* and *G2*, respectively. As with the males, feature statistics related to speaking rate did not play a significant role in separation but F_0 and energy related feature statistics (particularly EMS) showed good separation. Again it is clear that the use of other quantifying statistics besides the *average* provides accuracy levels that are, on a whole, better than the exclusive use of the *average* alone. While the maximum classification accuracy of 94% (1 misclassified controls) was produced utilizing both quantifying methods, generally the classification accuracy was better for quantifying statistics other than the *average*. Similar to the males, the classification errors are all control subjects and the only apparent difference among the different groupings (*G1*, *G2*) is in the type and number of feature statistics identified and used for classification.

VI. CONCLUSION

An improvement in separation accuracy using other observational quantifiers than the *average* is clearly seen. For the males, in every case at least one additional subject is correctly identified using statistics other than the *average* to quantify

TABLE III

COMPLETE ACCURACY COMPARISON (MALES-G1)

	Average			Study		
Category	***Intra-***	***Inter-***	**Acc**	***Intra-***	***Inter-***	**Acc**
F_0	RNG	Avg	67%	DRNG	Std	80%
	DRNG	Avg	80%	DRNG	Max	80%
				DRNG	Rng	80%
				IQR	Med	80%
				AVG_1	Std	87%
				AVG_1	Rng	80%
EDS	MED	Avg	67%			

TABLE IV

COMPLETE ACCURACY COMPARISON (MALES-G2)

	Average			Study		
Category	***Intra-***	***Inter-***	**Acc**	***Intra-***	***Inter-***	**Acc**
F_0	RNG	Avg	67%	DRNG	Med	80%
	DRNG	Avg	80%	DRNG	Max	80%
	AVG	Avg	67%	AVG_1	Std	87%
	MED	Avg	67%			

TABLE V
COMPLETE ACCURACY COMPARISON (FEMALES-G1)

	Average			Study		
Category	*Intra-*	*Inter-*	Acc	*Intra-*	*Inter-*	Acc
F_0	IQR	Avg	83%	IQR	Max	89%
	90P	Avg	83%	IQR	IQR	89%
	STD_1	Avg	83%	80P	Std	89%
				80P	Rng	89%
				80P	DRng	89%
				80P	IQR	89%
				a_0	Rng	89%
				a_0	DRng	89%
EMS	AVG	Avg	83%	MAX	Max	89%
	MAX	Avg	83%	STD	Max	89%
	STD	Avg	83%	RNG	Med	89%
	RNG	Avg	83%	RNG	Max	89%
	IQR	Avg	83%	MED_1	Min	89%
	STD_1	Avg	83%	MED_1	Rng	89%
	MR	Avg	94%	STD_1	Min	89%
				MR	Med	94%
				MR	Min	89%
EDS	AVG_1	Avg	83%	AVG_1	Std	89%
				AVG_1	Max	89%
				AVG_1	Rng	89%

TABLE VI
COMPLETE ACCURACY COMPARISON (FEMALES-G2)

	Average			Study		
Category	*Intra-*	*Inter-*	Acc	*Intra-*	*Inter-*	Acc
F_0	STD	Avg	83%	AVG	DRng	89%
	90P	Avg	83%	IQR	Std	89%
				IQR	IQR	89%
				40P	IQR	89%
				80P	Std	89%
				80P	Rng	89%
				80P	DRng	94%
				90P	Std	89%
				90P	Rng	89%
				RAP	Min	89%
				a_0	Rng	89%
				a_0	DRng	89%
EMS	AVG	Avg	83%	AVG	Std	89%
	MAX	Avg	83%	MED	Max	89%
	STD	Avg	83%	MAX	Rng	89%
	RNG	Avg	83%	STD	Med	89%
	IQR	Avg	83%	STD	Max	94%
	STD_1	Avg	83%	RNG	Med	89%
	MR	Avg	94%	RNG	Max	89%
				IQR	Max	89%
				MED_1	Min	94%
				STD_1	Med	89%
				STD_1	Min	89%
				MR	Min	89%
EDS	AVG_1	Avg	83%	AVG_1	Std	89%
				AVG_1	Max	89%

each observation. In some cases, as many as 2 additional subjects are correctly identified. For the females, the extended statistical model performed as well as, and in many cases, better than the exclusive use of the *average*. In general, one additional subject is correctly identified using statistics other than the *average* to quantify each observation. We believe the overall improvement we observed was a result of a more robust statistical model that allowed for temporal affective variation of the *intra-sentence* feature statistics within an observation. This is not to say that using the *average* as means of quantifying an observation is not effective, but we show the value of analyzing other statistics as a means of quantifying observations when studying the affective state of a speaker over time. Establishing better speech features for identifying depression and other types of mental deviancies could prove useful in a variety of ways including the development of early diagnosis tools, assisting on-call doctors unfamiliar with a patients normal behavior, and establishing additional objective patterns for evaluating improvement or digression.

ACKNOWLEDGMENT

The authors would like to thank the the students and volunteers at the Medical College of Georgia who made data collection possible.

REFERENCES

[1] H. Sato, Y. Mitsukura, M. Fukumi, and N. Akamatsu, "Emotional speech classification with prosodic parameters by using neural networks," in *7th Australian and New Zealand Intelligent Information Sys. Conf.*, Nov 2001, pp. 395–398.

[2] F. Dellart, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. of Fourth Int. Conf. on Spoken Lang. Processing*, vol. 3, 1996, pp. 1970–1973.

[3] C. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 240–243.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Mag.*, pp. 33–80, Jan 2001.

[5] K. Scherer, *Speech evaluation in psychiatry*. Grune & Stratton, 1981, ch. Speech and emotional states, pp. 189–214.

[6] K. Scherer and J. Pittman, *The handbook of emotions*. Guilford Press, 1993, ch. Vocal expression and communication of emotion, pp. 185–197.

[7] A. Nilsonne, J. Sundberg, S. Ternstrom, and A. Askenfelt, "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *Jrnl. Acous. Soc. Am.*, vol. 83, no. 2, pp. 716–728, Feb 1988.

[8] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, July 2000.

[9] A. Ozdas, D. M. Wilkes, R. G. Shiavi, S. E. Silverman, and M. K. Silverman, "Analysis of fundamental frequency for near term suicidal risk assessment," in *Proceedings, IEEE Int. Conf. on Systems, Man and Cybernetics*, vol. 3, 2000, pp. 1853–1858.

[10] B. F. Necioglu, "Objectively measurable descriptors of speech," Ph.D. dissertation, Georgia Institute of Technology, 1998.

[11] J. Markel, "The sift algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367–377, Dec 1972.

[12] I. R. Titze and N. B. Pinto, "Unification of perturbation measures in speech signals," *Jrnl. Acous. Soc. Am.*, vol. 87, no. 3, pp. 1278–1289, Mar 1990.