Using Prosodic and Spectral Features in Detecting Depression in Elderly Males

Michelle Hewlett Sanchez[1,2], *Dimitra Vergyri*[1], *Luciana Ferrer*[1], *Colleen Richey*[1], *Pablo Garcia*[3], *Bruce Knoth*[3], *William Jarrold*[4]

[1]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA 94025, U.S.A.
[2]Stanford University, Stanford, CA 94305, U.S.A.
[3]Robotics and Medical Systems Laboratory, SRI International, Menlo Park, CA 94025, U.S.A.
[4]Center for Mind and Brain, University of California Davis, Davis, CA 95616, U.S.A.

{mhewlett,dverg,lferrer,colleen}@speech.sri.com,
{pablo.garcia,bruce.knoth}@sri.com, wjarrold@ucdavis.edu

Abstract

As research in speech processing has matured, there has been much interest in paralinguistic speech processing problems including the speaker's mental and psychological health. In this study, we focus on speech features that can identify the speaker's emotional health, i.e., whether the speaker is depressed or not. We use prosodic speech measurements, such as pitch and energy, in addition to spectral features, such as formants and spectral tilt, and compute statistics of these features over different regions of the speech signal. These statistics are used as input features to a discriminative classifier that predicts the speaker's depression state. We find that with an N-fold leave-one-out cross-validation setup, we can achieve a prediction accuracy of 81.3%, where random guess is 50%.

Index Terms: Depression, Emotion Detection, Prosodic Features

1. Introduction

As automatic speech processing has advanced in the last few years, research attention has shifted away from linguistic problems, such as automatic speech recognition, to applications focusing on paralinguistic speech problems that aim to detect "beyond-the-words" information. Researchers have focused on automatically deriving speaker characteristics from speech and classifying speakers into categories ranging from age [1, 2], identity [3, 4], idiolect and sociolect [5] to truthfulness [6], cognitive health [7] and emotion.

Prosodic features in speech, such as speaking rate, pitch, energy or intensity, and pause duration have been used for emotion detection in previous work [8, 9, 10, 11, 12, 13]. In addition, other acoustic features such as voice quality [10, 11], spectral features [8, 12], and mel frequency cepstral coefficients (MFCCs) [12] have also been explored.

In an initial interdisciplinary study combining psychiatry and speech science, Darby et al. published work analyzing speech patterns of depressed patients before and after treatment with antidepressant medication [14, 15]. Dr. S. Silverman of Vanderbilt University studied decades of his recorded psychotherapy sessions and found that the speech of a depressed person undergoes a subtle shift as the person becomes near-term suicidal. In preliminary work at Vanderbilt, mean vocal jitter was found to distinguish suicidal individuals from treated patients who are no longer depressed [16, 17].

Other changes in acoustic features have also been explored in detecting depression and suicide risk. Prosodic features (pitch, energy, and speaking rate) [18, 19, 20, 21], spectral features (formants, their corresponding bandwidths, power spectral density (PSD), and spectral tilt) [18, 19, 21, 22, 23], and MFCCs have been found useful in identifying depression [21, 22].

Researchers at SRI have already been successful in applying acoustic analysis to tasks such as the detection of different types of emotions and deception. Ang et al. employed human-computer dialog strictly used for research making feigned air travel reservations in order to classify annoyed and frustrated versus neutral emotion [8]. Sanchez et al. used prosodic (pitch, energy, and speaking rate) features along with domain compensation to detect fear in real emergency 911 phone calls [13]. Graciarena et al. used both prosodic (pitch, energy, and duration patterns) and lexical features in detecting deceptive speech [6].

In this paper, we use prosodic and spectral features similar to those used in previous work to try to predict whether a speaker is depressed or not. In the following sections, we describe the dataset used for our experiments, present the prosodic features and classification algorithms used for this work, and analyze the experimental results on our dataset. We conclude with possible directions of future work.

2. Dataset

We used the Western Collaborative Group Study (WCGS) [24] dataset, collected by SRI International in 1988, which consists of 1182 older Caucasian men (mean age of 70). Although this study focused on personality types, it also contains data on the psychological health of the participants, including the Center for Epidemiologic Studies Depression Scale (CESD) [25] measure of depression. In our study, speakers with a CESD score of greater than or equal to 26 are assigned the category *severely depressed*, speakers with a CESD score of less than 26 and greater than 15 are assigned the category *mildly depressed*, and speakers with a CESD score of less than or equal to 15 are assigned the category of *nondepressed*. This categorization is in agreement with the clinical literature [26]. There were 16 severely depressed subjects, 52 mildly depressed subjects, and 1114 non-depressed controls in this dataset.

Our paper focuses on classifying the severely depressed speakers from the nondepressed controls. For our experiments, in order to have a similar number of severely depressed and nondepressed participants, we used only 16 nondepressed participants along with all 16 severely depressed participants, giving us a total of 32 male participants from ages 65 to 82. The nondepressed controls were chosen to match the severely de-

pressed subjects on certain factors such as age, marital status, and number of years of education.

The data in this study consists of mono-audio recordings of 15-minute structured interviews for type-A personality characteristics and associated CESD measure of depression. The recordings were digitized and downsampled to 16 kHz for use in our experiments. In order to have accurate times of the participant speech in the interviews, the entire recordings for the speakers that were chosen were manually transcribed.

Each interview was broken into utterances of the participant's voice. These utterances were obtained by a turn in the interview or a sentence of the participant. On average, there were 107 utterances per participant, where each utterance was on average 4.3 seconds.

3. Methodology

3.1. Features

As in past literature, we model acoustic features in our experiments like prosodic and spectral characteristics. The acoustic features are extracted using the SRI Algemy toolkit using the time alignments of the participants' speech from the manual transcriptions. Algemy contains a graphical user interface that allows users to easily read and program scripts for the calculation of acoustic features using modular algorithms as building blocks. These blocks are strung together in directed acyclic graphs to extract the desired features.

We explored acoustic features at both the speaker level and the utterance level. Speaker level features are overall statistics for each speaker, and utterance level features are the same statistics, which are extracted for each utterance of each speaker.

Our prosodic features consist of maximum, minimum, mean, standard deviation, and range (maximum-minimum) of the following measurements, computed for each speaker at the speaker level and each utterance at the utterance level:

- Pitch, extracted as the fundamental frequency (F0) on a log scale (*F0*)
- Pitch normalized by subtracting the overall mean of the speaker on a log scale (*F0Norm*)
- Pitch normalized by subtracting the overall mean of the speaker and dividing by the overall standard deviation of the speaker on a log scale (*F0ZNorm*)
- Pitch normalized by subtracting the mean of the speaker in the first five questions of the interview on a log scale (*F0NormFirst*)
- Pitch normalized by subtracting the mean of the speaker in the first five questions of the interview and dividing by the standard deviation of the speaker in the first five questions of the interview on a log scale (*F0ZNormFirst*)
- Energy, extracted as the root mean square (RMS) of the amplitude of the signal, normalized by dividing by the maximum energy during the interview on a log scale (*Eg*)
- Voiced energy, the energy over only the voiced regions or regions with speech, normalized by dividing by the maximum energy during the interview on a log scale (*VoicedEg*)
- Energy normalized by dividing by the maximum energy during the first five questions of the interview on a log scale (*EgFirst*)
- Voiced energy, the energy over only the voiced regions or regions with speech, normalized by dividing by the maximum energy during the first five questions of the interview on a log scale (*VoicedEgFirst*)

The pitch and energy signals are extracted using the *get_f0* function, which also contains whether the frame is voiced or not for use in finding the voiced energy features. We first normalized each speaker's pitch by the overall speaker mean (*F0Norm*) and by the overall speaker mean and standard deviation (*F0ZNorm*), but we were concerned that if a participant was depressed, then this would affect his overall mean and standard deviation. Therefore, we decided to also normalize based on the first five questions of the interview (*F0NormFirst*, *F0ZNormFirst*) where the participants are asked simple questions like "*May I ask your age?*" and "*Are you retired now?*", which in theory gives their *normal* pitch for more accurate normalization. After the first five questions, the participants are asked more probing questions that could bring out more depressive speech affecting their pitch like "*Are satisfied with your work situation?*" or "*In your work or career, have you accomplished most of the things that you wanted to accomplish?*". We also performed the same first five question normalization on the energy and voiced energy features (*EgFirst*, *VoicedEgFirst*) in case the speaker's depression level affected the energy features as well.

Our spectral features, extracted over only the voiced regions, consist of maximum, minimum, mean, standard deviation, and range of the following for each speaker at the speaker level and each utterance at the utterance level:

- Spectral Tilt, defined as the slope of the line that connects the values of the formants, extracted using Praat [27] (*SpecTilt*)
- First four formants, extracted on a log scale (*F1, F2, F3, F4*). The first five formants and their corresponding bandwidths were also extracted using Praat. The last formant extracted from Praat is usually a characteristic of the background noise. Therefore, if all five formants existed, the first four formants were used in the calculation of our feature statistics.
- Corresponding bandwidths to the first four formants, extracted on a log scale (*BW1, BW2, BW3, BW4*)

This gives us a total of 90 acoustic features for use in detecting depression.

3.2. Classification

Because of the small dataset size, we use N-fold leave-one-out cross-validation to maximize the use of our data, where N is the number of speakers in the experiment. The features are fed into a support vector machine (SVM) discriminative classifier. SVM Light with default parameters was used in all our classification experiments [28]. We performed normalization on the features so no one feature affects the objective function of the SVM more than any other. The features used in the training data are normalized to mean zero and variance one. The statistics of each feature, which are found from the training data, are then applied to the test data to normalize based on the same criteria.

3.3. Feature selection

Backward elimination was used to perform feature selection on the entire set of features. The statistics (max, min, mean, standard deviation, and range) of each feature are considered one group. A group was removed one at a time until a small subset

Features	Accuracy	Specificity	Sensitivity
Baseline	50%	0%/100%	100%/0%
F1 (5)	65.6%	87.5%	43.8%
BW2 (5)	65.6%	50%	81.3%
F0Norm (5)	65.6%	31.3%	100%
F0ZNorm (5)	68.8%	43.8%	93.8%
F0NormFirst (5)	68.8%	43.8%	93.8%
F0ZNormFirst (5)	71.9%	56.3%	87.5%
F0 (5)	75.0%	75%	75%
All (90)	65.6%	56.6%	75%
F0, F0Norm, VoicedEg, F1, BW2 (25)	**81.3%**	**75%**	**87.5%**

Table 1: *Results in detecting depression at the speaker level. The number in parentheses represents the number of input features used in each classifier.*

Features	Accuracy	Sensitivity	Specificity
Baseline	50%	0%/100%	100%/0%
BW2 (5)	65.6%	43.8%	87.5%
F0 (5)	68.8%	43.8%	93.8%
All (90)	59.4%	25%	93.8%
F0, F0NormFirst, F1, F4, BW2 (25)	**81.3%**	**68.8%**	**93.8%**

Table 2: *Results in detecting depression at the utterance level. The number in parentheses represents the number of input features used in each classifier.*

of features that gave the best overall performance accuracy and reduced the number of false negatives was found. Aside from obtaining the best overall performance, we wanted to minimize the number of false negatives or *misses*, which is when a depressed speaker is labeled as a nondepressed speaker. In the depression setting, it is better to mistake a patient for depressed, even when he is not depressed, and provide help for him for this depression (label a nondepressed speaker as depressed), than mistake a patient for not depressed, when he is actually depressed, and neglect providing assistance to the patient (label a depressed speaker as nondepressed).

Aside from the accuracy, the other performance metric that is the most useful in the depression setting is sensitivity[1]. Reducing the number of false negatives is equivalent to maximizing the sensitivity. When there are 0 false negatives, sensitivity is 100%. Therefore, the goal here is to maximize both performance accuracy and sensitivity.

4. Experiments and results

In all of the following experiments, each of the 18 groups of features consisting of maximum, minimum, mean, standard deviation, and range were tried individually. In addition, all features were tried together and the best set of features that were extracted using feature selection. There are 16 severely depressed speakers and 16 nondepressed speakers. Therefore, the priors are equal, giving a baseline accuracy of 50%.

4.1. Speaker level

For this experiment, each feature was extracted at the speaker level, and the speaker depression category was used as the target to train the SVM. This resulted in very few samples for training. In Table 1, we show the seven speaker level features that had the best overall performance individually: *F0*, *F0ZNormFirst*, *F0NormFirst*, *F0ZNorm*, *F0Norm*, *BW2*, and *F1*. These were the seven groups of features that performed as well as using all 90 features. Using all 90 features resulted in an overall performance accuracy of 65.6% with a specificity[2] of 56.6% and a sensitivity of 75%.

It is not surprising that training the SVM classifier on all 90 features performed worse than some features individually. Because there are only 32 samples (31 for training and 1 for testing), using a large number of features for training the classifier causes overfitting. Feature selection was performed to reduce the feature space and cause less overfitting. After performing feature selection, the best groups of features were found to be *F0*, *F0Norm*, *VoicedEg*, *F1*, and *BW2*. These five groups of features (25 features in total) resulted in an overall performance accuracy of 81.3%[3] with a specificity of 75% and a sensitivity of 87.5%. Even though the *VoicedEg* features did not perform well by themselves, they contain information additional to what is contained in the *F0*, *F0Norm*, *F1*, and *BW2* features so the *VoicedEg* features combine well with these features and improve performance over these features by themselves.

4.2. Utterance level

In this experiment, all the utterances of the interview for a particular speaker are assigned the depression label of each speaker, and each utterance is used as a sample to extract features and train the classifier. Although our SVM results are then produced per utterance, we combine the predictions per utterance for each speaker to get an overall depression label of severely depressed or nondepressed for that speaker. We tried various approaches to combine the utterance level predictions. The easiest and most basic method we tried was averaging the prediction values for each speaker (*mean*). We also tried using the median instead of the mean in order to reduce the effect of outliers (*median*). Last, we tried normalizing based on the length of the utterance or the length of the voiced regions in each utterance to see if the longer utterances gave better results than the shorter ones (*length*, *voiced_length*, respectively). All methods seemed to give similar results, so *mean* is the combination method used to report the results in Table 2.

As seen in Table 2, the two features that had the best overall performance were *F0* and *BW2*. Using all 90 features resulted in an overall performance accuracy of 59.4% with a specificity of 25% and a sensitivity of 93.8%. We were surprised that training the SVM classifier on all 90 features performed worse than at the speaker level even though there were a large number of samples both for training and testing. We believe we are modeling noisy changes in the statistics of the features that are not seen as much at the speaker level. We would like to improve the modeling of the utterance level features because they should contain more information about the emotional state of the speaker than the speaker level features.

After performing feature selection, the best groups of features were found to be *F0*, *F0Norm*, *F1*, *F4*, and *BW2*. These five groups of features (25 features in total) resulted in an overall performance accuracy of 81.3%[3] with a specificity of 68.8% and a sensitivity of 93.8%. Even though the *F0NormFirst*, *F1*, and *F4* features did not perform well by themselves, they con-

[1] $Sensitivity = \frac{TruePositives}{TruePositives+FalseNegatives}$

[2] $Specificity = \frac{TrueNegatives}{TrueNegatives+FalsePositives}$

[3] According the the Z-test with p-value=0.005

tain information in addition to what is contained in the *F0* or *BW2* features, so they combine well with these features and improve performance over the *F0* or *BW2* features by themselves.

5. Conclusions and future work

We have found that both prosodic and spectral speech features can be good indicators of the speaker's emotional state and, in particular, can be used to predict with high accuracy whether the speaker is suffering from depression. Since the utterance level and the speaker level results were the same, we would like to model features at an even smaller level - like at word or syllable regions - to explore whether we can obtain added improvement.

In future work, we will explore more acoustic features such as speaking rate, pause duration, power spectral density, and cepstral features, and also investigate the use of lexical features. We will also verify that the features found in this work can be useful with other datasets in identifying depressed patients or other emotional problems in individuals. In addition, we intend to perform feature selection in an unbiased way using $N-1$-fold leave-one-out cross-validation and use the selected features on the remaining speaker. We would repeat this procedure N times to obtain the overall performance.

6. Acknowledgments

We thank Harry Bratt and Martin Graciarena, for developing the software to extract the features and for their added help in using this software, Ruth Krasnow, for maintaining the collection of WCGS tapes and the associated data, and Prof. Robert M. Gray, for his valuable discussions. The WCGS data collection was supported by NIA Grant AG09341, and the work presented in the paper was supported by the Defense Advanced Research Projects Agency (DARPA)/Defense Sciences Office (DSO) and Space and Naval Warfare System Center Pacific (SSC Pacific) under Contract No. N66001-10-C-4055.

7. References

[1] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, 2002, pp. 137–140.

[2] C. Müller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," in *Interspeech*, Toulouse, France, 2006, pp. 1033–1036.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[4] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I: Fundamentals, Features, and Methods*, C. Müller, Ed., no. 4343 in *Lecture Notes on Artifical Intelligence*. Springer, 2007, pp. 241–259.

[5] T. Schultz, "Speaker characteristics," *Speaker Classification I: Fundamentals, Features, and Methods*, no. 4343 in Lecture Notes on Artificial Intelligence, pp. 47–74, 2007.

[6] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining prosodic, lexical and cepstral systems for deceptive speech detection," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Pittsburgh, PA, USA, 2006.

[7] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. Tempini, and J. Ogar, "Learning diagnostic models using speech and language measures," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, Vancouver, BC, 2008, pp. 4648–4651.

[8] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *International Conference on Spoken Language Processing*, Denver, CO, USA, 2002.

[9] M. W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," in *International Symposium on Circuits and Systems*, Vancouver, Canada, 2004.

[10] J. Liscombe, G. Riccardi, and D. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Interspeech*, Lisbon, Portugal, 2005.

[11] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, 2007.

[12] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, 2007.

[13] M. H. Sanchez, G. Tur, L. Ferrer, and D. Hakkani-Tur, "Domain adaptation and compensation for emotion detection," in *Interspeech*, Makuhari, Japan, 2010.

[14] J. Darby and H. Hollien, "Vocal and speech patterns of depressive patients," *Folia phoniat*, vol. 29, pp. 279–291, 1977.

[15] J. Darby, N. Simons, and P. Berger, "Speech and voice parameters of depression: A pilot study," *J. Commun. Disorders*, vol. 17, pp. 75–85, 1984.

[16] A. Ozdas, R. G. Shiavi, D. M. Wilkes, M. K. Silverman, and S. E. Silverman, "Analysis of vocal tract characteristics for near-term suicidal risk assessment," *Methods of Information in Medicine*, vol. 43, pp. 36–38, 2004.

[17] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, September 2004.

[18] 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 Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, July 2000.

[19] E. M. II, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, January 2008.

[20] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. la Torre, "Detecting depression from facial actions and vocal prosody," in *International Conference on Affective Computing and Intelligent Interaction*, 2009.

[21] L. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, USA, 2010, pp. 5154–5157.

[22] H. K. Keskinpala, T. Yingthawornsuk, D. M. Wilkes, R. G. Shiavi, and R. M. Salomon, "Screening for high risk suicidal states using mel-cepstral coefficients and energy in frequency bands," in *European Signal Processing Conference*, Poznan, Poland, 2007, pp. 2229–2233.

[23] T. Yingthawornsuk and R. G. Shiavi, "Distinguishing depression and suicidal risk in men using GMM based frequency contents of affective vocal tract response," in *International Conference on Control, Automation and Systems*, Seoul, Korea, 2008, pp. 901–904.

[24] R. H. Roseman, M. Friedman, R. Straus, M. Wurm, R. Kositchek, W. Hahn, and N. T. Werthessen, "A predictive study of coronary heart disease: The western collaborative group study," *Journal of the American Medical Association*, vol. 189, no. 1, pp. 15–22, 1964.

[25] L. S. Radloff, "The CES-D scale: A self-report depression scale for research in the general population," *Applied Psychological Measurement*, vol. 1, pp. 385–401, 1977.

[26] T. Furukawa, T. Hirai, T. Kitamura, and K. Takahashi, "Application of the center for epidemiologic studies depression scale among first-visit psychiatric patients: a new approach to improve its performance," *Journal of Affective Disorders*, vol. 46, pp. 1–13, 1997.

[27] "Praat: Doing phonetics by computer," *http://www.fon.hum.uva.nl/praat/*.

[28] "Svmlight support vector machine toolkit," *http://svmlight.joachims.org*.