INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Speech Features for Depression Detection

Saurabh Sahu [1] *and Carol Espy-Wilson* [1]

[1] Institute for Systems Research, Department of Electrical and Computer Engineering,
University of Maryland, College Park
ssahu89@umd.edu, espy@isr.umd.edu

Abstract

In this paper we discuss speech features that are useful in the detection of depression. Neuro-physiological changes associated with depression affect motor coordination and can disrupt articulatory precision in speech. We use the Mundt database and focus on six speakers in the database that transitioned between being depressed and not depressed based on their Hamilton depression scores. We quantify the degree of breathiness, jitter and shimmer computed from an AMDF based parameter. Measures from sustained vowels spoken in isolation show that all of these attributes can increase when a person is depressed. In this study, we focused on using features from free-flowing speech to classify the depressed state of an individual. To do so we looked at vowel regions that look the most like sustained vowels. We train an SVM for each speaker and do a speaker dependent classification of the test speech frames. Using the AMDF based feature we got a better accuracy (62-87% frame-wise accuracy for 5 out of 6 speakers) for most speakers than 13 dimensional MFCC along with its velocity and acceleration coefficients. Using the AMDF based feature, we also trained a speaker independent SVM which gave an average accuracy of 77.8% for utterance based classification.

Index Terms: clinical depression, articulatory control, speech features, SVM

1 Introduction

Suicide was the tenth leading cause of death in 2013 [1] and depression is the most common precursor to suicide [2]. Depression is very common in the young population also (age 13-20 years). It is estimated that a depressive episode affects between 14-30% of young females and 13-17% of males in a 12-month period, with 2.7-8.9% of females and 1.6-9.0% of males experiencing more severe depression [3]. These facts point to the need for better ways of diagnosing and monitoring depression. Our goal is to develop speech biomarkers that we will eventually combine with physiological signals and language analysis to come up with a robust system for detecting and monitoring depression.

Several studies have been conducted to find features that help distinguish depressed patients from non-depressed patients. One set of studies have looked at depressed patients relative to a control group. In Low et.al. [3] they used MFCCs and their deltas as features with a Gaussian Mixture Model (GMM) based classifier. They reported an accuracy of 51% for the depressed group and an accuracy of 61% for the control group. In [4], the same authors obtained a gender dependent accuracy of 78% (males) and 74.7% (females) for depressed subjects when they used a combination of Teager energy (including velocity and acceleration coefficients), F0, log energy, shimmer, spectral flux and spectral roll-off. Moore et al. [5] used statistical measures like mean, median, standard deviation, interquartile range on prosodic features like pitch, energy and speaking rate to capture emotional variations. They obtained 75% accuracy or higher. When they incorporated glottal and vocal tract features, the accuracy increased to at least 90%. Ozdas et al. [6] reported classification accuracy of 80% for control and suicidal classes using jitter as feature and an accuracy of 90% between control and depressed classes using spectral slope as feature.

In several other studies, researchers have used the Mundt database [7] which consists of only subjects who suffer from depression, some of whom transition from being depressed to being not depressed. These studies have looked at features like formants, formant bandwidths, jitter, shimmer, aspiration noise among others. Cummins et al. [10] reported a binary classification accuracy of ~70% (depressed and non-depressed classes) using energy and spectral features with GMMs. Quatieri and Malyska [9] used the sustained vowel sounds in the database to compute correlations between jitter, shimmer and aspiration noise and the Hamilton depression rating scale (HAM-D) scores [11,12]. They did not find a strong correlation of jitter with the HAM-D scores, but they did find a strong correlation between shimmer and the HAM-D scores.

In the current study, we work with the Mundt database to track the changes in the depressed states of the subjects that made a transition. We are looking at source features like shimmer and jitter, which are quantified in a more robust manner without explicit detection of perturbations in F0. In addition, we introduce a new breathiness measure that quantifies the amount of aspiration generated by incomplete closure of the vocal cords. The motivation for studying these features in particular is based on the fact that neuro-physiological changes associated with depression affect motor coordination and therefore the disruption of articulatory control and kinematics [8, 9, 10]. We believe that the articulatory imprecision can lead to changes in vocal cord vibration that will result in more jitter, shimmer and breathiness.

Our paper is organized as follows. In Section 2 we describe the details of the database we are using for this study. In Section 3, we describe the features that we are studying in this paper and the methods using which we are obtaining them. Section 4 describes the experiments we did and section 5 describes the results from these approaches. In section 6 we provide conclusions and directions for future work.

http://dx.doi.org/10.21437/Interspeech.2016-1566

2 Database

For this study, we used the Mundt database [7] which include data collected from 35 physician-referred patients undergoing treatment for depression. The patients were assessed weekly once over a period of 6 weeks.

The audio files collected over telephone were sampled at 8 kHz. For our study, we use four sustained vowel utterances (/a/, /i/, /u/, /ae/) and three or four utterances where the subjects talked freely about their emotional state, physical state and ability to function in the preceding week. The free speech utterances are 30 seconds to 2 minutes long while the sustained vowel sounds are 5-6 seconds in duration.

Severity of depression was measured using the HAM-D. It captures items such as the extent of depressed mood, psychomotor retardation and weight loss on a scale of 0-2 or 0-4. Subject sessions were labeled as depressed if their HAM-D score was 17 or greater, and as non-depressed if their score was 7 or lower, with scores of 8 to 16 excluded because their depression status is ambiguous [8]. Based on the above criteria, only 6 patients showed the transition from depressed state to non-depressed state during the course of their treatment. The patient ID's and the corresponding days are listed in Table 1.

Table 1. *List of patients and the days on which they were depressed or not depressed based on HAM-D scores*

Patient ID	Depressed	Not Depressed
101	Day 14	Day 42
111	Day 00	Day 14, 28, 47
119	Day 00	Day 31
123	Day 00	Day 42
127	Day 00	Day 42
128	Day 00	Day 27

3 Features

In this study we focused on the excitation parameters: jitter, shimmer and breathiness. Jitter is the cycle to cycle variability in the duration of the pitch period. For N consecutive glottal cycles the jitter factor is given by

$$Jitter = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}T_i} \qquad (1)$$

Shimmer is the cycle to cycle variability of the pitch period amplitude. For N consecutive glottal cycles, the shimmer factor is given by,

$$Shimmer = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}A_i} \qquad (2)$$

Breathiness is aspiration noise in the frequency range of F3 and above (> 2500 Hz in general) and is due to incomplete closure of the vocal folds [13]. Some degree of jitter, shimmer and breathiness occurs naturally in speech. However, due to psychomotor retardation which affects articulatory imprecision, the degree of variation can increase. To quantify these parameters, we used measures based on the Average Magnitude Difference Function (AMDF) that were developed for our Aperiodicity/Periodicity/Pitch Detector [14].

3.1 Average Magnitude Difference Function

AMDF computes the difference between the waveform and a lagged version of itself. If $\gamma_n(k)$ is the AMDF value for lag k, it is computed by the following equation

$$\gamma_n(k) = \sum_{m=-\infty}^{\infty}|y_n(m) - y_n(m-k)| \qquad (3)$$

where, $y_n(m)$ is the windowed version of signal $x(n)$ centered at n using a window $w(m)$, i.e.

$$y_n(m) = x(n+m)w(m) \qquad (4)$$

For a perfectly periodic signal, when k is equal to some multiple of the period, $\gamma_n(k)$ will be equal to zero and the AMDF dip will be equal to 1. For speech which is quasiperiodic $\gamma_n(k)$ will be greater than 0 and the AMDF dip will be less than 1 (see Fig. 1). To quantify the jitter, shimmer and breathiness, we divide the speech spectrum into 60 channels, compute the AMDF for each channel and then sum the dips across the channels. The AMDF dip profile for a frame from a vowel (left) and an unvoiced fricative (right) are shown in Fig. 1. Note that the dip profile for the vowel shows prominent clusters of dips at T_0, $2T_0$ and $3T_0$ where T_0 is the fundamental period of the vowel. The dip profile for the unvoiced fricative, on the other hand, shows no such clusters since the fricative is aperiodic.

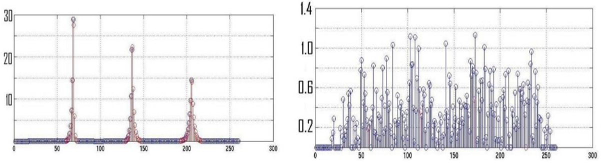

Figure 1: *Dip profile for a frame from a vowel (left) and unvoiced fricative (right)*

3.2 Measuring jitter, shimmer and breathiness from dip profiles

Jitter quantifies the variation in F0 across pitch periods. Hence, if its value is low, the signal will be more periodic and we get a dip profile with thin clusters, especially for the first cluster which is based on adjacent pitch periods. On the other hand, for a signal having a high jitter value, the cluster width will be larger. In Figure 2 we compare the dip profiles averaged over across all frames for the vowel /a/ the person produces on the days she is depressed (red) vs when she feels better (blue) and we can see that the cluster spread of the first peak is higher when she is depressed.

Shimmer quantifies the variation in amplitude across pitch periods. Hence, in case of less shimmer consecutive pitch periods will have similar amplitudes which leads to a higher value for the first cluster peak compared to the case when there is less shimmer. In Figure 2 we compare the dip profiles averaged over across all frames for a vowel the person produces on the days she is depressed (red) vs when she feels better (blue) and we can see that the cluster height of the first peak is smaller when she is depressed.

Hence, to measure jitter we measure the spread of the first dip cluster (blue arrows in fig 2). by calculating the standard deviation of all the indices (lag values) that lie within a certain tolerance region of the index corresponding to the first cluster peak and have a value > 0.2 times the height of the first cluster peak. To quantify shimmer, we measured the height of the first cluster (black arrows in fig 2). It should be kept in mind that larger cluster height indicates less shimmer. Since breathiness shows up as aspiration noise, to measure it we summed the value of the dips occurring outside a certain tolerance region of the cluster peaks (> 0.5ms but < 1ms). These are denoted by circled regions in Fig.2.

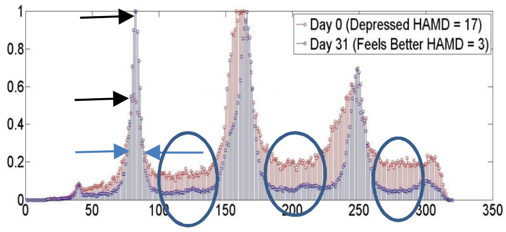

Figure 2*: Dip profiles for when speaker is depressed (red) and not depressed (blue)*

4 Experiments and Results

All the speech samples (vowels and free speech) were normalized to have zero mean and unit standard deviation before extracting the AMDF features.

4.1 Measurements in Sustained Vowels

In this study, we segment the sustained vowel sound into 20 ms segments with frame interval being 10ms. For each segment we compute the AMDF dip profile and quantify the values of jitter, shimmer and breathiness. In the figures below (3, 4 and 5) we have compared the temporal variation of these values for the same speaker producing the same vowel on the days when she is depressed (blue) and not depressed (red). It can be observed that for most of the time frames, the value of jitter, shimmer and breathiness is higher when the person is depressed. Note that for shimmer, we are plotting the cluster height which decreases as shimmer increases.

We calculated the mean of the cluster spread, cluster height and aperiodic energy across time for each of the sustained vowel utterances. We took the mean of these numbers obtained for the four vowel sounds (mentioned in section 2) to get a single value per person per day. The values are compared in Table 2 below.

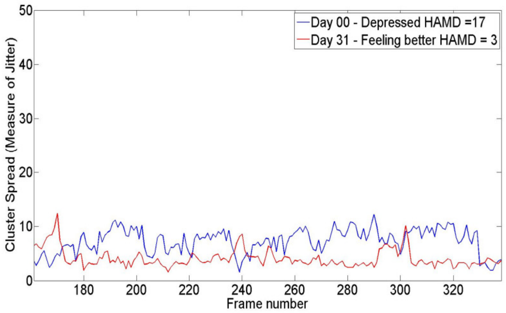

Figure 3: *Temporal variation of jitter*

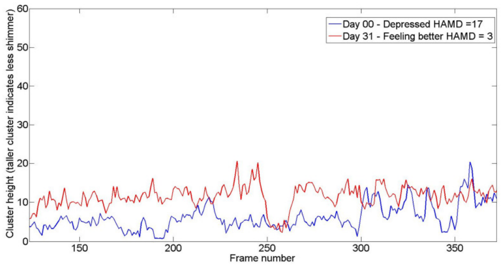

Figure 4: *Temporal variation of shimmer*

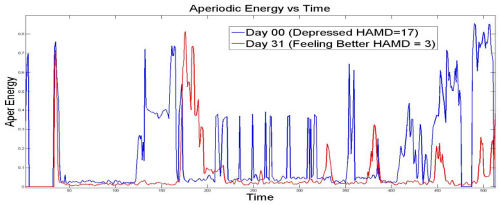

Figure 5: *Temporal variation of aperiodicity (breathiness)*

Table 2. *Mean values for different patients on days they were depressed (D) and not depressed (ND)*

ID	Day No.	Mean cluster spread	Mean cluster height	Breathiness
101	14(D)	9.75	11.74	11.16
	42(ND)	8.25	15.94	12.3
111	00 (D)	13.75	8.8378	10.87
	14(ND)	11.5	10.026	17.93
	28(ND)	10.5	13.33	9.09
	47(ND)	10.75	11.874	13.72
119	00(D)	16.75	4.305	19.84
	31(ND)	11.5	6.041	14.96
123	00(D)	9	13.51	8.18
	42(ND)	10.5	13.14	5.08
127	00(D)	7.5	14.38	9.84
	42(ND)	6.25	15.29	9.99
128	00(D)	10.25	8.0327	15.01
	27(ND)	8.25	8.4035	10.74

4.2 Measurement in Free Flowing Speech

Initially, to quantify jitter, simmer and breathiness in free flowing speech, we used the same procedures described in Section 4.1 for every periodic frame. However, we found that the results were not consistent due to high degree of variability in speech due to coarticulation. We also tried using every periodic frame for our SVM experiment mentioned below but the SVM training wouldn't converge. This may be due to the high variability seen by the SVM owing to co-articulation. So, we limited our analysis to vowel regions that resembled sustained vowels, i.e. the vowels were greater than 70 ms in duration and the pitch frequency changed by less than 15 Hz over the course of the vowel. These regions were selected manually. With these limitations, we were able to see similar results as those obtained for the sustained vowels for some of the speakers as shown in table 3.

4.3 Classification of Depressed and Non-depressed utterances in Free Flowing Speech

The dip profiles of the frames (input dimension of 320) from regions that resembled sustained vowels were used as features for binary support vector classification into depressed and non-depressed classes. Since, the amount of data is less; we used SVM instead of neural networks (linear kernel with Sequential Minimal Optimizer used to find the hyperplane). 70% of the frames for each speaker were used for training and the rest was used for testing. We also used the 13 dimensional MFCC along with its velocity and acceleration coefficients (39 coefficients in total) as features to compare the performances. Based on the results we obtain for the test frames, we decide whether an utterance was spoken in a depressed or a non-depressed state of mind. If majority of the frames from an utterance were classified as depressed then the utterance was also classified as depressed. The table below shows the number of utterances (not frames which are actually used for training and testing) used for training and testing. We used 26 utterances and 18 utterances for testing.

Table 3. *Mean values for different patients on days they were depressed (D) and not depressed (ND) computed from voiced segments in free speech*

ID	Day No.	Mean cluster spread	Mean cluster height	Breathiness
101	14(D)	19.8	7.84	8.17
	42(ND)	22.2	6.75	8.71
111	00 (D)	16	6.96	17.33
	28(ND)	14.57	10.26	11.43
119	00(D)	18.53	5.03	15.74
	31(ND)	14.59	5.27	15.64
123	00(D)	11.6	7.88	11.67
	42(ND)	16	10.28	10.04
127	00(D)	11.25	8.25	5.23
	42(ND)	15.17	8.83	5.38
128	00(D)	14.45	5.92	11.2
	27(ND)	13.08	9.38	6.47

Classification results using free utterances are shown in tables below. In the first stage we trained six SVMs for each of the six speakers and then tested on frames for that particular speaker (Speaker Dependent SVM). In the second stage we combined the training frames from all speakers and trained a single SVM (speaker Independent SVM).

Table 4. *Frame wise classification accuracy (%) for different patients (Speaker dependent classification)*

Patient ID	Dip Profiles	MFCC +vel+ acc	MFCC+vel +acc+dip profiles	MFC+ dip profiles
101	70.45	84.4	70.45	76.13
111 (Day 0,14)	47.83	33.33	60.87	65.21
111 (Day 0,28)	62.32	51.72	56.52	62.31
111 (Day 0,47)	56.18	57.27	56.18	57.3
119	80.9	57.54	85.71	85.03
123	70.62	100	81.12	81.12
127	45.07	44.2	45.07	45.07
128	87.5	72.26	91.07	89.3

Table 5. *Utterance wise classification accuracy (%) for different patients (Speaker dependent classification)*

Patient ID	Dip Profiles	MFCC + vel + acc	MFCC+ vel + acc + dip profiles	MFCC + dip profiles
101	100	100	100	100
111 (Day 0,28)	66.67	66.67	66.67	66.67
119	100	60	80	100
123	100	100	100	100
127	33.33	33.33	33.33	33.33
128	100	100	100	100

Table 6. *Utterance wise classification accuracy (%) for different patients (Speaker independent classification)*

Patient ID	Dip Profiles	MFCC + vel + acc	MFCC+ vel + acc + dip profiles	MFCC + dip profiles
101	66.67	33.33	66.67	66.67
111 (Day 0,28)	33.3	66.67	66.67	66.67
119	80	40	80	100
123	100	100	100	100
127	100	33.33	100	33.33
128	100	50	50	50

5 Conclusion and future work

Based on the comparison in Table 2 it can be seen that in 5 out of 6 cases there is an increase in jitter and shimmer when the person is depressed. Hence, it appears that these features can be useful in the detection and monitoring of depression. Three out of six cases show an increase in aperiodic energy or breathiness in the vowel sounds when depressed. For speakers 101 and 111 who show a reverse trend in breathiness, we found portions of vowel which are creaky on the days they were depressed (F0 was around 80Hz during creaky portions). Thus, we think depression can be associated with both a breathy and creaky voice quality.

We have used dip profiles and MFCC features to train SVM for classification. For some speakers, MFCC features work better than dip profiles and for some it's the other way round. Combining the features is a good idea but it deteriorates the accuracy in some cases (speaker 101 and 123). From table 4 it can be seen that for most speakers, combining dip profiles with MFCC along with its velocity and acceleration coefficients gave worse frame wise classification accuracy than combining dip profiles with MFCC. It can be observed the utterance level classification is 100% for 4 out of 6 speakers in case of speaker dependent SVC. Speaker 127 gives us poor classification result which is not surprising because the cluster spread (jitter) increases when the person is not depressed (reverse trend) and there is not much change in shimmer and breathiness. In case of speaker independent the average accuracy is ~80% (14 out of 18 utterances classified correctly) which shows us that the dip profiles contain important information about the depressed state of a person. Speaker 127 has a better accuracy when it comes to speaker independent system as compared to speaker dependent system. This might be because of the fact that we have more data to train the SVM with in case of speaker independent system.

In future, we plan to incorporate more features (e.g. OpenSmile features or may be a subset of them) for training. However, it might degrade our system because of few amounts of data. One way to counter that is instead of having a binary classification; we can have a multi class classification based on HAM-D scores. That way we can include more speakers and hence more training data and we might be able to do a more fine grained classification to track speaker changes. We can also try using a different classification scheme instead of SVM once we increase the amount of training data. We plan to incorporate speaking rate with our existing features as it gave us a good correlation with HAM-D score based on preliminary studies [15]. Apart from that, we can investigate other voice quality features like those related to vocal tract characteristics.

6 Acknowledgement

This research was supported by a University of Maryland Advance Seed Grant and a College of Engineering Seed Grant. We thank Dr. Jim Mundt for the database.

7 References

[1] Centers for Disease Control and Prevention (CDC). Web-based Injury Statistics Query and Reporting System (WISQARS) [Online]. (2013, 2011) National Center for Injury Prevention and Control, CDC (producer). Available from http://www.cdc.gov/injury/wisqars/index.html.

[2] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, 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, 2000.

[3] L.-S.A. Low, N. C. Maddage, M. Lech, N. Allen, "Mel Frequency Cepstral Feature and Gaussian Mixtures for Modeling Clinical Depression in Adolescents," *IEEE International Conference on Cognitive Informatics (ICCI)*, 2009, pp. 346-350.

[4] L.-S.A. Low, N. C. Maddage, M. Lech, L. Sheeber, N. Allen, "Influence of Acoustic Low-level Descriptors in the Detection of Clinical Depression in Adolescents", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2010, pp. 5154-5157.

[5] E. Moore II, M. Clements, J. Peifert and L. Weisser, "Analysis of Prosodic Variation in Speech for Clinical Depression", *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, 2003, pp. 2925-2928.

[6] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, D. 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, 2004.

[7] J. Mundt, P. Snyder, M. S. Cannizaro, K. Chappie, D. S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," *J. Neurolinguistics,* 20(1): 50-64, 2007.

[8] B. Helfer, T. Quatieri, J. Williamson, D. Mehta, R. Horwitz, B. Yu, "Classification of depression state based on articulatory precision", *INTERSPEECH 2013 - 14th Annual Conference of the International Speech Communication Association, September 6–10, Lyon, France, Proceedings*, 2013, pp. 2172-2176.

[9] T. F. Quatieri and N. Malyska, "Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity", *INTERSPEECH 2012 – Special Session: Analysis of Spoken Disorders in Health Applications, Portland, Oregon*, 2012.

[10] N. Cummins, J. Epps, E. Ambikairajah, "Spectro-Temporal Analysis of Speech Affected by Depression and Psychomotor Retardation", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7542-7546.

[11] N. Cummins, J. Epps, V. Sethu, M. Breakspear, & R. Goecke, "Modeling Spectral Variability for the Classification of Depressed Speech", Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech), Lyon, France, 2013.

[12] M. Hamilton., "A rating scale for depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, 23:56–62, 1960.

[13] D. H. Klatt, L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *Journal of the Acoustical Society of America*, Vol. 87, No. 2, pp. 820-857, 1990.

[14] O. Deshmukh, C. Espy-Wilson, A. Salomon and J. Singh, "Use of Temporal Information: Detection of the Periodicity and Aperiodicity Profile of Speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp. 776-786, 2005.

[15] S. Sahu, C. Espy-Wilson, "Effect of Depression on Syllabic rate of Speech", J. Acoust. Soc. Am. 138, 1781 (2015)