Speech-Based Assessment of PTSD in a Military Population using Diverse Feature Classes

Dimitra Vergyri[1], Bruce Knoth[1], Elizabeth Shriberg[1], Vikramjit Mitra[1], Mitchell McLaren[1], Luciana Ferrer[1,2], Pablo Garcia[1], Charles Marmar[3]

[1]SRI International, Menlo Park, CA
[2]CONICET and University of Buenos Aires, Argentina
[3]NYU Langone Medical Center, Department of Psychiatry, New York, NY

{dimitra.vergyri,bruce.knoth,elizabeth.shriberg,vikramjit.mitra,mitchell.mclarren}@sri.com, lferrer@dc.uba.ar, pablo.garcia@sri.com, Charles.Marmar@nyumc.org

Abstract

There is a critical need for detection and monitoring of Post-Traumatic Stress Disorder (PTSD) in both military and civilian populations. Current diagnosis is based on clinical interviews, but clinicians cannot keep up with the growing need. We examined the feasibility of using speech for assessment in a military population. We analyzed recordings of the Clinician-Administered PTSD Scale (CAPS) interview from military personnel diagnosed as PTSD positive versus negative. Three feature types were explored: frame-level spectral features, longer-range prosodic features, and lexical features. Results using gaussian backend, decision tree and neural network classifiers (for spectral and prosodic features) and boosting (for lexical features) showed an accuracy of 77% correct in split-half cross validation experiments, a figure significantly above chance (which was 61.5% for our dataset). Spectral and prosodic features outperformed lexical features, and feature combination yielded further gains. An important finding was that sparser prosodic features offered more robustness than acoustic features to channel-based variation in the interview recordings. Implications and future work are discussed.

Index Terms: PTSD assessment, mental health assessment.

1. Introduction

PTSD has significant impact on patients, their families and their communities. The ability to provide screening, diagnosis, and treatment to those at risk for PTSD has thus become an urgent concern for both military and civilian populations. PTSD poses a huge financial cost to the Department of Veterans' Affairs [1] and impacts troops' readiness. In addition an estimated seven percent of the U.S. civilian population will develop PTSD after exposure to traumatic events [2].

Even in a clinical setting, correct identification of PTSD is challenging. Few quantitative measures exist for diagnosis, which relies largely on self-reporting by patients during clinical interviews. The Clinician-Administered PTSD Scale (CAPS) [3] structured clinical interview is considered the gold standard for inferring a dichotomous PTSD diagnosis or a severity scale ranking for individual symptoms, symptom clusters, or the entire syndrome. CAPS was shown to have a 79% overall agreement with clinical diagnosis, with a kappa coefficient of .58 [4].. Nevertheless, the basis for CAPS is self-reporting, which may be degraded by distortions in memory and self-perception [5], or financial and social incentives [6]. Moreover, the interview requires a visit to the clinician's office, which some patients may not be willing or able to make. Thus, there is a need for a more objective, cost- and time-efficient means of PTSD assessment.

In this paper, we use automatic analysis of the patient's speech to determine PTSD status. Speech is natural, noninvasive, cheap, and can be obtained via phone for analysis at a distance. This work builds on prior research that focused on using speech signal to detect mental health [7] and emotion. For emotion detection, previous research used prosodic features such as speaking rate, pitch, energy or intensity, and pause duration [8-13] as well as other acoustic features such as voice quality [10,11], spectral features [8] and Mel frequency cepstral coefficients (MFCCs) [12].. Changes in prosodic features (pitch, energy, speaking rate), spectral features (formants, their corresponding bandwidths, power spectral density, spectral tilt), and MFCCs were also found useful in depression detection [14-19] and also recent PTSD-detection work [20, 21].

2. Data

We used recordings of interviews conducted at NYU, a site of the Systems Biology PTSD Biomarkers Consortium (SBPBC). PTSD is assessed via the CAPS interview and co-morbidities via the Structured Clinical Interview for DSM-IV, Patient edition (SCID-P) [22]. Participants were recruited among Operation Enduring Freedom/Operation Iraqi Freedom (OEF/OIF) veterans of age 20-60 with PTSD symptoms as a result of witnessing or experiencing a horrific or life-threatening event in combat, or veterans who were engaged in combat but do not have PTSD symptoms (to be part of the control group). After a phone screen interview, eligible candidates were invited for a longer interview that included CAPS, memory testing, magnetic resonance imaging (MRI) scan, and blood and urine samples. Exclusion criteria included current alcohol or drug dependence; history of any psychiatric disorder with psychotic features, bipolar disorder, or obsessive-compulsive disorder; currently suicidal or an attempt within the past year; neurologic disorder or major medical illness.

Each participant's structured interview (2-3 hours long) was recorded and scored per the protocol of the Biomarkers study. For our experiments, we used 39 male participants: 24 PTSD negative (PTSD-, CAPS score below 20) and 15 PTSD positive (PTSD+, CAPS score above 40). Participants with

Table 1. *Summary of frame-level features.*

Feat. Name	Dimension	Type
DOCC	13	Perceptually motivated spectral features
GCC	13	
NMCC	13	Perceptually motivated speech modulation information
MMeDuSA	16	
DeepTV	8	Articulator information
AP	13	Acoustic phonetic information
Kaldi-f0	2	Pitch and voicing information

score in between were excluded for this study since the CAPS score could not reliably conclude diagnosis (which was considered "moderate" PTSD). We manually time-marked the interviewer's and patient's speech in each interview and used only the patient's speech. After removing the interviewer's speech and short yes/no patient segments, we kept between 10-115 minutes of speech per speaker (average 38 mins/ speaker).

3. Method

3.1. Speech Feature Extraction

We examine the use of three speech feature types for PTSD detection: low-level spectral features, computed from short duration speech frames, e.g. every 20ms; high level temporal features, computed from long duration samples, e.g. at the utterance level; and lexical features.

Frame-level features (Table 1):

1. The Damped Oscillator Cepstral Coefficients (**DOCCs**) [23] model the dynamics of hair cells within the human ear as forced damped oscillators which detect the motion of incoming sound waves and create a spectral representation of the input to the auditory nerves.
2. The Normalized Modulation Cepstral Coefficients (**NMCCs**) [24] track the amplitude modulations (AM) of time-domain subband speech signals use the Discrete Energy Separation algorithm (DESA) [25] to create a modulation spectrum for feature extraction.
3. The Modulation of Medium Duration Speech Amplitudes (**MMeDuSA**) [26][27] track the subband AM and the overall summary modulation of speech using a medium duration analysis window. These are important for tracking speech activity and locating events such as vowel prominence, stress, etc. [27].
4. The Gammatone Cepstral Coefficients (**GCCs**) use a perceptually-motivated gammatone filterbank to compute a gammatone spectrum and extract cepstral features after performing DCT on root-compressed filterbank energies.
5. **DeepTVs** [28] are articulatory features obtained from a Deep Neural Network (DNN) which estimates constriction locations and degrees at various parts of the vocal tract, capturing information such as glottal and velic opening/closing, labial constriction, lip rounding, tongue tip and tongue blade constrictions.
6. The Acoustic Phonetic (**AP**) features [29], analyzed at a 5 msec frame rate with a 10 msec analysis window, represent information such as reflection coefficients, mean Hilbert envelope, periodic energy, aperiodic energy [30], nasal energy [31], etc.
7. The **Kaldi Pitch tracker** [32] is part of the kaldi pitch recognition toolkit [33] and provides a two-dimensional output: pitch tracks and a normalized cross-correlation function that gives an indication of voicing information.

Table 2. *Summary of longer-range prosodic features*

Feat. Name	Type	Extraction region in segment	Dimension
tilt	vocal effort	voiced frames	5
dev-onset	vocal effort	voiceless->voiced transitions	1
dev-offset	vocal effort	voiced->voiceless transitions	1
en-con	rhythmicity	200ms window	7
f0	pitch	frame	1
f0pk	pitch at peaks	frames at peaks	1
f0pk-stats	rhythmicity, rate, pitch	peak locations	9
int	intensity	frame	1
intpk	intensity at peaks	frames at peaks	1
intpk-stats	rhythmicity, rate, intensity	peak locations	9

Longer-range prosodic features (Table 2):

1. **Tilt** features, extracted from voiced frames, aim to capture vocal effort in a manner quasi-robust to extrinsic session variability. The first three, H2-H1, F1-H1 and F2-H1, reflect lower-order harmonics and formants given the microphone and room conditions. The last two are the spectral slope per frame and the difference between the maximum of the log power spectrum and the maximum in the 2kHz-3kHz range.
2. The **Dev** features target session-normalized vocal effort detection using the difference in log energy at the transition between voiceless and voiced speech. Dev-onset features are extracted at the boundaries from voiceless to voiced, and dev-offset features at the boundaries from voiced to voiceless.
3. Energy contour (**En-con**) features aim to capture rhythmicity by looking at the periodicity of energy peaks within each segment. They model the contour of 10-ms c0 and c1 MFCCs; each cepstral stream is mean-normalized over the utterance, making it robust to absolute differences over both full-session and within-session segments. A discrete cosine transform is then taken over a 200-ms sliding window with a 100-ms shift. Vector components comprise the first 5 and 2 bases from the DCT over each window of c0 and c1, respectively.
4. The **f0**, f0-peak (**f0pk**) and **f0pk-stats** are pitch-related features. F0 is computed for voiced regions using default parameter settings for the snack RAPT-style pitch tracker [34]. F0pk features record values found by a peak-picking algorithm; f0pk-stats include mean, max and standard deviation of peak pitches, as the temporal distribution of pitch-accented syllables in the segment.
5. Intensity-related features **int**, i**ntpk**, and **intpk-stats** are computed similarly to the pitch features. Intensity is computed using default intensity parameters in Praat [34]. Unlike pitch, raw intensity values reflect both the speaker and the recording session. Thus int, intpk, and mean, max and standard deviation in intpk-stats are expected to partially reflect extrinsic factors.

Lexical features

We ran our automatic speech recognition system [35] on the interview data. Since the data was insufficient for identification of word specific features, the recognized words

Table 3. *Examples of selected word features. The class {X} is named after a frequent within-class word.*

Class N-grams
- {EXPERIENCES} that I - not {ABLE} - {CAN} {ACCOMPLISH} {ANYTHING} - {CAN} {AFFECT} it - {AFTERWARD} it's causing - to {ANALYZE} it
Class definitions
{EXPERIENCES}={age, background, dreams, duties, experiences, faces, family, feelings, hobbies, homework, ideas, life, marriage, parents, relationship(s), …} {ABLE}={able, allowed, exposed, ready, trying, wanting…} {CAN}={can, can't, cannot, couldn't} {ACCOMPLISH}={accomplish, do, explain} {ANYTHING}={anything, something, whatever, whatnot} {AFFECT}={affect, bother, harm, …} {AFTERWARD}={afterward, afterwards, again, now, …}

were automatically clustered in classes based on their context (distributional statistics), The word classes were used as input to a classifier (AdaBoost) [36], which selected informative class N-grams to include as features in a PTSD prediction model. This classifier was tested using leave-one-speaker-out cross validation. Some examples of the common class N-grams selected across different cross-validation "folds" are listed in Table 3.

3.2. Audio Classification

Each audio file from a patient's interview was represented as a finite length vector by averaging the feature vectors over time[1]. We evaluated several classifiers and selected three that provided best performance across all features. These included a neural network (NN) with a single hidden layer, an extratrees decision tree (DT) classifier (from the sklearn python package [37]) which averages results from an ensemble of random trees classifiers trained on different subsets of the training data, and finally a Gaussian backend (GB) [38], which models data from each class as a Gaussian with class-specific means and shared covariance and has been commonly used for speaker identification. A single Gaussian mixture was used for the GB classifier.

A separate classifier was trained using each feature. We implemented a simple system/feature fusion by averaging the scores obtained from different systems. System/feature selection for fusion purposes was performed using exhaustive search across acoustic feature sets using the GB classifier. A more appropriate form of fusion such as logistic regression was attempted, but data limitations prevented splitting to a level appropriate for removing bias through this method of fusion.

4. Results

Each classifier was evaluated based on split-half accuracy: we split the data in two sets and used one for training and the other for testing, and then reversed their roles, reporting results in both cases. Chance results, based on the prior distribution in each split was 60% for splitA and 63% for splitB. Accuracy results by feature are reported in Table 4. Since in most cases the best results were obtained with the GB classifier, we used only that classifier for the final fusion experiments. Further investigation is needed to optimize parameter settings for the other classifiers. The lexical system performed around chance level. We believe a lexical based classifier needs more data to learn relevant features, across interviews.

Table 4. *Accuracy results for different features and different classifiers*

	GB classifier		DT classifier		NN classifier	
Features	splitA	splitB	splitA	splitB	splitA	splitB
Frame-level						
DOCC	0.50	0.68	0.45	0.58	0.40	0.42
NMCC	0.50	0.52	0.50	0.37	0.50	0.38
MMeDuSA	0.60	0.79	0.55	0.42	0.50	0.58
GCC	0.50	0.53	0.45	0.42	0.50	0.37
DeepTV	0.45	0.63	0.45	0.68	0.30	0.68
AP	0.50	0.37	0.35	0.32	0.50	0.42
Kaldi-F0	0.70	0.74	0.60	0.42	0.60	0.53
F0	0.60	0.58	0.45	0.37	0.60	0.58
Intensity	-	-	0.55	0.42	0.50	0.42
MFCC	0.55	0.78	-	-	-	-
Segment-level						
tilt	0.60	0.42	0.50	0.58	0.65	0.58
dev-off-stats	0.65	0.74	0.75	0.68	0.75	0.74
dev-on-stats	0.80	0.63	0.70	0.63	0.90	0.63
en-con	0.50	0.53	0.50	0.47	0.50	0.58
f0pk	0.35	0.42	0.45	0.42	0.40	0.42
f0pk-stats	-	-	0.35	0.42	0.40	0.42
Intpk	0.56	0.47	0.70	0.63	0.60	0.42
intpk-stats	-	-	0.35	0.42	0.50	0.53
Lexical	AdaBoost classifier					
class-ngrams	0.50	0.63				

Table 5. *Selected features for a 4-way fusion with results from both patient and clinician segments. The 4-best feature selection and evaluation was done using the GB classifier. Fusing the 4-best system with the lexical-based system did not give additional improvements.*

	Accuracy			
	Patient		Clinician	
4best features	Split A	Split B	Split A	Split B
dev-off-stats	0.65	0.74	0.60	0.58
f0	0.60	0.58	0.40	0.37
KALDI-f0	0.70	0.74	0.50	0.32
MMeDuSa	0.60	0.79	0.45	0.41
Fusion-4best features	0.75	0.79	0.34	0.37
Fusion-4best + Lexical	0.80	0.68	-	-

We found that the best system combination of multiple acoustic-feature-based systems used four features as shown in Table 5. These features were among the best individually performed features, but some high-performance features were not selected by the selection algorithm (e.g. the MFCC or the dev-on-stats features which had very good performance individually, are probably highly correlated with some of the other features, so they didn't help in combination). The overall accuracy for the fusion-4best system, averaging across the 2-splits, was about 77%, while the chance result (based on

[1] It should be noted that i-vectors were evaluated, however, due to lack of sufficient training data, even an i-vector subspace of around 30 dimensions did not produce good results. This will be re-visited as more data becomes available.

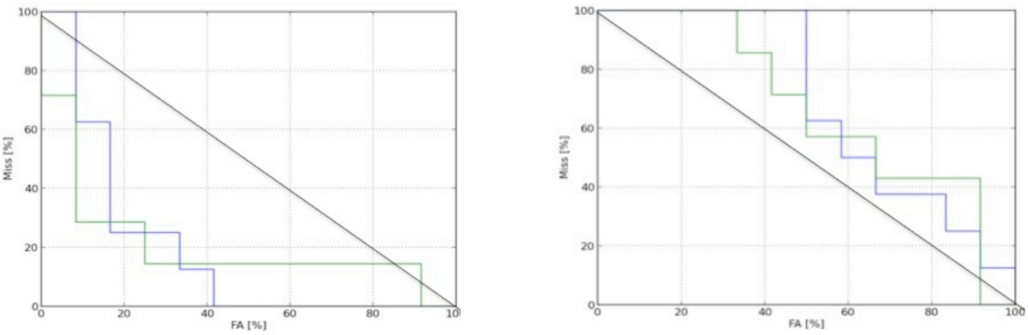

Figure 1. *DET curves demonstrating the 4best-feature system performance. The black line corresponds to the random guess while the green and blue lines correspond to the two splits. The left side shows the performance using the patient data, while the right (worse than random) shows the performance on the clinician data.*

the prior class distribution) is 61.5%. Thus our result is significantly above chance, for both splits and also overall, though more data is desirable to validate the reliability of our features for PTSD prediction.

Since the data for this experiment was quite limited, and the recordings were collected in a clinical setup over a period of a few months where recording conditions (e.g. room, background noises) might not be entirely fixed across all sessions, we wanted, for sanity purposes, to investigate whether our model was really learning the patient characteristics, or some other recording-condition relevant characteristics. For that purpose we run our classifiers also on the clinician channel, using only the first 10% from each recording where we had manually annotated clinician sentences, excluding any cross-talk from the patients. In Table 5 we show both individual features and fusion results for the 4-best selected features for both the patient and the clinician channels. With the exception of the “dev-off-stats” feature, the rest show worse than random results when using the clinician channel. In particular the system combination had the biggest difference in performance between the patient and clinician channels. Figure 1 shows the DET curves for that fused system (in both splits) for both the patient and clinician data and demonstrates clearly that only the patient results are significantly better than chance. After further investigation with some of the other individual-feature systems we observed that the MFCC-only system was the one with the best classification performance on the clinician channel. MFCCs are known to characterize well channel effects, unlike some of our other features (e.g. MMeDuSa and DOCC) that were designed to be more robust to channel characteristics. In particular the MMeDusa feature (one of the 4-best selected features) contains some degree of prosodic information (vowel prominence and stress) along with the spectral information (unlike MFCCs). Even though it is possible that there is some information about the patient’s condition in the clinician’s channel (i.e. the clinician may be responding to an indication he has about the subject’s state), we believe it is very unlikely that the MFCC system is the only one that can pick up this information. The most plausible explanation is that there is some acoustic bias in the recordings correlated with the condition of the speaker, and that is what the less channel-robust features are picking up. It is worth noting that in some recent work we found that systems based on the channel robust features performed well even across domains for the task of depression detection [39].

5. Discussion and Future Work

Psychiatry, along with many other fields in the health sciences, has seen the advent of highly data-intensive methods to augment the scientific and clinical picture. However, in the behavioral realm, there has been little in the way of revolutionary new technology. We believe that the rich, multidimensional sets of features obtained from speech analysis can constitute the high-data-intensive analog in the behavioral domain. Though we have limited training samples at this time, our results demonstrate that speech features have discriminative power for the prediction of PTSD. In the future we plan to explore more features, including prosodic speech measurements and re-evaluate the optimal rank and durations as more speakers become available.

An important potential confound in this study is extrinsic (not due to the speaker) variability across recordings. It is possible that microphone gains, distances between speakers, or other aspects of the set up could change, meaning assessment results could reflect these changes rather than the true behavioral changes we are after. This was the case when using the MFCC features. In real applications extrinsic variability will always be a factor, so we are studying a broad range of features assessing their robustness to outside factors.

Our research can potentially lead to the development of an objective, in-expensive non-invasive measure for PTSD. This measure can be used in a PTSD assessment tool as well as a measure for monitoring treatment effectiveness and will be beneficial to the public as well as the military.

6. Acknowledgments

This research and development project was conducted by SRI International and is made possible by a contract that was awarded and administered by the U.S. Army Medical Research \& Acquisition Activity (USAMRAA) and the Telemedicine \& Advanced Technology Research Center (TATRC), at Fort Detrick, MD, under contract: W81XWH-11-C-0004. The views, opinions and/or findings contained in this publication are those of the author(s) and do not necessarily reflect the views of the Department of Defense and should not be construed as an official Army position, policy, or decision unless so designated by other documentation.

7. References

[1] B. Roehr, “Cost of U. S. soldiers’ health care could reach $650bn,” British Medical Journal, vol. 335, no. 7628, pp. 1011–1011, 2007.

[2] R.C. Kessler, P. Berglund, O. Delmer,, R. Jin, K.R. Merikangas, and E.E. Walters, E.E. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Archives of General Psychiatry, 62(6): 593-602, 2005.

[3] D.D. Blake, F. W. Weathers, L.M. Nagy, D.G. Kaloupek, F.D. Gusman, D.S. Charney, and T.M. Keane, “The development of a Clinician-Administered PTSD Scale,” J. Trauma Stress, vol. 8, no. 1, pp. 75–90, 1995.

[4] J.E. Hovens,,H.M. van der Ploeg,,M.T.A. Klaarenbeek, I. Bramsen,, J.N. Schreuder and V.V. Rivero,, “The assessment of posttraumatic stress disorder: With the clinician administered PTSD scale: Dutch results,” J. Clin. Psychol., 50: 325–340, 1994

[5] P.D. Killworth and H.R. Bernand, “Informant accuracy in social network data,” Human Organization, vol. 35, pp. 269–286, 1976.

[6] C. Martinelli and S.W. Parker, “Deception and misreporting in a social program,” Journal of the European Economic Association, vol. 7, pp. 886–908, 2009.

[7] B. Peintner, W. Jarrold, D. Vergyri, C. Richey, M. Gorno Tempini, and J. Ogar, “Learning diagnostic models using speech and language measures,” in Proc. IEEE Engineering in Medicine and Biology Society, 2008.

[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 Proc. ICSLP, 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, 2004.

[10] M. Lugger and B. Yang, “The relevance of voice quality features in speaker independent emotion recognition,” in Proc. ICASSP, 2007.

[11] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, “Towards more reality in the recognition of emotional speech,” in Proc. ICASSP, 2007.

[12] M. Hewlett-Sanchez, G. Tur, L. Ferrer, and D. Hakkani-Tur, “Domain adaptation and compensation for emotion detection,” in Proc. Interspeech, 2010.

[13] 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, 2000.

[14] E.M. Moore, 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, 2008.

[15] 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 Proc. Int. Conf. on Affective Computing and Intelligent Interaction, 2009.

[16] 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 Proc. ICASSP, 2010.

[17] T. Yingthawornsuk and R.G. Shiavi, “Distinguishing depression and suicidal risk in men using GMM based frequency contents of affective vocal tract response,” in Proc. International Conference on Control, Automation and Systems, 2008.

[18] J.R. Williamson, T.F. Quatieri, B.S. Helfer, R. Horwitz, B. Yu and D. D. Mehta, “Vocal biomarkers of depression based on motor incoordination,” in Proc. of the 3rd ACM international workshop on Audio/visual emotion challenge, 2013.

[19] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps and T.F. Quatieri “A review of depression and suicide risk assessment using speech analysis”, Speech Communication, 71 (2015): 10-49.

[20] E.L. van den Broek, F. van der Sluis, and T. Dijkstra. "Telling the story and re-living the past: How speech analysis can reveal emotions in post-traumatic stress disorder (PTSD) patients." Sensing Emotions. Springer Netherlands, 2011. 153-180.

[21] S. Scherer, G. Stratou, J. Gratch, and L.P. Morency “Investigating voice quality as a speaker-independent indicator of depression and PTSD. In Proc. Interspeech, 2013..

[22] M.B. First, R.L. Spitzer, M.Gibbon, and J.B.W. Williams, “Structured clinical interview for DSM-IVTR axis I disorders, research version, patient edition. (SCID-I/P),” November 2002.

[23] V. Mitra, H. Franco and M. Graciarena, “Damped Oscillator Cepstral Coefficients for Robust Speech Recognition,” in Proc. Interspeech, 2013.

[24] V. Mitra, H. Franco, M. Graciarena and A. Mandal, “Normalized Amplitude Modulation Features for Large Vocabulary Noise-Robust Speech Recognition,” in Proc. ICASSP, 2012.

[25] P. Maragos, J. Kaiser and T. Quatieri, “Energy Separation in Signal Modulations with Application to Speech Analysis,” IEEE Trans. Signal Processing, Vol. 41, pp.3024–3051, 1993.

[26] V. Mitra, M. McLaren, H. Franco, M. Graciarena and N. Scheffer, “Modulation Features for Noise Robust Speaker Identification,” in Proc. Interspeech, 2013.

[27] V. Mitra, H. Franco, M. Graciarena and D. Vergyri, “Medium duration modulation cepstral feature for robust speech recognition,” in Proc. ICASSP, 2014.

[28] V. Mitra, G. Sivaraman, H. Nam, C. Espy-Wilson, E. Saltzman, “Articulatory features from deep neural networks and their role in speech recognition,” in Proc. ICASSP, 2014.

[29] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks", PhD thesis, University of Maryland College Park, December, 2004.

[30] O. Deshmukh, J. Singh, C. Espy-Wilson. 2004. "A novel method for computation of periodicity, aperiodicity and pitch of speech signals," in Proc. ICASSP, 2004.

[31] T. Pruthi and C. Espy-Wilson, “Acoustic parameters for the automatic detection of vowel nasalization,”in Proc. Interspeech, 2007.

[32] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal and S. Khudanpu "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," in Proc. ICASSP, 2014.

[33] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., "The kaldi speech recognition toolkit," in Proc. ASRU, 2011.

[34] P. Boersma, D. Weenink, “Praat: doing phonetics by computer,” Version 5.1.05, url: http://www.praat.org/, 2009.

[35] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Gr´ezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, “Further progress in meeting recognition: The ICSI-SRI Spring 2005 speech-to-text evaluation system”, Machine Learning for Multimodal Interaction: Second International Workshop, vol. 3869 of Lecture Notes in Computer Science, pp. 463–475, 2006.

[36] Y. Freund and R.E. Schapire. “A decision-theoretic generalization of on-line learning and an application to boosting”, Journal of Computer and System Sciences, 55(1), 119–139, 1997.

[37] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python”, The Journal of Machine Learning Research, 12, p.2825-2830, 2011.

[38] D.G. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, “Language recognition in iVectors space,” in Proc. Interspeech, 2011.

[39] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth and R.M. Salomon. “Cross-corpus depression prediction from speech.” in Proc. ICASSP, 2015.