Automatic Detection of Depression in Speech using Gaussian Mixture Modeling with Factor Analysis [1]

Douglas Sturim, Pedro Torres-Carrasquillo, Thomas F. Quatieri, Nicolas Malyska, and Alan McCree

MIT Lincoln Laboratory, Lexington MA

[sturim,torres,quatieri,nmalyska,mccree]@ll.mit.edu

Abstract

Of increasing importance in the civilian and military population is the recognition of Major Depressive Disorder at its earliest stages and intervention before the onset of severe symptoms. Toward the goal of more effective monitoring of depression severity, we investigate automatic classifiers of depression state, that have the important property of mitigating nuisances due to data variability, such as speaker and channel effects, unrelated to levels of depression. To assess our measures, we use a 35-speaker free-response speech database of subjects treated for depression over a six-week duration, along with standard clinical HAMD depression ratings. Preliminary experiments indicate that by mitigating nuisances, thus focusing on depression severity as a class, we can significantly improve classification accuracy over baseline Gaussian-mixture-model-based classifiers.

Index Terms: major depressive disorder, Gaussian-mixture models, joint factor analysis, nuisance mitigation

1. Introduction

MAJOR DEPRESSIVE DISORDER (MDD) is the most widely affecting of the mood disorders; the lifetime risk has been observed to fall between 10-20% for women and 5-12% for men [5]. Accurate diagnosis of MDD requires intensive training and experience. Thus the growing global burden of depression suggests that an automatic means to monitor depression severity would be a beneficial tool for patients, clinicians, and healthcare providers. For example, such a tool would be useful in monitoring the effects of new treatments. Reliable classifiers could also be used as a tool to aid in the standardization of depression ratings. One such approach relies on the extraction of biomarkers to provide reliable indicators of depression.

A class of biomarkers of growing interest is the group of vocal features observed to change with a patient's mental condition and emotional state. Examples that correlate with the particular condition of depression include vocal characteristics of prosody (e.g., pitch and speech rate), spectral features, and glottal (vocal fold) excitation patterns [4][7][8][9][10][17]. These vocal features have been shown to have statistical relationships with presence and severity of depressive conditions, and, in a number of cases, have been applied towards developing automatic classifiers.

In this paper, we introduce a specific set of classifiers based on Gaussian Mixture Models (GMM) and Latent Factor Analysis (LFA) toward recognizing levels of depression severity. We examine the accuracy of these classifiers in predicting the standard clinical HAMD [5] ratings by designing a multi-state classifier where rating level intervals are set to their own class. To assess our measures, we use 3-6 minutes of audio from a 35-speaker free-response speech database of subjects treated for depression over a six-week duration, along with standard clinical HAMD depression ratings. In designing classifiers of this type, and with such limited data, it is important to acknowledge the possibility that classes may reflect unwanted 'nuisances', i.e., undesired class influences.

An important contribution of our work in classifying depression severity is the *mitigation of nuisances due to speaker and channel effects, thus focusing on depression severity as a class distinction*. Our preliminary experiments indicate improved classification accuracy over standard baseline GMM classifiers using a variation of factor analysis for nuisance mitigation referred to as *nuisance mitigation with Wiener filtering* [11]. Features are based on two different spectral representations of speech, mel-cepstra and shifted-delta cepstra, that are standard static- and dynamic-based feature methodologies used typically in speaker and language identification, respectively.

Previous classifier work in this area includes, for example, Ozdas et al. [10] who investigated the use of two vocal features, vocal-fold jitter and the glottal flow spectrum, for differentiating between control, MDD, and near-term suicidal risk subjects. Depressed and near-term suicidal patients showed increased vocal-cord jitter and glottal spectral slope. Moore et al. [9] also investigated vocal glottal excitation and prosodic characteristics, in addition to spectral-based features, using 3-5 minutes of audio from 15 Major Depressive Disorder/18 control subjects. A large variety of statistical measures were then utilized to construct classifiers for distinguishing control from depressed patient groups; these classifiers were used to infer the most differentiating feature-statistic combinations for their dataset. Another example of a prosodic-based approach derives from a phonological framework to classify depression severity, exploiting the phoneme-dependence of speaking rate [17].

State-of-the-art classifier development most relevant to our work in this paper is by Low et al. who in initial work [16] developed GMM-based classifiers using mel-cepstral (plus delta versions) in identifying controls versus depressed adolescents (139 adolescent subjects with 68 Major Depressive Disorder/73 control). These classifiers achieved ~50% classification accuracy for depressed and ~60% accuracy for normal (where accuracy equals percentage of correctly identified states for each class). In later work [8], Low et al. combined prosodic, spectral, and the first and

[1] This work was sponsored by the Air Force under contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

second derivatives of the mel-cepstra features to classify the same control and clinically depressed adolescents, using a Gaussian Mixture Model (GMM)-based classifier. With a combination of vocal features (including mel-cepstra plus delta versions, prosodic features and Teager energy, among others) classification achieved ~75% accuracy.

Our paper is organized as follows. In Section 2, we setup the depression classification problem and review briefly the baseline GMM classifier. We also in this section, describe our particular 35-subject depression database collected by Mundt et al. [4] and give performance of our baseline GMM classifiers on this database. In Section 3, we introduce the concept of nuisance mitigation and review the rendition used in our study: *nuisance modeling with Wiener filtering*. In Section 4, we describe results with our more advanced classifiers, and finally in Section 5 provide conclusions and projections to future work.

2. Baseline Classification

In this section, we briefly review our GMM classifier foundation, describe the database and experimental setup, and summarize key baseline results.

2.1 GMM-based Classifiers

To address the depression detection problem we draw upon work in the language identification (LID) and speaker identification (SID) areas. The state of the art techniques in these areas rely upon statistical approaches to build classifiers which in turn act as detectors operating on a putative utterance.

In this study the baseline classifiers rely on a foundational Gaussian-mixture-model system [15] trained discriminatively with the maximum-mutual-information criterion (GMM-MMI) [14]. In Section 4, we review a variant of these classifiers using feature-domain nuisance mitigation [11].

Two cepstral features sets were used: 1) A standard speaker identification mel-frequency cepstral coefficient (MFCC) frontend with 19 cepstral coefficients and deltas to produce a 38 dimensional feature vector [15] and 2) A language identification frontend with MFCCs concatenated into a 56-dimensional feature vector composed of 7 static coefficients and stacked with the set of shifted delta cepstral (SDC) features produced by applying a 7-1-3-7 SDC scheme [14]. Channel compensating RASTA is then applied to both feature streams [15].

2.2 Dataset and experimental setup

The data used in this analysis was originally collected by Mundt et al. [4] for a depression-severity study, involving both in-clinic and telephone-response speech recordings. Thirty-five physician-referred subjects (20 women and 15 men, mean age 41.8 years) participated in this study. The subjects were predominately Caucasian (88.6%), with four subjects of other descent. The subjects had all recently started on pharmacotherapy and/or psychotherapy for depression and continued treatment over a 6-week assessment period. Speech recordings (sampled at 8 kHz) were collected at weeks 0, 2, 4, and 6 during an interview and assessment process that involved HAMD scoring. Additionally, we used only data from subjects that completed the entire longitudinal study. This resulted in approximately 3-6 minutes of speech per session (i.e., per day). More details of the collection process and feature-correlation studies on this database are given in [4].

Within this database, the standard method of evaluating levels of MDD in patients was invoked using the clinical 17-question HAMD assessment [5]. To determine the overall or

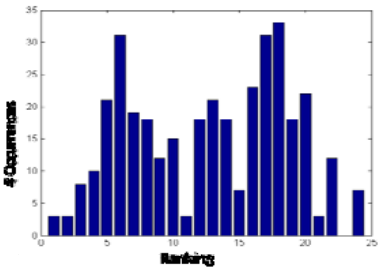

Figure 1: *Distribution of HAMD scores from 35-subject database.*

total score, individual ratings are first determined for symptom sub-topics (such as mood, guilt, psychomotor retardation, suicidal tendency, etc.) with scores for component sub-topics (17 symptom sub-topics) having ranges of (0-2), (0-3), or (0-4). The total score is then the aggregate of the ratings for all sub-topics. The distribution of the HAMD scores is given in Figure 1. These scores are used as our truth markings in defining classes for training and testing scenarios. Although the HAMD assessment is a standard evaluation method, there are some concerns about its reliability [6]. Nevertheless, addressing this concern is outside our scope.

In forming depression classifiers, we consider two different class problems for the HAMD total score. The data was separated into two or five categories. The 5-class case is divided into the ranges 0-5, 6-10, 11-15, 16-20, and 21-27, providing a means to monitor state. We also implemented the 2-class problem using the ranges: 0-17 and 18-27, thus providing more data for each class and a binary decision. Since the dataset was limited, we construct a cross validation experiment. All talker utterances except one held out talker were used as training data for the

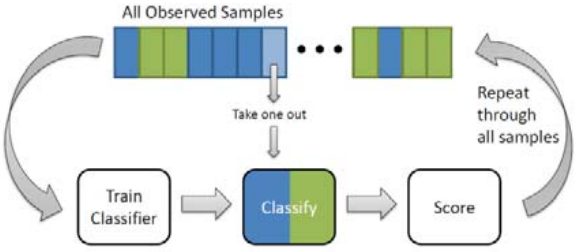

Figure 2*. Illustration of the leave-one-out cross-validation approach for the 2-class problem, depicted as green vs. blue. Each unique subject-session pair in our dataset is an 'observed sample' that is described by its feature vector. For cross validation, we take one sample out, train the classifier on the remaining samples, classify the excluded sample and record the performance. The process is repeated until all of the observed samples have been tested.*

depression classifier. This was then repeated exhaustively for all the talkers of the dataset. This leave-one-out cross validation scheme is illustrated schematically for the 2-class case in Figure 2.

2.3 Baseline Results

The results for the baseline GMM-MMI systems are presented in TABLE 1. When comparing results across feature options, there appears to be no one feature set that stands out across class options with respect to detection performance.

Nevertheless, even with the small dataset used, there are some significant feature-specific gender differences in performance with an error range for 95% confidence (5 classes: -4.7% < x < 5.7%; 2 classes: -5.0% < x < 6.0%).

System	Features	Number of Categories	Equal Error Rate: M/F
MMI	Cepstral+ Delta	2	40/42
MMI	SDC	2	30/48
MMI	Cepstral+ Delta	5	35/45
MMI	SDC	5	51/40

TABLE I. *Baseline MMI with MFCC and SDC features. Results for 2- and 5-class detection categories. EERs (rounded to nearest integer) are presented for Males/Females.* Error ranges: *5 Class: -4.7% < x < 5.7%; 2 Class: -5.0% < x < 6.0% for 95% confidence.*

Alluded to in this paper's introduction is previous classifier development most relevant to our work by Low et al. [16] who developed GMM-based classifiers using MFCCs only (plus delta versions) in identifying controls versus depressed adolescents (139 adolescent subjects with 68 Major Depressive Disorder/73 control). These classifiers achieved ~50% accuracy for depressed and ~60% accuracy for normal. Although these experiments differ from our work in many ways (the database, control/depressed categories, the specific GMM-based classifier, and performance metric), they do provide a reference point for state-of-the-art in GMM/MFCC-based classification for detection of depression.

3. Nuisance Mitigation

In this section we outline our approach to Gaussian mixture modeling with factor analysis for nuisance mitigation. We refer to this specific approach as *nuisance mitigation with Wiener filtering*.

3.1 Concept of Nuisance Mitigation

In recent years nuisance mitigation has become part of the state-of-the-art algorithms for both speaker and language identification systems. Two basic frameworks have emerged: 1) Joint Factor Analysis (JFA) with generative classifiers [13] and 2) Nuisance Attribute Projection (NAP) with discriminative classifiers [12]. In this study we will employ a version of Joint factor Analysis with a Wiener filtering framework.

3.2 Feature domain nuisance modeling with Wiener filtering (fWiener)

Wiener filtering for nuisance modeling was first proposed in the 2009 NIST Language Identification Evaluation [14]. The Wiener filtering approach is an alternative perspective on the Factor Analysis for modeling unwanted variation [1][2]. In the application of this paper, the detection category is depression state and the nuisances are due to speaker and channel variation. A more complete presentation of the Wiener filtering for Factor Analysis may be found here [11].

Consider a set of features that suffer from measurement noise and speaker/channel variation. Our goal is to reduce the effect of these disturbances. Given a GMM model associated with this feature set, the current approach forms a "supervector" from the stacked mean vectors of all components of the Gaussian mixture distribution. Let $\boldsymbol{m} = [\boldsymbol{m}_1^T \quad \cdots \quad \boldsymbol{m}_K^T]$ be a column vector in which $\boldsymbol{m}_i$ is the set of means for the *i*-th Gaussian component of *K* total mixtures. As a first step, we propose that the observed GMM statistics are modeled in part with independent-additive measurement noise:

$$\bar{\boldsymbol{x}} = \boldsymbol{m} + \boldsymbol{n} \qquad (1)$$

where $\boldsymbol{m}$ is the underlying model (set of GMM means), which in this case is the model for a depressive state. The measurement noise is vector $\boldsymbol{n}$**.** However, it is also desired to model the variation of nuisances (speaker and channel for our problem), since experience shows that nuisances are a major factor in degrading system performance. Equation (1) can be formed into a new model by adding a supervector $\boldsymbol{c}$ that will be the portion of the variability that is due to the nuisances:

$$\boldsymbol{y} = \boldsymbol{m} + \boldsymbol{n} + \boldsymbol{c} = \bar{\boldsymbol{x}} + \boldsymbol{c}. \qquad (2)$$

We then enforce a minimum mean-squared error performance criterion which results in [11]

$$\hat{\boldsymbol{x}} = \boldsymbol{\Sigma_m}(\boldsymbol{\Sigma_m} + \boldsymbol{\Sigma_n} + \boldsymbol{\Sigma_c})^{-1}\boldsymbol{y} \qquad (3)$$

with $\boldsymbol{\Sigma}$ defined as a correlation matrix**.** This forms a Wiener filter for removing component measurement noise and speaker and channel variation.

The estimated supervector $\hat{\boldsymbol{x}}$ is the stacked means of a Gaussian mixture model and is in the model domain. It represents a compensated form of the original supervector $\boldsymbol{y}$. Our final classifier will be a discriminative algorithm: GMM with Maximal Mutual Information (GMM-MMI) [14]. In order to feed into the MMI algorithm we project the supervector, $\hat{\boldsymbol{x}}$, back to the feature domain using a technique proposed by [3].

Consider a feature vector $\boldsymbol{f}(t)$at time step *t*. We can form a new corrected feature vector $\hat{\boldsymbol{f}}(t)$ by projecting the compensated supervector $\hat{\boldsymbol{x}}$ back into the feature domain through a weighted sum of the channel compensation offset values. This can be subtracted from the original observation feature:

$$\hat{\boldsymbol{f}}(t) = \boldsymbol{f}(t) - \sum_{\boldsymbol{p=1}}^{K} \gamma_{\boldsymbol{p}}(t) U_{\boldsymbol{p}} \hat{\boldsymbol{x}} \qquad (4)$$

where $\gamma_p(t)$ is the Gaussian occupation probability for the sufficient statistics from the universal background model [3]. U_p is the channel compensation offset related to the *p-th* Gaussian. This forms our feature-domain nuisance modeling with Wiener filtering or *fWiener*.

3.3 Results

Our depression detection systems are first run without feature-domain nuisance mitigation (referred to earlier as baseline). We presented these results in Table 1 with mel-frequency-cepstral coefficient (MFCC) and shifted-delta-cepstral (SDC) features.

Table 2 shows results with the feature-domain nuisance mitigation activated. Generally, we see a consistent improvement over the baseline results of Table 1. For the 2-class problem, with cepstral features, we obtain ~21% and ~29% ERR absolute reduction for males and females, respectively, while with SDC features, we obtain ~12% and ~31% absolute EER reduction for males and females, respectively. For the 5-class problem, with cepstral features,

we obtain ~10% EER absolute reduction for males and with SDC features, we obtain ~9% absolute reduction for females. These results are significant even with our small dataset and with the large 5-6% error range for 95% confidence. For the other two SDC cases, there was no change in performance given the error range for 95% confidence. The smaller or equivalent performance for the 5-class case is likely due to the scarcity of data for each class relative to the 2-class problem. Nevertheless, nuisance mitigation overall provides a significant gain.

System	Features	Number of Categories	Equal Error Rate: M/F	Gain
fWiener MMI	Cepstral+ Delta	2	19/13	**21/29**
fWiener MMI	SDC	2	18/17	**12/31**
fWiener MMI	Cepstral+ Delta	5	37/35	**-2/10**
fWiener MMI	SDC	5	42/38	**9/2**

TABLE 2*. fWiener MMI with MFCC and SDC features. Results for 2 and 5 detection categories.* Error range: *5 Class: -4.7% < x < 5.7%; 2 Class: -5.0% < x < 6.0% for 95% confidence.*

4. Conclusions and Future Work

In this paper, motivated by earlier work by Low et al. [16], we introduced a specific set of automatic classifiers based on Gaussian Mixture Models (GMM) and Latent Factor Analysis (LFA) toward recognizing different levels of depression severity. Our classifiers addressed the possibility that classes may reflect unwanted *nuisances*, i.e., undesired class influences such as speaker and channel effects. We first provided a set of experimental results using standard GMM-MMI classifiers and noted that, although based on a different experimental framework, our baseline EER performance is consistent with reference classification results by Low et al. [16].

An important contribution of our work in depression classification is that by mitigating nuisances, thus focusing on depression severity as the class distinction, we significantly improved classification accuracy over standard baseline GMM classifiers. Although performance comparisons were limited by data size (35 subjects with about 3-6 minutes per subject per session over a six week period), even with a ~5-6% error range for 95% confidence, we saw a large upward performance trend with nuisance mitigation and provided a foundation for future comparative studies of nuisance mitigation in classifying depression severity. Specifically, we saw large absolute gains of 20-30% EER for the 2-class problem and smaller, but significant, ~10% ERR gains in two of the four 5-class cases.

A more exhaustive classification study requires a larger, more comprehensive database and investigation of a broader suite of speech rate features. Specifically, plans to extend our preliminary study include the use of both alternate classification schemes, such as the use of Support Vector Machines (SVMs), fine-tuning our MFCC and SDC features for the depression problem, and expanding our feature set to include prosodic and linguistic speech patterns

Acknowledgement

The authors acknowledge Dr. James Mundt for helpful discussions on details and advice on the use of his data collection and the National Institute of Mental Health who supported Dr. Mundt in the collection.

References

[1] M.E. Tipping, C.M. Bishop, "Mixtures of Probabilistic Principal Component Analyzers," Tech. Rep. NCRG/97/003, Aston Univ., U.K., July 1998.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2008.

[3] V. Claudio, D. Colibro, F. Castaldo, E. Dalmasso, P. Laface, "Channel Factors Compensation in Model and Feature Domain for Speaker Recognition," *Proc Speaker Odyssey Workshop*, 2006.

[4] 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," *Journal of Neurolinguistics,* 20(1): 50-64, 2007.

[5] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, Text Revision, Washington, DC, American Psychiatric Association, 2000.

[6] R.M. Bagby, A. G. Ryder, M.A., D. R. Schuller, M. B. Marshall, "The Hamilton Depression Rating Scale: Has the Gold Standard Become a Lead Weight?", *Am J Psychiatry*, 161:2163-2177, December 2004.

[7] D. France, R. Shiavi, et al., "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions on Biomedical Engineering* 47(7): 829.

[8] L.A. Low, T. Maddage, M. Lech, L. Sheeber, N. Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression in adults," *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010*.

[9] E. Moore II, M. Clements, J. Peifer, L. Weisser, "Analysis of prosodic variation in speech for clinical depression," *Proceedings of the 25th Annual International Conference of the IEEE EMBS*, 2003.

[10] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, D. Mitchell, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering* 51(9), 2004.

[11] A. McCree, D. Sturim, D. Reynolds, "A New Perspective on GMM Subspace Compensation Based on PPCA and Wiener Filtering", *Interspeech 2011*, Sept, 2011.

[12] W.M. Campbell, D.E. Sturim, D.A. Reynolds, A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," *ICASSP 2006 Proceedings*, vol.1, no.1, pp. 14-19, May 2006.

[13] N. Dehak P. Kenny, R. Dehak, V. Gupta, P. Dumouchel, "The role of speaker factors in the NIST extended data task," *Proc. IEEE Odyssey*, 2008.

[14] P.A. Torres-Carrasquillo, E. Singer, T. Gleason, A. McCree, D.A. Reynolds, F. Richardson, D. Sturim, "The MITLL NIST LRE 2009 language recognition system," *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, March 2010.

[15] D. Reynolds, T.F. Quatieri, R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, 2000.

[16] L.A. Low, T. Maddage, M. Lech, L. Sheeber, N. Allen, "Mel Frequency Cepstral Feature and Gaussian Mixtures for Modeling Clinical Depression in Adolescents," *Proc. IEEE 8th Int. Conf. on Cognitive Informatics*, 2009.

[17] A. Trevino, T.F. Quatieri, N. Malyska, "Phonologically-Based Biomarkers for Major Depressive Disorder," to be published, *EURASIP Journal on Advances in Signal Processing: Special Issue on Emotion and Mental State Recognition from Speech*, 2011.