APPLICATION OF LANGUAGE IDENTIFICATION METHODOLOGY FOR THE TASK OF MENTAL PATHOLOGY DISCRIMINATION

Yonatan Vaizman¹, Oldřich Plchot² and Lukáš Burger²

1. The Interdisciplinary Center for Neural Computation and Department of Computer Science, the Hebrew University in Jerusalem Israel, 2. Brno University of Technology, Speech@FIT group, Brno, Czech Republic.

ABSTRACT

The task of discriminating subjects with different mental conditions using speech recordings can be defined similarly to the problem of language identification (LID). Both problems involve categorical classification based on speech samples, when each class has multiple speech segments from multiple subjects. In this work LID methodology, including the extracted features, statistical models and learning methods, was applied on speech samples of Healthy, Schizophrenia and Depressed subjects, for the task of discriminating these three mental conditions. The experiments show relative success for the discrimination of Depression patients (25% error) and Schizophrenia patients (17% error) from healthy subjects. These results demonstrate that local spectral acoustic features carry relevant information about mental pathology, and suggest a new way to model variability of speech properties among different mental conditions.

Index Terms—Mental state detection, Speech analysis, language identification, depression, schizophrenia

1. INTRODUCTION

Diagnosis and monitoring of mental disorders such as Schizophrenia and Depression involve assessment of characteristics of speech. For instance, “flat”, non-emotionally-expressive speech is one property associated with Schizophrenia, and slow, sluggish speech is associated with Depression. Supplying automated, objective measurements of speech, relevant to the mental pathology, can be useful for screening and monitoring of mental pathologies. In previous work, we addressed this task by extracting prosodic features, such as pitch and power variability [1]. Those features were designed in a top-down manner, to characterize the prosodic nature of speech, perceived by the listener (the experienced psychiatrist). In the present work we experimented with bottom up features and methods, already applicable for the parallel task of Language Identification (LID), and used them for the mental pathology discrimination task.

2. METHODS

2.1. Data

The data was collected at the Schizophrenia and Bipolar program in McLean Psychiatric Hospital in Belmont, USA. It is comprised of audio recordings of subjects of 3 classes: Healthy (HL), Schizophrenia (SZ) and Depression (DP). The control subjects (HL) were carefully collected to match the SZ and DP groups, covering similar distributions of age, gender and education. Subjects performed two different speech tasks:

1. Semi structured interview (SCID) conducted by an experienced psychiatrist. The 62 subjects included: n(HL)=20, n(SZ)=22, n(DP)=20. The median duration of an interview session was 50 minutes. Some subjects participated in multiple sessions. Interviews were segmented to 2 minute chapters, and each chapter was analyzed as one data point. Total recording time was ~4400 minutes, including silence.

2. Reading a list of words. The 53 subjects included: n(HL)=16, n(SZ)=17, n(DP)=20. The task took ~2 minutes per subject. Total recording time was ~107 minutes, including silence. For this task all subjects read the same standard text: North American Adult Reading Test (NAART), which is used to assess cognitive function and contains semantically independent words. All subjects (for both tasks) spoke English as a first language, and spoke only English during the sessions. Audio was sampled in 44,100 Hz.

2.2. Feature extraction

Each file was segmented to speech and silence segments. For the speech segments vocal tract length normalization was performed [2]. Mel Frequency Cepstral Coefficients (MFCCs) were calculated. Subsequently Shifted Delta Cepstra (SDC) of the MFCCs were calculated [3]. As previously applied and validated in LID experiments [4], we used SDC parameters 7-1-3-7. For each time-frame of 10
msec 7 MFCCs (including C0) and 49 SDC coefficients were concatenated to create a 56-dimensional feature vector.

2.3. Generative model

To model the variability of the features the Joint Factor Analysis (JFA) approach was used [5]. We applied it in a way similar to the way it was used for the LID task [6]. This model assumes that every data point has been generated in a hierarchical manner: Each observed data point (speech segment) is assumed to be generated from a different Gaussian Mixture Model (GMM). The segment-specific GMM generates a 56D feature vector independently for each time-frame of the speech segment. We used mixtures of 256 Gaussians. The model also describes the variability between classes and the variability within classes. A Universal Background Model (UBM) is a GMM that acts as a reference model, around which the different classes are spread. We assume that all Gaussians have diagonal covariance matrices, and we further simplify the model by assuming that all GMMs share the same covariance matrix and Gaussian weights - the covariance and weights of the UBM. Each segment can then be represented as the mean supervector of its GMM - \(m_s\), which is a concatenation of the 256 mean vectors of the Gaussians, resulting in a supervector of length 256x56=14,336. This mean supervector can be modeled as:

\[m_s = m + Dz + Ux\]

Where \(m\) is the mean supervector of the UBM, \(D\) is a diagonal matrix describing the between-classes variability, \(z\) is a "location supervector" describing the location of a class around the UBM, assumed to be a standard normal random vector, \(U\) is a low rank matrix describing the within-class variability (sometimes referred to as channel variability) and \(x\) is a low dimension vector indicating the location of the segment-specific GMM around the mean of the class. As initial experiment, here we simplified the model even more and disregarded the within-class variability terms, simply modeling the mean-supervector of a segment-specific GMM as the mean-supervector of the class the segment belongs to:

\[m_s = m + Dz\]

This simplified version of the model is actually equivalent to "relevance Maximum A Posteriori (MAP) adaptation". The UBM is the prior, and the location vectors, learned from the data, are combined with the prior to give the MAP parameters. Further experiments will also include the channel variability terms and perform channel-compensation, to better utilize the full strength of the model.

Using constant length sufficient statistics for each speech segment (the speech segment can have arbitrary length), it is possible to approximate the GMM likelihood in a convenient and efficient way [6].

The UBM used was pre-trained over completely different speech data: the CALLFRIEND data set (from the Linguistic Data consortium – LDC: [http://www.ldc.upenn.edu/]). This data set contains recordings of telephone conversations, sampled at 8000 Hz, from 15 dialects of 12 different languages (including English), 60 conversations from each dialect, ranging from 5 to 30 minutes per conversation. Granted, this dataset is very different from the data analyzed here, but as part of the experiment, we modeled the 3 mental classes in our data (HL, SZ and DP) using this foreign reference UBM. One reason for using a UBM trained on external data was the small number of subjects in our dataset, and the need to apply jackknifing for evaluation. Having a constant common UBM is helpful.

As heuristic, the values of the diagonal \(D\) matrix were asserted to be 0.5 of the standard deviations of the Gaussians in the UBM. Lower values for \(D\) would cause the location vectors of the classes (\(z\)) to be too close to the mean of the UBM (\(m\)). Since this UBM was trained over a different type of data, we wanted to give more weight to our data, when training the class location vectors, and rely less on the UBM. Using the "Joint Factor Analysis Matlab Demo" ([http://speech.fit.vutbr.cz/en/software/joint-factor-analysis-matlab-demo]), sufficient statistics were collected for each data point, and point MAP estimates of the location supervectors were calculated for each class \((z_{HL}, z_{SZ}, z_{DP})\) over the training (enrollment) data. A linear scoring method was performed on the test data, to approximate the log likelihood of each data point to be generated from each of the class models [7].

2.5. Performance evaluation

Due to the small number of subjects in our dataset, we didn't extract a single portion of the data as a constant evaluation set. Instead, a jackknifing approach was applied. We evaluated performance using leave one out (LOO) cross-validation: for each partition of the data, all the data points of a single subject were extracted for test set, and the rest of the data was used as the training (enrollment) set, over which the location vectors of the three classes were trained. Binary classification was done to each pair of classes (e.g. HL vs. SZ), with pseudo log likelihood ratio test, using the linear scorings mentioned above as log likelihoods. For each pair of classes, and for each partition of the data, the rate of choosing the first class was saved, and was later averaged over all partitions of the data, in which the test subjects belong to the same class.

3. RESULTS

Detection Error Tradeoff (DET) curves are presented in figures 1-3. Each point represents a different threshold for the log likelihood ratio tests. Circles mark the approximated
Equal Error Rate (EER) points – points where the miss rate is closest to the false alarm rate. The miss rates of these points are presented as EER in the legend. Squares mark the Minimum Error Rate (MER) points – points where the average of miss rate and false alarm rate was minimal. For comparison, the performance of our previous system (with prosodic features) is also presented: up-facing triangles indicate the miss and false alarm rates over the interviews, and down-facing triangles indicate the miss and false alarm rates over the list of words.

As seen in our previous work [1], these results suggest that both speech tasks, being interviewed (interview) and reading a list of words (NAART), carry information relevant to different classification tasks. For instance, classification of HL vs. SZ was better using the NAART data than using the interviews. Notable, good performance was achieved for HL vs. SZ (with the NAART data) - MER = 17%, and for HL vs. DP (mostly with the interviews data) - MER = 25%. In most scenarios the prosodic system achieved better classification than the local spectral acoustic system (where the triangle is presented below the relevant DET curve – closer to the <0,0> error rates point). An improvement was seen for the HL vs. SZ task, using the NAART, where the prosodic system had 27% miss rate and 25% false alarm rate, and the local spectral acoustic system reached a point of 15% miss rate and 19% false alarm rate (MER = 17%). The SZ vs. DP classification task didn't achieve as good performance as the other two (HL vs. SZ, HL vs. DP): relatively high MERs of 35% (interview) and 32% (NAART). Better performance was already achieved with the prosodic system: MERs of 23% (interview) and 28% (NAART).

4. DISCUSSION

Mental disorders, such as Schizophrenia and Depression, are complex phenomena. The disorders themselves are not well understood, and the diagnosis and monitoring of the disorders involve many factors. During a structured clinical interview, there are many factors the psychiatrist takes into account, such as body language, facial expressions and the semantic content of what the patient is saying. The acoustic characteristics of speech is only one factor. This study demonstrates that this factor in isolation carries information about the mental disorder. Here we show that basic low-level analysis of the speech sound (with local spectral acoustic features) can be used to discriminate control subjects from SZ and from DP patients. Given that the mental pathology classification here is based solely on speech, the results presented are good.
Such technology can be used to supply clinicians with automated and objective measures of speech acoustic characteristics, instead of relying on subjective ratings of perceived prosody. Applications of such a system include automated large scale screening for a mental disorder and monitoring its progression.

The task of discriminating SZ vs. DP is important in the clinical domain. The two disorders share common characteristics, and in some cases it is unclear at first whether a patient suffers from depression or exhibits negative symptoms of schizophrenia. Hence, capturing the discriminating features between SZ and DP is an important challenge, which is achieved to some level using prosodic features, as described in [1]. It might be possible to perform fusion of the prosodic system with the spectral acoustic system described here, to improve the overall discrimination between the groups.

The use of a UBM trained on external data was done to overcome the limitation of the small sample in our pathologic dataset. It also adds a level of generalization to the analysis, since it is a constant UBM, used for all LOO partitions, and used for both speech tasks. Although this UBM was trained over telephone conversations, we treat it as if it represents a more general model of speech from the general population. This means that it might incidentally include some subjects with a mental pathology, which will contribute to the generality of this model and its validity as a background model, around which the classes are spread. This UBM is not specific to a reading task or interview task, and it even includes variability of language. It is also based on recordings from telephone, and our data was recorded over a different channel – a microphone in a lab room. We aim towards a characterization and classification system that is speech-task independent or even channel-independent, and can be applied over speech segments taken from different sources, contents and channels. Using an external UBM as reference is one step in this direction. Further steps include training mental class location vectors on one speech task and applying them on segments from another speech task, or training a UBM on a dataset of different types of speech.

5. ACKNOWLEDGEMENTS

This work was partially supported by the European Union under DIRAC integrated project IST-027787. BUT researchers were partially supported by Grant Agency of Czech Republic project No. 102/08/0707, and by Czech Ministry of Education project No. MSM0021630528.

6. REFERENCES


