MULTI-LINGUAL DEPRESSION-LEVEL ASSESSMENT FROM CONVERSATIONAL SPEECH USING ACOUSTIC AND TEXT FEATURES

A Thesis

by

Yasin Serdar Özkanca

Submitted to the
Graduate School of Sciences and Engineering
In Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in the
Department of Electrical and Electronics Engineering

Özyeğin University
December 2018

MULTI-LINGUAL DEPRESSION-LEVEL ASSESSMENT FROM CONVERSATIONAL SPEECH USING ACOUSTIC AND TEXT FEATURES

Approved by:

Asst. Prof. Cenk Demiroğlu, Advisor
Department of Electrical and Electronics Engineering
Özyeğin University

Asst. Prof. Reyhan Aydoğan
Department of Computer Science
Özyeğin University

Assoc. Prof. Ümit Güz
Department of Electrical and Electronics Engineering
Işık University

Date Approved: 19 December 2018

To friends, and my family

ABSTRACT

Depression is a common mental health problem around the world with a large burden on economies, well-being, hence productivity, of individuals. Early diagnosis and detection of depression can aid treatment, but diagnosis typically requires an interview with a health provider or structured diagnostic questionnaire. Thus, unobtrusive measures that might be able to monitor depression symptoms in daily life could have great utility in monitoring depression for clinical treatment. Vocal biomarkers of depression are a potentially effective method of assessing depression symptoms in daily life, which is the focus of the current research. Although there have been efforts to automatically assess depression levels from audiovisual features, use of transcriptions along with the acoustic features has emerged as a more recent research venue. Moreover, difficulty in data collection and the limited amounts of data available for research are also challenges that are hampering the success of the algorithms. One of the novel contributions in this thesis is to exploit the databases from multiple languages for feature selection. Since a large number of features can be extracted from speech, and given the small amounts of training data available, effective data selection is critical for success. Our proposed multi-lingual method was effective at selecting better features and significantly improved depression assessment accuracy. In addition, text-based features were used for assessment and a novel strategy to fuse the text- and speech-based classifiers were proposed, which further boosted the performance.

ÖZETÇE

Depresyon, bireylere ekonomik, refah düzeyi, dolayısıyla üretgenlik açısından büyük bir yük olan, yaygın bir zihinsel sağlık sorunudur. Erken tanı ve depresyonun tespiti tedaviye yardımcı olabilir, ancak tanı genellikle bir sağlık kuruluşu ile iletişim veya yapılandırılmış tanısal bir anket gerektirir. Bu nedenle, günlük hayatta depresyon belirtilerini izleyebilecek göze batmayan önlemler, klinik tedavi için depresyonun izlenmesinde büyük yarar sağlayabilir. Depresyonun vokal biyobelirteçleri, güncel araştırmaların odağı olan, günlük hayatta depresyon belirtilerini değerlendirmede potansiyel olarak kullanılabilecek etkili bir araçtır. Görsel-işitsel özelliklerden depresyon düzeylerini otomatik olarak değerlendirme çabalarına rağmen, akustik özellikler ile birlikte yazılı metin kullanımı daha yeni bir araştırma alanı olarak ortaya çıkmıştır. Ek olarak, veri toplanmasındaki zorluk ve araştırmaya açık sınırlı miktarda veri de algoritmaların başarısını engelleyen zorluklardandır. Bu makalenin sunduğu katkılardan biri, öznitelik seçimi için veritabanı olarak birden çok dil kullanmaktır. Etkili bir öznitelik seçimi, az sayıda konuşma verisinden çok sayıda öznitelik elde edilebildiğinden dolayı, başarılı bir çözüm için çok önemlidir. Önerilen çok dilli yöntemimizin daha iyi öznitelikler seçmede etkili olduğu ve depresyon değerlendirme doğruluğunu önemli ölçüde geliştirdiği gözlemlendi. Ayrıca, değerlendirme için metin tabanlı öznitelikler de kullanıldı ve performansı arttırması adına metin ve konuşma temelli sınıflandırıcıları birleştiren bir strateji önerildi.

ACKNOWLEDGEMENTS

First I would like to thank my advisor Assoc. Prof. Cenk Demiroğlu for his support through my research period. Also, I am thankful to my friends and family, for not leaving me alone at my thesis presentation.

Also, I have to thank the staff of Özyeğin University for making the university home for me for 7 years. It is also important for me to mention the financial aid that my university provided for my conference trip to the Hyderabad, India. Which made my further career possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Depression is a vital problem that affects a large portion of the population. It affects the well-being and productivity of individuals as well as being a heavy economic burden for the society [3]. Compared to other diseases, depression and many other mental illnesses affect humans that are working and contributing to the economy, hence it decreases the efficiency of productivity of the works being done [4]. Depression alone accounts for 10% of all disability due to physical and mental health problems globally, which shows that it affects the economic well-being of the nations too. Moreover, it is the primary reason for suicide, estimated to be responsible for 1.4% of all deaths around the world [5]. It is also predicted to be the leading cause of disease burden by 2030 [6]. However, better diagnosis of depression followed by successful treatment was shown to be effective in mitigating the symptoms and decreasing suicide rates [7]. Thus, inexpensive and accurate diagnosis with the help of technology is an increasingly important research challenge [2].

It is well known that getting psychological help from an expert is expensive for most people. Thus, creating an automated system that can identify one's psychological condition would lower the costs of getting help. This thesis shows, only from one's voice, we can predict depression severity level of a person. Which sheds light on further studies that can combine voice with face features, movement inputs or even biological signals. This study is a part of behavioral signal processing (BSP), which becomes more common with the ability to collect data is increased. It is being employed by multiple topics like spouse therapies [8, 9], addiction counseling [10] or autism [11].

The distinguish of quality in speech, between depressed and non-depressed subjects were a known fact for years [12, 13], but the mathematical proof of whether depressive subjects have differences in their voice parameters was not present until mid 70s [14, 15]. The speech pause times and different acoustic indicators were counted as biomarkers for clues of depressive voice [16, 17]. The studies referenced above were in the English language, the generalization of these works in different languages was proved in [18]. Moreover, the change in speech pitch has been a topic of this study [19]. It is also indicated that depression alters the way humans use their language, in [20, 21], it is stated that depressive people tend to use the word ”I” more frequently than healthy subjects. Moreover, [22] suggests that the narrative structure of the conversations and the syntax that is being used by depressive people are meaningful discriminators that can help assess depression. In addition to that, there are non-verbal hints that have been explored by researchers that indicate depression, such as lack of smiles and withdrawing gestures [23, 24].

Use of antidepressant treatments and attaining clinical improvements were modeled, employing multivariate equations of voice acoustic parameters in 90s [25–27]. The use of speech-related features of depression severity and healing response of clinical trials have needs like quality recording devices, software, and technical skills to analyze the speech data, and because of it, the studies were limited and expensive. However, advancements and automation in collecting data were much easier after powerful devices like smart phone became accessible and widely used. Speech samples were collected over telephones and were used as clinical data [28]. These innovations and progress made the studies cheaper and much more common.

The main goal of this study, is to lower the costs of people getting psychological help from professionals, by bypassing the necessity of expensive clinical examinations.

1.1 Related Work

It has been shown that speech signal carries significant amount of information about the mental health of the speakers [29–31]. France et al. [31] show that features that are coming from power spectral density and formant measurements were the best indicators of depression and suicidal risk for both male and female groups. Moreover, the paper claims that features that are derived from F_0 do not perform well as discriminators.

In [32], phase distortion deviation that is used for voice quality examinations is found to be helpful for detecting depression. In [33], distortions in formant trajectories were used to detect depression. Moreover, principal component analysis (PCA) was used for feature selection. In [34], degradation in spectral variability was used. In addition, they claim that affect on motor control in depression may lead to degradation in acoustic variations.

In [35], gender-dependent feature extraction was found to improve the detection performance. In addition, [36] also claims that different genders that are affected by depression might show differences in formant features. The conclusion in [36] is backed up by studies which indicated the advantage of gender-dependent classification when the use of formant and spectral features were implemented [37, 38].

In [39], voice quality features like i-vectors and MFCCs were found to be helpful for depression detection. They have also used a score-based fusion algorithm which improved the performance of their system. Moreover, the system in [39] found to work well when the utterances were shortened to 10 seconds.

Various studies show that absence of spectral is correlated with the depression level [40–42]. Studies in [29, 36] state that prosodic features are robust and reliable discriminators for the use of depression level assessment. Speech segments with higher articulation effort were found to be more informative for depression detection in [30].

Besides acoustics-only methods, there are also multi-modal approaches for detecting depression. In [43], face analysis and speech prosody are used for depression detection. Similarly, audio-visual features are used in [44–47]. In [44], the fusion algorithm that implemented was used to fuse the vocal quality features and visual features. Study in [48] investigated the affiliation between audio-visual features and depression with the use of Canonical Correlation Analysis (CCA).

Retardation in motor control due to depression causes changes in coordination and timing of speech and face movements, which were used for audio-visual detection in [49–51]. Moreover, Syed et al., [52] proposed a set of temporal features and demonstrated the usefulness of those features. Further, they explored craniofacial movements are ominous of psychomotor retardation, thus the depression. In [53], low and high-level features were used for each modality, audio, video, and text. They also performed gender dependent and independent assessment. In addition, for text features, they used speech-rate and semantic content. Nasir et al. [54] are also studied on an audio-visual multi-modal system. For the audio part they have used Teager energy cepstral coefficients (TECC), and for the video features, they have succeeded with polynomial parameterization of facial landmark features.

Besides face features, text analysis of transcriptions has also been used as another form of information [2]. In [55], transcription-derived features were employed in addition to the speech features. Furthermore, sentiment analysis was conducted on text and sentiment features were used to develop an independent detector. Then, score fusion was used to combine acoustic and text-based system scores. Syntactic and semantic features were derived from transcriptions in [56] and proved to be effective indicators of depression. They also declared that negatively-draped and pronoun words are in accord with depression. In [57], biomarkers that are originated from facial coordination and timing features were used together with vocal cues and semantic features from dialogue content employing a sparse coded lexical embedding space.

They also practiced contextual cues like past or present of the patient's depression state.

In [58], the study claims depression and narcissism can be foretold from the usage of words in personal narratives. The paper declares that depressive individuals use less social words and more anxiety-related words, where narcissist do the opposite. Recent study [59] manifests an automated depression-detection algorithm that models the interaction between the depressed individual and the computer agent without the need for explicit topic modeling of the content. They used Long-Short Term Memory (LSTM) neural network and fed it with audio and text features.

In depression detection, another research challenge is to use speech data from other languages/cultures to train models. This approach is not only necessary for understanding universal cues of depression across different cultures/languages but also it enables the use of data from other languages, which is important given the typically small amounts of data available in the public databases. In [60], prediction models built with a German database were shown to produce prediction scores in English that were correlated with self-assessment scores. In [61], the combination of datasets in different languages was shown to yield high accuracy whereas if the train and test data are in different languages, performance was found to be lower.

Conversations with patients can be designed in a way to obtain data that is more indicative of depression, as opposed to a regular conversation. In [62], type of questions (positive and negative stimulus) during conversations has been shown to impact voice quality parameters in psychologically distressed subjects.

1.2 Contributions

This study has two contributions. One of the contributions is novel algorithms for feature selection which was not explored as much in the literature. We offer a multi-lingual feature selection where three databases, Turkish, German and English were used concurrently. Moreover, multi-lingual and single-lingual methods to enhance redundancy and relevance computations in the case of data sparsity are introduced. In addition, it has been confirmed that there are common features among languages that can be essential to predict depression from conversational speech. The second contribution is a novel feature fusion technique where transcription-derived model predictions were used to adjust the predictions of the acoustic-only model when their predictions are highly conflicting. Significant improvements are obtained for the Turkish, German and English databases using the proposed techniques. The study aims to improve the automatic detection of a mental illness to reduce the costs of professional diagnosis.

CHAPTER II

METHODS

2.1 Feature selection

As most of the statistical applications, machine learning algorithms are also vulnerable to outliers and noisy data. However, this is much more effective when the data is scarce and it is not enough to eliminate the noisy information comes from the depths of the data. Thus, there are statistical ways that can only choose features that can improve the effectiveness of the model that is being built. Not only the accuracy, but these feature selection algorithms also can improve the computation time and efficiency of the systems.

The motivations to use the feature selection algorithms are;

- **Computation Time;** Eliminates the unnecessary computations,

- **Complexity of a Model;** Makes the models simpler,

- **Accuracy;** Most of the times it improves the accuracy,

- **Overfitting;** It is a good way to fight against overfitting,

2.1.1 Minimum redundancy maximum relevance (MRMR) feature selection

A large number of features can be derived from conversational speech to detect depression. However, building models with those features are challenging because of the curse of dimensionality especially given the typically small amounts of training data available in depression studies.

One way of reducing the dimensionality features is to use feature selection where features that are most relevant for the classification task and least correlated among themselves are selected. To that end, ”Minimum Redundancy Maximum Relevance” (MRMR) algorithm is commonly used [63–65]. MRMR is widely used for bio-informatics tasks. The strongest feature of the MRMR is that it does not only select the most important features, it also eliminates the ones that are redundant.

In the MRMR approach, for maximizing the relevance of selected features for the classification task, F-statistic is used.

$$F(g_i) = [\sum_k n_k(\bar{g_k} - \bar{g})^2/(K-1)]/\sigma^2, \tag{1}$$

$\bar{g_k}$ is the mean of the g_i, within the kth class. $\bar{g}$ is the global mean of whole feature set. The number of classes denoted by K and σ^2 is the pooled variance:

$$\sigma^2 = [\sum_k (n_k - 1)\sigma_k^2]/(n-K), \tag{2}$$

where for each class, n_k and the σ_k are the size and the variance of those classes. Relevance of the feature set S is then defined as

$$max V_F, \qquad V_F = \frac{1}{|S|}\sum_{i \in S} F(i). \tag{3}$$

Redundancy is defined using the Pearson correlation for every possible feature combination:

$$min W_c, \qquad W_c = \frac{1}{|S|^2}\sum_{i,j} |c(i,j)|, \tag{4}$$

where absolute value of the correlation $c(i,j)$ is used. Finally, the optimization criteria for MRMR is

$$max(V_F - W_c). \tag{5}$$

Figure 8 roughly shows how the process of MRMR looks when we have various methods, including ml-MRMR.

2.2 Proposed feature selection algorithms

We propose several algorithms to improve the performance of the MRMR method for the depression detection problem where data is typically limited and, therefore, computation of F-statistic and correlation is unreliable.

2.2.1 Multi-lingual computation of relevance

The F-statistic computation in Eq.(1) assumes that there is enough data for each class to compute the mean and the variance of each class reliably, which is not the case when the number of classes is large and the data is limited. For instance, when there is only one sample for a class, the variance of that class becomes 0, which is problematic for the computation of the F-statistic. In the multi-lingual MRMR (ml-MRMR) approach, the core idea is to use the data, that is collected from the subjects that are talking in different languages and use it for relevance computation and make the process much more robust to scarcity situation of data.

Figure 1 briefly explains the process of the ml-MRMR algorithm. Small boxes represent each of every class and inside the boxes is the amount of sample the class has. As can be seen from the figure, the classes that don't have enough samples, tagged as yellow boxes, will be reinforced by the algorithm, which transports samples from other languages and fills it till the N_{min} constraint is satisfied.

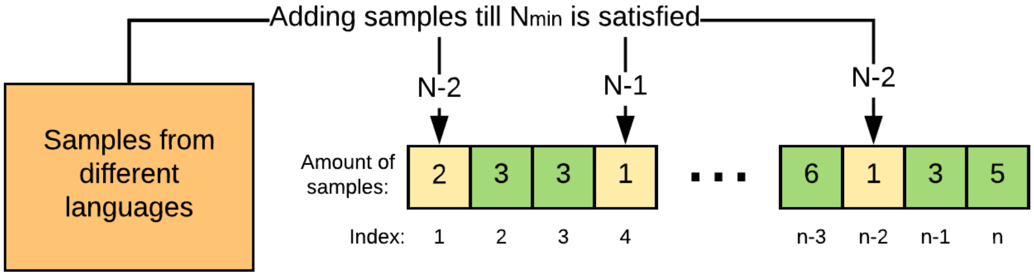
Adding samples till Nmin is satisfied
Samples from different languages
N-2
N-1
N-2
Amount of samples:
2
3
3
1
6
1
3
5
Index:
1
2
3
4
n-3
n-2
n-1
n

Figure 1: A flowchart that explains how the ml-MRMR algorithm works.

In order to increase the number of available samples for each class, hence improve the computation of relevance, we exploit the samples available in a different language for the same or neighboring classes with reduced weights assigned to the samples as the neighbors are further away on the depression scale. To that end, we have changed the computation of $\bar{g_k}$ and n_k. The weight parameter γ is defined as

$$\gamma_t = e^{-t^2}. \tag{6}$$

where t indicates how close the neighbors are on the depression scale. The number of samples in class k, $\hat{n}_k$ is adjusted using the parameter γ, the amount of adjustment depends on how much we need to satisfy the N_{min} constraint.

$$\hat{n}_k = \sum_{j=-J_k}^{+J_k} \gamma_j n_{k+j} \tag{7}$$

J is set such that $\hat{n}_k > N_{min}$. Thus, by including data from the same and neighboring classes in a different database, we ensure that there are at least N_{min} samples for each class in the target database. The adjusted mean of each class k, $\bar{g_k}'$, is then

$$\bar{g_k}^{'} = \frac{1}{\hat{n}_k} \sum_{j=-J_k}^{+J_k} \sum_{s=0}^{n_{k-j}-1} \gamma_j g_{k-j}(s)k-j(s) \tag{8}$$

where $g_{k-j}(s)$ is sample s in class $k-j$. Thus, the final equation to compute F-score becomes:

$$F(g_i) = [\sum_k \hat{n}_k(\bar{g_k}^{'} - \bar{g})^2/(K-1)]/\sigma^2, \tag{9}$$

Figure 2 presents the $\bar{g_k}$ for the baseline MRMR and ml-MRMR algorithm with $N_{min} = 3$ German, it can be seen that ml-MRMR's curve is much more close to the normal distribution with respect to the baseline MRMR algorithm's, which means our ml-MRMR algorithm tries to minimize the outliers of the dataset with enriching the samples for classes that lacks.

2.2.2 Clustering approach

Even though the Beck depression scale is from 1 to 63 with a step size of 1, given randomness in the responses to Beck questionnaire, the resolution is expected to be lower than that. Hence, the difference between a person with a score of 3 or 4 may not be as significant to warrant different classes for those two cases especially gave the very limited training data available.

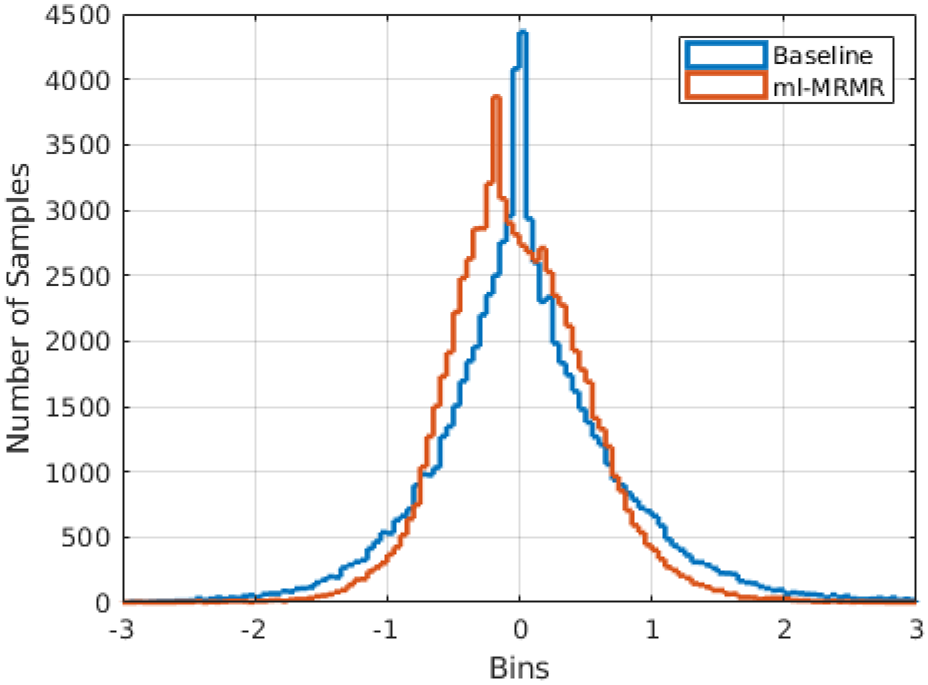

Figure 2: $\bar{g}_k$ histograms for the baseline Turkish MRMR and ml-MRMR.

In the clustering approach, we clustered the depression classes and reduced the number of classes in the MRMR training process to improve the feature selection performance by increasing the data available for each class. In this approach, data is split uniformly into N_{clus} classes.

2.2.3 Weighted F-Statistic

After the calculation of the F-Statistic for the 3 languages, we have mixed them using a linear approach that multiplies every language's F-Statistic with weight parameters. Thus, calculating the F-Statistic in a multi-lingual manner. The weights are shown in the Eq. 10 as a, b, and c. The motivation here was creating a multi-lingual algorithm without including within class calculations.

$$F_{final} = a * (F_{Tr}) + b * (F_{Ger}) + c * (F_{Eng}) \tag{10}$$

2.2.4 Robust computation of redundancy (RCR)

Class labels are not required for the computation of redundancy, as shown in Eq.(4). Thus, large amounts of unlabeled, i.e. depression scores not available, speech data can be exploited for computing the redundancy. In this approach, we propose using unlabeled speech databases to compute redundancy when the amount of labeled data is limited.

2.3 Fusion with text-based features

2.3.1 Description of text-based features

Sentiments in questions and patient responses were manually tagged for the Turkish database. Examples of positive, negative and neutral questions and answers are presented at Table 1. Counts of combinations of question and answer sentiments were used as a feature vector. Because we have three sentiments for questions and three sentiments for answers, a total of 9-dimensional sentiment feature vector was generated for each conversation.

Table 1: Example of an interview in the Turkish database. Sentiment labels of both questions and answers are shown.

Turn	Phrase	Sentiment
Question:	Can you tell us a happy moment lately?	Positive
Answer:	I don't have one for a long time.	Negative
Question:	Can you tell us an unhappy moment lately?	Negative
Answer:	Everything goes well lately.	Positive
Question:	What is your favorite food?	Neutral
Answer:	I like stuffed peppers.	Neutral

Using the timing information in the transcriptions, length of the utterances and the rate of speech were computed for each patient. Next, the average length and average rate of speech are computed for each sentiment. Thus, a total of 6 features were obtained for those two features. Concatenating them with the 9 features described above, a total of 15 features were derived from the transcriptions.

Table 2: Descriptions of text-features, a total of 15 dimensional feature set.

Feature	Description
Average Length of the Utterances	The length of the subject's answers for negative, positive, and neutral answers separately. 3 dimensional feature.
Rate of Speech	The length over time of the utterances for negative, positive, and neutral answers separately. 3 dimensional feature.
Sentimental Text Features	The sentiment of the answers and the questions taken into consideration and for each possible combination a feature created. 9 dimensional feature.

2.3.2 The fusion algorithm

The fusion algorithm is designed based on the observation that acoustics-only system often overestimates the depression level for the regression task. Those overestimations significantly impact the overall performance of the system and reduce its reliability.

In our proposed approach, instead of performing commonly used score or feature fusion methods, we used a co-training algorithm to adjust the scores produced by the acoustics-only system. In this approach, we first divided the data into two classes. Any patient with BDI-II score above 30 is tagged as class-1 and any patient with the score below 18 is tagged as class-2.

If the acoustic-only system generates a depression level prediction that is above 30 or below 18 and if the text-only system also produces a score in the same class (agreement case), then the score from the acoustic-only system is used. If they are in disagreement, i.e., one of the systems produces a prediction that is in class-1 and the other produces a prediction that is in class-2, the final prediction is computed by fine-tuning the acoustic-only prediction by getting it closer to the opposite class. If the prediction of the acoustic system is $p_{acoustic}$, final prediction p_{final} is computed by

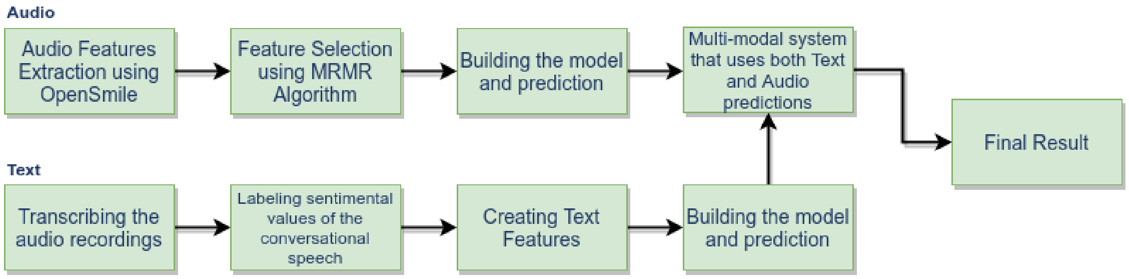
Audio
Audio Features Extraction using OpenSmile
Feature Selection using MRMR Algorithm
Building the model and prediction
Multi-modal system that uses both Text and Audio predictions
Final Result
Text
Transcribing the audio recordings
Labeling sentimental values of the conversational speech
Creating Text Features
Building the model and prediction

Figure 3: A flowchart that summarizes the general working process of the system.

the linear model:

$$p_{final} = \alpha p_{acou} \pm \Gamma \tag{11}$$

where α and Γ are constant parameters. α and Γ was found using GridSearch Algorithm.

2.4 Baseline system

In the baseline system, the MRMR feature selection method was first applied [66] to reduce the number of acoustic features. Support Vector Regression (SVR) was used for regression and SVM were used for classification. Because the amount of training data is small, the leave-one-out method was used for the Turkish and German experiments. There are enough training and test data for the English tasks, we did not use leave-one-out for it. The training and development partitions were used. The training set has 107 samples and the test has 35.

CHAPTER III

EXPERIMENTS

3.1 Experiment setup

3.1.1 Databases

3.1.1.1 Turkish Database

The Turkish database was collected at a hospital in Istanbul. It consists of 70 subjects. Mean age of the patients is 34. 14 of them are male and the rest is female. Beck scores of all subjects are available using the depression questionnaire, the Beck Depression Inventory-II (BDI-II) [67]. The average BDI-II score of the patients is 23.45 with a standard deviation of 11.01.

The Turkish database consists of interviews with the patients. Three types of questions were directed to the patients: neutral, positive and negative questions. Each question type refers to the sentiment that we expect to invoke in the patient. Similarly, the sentiment of the responses from the patients was manually-tagged by three independent evaluators. Majority voting was used for the final sentiment label of each response. Examples for sentiment labels are shown in Table 1.

The interview consists of 16 questions. Mean length of the conversations is approximate to 5 minutes. The total length of the recordings is 6 hours. They were recorded using a headphone microphone connected to a built-in sound card of a laptop with a sampling rate of 48 kHz.

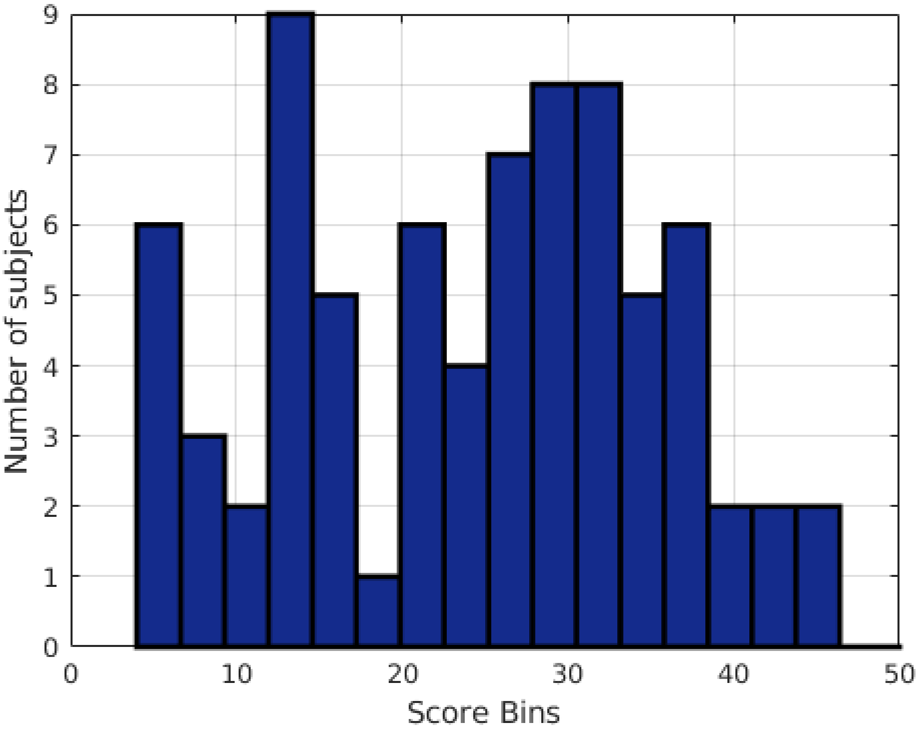

Figure 4: The distribution of BDI-II scores for Turkish database.

3.1.1.2 Structure of Turkish Database

Table 3.1.1.2 shows the top node of the Turkish database structure that is created.

First row indicates the name of the interview as, *XX_#_G_int* . *XX* stands for the first two letters of the patient's name and surname respectively. # indicates the number of recordings in case the patient has multiple recordings. *G* shows the gender of the patient, *int* basically means interview.

Table 3: The top node of the structure for Turkish.

Name	XX_#_G_int
TRScript	$Lx4$
ENGScript	$Lx1$
INTScript	$Mx2$
DepScore	D
AnxScore	A

Second row from the Table 3.1.1.2, *TRScript*, has the Turkish transcription of the patient's answers from the interview, the start and end time of the sentences that the patient created, and lastly the sentiment of the sentence of that turn. L indicates the total number of sentences of the answers of the patient in the interview. In Table 4, an example of a *TRScript* node has been presented. The first column shows example sentences, each row is the consecutive turn that patient answers the question of the interviewer. The second column shows the start time of that particular sentence, the third column is the end time of it. The fourth column is the sentimental value of the answer. Later the start-end time, and sentimental label information, which was used to create the text features.

Table 4: *TRScript* node of the main structure node.

Sentence	**Start Time**	**End Time**	**Sentimental Label**
Tavuk içeren her türlü yemeği seviyoru...	4.9715	9.6178	0
Yurtdışına gezmeye çıkmıştım üniversit...	14.7852	21.1630	1
Yani genelde annemle babamın yaşadı...	26.9325	34.5522	-1
Olay yaşandığı zaman mutsuz oluyorum.	34.5522	38.1451	-1
. .	.		.
. .	.		.
. .	.		.
. .	.		.

Third row of the Table 3.1.1.2, *ENGScript*, includes the English translations of the *TRScript*. The translations are made manually and used for automatic sentimental analysis using open-source tools, however, it has not been used, instead, we labeled the sentences manually using 3 volunteers, and majority voting was used to decide on the final label.

The fourth row of the Table 3.1.1.2, *INTScript*, has the Turkish transcriptions for the interviewer's questions turn by turn, M stands for the number of questions in the session. The second column of the *INTScript* is the sentiment information of the question. Later this information is being used for text features.

Last 2 rows show the depression score and anxiety score respectively. Depression scores are used for model training and predictions, however, anxiety scores are not part of this study.

3.1.1.3 German Database

The German database, distributed as part of the AVEC 2014 challenge [68], consists of conversations with 84 patients, some patients are in multiple recordings with a period of two weeks. Beck scores of the 100 recordings in the training and development data are available. However, they are not available for the 50 recordings in the test data. The mean age of German database subjects is 31.5. Duration of the recordings ranges from 6 seconds to 4 minutes. All recordings below 20 seconds were removed from the experiments and eventually 98 recordings are left for processing. The average BDI-II score of the patients is 15.0 with a standard deviation of 12.30.

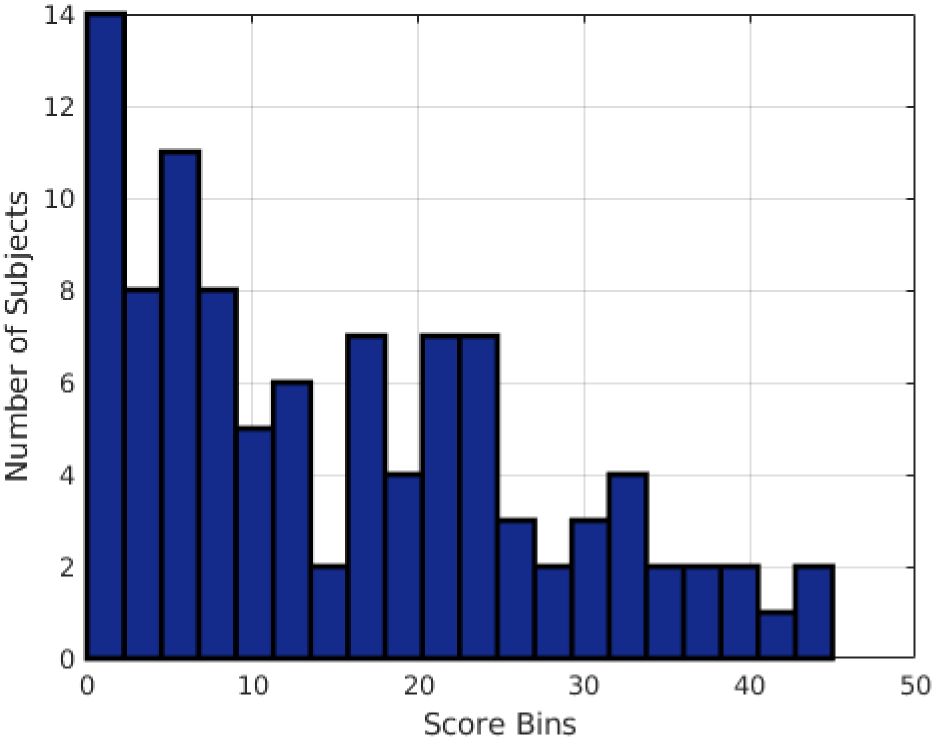

Figure 5: The distribution of BDI-II scores for German database.

3.1.1.4 English Database

The English database is part of The Distress Analysis Interview Corpus (DAIC) [69]. It contains clinical interviews designed to help diagnose psychological distress conditions such as anxiety, depression or post-traumatic stress disorder. The depression part of the corpus is the Wizard-of-Oz interviews, which are conducted by a virtual interviewer. The depression score of the patients was calculated using the PHQ-8 depression inventory [70], which differs from the German and Turkish databases. The average depression severity of the training and development data is 6.67, and the standard deviation is 5.75. Total of 189 recordings from 189 patients is available.

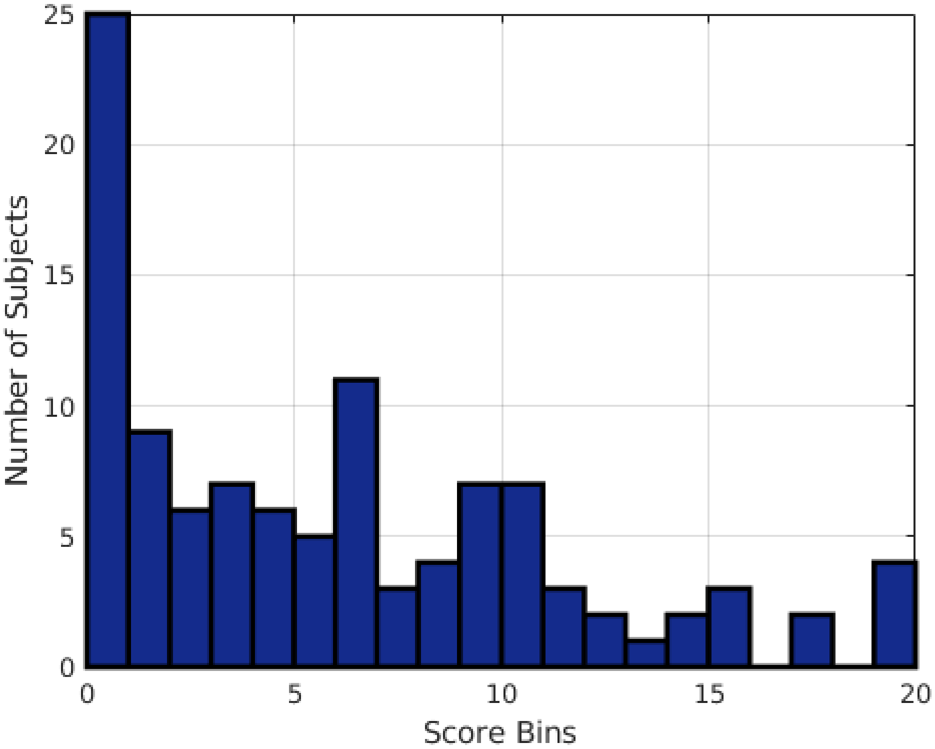

Figure 6: The distribution of PHQ-8 scores for English database.

3.1.2 Acoustic features extraction

The open-source toolkit OpenSMILE [71] was used for acoustic feature extraction. OpenSmile is a toolkit that can extract features based on the selected script. The backend system is written in C++, which makes it fast and efficient. It can be used on Linux, Windows or MacOS. The system supports writing data in CSV (Comma Separated Value) and ARFF (Weka Data Mining).

The AVEC 2013 and GeMAPS feature extraction scripts were used for our study. Feature vectors for AVEC 2013 includes 32 energy and spectral related low-level descriptors (LLDs) and their functionals [68] such as statistical functionals (maximum, mean, skewness, flatness etc.), regression functionals (linear regression slope, quadratic regression coefficient *a* etc.) and local minima/maxima related functionals (mean and standard deviation of rising and falling slopes etc.). 2268 dimensional features were extracted per speaker. Table 5 presents all the LLDs and functionals that are included in Avec 2013 script. Functionals were computed over 20 seconds time windows and averaged over the recording.

GeMAPS [72] has 18 low-level descriptors. Only the first 4 MFCC features are used because those are more crucial for affect and paralinguistic voice analysis studies [72]. In addition, jitter, shimmer, loudness, and spectral slope were used. Similar to AVEC 2013, functionals of those low-level descriptors were also computed. The dimensionality of the final feature set is 62.

3.1.3 Depression scores

3.1.3.1 Beck Depression Inventory II (BDI-II)

BDI-II is a revised version of the BDI questionnaire, which has been used by clinics for over 35 years to identify depression severity. An important problem with the early version of BDI was being gender biased and it could not keep up with the change of times. BDI-II addressed these issues and it has become a successful depression assessment questionnaire. It has 21 questions with 4 possible answers for each of them, the subject can choose possible answers from not present (0) to severe (3). BDI-II is a short questionnaire which only takes 5 to 10 minutes to complete, it also has clear guidelines which helps subjects to follow the task easily. The score ranges

Table 5: Low-level descriptors and functionals of Avec 2013 feature set.

Low-level Descriptors
Loudness, Zero Crossing Rate, Energy in Bands from 250-650 Hz, 1 kHz - 4 kHz, Spectral Roll-off Points, Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Psychoacoustic sharpness, Harmonicity, Flatness, MFCC 1-16, F_0, Probability of voicing, Jitter, Shimmer (local), Jitter of Jitter, LogHNR
Functionals
Arithmetic mean, Root quadratic mean, Standard deviaton, Flatness, Skewness, Kurtosis, Quartiles, Inter-quartile ranges, 1%, 99% Percentile, Percentile range 1%-99%, Up-level time: 25%, 50% and 90%, Maximum, Mean, Minimum segment length, standard deviation of segment length, Linear regression slope, Quadratic regression coefficient a, Mean and standard deviation of rising and falling slopes, Mean and standard deviation of inter maxima distances, Amplitude mean of maxima, Amplitude range of minima and maxima

of BDI-II is 0 to 63. The meaning of the score ranges are;

- **0-13** Minimal depression,

- **14-19** Mild depression,

- **20-28** Moderate depression,

- **29-63** Severe depression,

For the classification task, the scores were split into binary classes. For BDI-II scores that were used in the Turkish and German databases subjects that have scores below 18 were classified as non-depressed. Other patients were classified as depressed.

3.1.3.2 Personal Health Questionnaire Depression Scale (PHQ-8)

The PHQ-8 questionnaire has 8 questions, each question has 4 possible answers and the scoring goes from 0 to 3. At the end of the questionnaire, the corresponding scores for the answers are summed up and in total it makes minimum 0 and maximum 24.

This questionnaire mainly made for the use in the United States of America. It has been tested on a large corpus to prove it's reliability. The meaning of score ranges defined as;

- **0-4** No significant depressive symptoms,
- **5-9** Mild depressive symptoms,
- **10-14** Moderate depressive symptoms,
- **15-19** Moderately severe depressive symptoms,
- **20-24** Severe depressive symptoms,

The PHQ-8 scoring system was used in the English dataset we have used, subjects that have scores below 10 were classified as non-depressed and other patients were classified as depressed [70].

In regression, ml-MRMR algorithm requires databases to have same depression scales to compute within class statistics. Because the English database has PHQ-8 scores that range from 1 to 24 and the German and Turkish databases have Beck scores ranging from 1 to 63, we converted the Beck scores to PHQ-8 scores. During conversion, we uniformly split the 1-63 range into 24 segments and the Beck score in each segment is mapped to the scale of PHQ-8.

3.1.4 Significance test

Significance test is the answer to the question; "What is the likelihood that a random sample that was chosen is not a part of the population have?". The test also has significance level threshold, for this study we have selected 0.05. Basically, if the p value is lower than 0.5, the Null Hypothesis is rejected and it means the sample we have provided is significant.

3.1.4.1 T-test

There are various setups for this statistical approach. T-test assumes that the data is normally distributed, for our case we conducted right-tailed t-test. This is because our distribution is more likely fits the right-tailed assumption. The t-test is basically comparing two means and the standard deviations of the data. For regression tasks for all databases, T-test was used, which assumes that if $p < 0.05$ the regression results are significant.

3.1.4.2 McNemar Test

A statistical test which is being used on a paired nominal data is called McNemar's test. It is applied to a 2x2 contingency table. Table 6 shows an example of how a contingency table looks like.

Table 6: A 2x2 contingency table that shows the output of two tests.

	Test 2 Positive	**Test 2 Negative**
Test 1 Positive	a	b
Test 2 Negative	c	d

$$\chi^2 = \frac{(b-c)^2}{(b+c)} \tag{12}$$

The results from the Eq. 12 determines if the difference between the two models is significant or not. For the classification test we have used McNemar test, the confidence level was $p < 0.05$.

3.1.5 Machine-learning algorithm

In this study, all machine-learning applications, both classification and regression tasks were done by using Support Vector Machine (SVM). SVM is a member of the supervised machine-learning family, which can be used for classification and regression applications intuitively. The main idea of the SVM is to locate the best hyperplane that distinguishes the classes of the data.

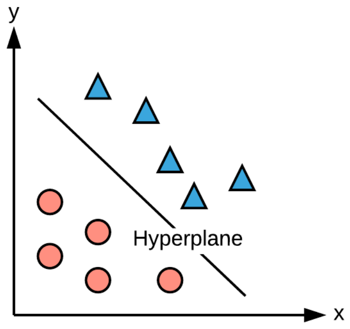

Figure 7: A basic representation of classification on a 2-D data.

To maximize the distances between hyperplane and the data points SVM algorithm uses distances called support vectors. These support vectors are the nearest distances that a data point has to the hyperplane. Thus, the algorithm trying to optimize these support vectors in order to maximize the separation of the classes.

Above explanation is the situation when we have data that can clearly separable with linear hyperplanes. However, that is not the case for most situations. When we have data that cannot be classified with linear methods, we are using kernels that expand our inventory of solutions.

In our study, we are using Radial Basis Function (RBF) kernels for all SVM applications. SVM also has *gamma* and C parameters which can be altered. C parameter can adjust the decision function's margin, which means when C gets bigger the accepted decision function will have a smaller margin. Otherwise, if the C value gets smaller the margin is encouraged to be bigger. Thus, the regularization parameter of the SVM is C. *gamma* is controlling the impact of a single training sample on the model.

For our study, the best *gamma* and C values are decided with using GridSearch algorithm for every machine learning application.

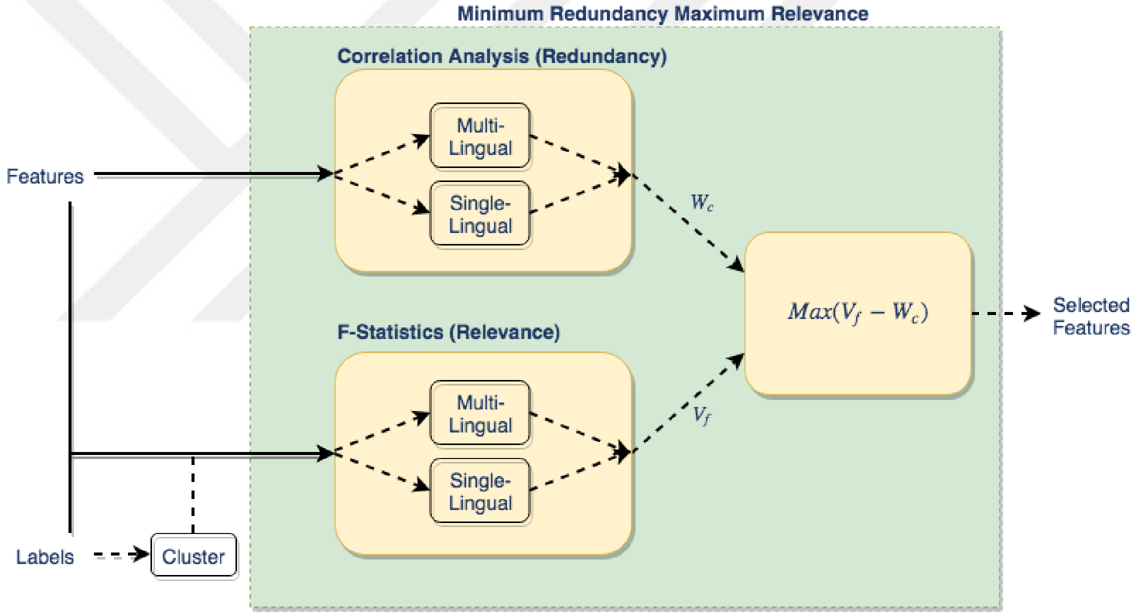

Figure 8: A flowchart that explains how MRMR works.

CHAPTER IV

RESULTS AND DISCUSSION

Two sets of experiments were conducted. In the first set, the proposed feature selection algorithms were tested and compared with the baseline MRMR algorithm for the German, Turkish and English tasks. The RCR algorithm proposed for redundancy computation in Section 2.2.4 was used only for the German and English tasks since unlabeled data are not available in the Turkish database. In the second set, text-based features were tested only with the Turkish database because the transcriptions were not available for the German database, and for the English database, the interviews were not in the question/answer format but rather a free-form talk between human and computer.

The evaluation criteria for all regression experiments were the root mean square error (RMSE), which is also used in the AVEC challenges [1, 2, 68, 73]. Statistical significance of the results were tested using the t-test with $p < 0.05$.

The evaluation criteria for all classification systems were F1-score, precision, and recall for both depressed and non-depressed classes. For the classification tasks, the statistical significance of results was measured using McNemar's test with $p < 0.05$.

4.1 Performance of the ml-MRMR feature selection

4.1.1 Turkish task

Table 7 and Table 8 show the results with the baseline and the ml-MRMR algorithms for the Turkish task. Best result with the regression model was 9.36 with the Turkish-English ml-MRMR $N_{min} = 3$ algorithm and the improvement compared to the baseline was statistically significant. Similarly, Turkish-German ml-MRMR algorithm performed better than the baseline and the difference was statistically significant. Gemaps feature set performed worse than the ml-MRMR algorithms, which has the result of 11.48. It is also can be seen that when we increase the threshold of N_{min} the result got worse.

The ml-MRMR algorithm cannot be applied directly in the case of classification because there are only two classes and each class has enough samples, 27 non-depressed and 50 depressed. Still, we used the ml-MRMR algorithm to enrich each class with samples from other languages by treating the problem as if it was a regression problem during the selection of cross-lingual samples. After each class (from 1 to 45) is enriched with those samples, training data is split into two classes.

Classification results are shown in Table 8. Even though ml-MRMR algorithm improves the performance, the improvement was not found to be statistically significant. Thus, in the classification case where there is enough data in each class for training the classifier, the ml-MRMR algorithm was not as effective. Still, the system trained with the text-based features significantly outperformed the other systems which are exploited in our fusion-based system described in Section 2.3.2.

4.1.2 English task

Table 9 shows regression results for English task. Best result obtained by using the ml-MRMR algorithm with Turkish $N_{min} = 5$. ml-MRMR using German database did not perform well, although there are slight improvements it is not significant.

Table 10 shows the results for the English classification task. The ml-MRMR algorithm with Turkish using $N_{min} = 5$ outperformed the baseline model, but they are not statistically significant. However, when German data was used, performance did not change. The baseline results from the Avec 2016 [2] are also reported in Table 10 and the improvement with ml-MRMR $N_{min} = 5$ Turkish is statistically significant compared to the Avec 2016 baseline result. The improvements with the ml-MRMR

Table 7: Regression performance of the ml-MRMR Methods for the Turkish task when the minimum occurrence threshold (N_{min}) is set 3 and 5. In the underlined bold case, improvement is significant compared to the baseline system. Result with Gemaps feature set was: 11.48

Dim	Baseline	$N_{min} = 3$ (Tr+Ger)	$N_{min} = 5$ (Tr+Ger)	$N_{min} = 3$ (Tr+Eng)	$N_{min} = 5$ (Tr+Eng)
2	13.30	10.84	11.98	10.79	10.88
3	12.30	**10.51**	11.26	10.40	11.94
4	12.45	10.85	**10.74**	9.85	12.68
5	12.56	10.58	11.23	**9.36**	13.65
10	12.45	10.82	12.13	11.87	13.93
15	12.08	11.12	12.00	11.99	13.30
20	12.87	11.91	11.46	10.92	12.00
40	13.28	12.67	11.98	10.93	10.31
80	11.58	12.28	13.06	10.80	10.50
100	11.75	11.95	13.08	10.88	**10.23**
200	11.32	11.55	12.14	11.05	11.06
400	11.42	11.72	12.00	10.99	11.23
800	11.31	11.39	11.35	11.10	11.08

Table 8: Best classification results for Turkish Task. Avec 2013 feature set used for all results except Gemaps tab. The distribution of the scores was 50 Depressed, 27 Non-Depressed. Best F1-score results are shown in bold. Only the classification with text features tab is statistically significant compared to baseline.

Method	Classes	Precision	Recall	F1-score
Baseline MRMR:3	**Non-Depressed**	0.61	0.40	0.48
	Depressed	0.72	0.86	0.78
	Average	0.67	0.63	0.63
English N_min = 5 ml-MRMR:400	**Non-Depressed**	0.59	0.48	**0.53**
	Depressed	0.74	0.82	0.78
	Average	0.66	0.65	0.66
German N_min = 3 ml-MRMR:100	**Non-Depressed**	0.42	0.29	0.34
	Depressed	0.67	0.78	0.72
	Average	0.54	0.54	0.53
Only Text Features MRMR:7	**Non-Depressed**	0.78	0.40	0.53
	Depressed	0.74	0.94	**0.83**
	Average	0.76	0.67	**0.68**
GEMAPS	**Non-Depressed**	0.38	0.37	0.37
	Depressed	0.66	0.68	0.67
	Average	0.52	0.53	0.52

Table 9: Regression performance of the ml-MRMR Methods for the English task when the minimum occurrence threshold (N_{min}) is set 5. In the bold case, improvement is insignificant compared to the baseline system, however it is significant compared to the baseline of the challenge paper [1]. Result with Gemaps feature set was: 6.72

Dim	**Baseline**	$N_{min} = 5$	$N_{min} = 5$
		(Ger+Eng)	**(Tr+Eng)**
3	6.85	6.66	6.66
4	6.87	7.21	7.20
5	7.31	7.78	7.78
10	7.85	7.05	7.05
15	8.08	7.94	7.70
20	7.89	7.63	7.46
40	7.25	7.12	6.63
80	7.67	6.26	6.37
100	7.57	6.38	**6.15**
200	7.13	6.42	6.70
400	6.95	6.72	6.73
800	6.92	6.74	6.66

algorithm are obtained on both for F1 scores of both depressed and non-depressed subjects.

4.1.3 German task

Regression performance of the baseline and the proposed ml-MRMR and RCR feature selection algorithms for the German task are shown in Table 11. In the multi-lingual approach, Turkish database was used to supplement additional features for each depression class in the German database when the number of samples is less than N_{min} as described in Section 2.2.1. Even though performance improved for $N_{min} = 3$, the improvement was not significant. Improvement with $N_{min} = 5$ was found to be significant only when the RCR algorithm was also used. RCR algorithm was not effective when it was used by itself.

For the classification task, Table 12 shows the results. The baseline feature selection algorithm has the best results for non-depressed recall and depressed precision with the 10-dimensional feature set. However, the ml-MRMR algorithm with Turkish

Table 10: Best classification results for English task Development set. Multi-Lingual methods annotated with M-L. Avec 2013 feature set used for all results except GeMaps tab. Best F1-score results are shown in bold. Results are not statistically significant compared to baseline MRMR, however they are statistically significant compared to the baseline results from Avec 2016 [2].

Method	Classes	Precision	Recall	F1-score
Baseline MRMR:20	Non-Depressed	0.76	0.96	0.85
	Depressed	0.83	0.42	0.56
	Average	0.79	0.69	0.71
Baseline Avec 2016 Results from [2]	Non-Depressed	0.93	0.54	0.68
	Depressed	0.31	0.85	0.46
	Average	0.62	0.69	0.57
Turkish	Non-Depressed	0.79	1	**0.88**
N_min = 5	Depressed	1	0.50	**0.66**
ml-MRMR:100	Average	0.89	0.75	**0.77**
German	Non-Depressed	0.76	0.96	0.85
N_min = 5	Depressed	0.83	0.42	0.56
ml-MRMR:20	Average	0.79	0.69	0.71
GEMAPS	Non-Depressed	0.70	0.82	0.76
	Depressed	0.50	0.33	0.40
	Average	0.60	0.58	0.58

using $N_{min} = 3$ outperformed on the rest of the indicators. ml-MRMR algorithm with English performed worse than the baseline.

4.2 Performance of clustering algorithms

The clustering algorithm proposed in Section 2.2.2 for the Turkish database was used with 2, 9, and 15 clusters instead of the 63 distinct classes available in the Beck scores. Results are shown in Table 14. Even though the system with 15 clusters significantly outperformed the baseline system, the improvement was still below what was obtained with the multi-lingual MRMR approach.

4.3 Performance of weighted F-statistic

The weighted F-statistic algorithm mixes three languages into one F-statistic computation. The detailed explanation of the algorithm can be found in Section 2.2.3. Table 13 shows the results for the weighted F-statistic algorithm. The tests were done

Table 11: Regression performance of the ml-MRMR Methods for the German database when the minimum occurrence threshold N_{min} is set 3 and 5. Results are shown both when the RCR algorithm is used and not used. Best results are shown in bold together with their statistical significance using t-test. Result with GEMAPS feature set was: 10.14.

Dim	**Baseline**	$N_{min} = 5$ **(Ger+Tr)**	$N_{min} = 5$ **and RCR** **(Ger+Tr)**	$N_{min} = 3$ **(Ger+Tr)**	$N_{min} = 3$ **and RCR** **(Ger+Tr)**	**RCR**
10	9.90	9.97	9.99	10.37	12.39	10.02
15	9.81	10.21	**9.43** $p(0.01)$	10.13	12.12	10.08
20	9.86	10.32	9.52	9.84	10.68	9.74
40	10.25	10.35	10.45	9.73	11.36	10.22
80	10.69	9.93	9.88	**9.42** $p(0.47)$	10.93	10.06
100	10.48	9.93	9.74	9.50	10.54	10.17
200	10.12	10.00	10.38	9.69	10.28	10.44
400	10.14	9.79	10.21	9.58	10.29	10.13
800	10.08	9.86	10.11	9.91	10.11	9.89
1000	10.02	9.85	10.14	9.79	10.16	9.98

Dim	$N_{min} = 3$ **(Ger+Eng)**	$N_{min} = 3$ **and RCR** **(Ger+Eng)**	$N_{min} = 5$ **(Ger+Eng)**	$N_{min} = 5$ **and RCR** **(Ger+Eng)**
10	11.16	11.24	12.75	13.40
15	11.54	11.54	12.44	12.49
20	11.76	11.70	12.27	13.15
40	11.47	11.57	12.35	11.48
80	9.79	9.79	11.01	10.89
100	10.41	10.37	10.83	10.38
200	9.96	9.96	10.55	9.95
400	9.66	**9.60** $p(0.22)$	10.43	10.00
800	9.64	9.71	10.34	10.32
1000	9.71	9.68	10.41	10.28

Table 12: Best classification results for German Task. Multi-lingual methods annotated with M-L. Avec 2013 Feature set used for all results except GeMaps Tab. Best F1-score results are shown in bold. Although there are improvements, results are insignificant compared to the baseline MRMR predictions.

Method	Classes	Precision	Recall	F1-score
Baseline MRMR:10	**Non-Depressed**	0.76	0.87	0.81
	Depressed	0.80	0.65	0.72
	Average	0.78	0.76	0.77
Turkish N_min = 3 ml-MRMR:40	**Non-Depressed**	0.83	0.79	**0.81**
	Depressed	0.74	0.80	**0.77**
	Average	0.79	0.80	**0.79**
English N_min = 5 ml-MRMR:800	**Non-Depressed**	0.73	0.78	0.75
	Depressed	0.70	0.63	0.66
	Average	0.72	0.71	0.71
GEMAPS	**Non-Depressed**	0.74	0.70	0.72
	Depressed	0.64	0.68	0.66
	Average	0.69	0.69	0.69

on the Turkish database. The best result is obtained by using 0.3 Tr, 0.4 Eng, and 0.3 Ger, 9.62, which is also statistically significant compared to the baseline.

4.4 Performance of score fusion

Table 15 shows results when speech-based features were fused with text-based features using the proposed approach described in Section 2.3.2. Fusion algorithm significantly improved the performance (p-value=0.00006) compared to the baseline case by reducing the error more than 25% using ml-MRMR with English and Turkish ($N_{min} = 3$). Spread of the prediction errors is substantially reduced after fusion as shown in Figure 9.

Fifth column shows the results obtained with the clustering approach together with the fusion method. That algorithm not only outperformed the baseline but also outperformed the base-fusion algorithm significantly.

Comparison of real and predicted scores are shown in Figure 10 for the baseline and the best ml-MRMR algorithm with fusion. Predictions get closer to the true scores and errors significantly decrease with the proposed fusion method as shown

Table 13: Regression results with weighted feature selection. The weight parameters a,b, and c are changing. Turkish database is used. Statistically significant ($p < 0.05$) improvement is shown in bold.

Dim	Baseline	0.8 Tr- 0.1 Eng 0.1 Ger	0.7 Tr- 0.2 Eng 0.1 Ger	0.6 Tr- 0.3 Eng 0.1 Ger	0.5 Tr- 0.4 Eng 0.1 Ger	0.7 Tr- 0.1 Eng 0.2 Ger	0.6 Tr- 0.1 Eng 0.3 Ger
3	12.30	11.33	11.33	11.33	11.33	11.33	11.02
4	12.30	10.03	10.04	10.04	10.04	10.04	11.65
5	12.45	10.11	10.12	10.12	10.12	10.12	10.93
10	12.56	10.53	10.54	10.54	10.54	11.50	11.50
15	12.08	10.54	10.54	10.55	10.55	10.35	10.35
20	12.87	10.73	10.74	10.74	10.74	10.73	10.58
40	13.28	10.73	10.74	10.74	10.49	10.54	10.54
80	11.58	10.37	10.02	10.00	10.05	10.02	9.95
100	11.75	10.35	10.02	10.02	10.12	10.03	10.06
200	11.32	10.48	10.49	10.62	10.63	10.50	10.73
400	11.42	11.22	11.27	11.31	11.31	11.37	11.48
800	11.31	11.45	11.52	11.50	11.62	11.48	11.47

Dim	Baseline	0.5 Tr- 0.1 Eng 0.4 Ger	0.6 Tr- 0.2 Eng 0.2 Ger	0.5 Tr- 0.2 Eng 0.3 Ger	0.4 Tr- 0.2 Eng 0.4 Ger	0.5 Tr- 0.3 Eng 0.2 Ger	0.4 Tr- 0.3 Eng 0.3 Ger
3	12.30	11.02	11.33	11.02	11.02	11.33	11.60
4	12.30	11.65	10.04	11.65	11.65	10.04	11.65
5	12.45	10.93	10.12	10.93	10.93	10.12	10.93
10	12.56	10.49	11.50	11.50	10.49	11.50	11.50
15	12.08	10.13	10.35	10.35	10.13	10.35	10.33
20	12.87	10.00	10.73	10.20	10.00	10.73	10.20
40	13.28	11.05	10.54	10.56	11.05	10.54	10.56
80	11.58	10.05	10.02	10.01	10.06	9.95	10.03
100	11.75	10.34	9.92	10.18	10.34	9.92	10.20
200	11.32	10.65	10.43	10.62	10.61	10.55	10.68
400	11.42	11.49	11.42	11.46	11.48	11.47	11.52
800	11.31	11.54	11.45	11.50	11.67	11.50	11.71

Dim	Baseline	0.3 Tr- 0.3 Eng 0.4 Ger	0.4 Tr- 0.4 Eng 0.2 Ger	0.3 Tr- 0.4 Eng 0.3 Ger	0.2 Tr- 0.4 Eng 0.4 Ger
3	12.30	11.02	11.33	11.60	11.60
4	12.30	11.65	11.14	11.65	11.65
5	12.45	10.93	10.28	10.93	10.93
10	12.56	10.49	11.50	11.23	10.49
15	12.08	10.13	10.35	**9.62**	10.60
20	12.87	10.32	10.73	10.20	10.45
40	13.28	10.90	10.71	10.44	11.11
80	11.58	10.06	10.04	10.06	10.05
100	11.75	10.66	10.24	10.32	10.70
200	11.32	10.59	10.62	10.62	10.66
400	11.42	11.60	11.46	11.64	11.57
800	11.31	11.66	11.70	11.67	11.67

Table 14: Regression results with feature selection using the clustering approach with 2, 9, and 15 clusters. Turkish database is used. Statistically significant ($p < 0.05$) improvement is shown in bold.

Dim	Baseline	2-Cluster	9-Cluster	15-Cluster
5	12.56	11.35	13.14	11.99
10	12.45	10.95	13.42	12.25
15	12.08	11.13	13.07	11.75
20	12.87	11.74	13.23	12.95
40	13.28	12.33	13.73	12.06
80	11.58	12.72	13.33	**10.83**
100	11.75	13.22	13.09	10.97
200	11.32	11.72	12.66	11.50
400	11.42	11.83	12.00	11.40
800	11.31	11.62	11.70	11.64

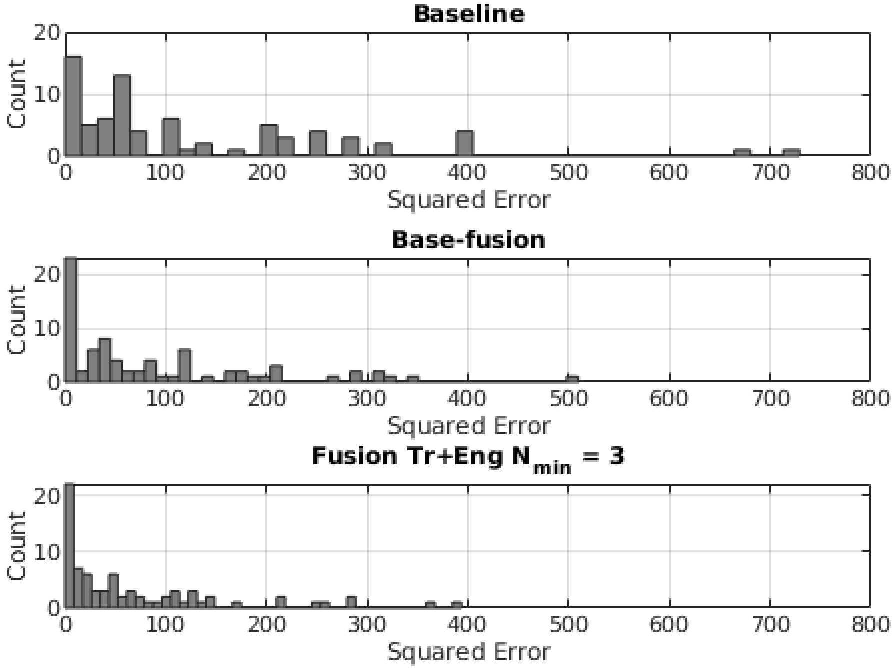

Figure 9: Distribution of squared errors for the baseline MRMR case is shown in the top figure. The middle figure shows the squared error distribution for the Baseline MRMR after fusion. Bottom figure shows the squared error distribution with $N_{min} = 3$ with English and after fusion.

Table 15: Regression results after fusing with text classification predictions. Turkish database was used. Baseline acoustic system predictions are used in base-fusion. Bold results show cases where the improvement is significant compared to the baseline case but not to the base-fusion case. In the underlined bold case, improvement is significant both compared to the baseline system and the base-fusion system.

Dim	**Baseline**	**Base-fusion**	**Fusion Tr+Ger**	**Fusion**	**Fusion Tr+Eng**
			$N_{min} = 3$	**15 Clus.**	$N_{min} = 3$
3	12.30	10.71	9.76	9.75	9.43
4	12.45	10.68	**9.54**	9.74	8.88
5	12.56	10.69	9.61	9.78	**8.30**
10	12.45	10.91	9.83	10.05	9.74
15	12.08	10.44	9.63	9.95	10.12
20	12.87	10.91	9.65	**9.66**	9.62
40	13.28	11.33	10.45	10.88	9.55
80	11.58	10.12	10.67	9.99	9.37
100	11.75	10.25	10.77	10.24	9.45
200	11.32	**10.03**	9.98	10.40	9.76
400	11.42	10.29	10.00	9.78	9.60
800	11.31	10.31	10.19	9.88	9.79

in Figure 10. The best RMSE is 8.30, which is interestingly obtained with only 5 features. Three of the 5 selected features are MFCC related: peak standard deviation of MFFC-5, amplitude mean of maxima for MFCC-5 and mean segment length of MFCC-14. The other two is mean of the rising slope for spectral harmonicity and up-level time(25) of spectral flatness. Those features are described in detail in Table 17.

Fourth column in Table 15 shows the results for the ml-MRMR algorithm with German and Turkish when $N_{min} = 3$. Even though that approach worked well compared to the baseline, it did not perform as well as the Turkish and English case. Moreover, its performance was not significantly different from the base-fusion. These results are aligned with the results reported in Table 7 where performance with Turkish and English datasets was better compared to the Turkish and German datasets.

Fifth column shows the results obtained with the clustering approach together with the fusion method. That algorithm not only outperformed the baseline but also outperformed the base-fusion algorithm significantly.

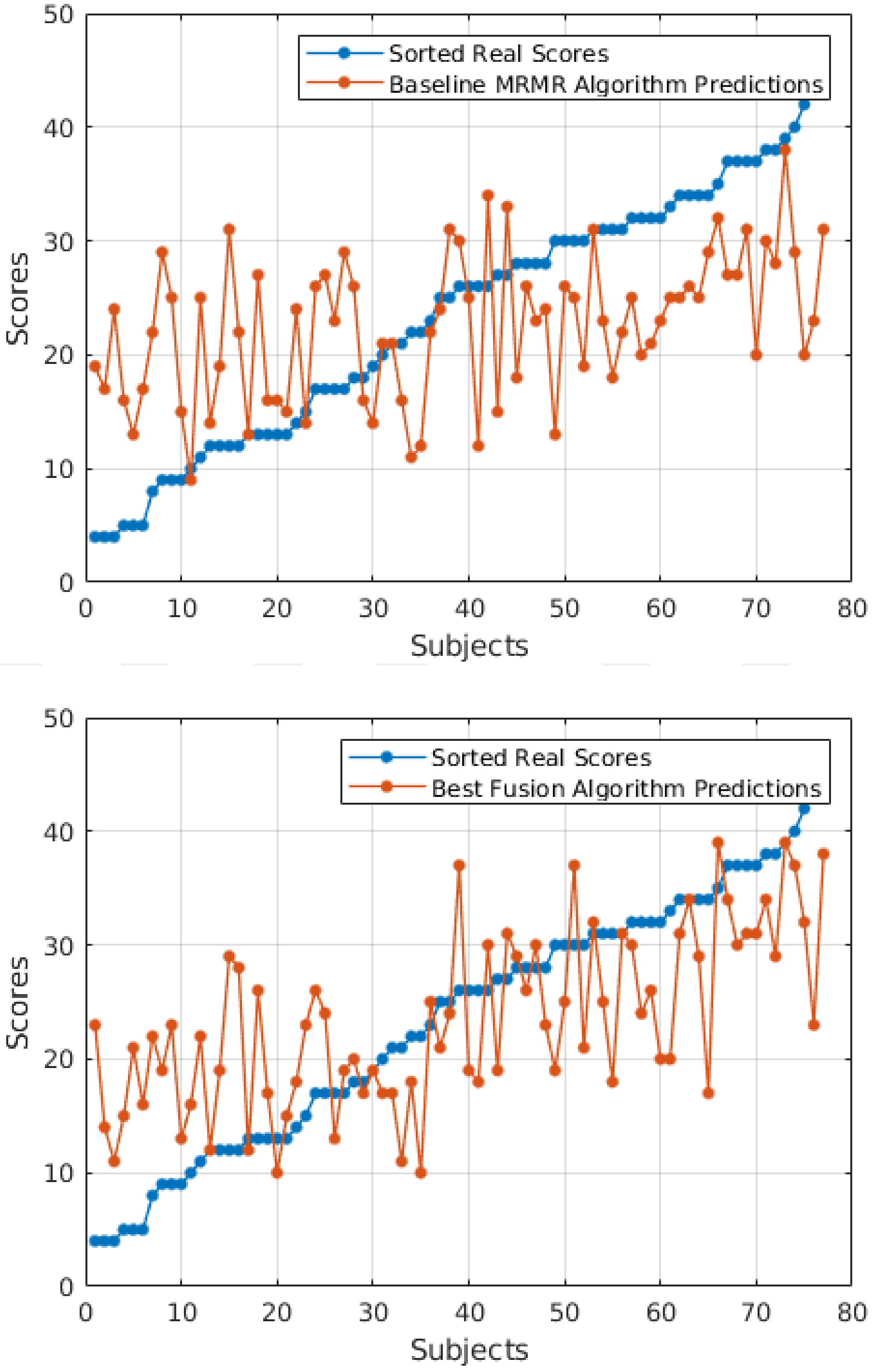

Figure 10: The sorted real scores and the baseline MRMR predictions at the top. Bottom figure shows the real scores vs. $N_{min} = 3$ with Tr+Eng predictions.

Table 16: Classification results after fusing with text classification. Turkish database was used. Best F1-score results are shown in bold. The results are statistically significant except $N_{min} = 3$ with German.

Method	Classes	Precision	Recall	F1-score
Baseline MRMR:3	**Non-Depressed**	0.61	0.40	0.48
	Depressed	0.72	0.86	0.78
	Average	0.67	0.63	0.63
Fusion Eng	**Non-Depressed**	0.71	0.63	**0.67**
$N_{min} = 3$	**Depressed**	0.81	0.86	0.83
ml-MRMR:100	**Average**	0.76	0.75	**0.75**
Fusion Ger	**Non-Depressed**	0.75	0.55	0.63
$N_{min} = 3$	**Depressed**	0.78	0.90	**0.84**
ml-MRMR:4	**Average**	0.76	0.73	0.73
Only Text	**Non-Depressed**	0.78	0.40	0.53
Features	**Depressed**	0.74	0.94	0.83
MRMR:7	**Average**	0.76	0.67	0.68

For the classification in Table 8 text-only model has the best results for the precision of non-depressed 0.78, and the recall of depressed 0.94. However, when the text-only model used for fusion algorithm with acoustic predictions, the f1-scores outperformed relative models, which is 0.67 for non-depressed and 0.83 for depressed when $N_{min} = 3$ English is used. The result is statistically significant with p-value of $p = 0.02$. In addition, when $N_{min} = 3$ German is used the f1-score for depressed is better than the English used ml-MRMR. Although it is not statistically significant, with German, only 4 features were used in comparison to 100 features of English-Turkish.

4.5 Common Selected Features Among Languages

Using the MRMR feature selection algorithm across the languages and selecting the features that are most relevant has shown which acoustic features are important for predicting depression. Across the three languages, we have examined the common selected features. The first row of Table 17 shows the 150 most important selected features that are overlapping between Turkish and English, the second row presents it for Turkish and German, the third row is for English and German.

As can be seen from Table 17, different languages have common acoustic cues that might lead to observe depression. Even mostly phonetically distant languages like German and Turkish or English and Turkish shares spectral information, which is tied to the phonetic structure of the languages.

The common features between English and German are mostly frequency related. Jitter is the descriptor of disruption of the frequency, and German and English languages have 4 common Jitter based features. This indicates the similarity of prosody between western languages. Also, Table 17 shows, higher order MFCCs are effective for the depression prediction for both German and English languages.

The common acoustic features for Turkish and German languages are mostly higher-order MFCCs, which is a widely used acoustic feature for voice quality classification tasks and it is also effective for distinguishing healthy and unhealthy speech. Lastly, for Turkish and English the common features are mostly energy related.

The details about the selected common low-level descriptors can be found at Table 18, and the functionals can be found at Table 19.

Table 17: For all the languages we studied, these are the common selected features among languages that are in top 150.

Language Combination	LLD & Functionals
Turkish-English	MFCC-13 - Relative Mean of Peaks
	Energy in Band 1000-4000 - Kurtosis
	Spectral Harmonicity - Up-level Time:90
	Spectral Harmonicity - Up-level Time:50
	Energy in Band 1000-4000 - Up-level Time:25
	Energy in Band 1000-4000 - Relative Mean of Peaks
	Energy in Band 1000-4000 - Minimum Segment Length
Turkish-German	Spectral Skewness - IQR 2-3
	Spectral Skewness - IQR 1-3
	MFCC-4 - Mean Distance Between Peaks
	MFCC-9 - Rise-time
	MFCC-11 - Rise-time
	MFCC-13 - Skewness
	MFCC-14 - Flatness
	MFCC-14 - Kurtosis
English-German	Energy in Band 250-650 - Quartile 1
	Energy in Band 250-650 - Percentile 1.0
	Spectral Harmonicity - Up-level Time:25
	JitterLocal - Mean
	JitterLocal - Standard Deviation
	MFCC-10 - Mean Segment Length
	MFCC-14 - Up-level Time:90
	MFCC-16 - Mean Distance Between Peaks
	JitterDDP - Up-level Time:50
	JitterDDP - Up-level Time:90

Table 18: The descriptions for the common LLDs across 3 languages.

Low-level Descriptor	Description
MFCC 1-16	Mel Frequency Cepstral Coefficient is a common used automatic speech recognition (ASR) feature, in the Avec 2013 feature set 16 dimension were used.
Energy	Sum squares of amplitudes of a signal.
Spectral Harmonicity	Aspect and number of the harmonics in a signal.
Spectral Skewness	The third order moment of the power spectrum.
Jitter (Local)	Length variation of the fundamental period from a single period towards the next one.
Jitter (DDP)	Delta period-to-period jitter can be defined as ”Jitter of the Jitter”. It is explained as the change between two successive period-to-period jitters.

Table 19: The functionals that has been used with the LLDs.

Statistical Functionals	**Description**
Relative Mean of Peaks	Proportion of the mean of the peak amplitudes to the mean of windowed signal.
Kurtosis	A higher order statistical moment, fourth standardised moment.
Skewness	A higher order statistical moment, third standardised moment.
Up-level Time	Time or number of frames that the signal is above a threshold, the possible percentiles for the threshold are 25,50,75,90.
Minimum Segment Length	Minimum length of a particular segment.
Mean Segment Length	Arithmetic mean of a particular segment.
Inter Quartile Range 1-2-3 (IQR)	The range between two percentile. The possible combination of quartiles are 1-3, 1-2 and 2-3.
Mean distance between peaks	The mean of distances between the peaks, where a signal reaches the highest.
Rise-time	The time where the signal contour is rising.
Percentile 1.0	The minimum value of a signal.

CHAPTER V

CONCLUSION

5.1 Conclusion and future work

This thesis is an expanded version of the study that we presented at the Interspeech 2018 conference in Hyderabad, India. We have added the database that was presented at the AVEC 2016 conference, which is in the English language. The addition has shown us that the proposed multi-lingual algorithm was applicable to more languages other than German and Turkish. In addition, we have seen Turkish-English multi-lingual feature selection performed better than the Turkish-German setup.

We investigated using multi-lingual databases for feature selection in the context of depression assessment, which was found to be effective. This result is significant not only because it is a step towards using larger multi-lingual databases for depression detection, but also it indicates that there are similarities between entirely different languages in the way they manifest depression.

As a second contribution, we proposed novel features derived from transcriptions of the Turkish database and fused them with the acoustic features which significantly improved the performance. The study has shown an RMSE improvement of nearly 25% when the fusion algorithm was used.

In the study, all the results that were reported have comparisons with the baseline results that were reported by the database providers. We have outperformed these results and also some of the findings that we reported are slightly better than the state-of-art studies. Especially, the result we reported for English database classification task is remarkable compared to the recent studies.

In possible future work, more languages could be added to the database and continue to improve the feature selection process. Moreover, we believe that our text features are also language-independent and we will investigate fusion algorithms in a multi-lingual setting.

Bibliography

[1] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, “Avec 2017: Real-life depression, and affect recognition workshop and challenge,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3–9, ACM, 2017.

[2] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, “Avec 2016: Depression, mood, and emotion recognition workshop and challenge,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, ACM, 2016.

[3] A. Halfin, “Depression: the benefits of early and appropriate treatment.,” *The American journal of managed care*, vol. 13, no. 4 Suppl, pp. S92–7, 2007.

[4] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, “Cost of lost productive work time among us workers with depression,” *Jama*, vol. 289, no. 23, pp. 3135–3144, 2003.

[5] W. H. Organization *et al.*, “Causes of death 2008: data sources and methods,” *World Health Organization: Geneva, Switzerland*, 2011.

[6] W. H. Organization *et al.*, “The global burden of disease: 2004 update. geneva, switzerland: Who, 2008,” *World Health Organization*, 2011.

[7] Z. Rihmer, “Can better recognition and treatment of depression reduce suicide rates? a brief review,” *European Psychiatry*, vol. 16, no. 7, pp. 406–409, 2001.

[8] M. P. Black, A. Katsamanis, B. R. Baucom, C.-C. Lee, A. C. Lammert, A. Christensen, P. G. Georgiou, and S. S. Narayanan, “Toward automating a human behavioral coding system for married couples interactions using speech acoustic features,” *Speech communication*, vol. 55, no. 1, pp. 1–21, 2013.

[9] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, “Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions,” *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[10] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, “Modeling therapist empathy and vocal entrainment in drug addiction counseling.,” in *INTERSPEECH*, pp. 2861–2865, 2013.

[11] P. K. Kuhl, S. Coffey-Corina, D. Padden, and G. Dawson, “Links between social and linguistic processing of speech in preschool children with autism: behavioral and electrophysiological measures,” *Developmental science*, vol. 8, no. 1, 2005.

[12] P. J. Moses, “The voice of neurosis.,” 1954.

[13] J. K. Darby and H. Hollien, “Vocal and speech patterns of depressive patients,” *Folia Phoniatrica et Logopaedica*, vol. 29, no. 4, pp. 279–291, 1977.

[14] E. Szabadi, C. Bradshaw, and J. Besson, “Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression,” *The British Journal of Psychiatry*, vol. 129, no. 6, pp. 592–597, 1976.

[15] J. F. Greden, A. A. Albala, I. A. Smokler, R. Gardner, and B. Carroll, “Speech pause time: a marker of psychomotor retardation among endogenous depressives.,” *Biological Psychiatry*, 1981.

[16] J. F. Greden and B. J. Carroll, “Decrease in speech pause times with treatment of endogenous depression.,” *Biological Psychiatry*, 1980.

[17] J. Greden, “Biological markers of melancholia and reclassification of depressive disorders.,” *L’Encephale*, vol. 8, no. 2, pp. 193–202, 1982.

[18] R. Jouvent, D. Widlöcher, *et al.*, “Speech pause time and the retardation rating scale for depression (erd): Towards a reciprocal validation,” *Journal of affective disorders*, vol. 6, no. 1, pp. 123–127, 1984.

[19] Å. Nilsonne, “Acoustic analysis of speech variables during depression and after improvement,” *Acta Psychiatrica Scandinavica*, vol. 76, no. 3, pp. 235–245, 1987.

[20] W. Weintraub, *Verbal behavior: Adaptation and psychopathology.* Springer Publishing Company, 1981.

[21] S. Rude, E.-M. Gortner, and J. Pennebaker, “Language use of depressed and depression-vulnerable college students,” *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004.

[22] J. Zinken, K. Zinken, J. C. Wilson, L. Butler, and T. Skinner, “Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression,” *Psychiatry research*, vol. 179, no. 2, pp. 181–186, 2010.

[23] P. Waxer, “Nonverbal cues for depression.,” *Journal of Abnormal Psychology*, vol. 83, no. 3, p. 319, 1974.

[24] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, “Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses,” *Image and vision computing*, vol. 32, no. 10, pp. 641–647, 2014.

[25] C. Sobin and M. Alpert, “Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy,” *Journal of psycholinguistic research*, vol. 28, no. 4, pp. 347–365, 1999.

[26] H. Stassen, S. Kuny, and D. Hell, "The speech analysis approach to determining onset of improvement under antidepressants," *European Neuropsychopharmacology*, vol. 8, no. 4, pp. 303–310, 1998.

[27] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de Heyning, "The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 3, pp. 796–809, 2000.

[28] M. S. Cannizzaro, N. Reilly, J. C. Mundt, and P. J. Snyder, "Remote capture of human voice acoustical data by telephone: A methods study," *Clinical linguistics & phonetics*, vol. 19, no. 8, pp. 649–658, 2005.

[29] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.

[30] B. Stasak, J. Epps, and R. Goecke, "Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect," in *Proc. Interspeech 2017*, pp. 834–838, 2017.

[31] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and 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.

[32] O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, and M. Cooke, "Glottal source features for automatic speech-based depression assessment," in *Proc. Interspeech 2017*, pp. 2700–2704, 2017.

[33] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision.," in *Interspeech*, pp. 2172–2176, 2013.

[34] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression.," in *Interspeech*, pp. 1238–1242, 2014.

[35] B. Vlasenko, H. Sagha, N. Cummins, and B. Schuller, "Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition," in *Proc. Interspeech 2017*, pp. 3266–3270, 2017.

[36] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Automatic modelling of depressed speech: relevant features and relevance of gender," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[37] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2011.

[38] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker, *et al.*, "From joyous to clinically depressed: Mood detection using spontaneous speech.," in *FLAIRS Conference*, 2012.

[39] A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, and A. Alwan, "Effectiveness of voice quality features in detecting depression," *Proc. Interspeech 2018*, pp. 1676–1680, 2018.

[40] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-temporal analysis of speech affected by depression and psychomotor retardation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7542–7546, IEEE, 2013.

[41] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech.," in *Interspeech*, pp. 857–861, 2013.

[42] T. F. Quatieri and N. Malyska, "Vocal-source biomarkers for depression: A link to psychomotor activity," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[43] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–7, IEEE, 2009.

[44] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," *depression*, vol. 1, no. 1, 2014.

[45] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 87–91, ACM, 2014.

[46] R. Gupta and S. S. Narayanan, "Predicting affective dimensions based on self assessed depression severity.," in *INTERSPEECH*, pp. 1427–1431, 2016.

[47] R. Gupta, S. Sahu, C. Espy-Wilson, and S. S. Narayanan, "An affect prediction approach through depression severity parameter incorporation in neural networks," in *Proc. Interspeech 2017*, pp. 3122–3126, 2017.

[48] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "Cca based feature selection with application to continuous depression recognition from acoustic speech features," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 3729–3733, IEEE, 2014.

[49] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, “Vocal and facial biomarkers of depression based on motor incoordination and timing,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 65–72, ACM, 2014.

[50] A. C. Trevino, T. F. Quatieri, and N. Malyska, “Phonologically-based biomarkers for major depressive disorder,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 42, 2011.

[51] 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 *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 41–48, ACM, 2013.

[52] Z. S. Syed, K. Sidorov, and D. Marshall, “Depression severity prediction based on biomarkers of psychomotor retardation,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 37–43, ACM, 2017.

[53] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, *et al.*, “Depression assessment by fusing high and low level features from audio, video, and text,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 27–34, ACM, 2016.

[54] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, “Multimodal and multiresolution depression detection from speech and facial landmark features,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 43–50, ACM, 2016.

[55] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, “Multimodal prediction of affective dimensions and depression in human-computer interactions,” in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 33–40, ACM, 2014.

[56] M. R. Morales and R. Levitan, “Speech vs. text: A comparative analysis of features for depression detection systems,” in *Spoken Language Technology Workshop (SLT), 2016 IEEE*, pp. 136–143, IEEE, 2016.

[57] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, and T. F. Quatieri, “Detecting depression using vocal, facial and semantic communication cues,” in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 11–18, ACM, 2016.

[58] E.-M. Rathner, J. Djamali, Y. Terhorst, B. Schuller, N. Cummins, G. Salamon, C. Hunger-Schoppe, and H. Baumeister, “How did you like 2017? detection of language markers of depression and narcissism in personal narratives,” *Future*, vol. 1, no. 2.58, p. 0, 2018.

[59] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews,"

[60] V. Mitra, E. Shriberg, D. Vergyri, B. Knoth, and R. M. Salomon, "Cross-corpus depression prediction from speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 4769–4773, IEEE, 2015.

[61] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. Cohn, "Cross-cultural depression recognition from vocal biomarkers," *Interspeech 2016*, pp. 1943–1947, 2016.

[62] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and ptsd.," in *Interspeech*, pp. 847–851, 2013.

[63] B.-Q. Li, L.-L. Hu, L. Chen, K.-Y. Feng, Y.-D. Cai, and K.-C. Chou, "Prediction of protein domain with mrmr feature selection and analysis," *PLoS One*, vol. 7, no. 6, p. e39308, 2012.

[64] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mrmr feature selection and analysis," *Amino acids*, vol. 42, no. 4, pp. 1387–1395, 2012.

[65] M. Pal and G. M. Foody, "Feature selection for classification of hyperspectral data by svm," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2297–2307, 2010.

[66] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.

[67] A. T. Beck, R. A. Steer, and G. K. Brown, "Beck depression inventory-ii," *San Antonio*, vol. 78, no. 2, pp. 490–8, 1996.

[68] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, ACM, 2014.

[69] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, *et al.*, "The distress analysis interview corpus of human and computer interviews.," in *LREC*, pp. 3123–3128, Citeseer, 2014.

[70] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The phq-8 as a measure of current depression in the general population," *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[71] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, ACM, 2010.

[72] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[73] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 3–10, ACM, 2013.

VITA

Name: Yasin Serdar Özkanca

Data of Birth: 28/06/1993

Birth Place: Coventry, UK

Languages: Turkish, English

Education

- MS: Özyeğin University Electric Electronics Engineering 2018
- BS: Özyeğin University Electric Electronics Engineering 2016

Work Experience

- Türk Telekom 2015 Summer Internship

Publications

1. Özkanca, Y., Demiroglu, C., Besirli, A., Celik, S. (2018). Multi-Lingual Depression-Level Assessment from Conversational Speech Using Acoustic and Text Features. Proc. Interspeech 2018, 3398-3402.
2. Wroge, T. J., Ozkanca, Y., Demiroglu, C., Si, D., Atkins, D. C., Ghomi, R. H. Parkinsons Disease Diagnosis Using Machine Learning and Voice.

Honors and Awards

- 100 % Scholarship for undergrad at Özyeğin University
- The top 3nd BS alumnus in the entire graduating class of 2016 at Ozyegin University with a CGPA of 3.16 out of 4.00

Test Information:

- GRE Quantitative (164/170) 01.2018

- TOEFL (105/120) 02.2018