Model Fusion for Multimodal Depression Classification and Level Detection

Mohammed Senoussaoui, Milton Sarria-Paja, João F. Santos, and Tiago H. Falk
Institut National de la Recherche Scientifique, Centre EMT
800 rue de la Gauchetière Ouest, Suite 6900, Montreal, QC, Canada H5A-1K6
{mohammed.senoussaoui, sarria, jfsantos, falk}@emt.inrs.ca

ABSTRACT

Audio-visual emotion and mood disorder cues have been recently explored to develop tools to assist psychologists and psychiatrists in evaluating a patient's level of depression. In this paper, we present a number of different multimodal depression level predictors using a model fusion approach, in the context of the AVEC14 challenge. We show that an i-vector based representation for short term audio features contains useful information for depression classification and prediction. We also employed a classification step prior to regression to allow having different regression models depending on the presence or absence of depression. Our experiments show that a combination of our audio-based model and two other models based on the LGBP- TOP video features lead to an improvement of 4% over the baseline model proposed by the challenge organizers.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Applications; J [**Computer Applications**]: Miscellaneous

Keywords

Depression, i-vectors, Support Vector Machine, Support Vector Regression, Generalized Linear Models.

1. INTRODUCTION

Natural speech contains both linguistic and emotional (non-linguistic) information. Acoustic biomarkers in speech have also been shown to be useful in detecting neurophysiologic and mental conditions, being an important complement to motor expression related features, such as facial expression and gestures/body movement [10]. Audio-visual emotion and mood disorder cues have been recently explored to develop tools to assist psychologists and psychiatrists in evaluating a patient's level of depression. In the context of the Audio-Visual Emotion recognition Challenge (AVEC 2014) [19], affect and depression are to be classified and ranked based on data from the AVEC2013 audio-visual depression corpus in two separate sub-challenges. This paper focuses on the Depression Recognition Sub-Challenge (DSC), where the objective is to predict the level of self-reported depression as indicated by the Beck Depression Index-II (BDI) for each multimedia file in the corpus [19].

In the literature, different types of acoustic features were explored for the characterization of depressed speech [2, 10, 4, 21]. However, in most of the cited works, the combination of short term Mel-Frequency Cepstral Coefficients (MFCC) features with Gaussian Mixture Models (GMM) was found to be very efficient for depression detection based on speech. With the purpose of reducing the nuisances due to undesirable data variabilities, a sophisticated GMM-based approach was recently adopted for depressed speech modelling, namely, the Joint Factor Analysis (JFA) [17]. Consequently, an improved version of the JFA model, known as the i-vector representation, was successfully applied in the depressed speech detection task [12, 3].

Our approach to the DSC sub-challenge is based on mapping short-term acoustic features to the i-vector space, which can be seen as a more compact representation of the acoustic features vector. While this approach is common in other tasks such as speaker identification/recognition, it has only recently been applied in depression classification [12][3]. Here, we present two different approaches. In the first approach, we apply the i-vector representation of depressed speech in a system comprised of two stages: a depression classifier and a depression level regression model. In the second approach, we combine a single regression model based on the acoustic i-vectors to two other regression models using the baseline video features (based on the dynamic appearance descriptor LGBP-TOP [19]) to compute a final depression level score.

The remainder of this paper is organized as follows. Section 2 describes the audio-visual features used in our models. Section 3 describes the depression level prediction model and the general regression model architectures employed in this work, while Section 4 describes the specific classification and regression methods used as building blocks in these architectures. Section 5 presents the experimental setup. In Section 6 we present and discuss the results obtained for all the different architectures. In Section 7, we conclude the paper with some final considerations.

AVEC'14, November 7, 2014, Orlando, FL, USA.
Copyright 2014 ACM 978-1-4503-3119-7/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661806.2661819.

2. SIGNAL REPRESENTATIONS

2.1 Short term acoustic features

Prior to i-vector feature extraction, each speech recording was downsampled to 16 kHz and normalized to -26 dBov (dB overload). After silence removing, the speech recordings are represented by a 60-dimensional feature vector which contains 20 static Mel Frequency Cepstral Coefficients (MFCC) and 40 dynamic MFCC coefficients to convey temporal dynamics information (i.e. Δ and $\Delta\Delta$ derivatives of MFCC). The MFCC were computed on a per-window basis including the 0–th order cepstral coefficient (log-energy), using a 25 ms window with 40% overlap and 20 triangular bandpass filters. The Δ and $\Delta\Delta$ coefficients were computed by means of an anti-symmetric Finite Impulse Response (FIR) filter of length 5 to avoid phase distortion of the temporal sequence.

2.2 I-vector space

In the last five years, representation of the complex speech signal by means of a simple vector [5] of moderate dimensions (typically in the range of hundreds) has become commonly adopted in many speech-based technologies [6, 16, 15]. Within the Speaker Recognition community where this representation was first proposed, it is well known as the *i-vector* space representation. The most important characteristics of this representation are its moderate size (one vector) and its richness in terms of modeled information. These characteristics made the i-vector paradigm suitable for many fields. An i-vector can be defined simply as the mapping, using the Factor Analysis or the Probabilistic Principal Component Analysis, of a high dimensional supervector to a low-dimensional space called the *total variability* space (here the word *total* is used to refer to both speaker and channel variabilities). A supervector is a high dimensional vector, usually obtained by the concatenation of the component mean vectors of a Gaussian Mixture Model (GMM) of the short-term acoustic features.

The i-vector paradigm is a sophisticated version of the well-known GMM-UBM model (where the acronym UBM refers to the Universal Background Model) [13]. The idea is to train the UBM, which is a GMM model, with a large number of speech recordings coming from different speakers and in the presence of a relatively small amount of a given speaker-dependent speech, the corresponding GMM model is obtained by the Maximum a Posteriori (MAP) adaptation of the parameters of the UBM model. A more detailed mathematical description is provided next.

2.2.1 Mathematical formulation

The hidden variable based generative model of an i-vector extractor can be mathematically expressed as follow:

$$\mathbf{X} = \mathbf{M} + \mathbf{T}\mathbf{x} \tag{1}$$

where $\mathbf{X}$ is a speaker- and channel-dependent supervector of dimension $(1 \times NF)$, $\mathbf{M}$ is the $(1 \times NF)$ speaker- and channel-independent supervector obtained by the concatenation of the mean vectors of the N components of the Universal Background Model (UBM), $\mathbf{T}$ is a $(NF \times D)$ rectangular matrix which the columns span the Total Variability space and $\mathbf{x}$ is a $(D \times 1)$ hidden vector having a standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. A point estimate of the hidden vector $\mathbf{x}$ is what is called the i-vector. The variables F and D represent, respectively, the dimension of the short-term acoustic vector space and the dimension of the total variability space (i.e. the dimension of the i-vector space).

The training of the generative model (i.e. the i-vector extractor) given by (1) consists of estimating the total variability matrix $\mathbf{T}$. In practice, an EM-like algorithm is usually used in order to estimate the $\mathbf{T}$ matrix [11]. In this work, the i-vector representation was used to represent individual speech signals. Optimal configurations of the i-vector extractor (i.e. the number of the UBM components and the dimension of the i-vectors) were explored and chosen empirically using development data, as described in Section 5.2.

2.3 Baseline video features

Video features were provided by the challenge organizers and were based on the local dynamic appearance descriptor LGBP-TOP [19]. Video frames were preprocessed using face localization and segmentation prior to the LGBP- TOP feature extraction. Since features were extracted for frame blocks and we needed to predict a single label per multimedia file, we adopted the same approach as in the baseline system and used a single mean video feature vector taken across all feature vectors in each file. For more details on the LGBP- TOP features, the reader is invited to consult the official baseline paper [19].

3. DEPRESSION LEVEL PREDICTION

In this work, all our experiments are performed in the framework of the AVEC 2014 challenge[1]. We specifically focus on the Depression recognition Sub-Challenge (DSC). Thus, our main task is to use an audiovisual recording of a given speaker in order to predict a number representing his/her depression level according to the Beck Depression Inventory- II (BDI-II) [1, 19]. The BDI-II scores range from 0 to 63 and are grouped into 4 depression classes as follows:

- From 0 to 13: *no* or *minimal* depression.
- From 14 to 19: *mild* depression
- From 20 to 28: *moderate* depression.
- From 29 to 63: *severe* depression.

In addition to our purpose of predicting the depression level, which is a regression problem, we are also interested in the depression classification problem as a tool to improve our regression models, as detailed in the following section.

3.1 Depression level classification

The simplest way of designing a classification problem in the BDI-II context is by considering the originally proposed class splitting of the scale into 4 classes, namely, *minimal*, *mild*, *moderate* and *severe* (see above). However, since we have a small amount of training data (50 samples spread over the four classes) we reduced the number of classes from 4 to only 2 classes, i.e., *absence/presence* of depression in a given individual recording. In the work presented in [12] and reported using AVEC13 training and development datasets, *absence* of depression is considered when the minimal or mild symptoms are detected for a given person, otherwise (i.e. in the presence of *moderate* or *severe* symptoms) the

[1] http://sspnet.eu/avec2014/

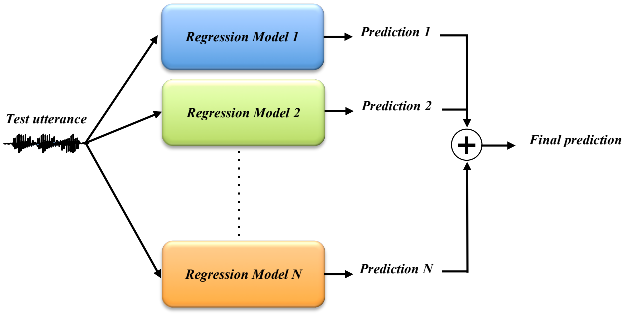


Figure 1: Multiple regression models based architecture. The final score is obtained by simply averaging the elementary predictions.

system considers *presence* of depression for the given recording. Since misdetecting the presence of depression symptoms could have serious consequences in one's diagnostic, we decided to apply a different split to the *absence*/*presence* of depression classifier, as follows: the *absence* is signaled only in case of the *no* or *minimal* depression symptoms, otherwise (i.e. in the presence of *mild*, *moderate* or *severe* symptoms), the *presence* of depression is signaled by the classifier.

3.2 Depression level regression

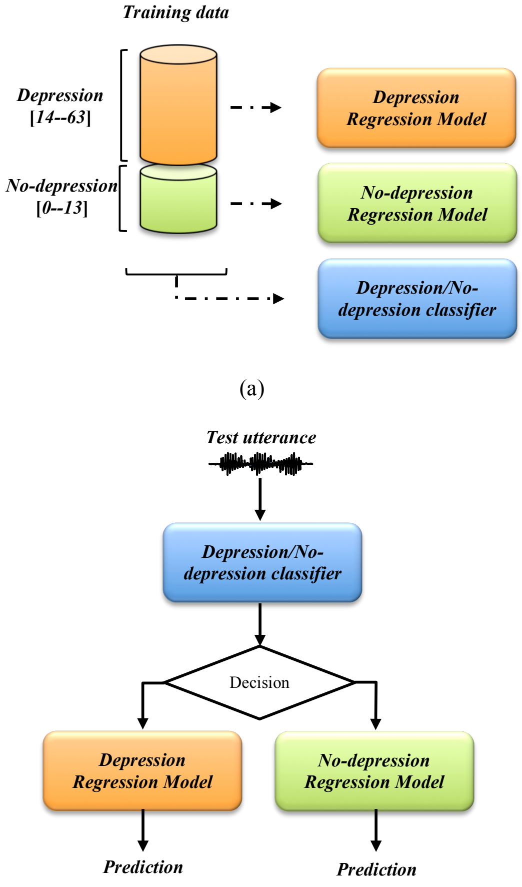
As stated previously, depression classification is not a goal in itself in the work presented in this paper. However, our aim is to develop an efficient classification system that could contribute to the improvement of the regression models for depression level prediction.

3.2.1 *Single-model regression problem*

In the most straightforward approach to a regression problem, one can simply train one regressor using the training data corresponding to the set of all depression levels (i.e. from 0 to 63 levels in case of the BDI-II inventory) and use this model to predict all unseen data during test stage, i.e., a *single-model* is used to predict all possible depression levels. More generally, in this stage several regression models with different approaches can be used for regression, then the final decision can be obtained by weighting each output and averaging the predictions. Since none of these models depend on subclasses, in this case a classifier is obviously not needed. A diagram depicting the general architecture for solving this kind of problem is depicted in Figure 1.

3.2.2 *Two-model regression problem*

The idea consists of considering the splitting of the BDI-II into two classes as discussed in the section 3.1 and build a binary classifier able to distinguish between these two classes. Then, for each class we train a regressor able to predict the depression levels within its specific class (see Figure 2 (a)). During the test stage, a given test utterance should be assigned (based on a soft or a hard decision) by the classifier to one of the two classes and, according to this decision, the corresponding regressor is used to predict the depression level (see Figure 2 (b)). Note that in the case of a soft decision based classifier, we can combine, based on a weighted average, the predictions (for a given test utterance) of both regressors using the output classifier probabilities as weights.



(a)

(b)

Figure 2: The combination of 2-classes classifier with two class-dependent regression models. the training stage is illustrated in (a) and the test stage in (b).

Unlike the single model regression problem, the one with two models has the advantage of making predictions over smaller and more homogeneous scales. However, this strategy has a drawback that the performance of the whole system is dependent on the performance of the classifier and the performance of the two regressors, especially if the parameters of the two regressors are completely different.

4. STATISTICAL MODELS

4.1 Support Vector Machines (SVM)

The Support vector machine (SVM) is a binary classifier which models the decision boundary between two classes as a separating hyperplane. In our experiments, two classes are involved: a "positive class", i.e., individuals with certain depression level and a "negative class", i.e., individuals with no depression. By using labelled training vectors, the SVM optimizer finds a separating hyperplane that maximizes the separation between these two classes. The discriminant function is given by:

$$f(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j c_j K(\mathbf{x}, \mathbf{x}_j) + b \quad (2)$$

where $c_j \in \{+1, -1\}$, are the labels for the training vectors. The kernel function $K(\cdot, \cdot)$ is constrained to have certain properties (the Mercer condition). The support vectors x_j, their weights α_j and the bias term b, are determined during training [20]. The decision score is calculated by comparing the test vector with the SVM discriminant function established in the SVM training step, and a decision is made based on thresholding ("hard classification") [20]. In [8], the authors present an extension of the method to the regression problem, called Support Vector Regression (SVR).

4.2 Relevance Vector Machines (RVM)

The Relevance Vector Machines are a Bayesian treatment of the SVM which does not suffer from the common limitations of the standard SVM approach, i.e., non-probabilistic predictions, kernel functions having to satisfy Mercer's condition, and need of a cross-validation procedure to set free parameters [18]. As in SVM, consider a set of M examples of input vectors $\{\mathbf{x}_n\}_{n=1}^{M}$ along with the corresponding targets $\{t_n\}_{n=1}^{M}$, which may be real values for a regression task or class labels for a classification task. The predictions are based upon some function $y(\mathbf{x})$ defined over the input space; a flexible and popular choice for this function is one of the form presented in 2. Relevance vector machines make use of Bayesian estimation theory for learning the model parameters and making predictions, which allows them to have output predictions that are probabilistic [18].

4.3 Generalized Linear Models (GLM)

Generalized linear models (GLMs) are a large class of statistical models for relating targets or labels to linear combinations of predictor variables, including many commonly encountered types of dependent variables and error structures as special cases. The GLM approach is attractive mostly because it provides a general theoretical framework for many common statistical models. A generalized linear model consists of three components: (*i*) a random component, specifying the conditional distribution of the target variable which is a member of an exponential family, such as the Gaussian (normal), binomial, Poisson, gamma, or inverse-Gaussian families of distributions, (*ii*) a linear predictor - that is a linear function of regressors and (*iii*) a smooth and invertible linearizing link function $g(\cdot)$, which transforms the expectation of the response variable, to the linear predictor, this is a link between the random and the linear components [7]. For our experiments the normal distribution was chosen as the random component and logit as link function.

5. EXPERIMENTAL SETUP & DATASETS

5.1 AVEC 2014 challenge

Two sub-challenges were proposed for the AVEC14' participants, namely, the recognition of 3 continuous affect dimensions (i.e. *affective*, *valence* and *arousal*) in the Affect recognition Sub-Challenge (ASC) and the prediction of the self-reported depression indicator BDI-II in the Depression recognition Sub-Challenge (DSC) [19]. As opposed to AVEC13, only two tasks were selected for the 2014 challenge:

- Northwind: participants were asked to read aloud an excerpt.
- Freeform: participants were asked to answer to one of a number of specific questions.

The participants were native German speakers and both tasks were performed in German. The challenge organizers provided 150 Northwind-Freeform pairs of recordings separated in three subsets (training, development, and testing) with Northwind and Freeform files for 50 different participants each. Labels for the training and development sets were provided. Note that the provided data is audiovisual and one could use audio signals only, video signals only, or audio-visual data for the sub-challenges.

5.2 I-vector configuration

5.2.1 UBM

In order to find the best configurations of the i-vector extractor we built six GMM gender-independent UBMs containing respectively: {16, 32, 64, 128, 256, 512} Gaussians. In the speaker verification field, it is widely known that using a large amount of data for training the UBMs is key component for getting a useful i-vector representation. Therefore, we trained our UBMs using the TIMIT database, which is comprised of 6300 recordings from 630 different speakers. Each utterance is around 3 seconds duration, recorded using 16 bits precision and a sampling rate of 16 kHz [9]. The training set of the AVEC14 data is also used in addition to TIMIT data to train these UBMs. For consistency with the TIMIT sampling rate, all speech signals in the AVEC14 datasets were down-sampled to 16 kHz before using them for training the UBMs and extracting i-vectors.

5.2.2 Total variability matrix

In addition to the different UBM configurations, we have also trained six total variability matrices (i.e. the **T** matrices) of dimensions: {40, 60, 80, 100, 150, 200}, using each of the UBM configurations. Thus, we have a total of 36 i-vector extractor configurations, obtained by combining the 6 UBMs with the 6 **T** matrices. The best configurations for classification and for regression were empirically selected using the AVEC14 development data. The same datasets used to train the UBMs were used to train the variability matrices (i.e. the whole TIMIT dataset and the AVEC14 training dataset). The MATLAB implementation of the i-vector extractor described in [14] was used in this paper.

5.2.3 I-vector extraction strategies

We adopted two different strategies for extracting the i-vectors from a given audio recording, namely, a *per-file* and a *per-frame* strategy. In the *per-file* case, the idea is to extract a single i-vector for the entire audio recording. On the other hand, in the *per- frame* strategy, we adopt a dynamic scheme based on a sliding window (with a length of 5 s and 50% overlap), and extract an i-vector from each frame, resulting in a sequence of i-vectors for each audio recording.

5.3 Proposed classification models

As mentioned previously, a binary depression classifier was used in this work (see section 3.2.2). The first class represents the *absence* of depression (and it is expressed by the *minimal* depression levels on the BDI-II scale), while the

Table 1: Configuration of the proposed depression level regression models. In the table, *Mix* refers to the number of Gaussians in the GMM model, *Dim* is the dimension of the total variability matrix, and *Cum. var* is the cumulative percent of variance for the PCA projection. The systems are named according to the type of input: AO refers to Audio Only, AV to Audio and Video and VO refers to Video Only.

Model	Raw features	PCA	Classification		Regression	
			Method	Hyper param.	Method	Hyper param.
AO-S1	per-file i-vectors (Mix = 64 Dim = 80)	From 80 to 10 (Cum. Var = 45.11%)	—	—	GLM	—
AO-S2	per-frame i-vectors (Mix = 16 Dim = 80)	From 80 to 57 (Cum. Var = 98.82%)	—	—	GLM	—
AO-S3	per-file i-vectors (Mix = 64 Dim = 80)	From 80 to 1 (Cum. Var = 7.03%)	SVM on audio i-vectors	$\sigma = 15/C = 1$	GLM	—
AV-S4	per-file i-vectors (Mix = 64 Dim = 80)	From 80 to 3 (Cum. Var = 18.41%)	SVM on LGBP-TOP features	$\sigma = 90/C = 340$	GLM	—
VO-S5	LGBP-TOP features	From 16992 to 62 (Cum. Var = 99.67%)	—	—	RVM	—
VO-S6	LGBP-TOP features	—	—	—	SVR	$\epsilon = 0.001/C = 1$

second one represents the *presence* of depression (and it is expressed by the grouping of the three other classes of BDI-II scale, namely, *mild*, *moderate* and *severe* depression classes). Video and audio data were used separately in the depression classification task. Each audio recording is represented by a *per-file* i-vector of dimension 150 (obtained from an UBM with 16 components) and for all our experiments (classification / regression) we have used the baseline features for video data. For both types of data (i.e. audio and video), SVMs were used as the classifier. The SVM implementation from the MATLAB Statistics Toolbox was used for training the classifier and predicting the output class labels for each multimedia file.

Table 2: Video and audio 2-classes classification results on the development set of AVEC14 data. The σ parameter of the Gaussian Radial Basis (RBF) and the C parameter for the soft margin are given.

	Parameters (σ/C)	Accuracy
Audio	15 / 1	82%
Video	90 / 340	82%

5.4 Proposed regression models

Table 1 summarizes the different models used for depression level regression in this paper. For our five submissions we use the systems described in Table 1. More specifically, for *submission 1* predictions of test set were computed by using the system AO-S1 which is an i-vector based system using GLM, where after i-vector extraction the feature vectors are projected to a 10-dimensional space using PCA. For submission 2, the system AO-S3 was used; in this case, we included a classification stage which allows to use two GLM regressors in the prediction stage. One predictor is used in case the recording is classified as “non-depressed” and a second predictor if the recording is classified as “depressed”. For submission 3, we used a combination of two systems, i.e., systems AO-S1 and VO-S5, where we average the predictions of each system to get a final prediction. It is important to note that in this case for the system VO-S5 only Freeform was used for training and testing and the LGBP-TOP features were projected to a 38-dimensional space using PCA. For submission 4, a combination of 3 systems was used in a similar way as in submission 3: the predictions from system AO-S2, VO-S5 and VO-S6 were averaged to get the final prediction. For submission 5, a combination of 3 different systems was used, i.e., AV-S4, VO-S5 and VO-S6.

For our RVM depression regression model (VO-S5), we used the implementation in the open-source MATLAB Pattern Recognition Toolbox (http://newfolder.github.io/). As discussed in section 4.2, there are no hyperparameters to choose or optimize in this algorithm. A PCA was used to reduce the dimensionality from 16992 to 62, accounting for more than 99% of the total variance.

Finally, the system VO-S6 is similar to the baseline model described in [19]. However, here we trained the model on the Freeform and Northwind samples separately. For predicting the depression level for a given subject, we use the model to find a prediction for each of the two multimedia files and then average the results to yield a final prediction. We experimented using a PCA to improve the system performance but, unlike our finding for the RVM-based model, in this case reducing the dimensionality was not helpful. A grid search over the hyperparameters C (tested values: [0.01, 0.05, 0.1, 0.5, 1, 2]) and ϵ ([0.0001, 0.001, 0.05, 0.01, 0.1, 0.5]) was also done, which has shown the parameters used in the baseline paper were optimal [19].

6. RESULTS AND DISCUSSION

6.1 Classification results

The obtained results, presented in Table 2, show that the accuracy of the audio and the video classifiers are equivalent. The work presented in [12] presents a similar depression classification approach using i-vectors; however, their class definition is different of the one presented here, as discussed in Section 3.1. In the referred work, the training and the development set of the AVEC13 challenge were used, which is more than 26 hours of audio data. The best accu-

Table 3: Performance comparison of proposed regression models on the development set.

Model	Freeform		Northwind		Both	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
AO-S1	10.50	8.91	11.79	9.67	11.08	9.28
AO-S2	9.37	7.63	11.44	9.43	10.04	8.44
AO-S3	9.93	7.34	10.29	7.59	10.09	7.41
AV-S4	9.17	6.77	9.17	6.76	9.14	6.74
VO-S5	8.12	6.37	10.25	8.46	8.52	6.95
VO-S6	9.05	7.32	9.28	7.66	8.88	7.24
Fusion: AO-S1 and VO-S5	8.47	7.22	–	–	–	–
Fusion: AO-S2, VO-S5, and VO-S6	–	–	–	–	7.91	6.57
Fusion: AV-S4, VO-S5, and VO-S6	–	–	–	–	7.90	5.92

racy reported in [12] was 70%, which is significantly below our scores (82%).

6.2 Regression results

Predictions for the audio-based models were performed only for Freeform files in the test set, as they were shown to perform better in the development set (see Table 3). Performance for AO-S1 alone was found to be slightly higher than that of the baseline for the development set but lower for the test set, which may indicate overfitting. However, results for the test set were in line with those of the baseline video model, even though the dimensionality of the features was significantly smaller (10 variables against 16992).

The nature of the task performed by the subject had no significant difference for video-based models as it had for audio-based models. A performance difference was observed in VO-S5, where RMSE values were almost 2 points higher for Northwind. This improvement led us to test a combination of AO-S1 and VO-S5. In this case, we adjusted the configuration of VO-S5 to use only the first 38 PCA coefficients instead of 62, as it was empirically found to perform better on the development set using this configuration. However, the results were lower than for AO-S1 alone in the test set.

Employing the binary classifier on audio features to build a different regression model for each class (AO-S3) was shown to be useful in the development set but not in the test set. Even though both classifiers have similar performance, the one based on the video features (AV-S4) led to better performance than the one based on audio features when tested on the development set. AO-S3 had the lowest performance of all proposed models on the test set. Due to the limited amount of trials in the test set, we did not test the system with video-based classification so we are not able to compare the results. AV-S4, which uses video-based classification, was fused with VO-S5 and VO-S6, which resulted in the best performance on the development set, but similar to the baseline on the test set.

Finally, results show that combining the baseline model (VO-S6) with some of our proposed models leads to improved performance over the baseline. Our fusion of AO-S2, VO-S5, and VO-S6 yielded the best results in the test set. The three different models seem to extract complementary information from the available data, even though two of the models use the same feature set (LGBP- TOP).

Table 4: Performance comparison of submitted regression models on the test set.

Model	RMSE	MAE
Baseline	10.86	8.86
AO-S1	11.03	8.89
AO-S3	12.71	9.82
Fusion: AO-S1 and VO-S5	12.02	9.53
Fusion: AO-S2, VO-S5, and VO-S6	10.43	8.33
Fusion: AV-S4, VO-S5, and VO-S6	10.83	8.69

7. CONCLUSION

In this paper, we presented a number of different multimodal depression level predictors using a model fusion approach for the AVEC14 challenge. First, we proposed a model based on a total variability representation of short-term audio features, which was shown to perform in line with the baseline model but with a significantly smaller number of independent variables. The fusion of this model with two other models based on the LGBP- TOP features yielded an improvement of 4% in the RMSE compared to the baseline model on the test set. We should also highlight that all proposed models have higher performance than the baseline model in the development set, but some models were not able to generalize the prediction on the test set. We have also presented two different systems for depression classification, one based on the i-vector representation and another on the LGBP-top features. Both systems had a similar accuracy of 82%.

8. ACKNOWLEDGMENTS

MS and THF acknowledge funding from NSERC Canada. MSP and JFS are funded by PhD scholarships from Colciencias and FQRNT, respectively.

9. REFERENCES

[1] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri. Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of personality assessment*, 67(3):588–597, 1996.

[2] N. Cummins, J. Epps, M. Breakspear, and R. Goecke. An investigation of depressed speech detection: Features and normalization. In *INTERSPEECH*, pages 2997–3000. ISCA, 2011.

[3] N. Cummins, J. Epps, V. Sethu, and J. Krajewski. Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 970–974, May 2014.

[4] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps. Diagnosis of depression by behavioural signals: A multimodal approach. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, pages 11–20, New York, NY, USA, 2013. ACM.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, 2011.

[6] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak. Language recognition via i-vectors and dimensionality reduction. In *INTERSPEECH*, pages 857–860, 2011.

[7] A. Dobson. *An Introduction to Genelarized Linear Models*. Chapman & Hall/CRC; 3 edition, 2008.

[8] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997.

[9] J. S. Garofolo, L. D. Consortium, et al. TIMIT: acoustic-phonetic continuous speech corpus, 1993.

[10] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt. On the relative importance of vocal source, system, and prosody in human depression. In *Body Sensor Networks (BSN), 2013 IEEE International Conference on*, pages 1–6, May 2013.

[11] P. Kenny. A small foot-print i-vector extractor. In *Proc. Odyssey*, 2012.

[12] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo. A study of acoustic features for the classification of depressed speech. In *Proceedings of the International Convention Mipro Conference On Intelligent Systems (CIS), Special Session on Biometrics & Forensics & De-identification and Privacy Protection (BiForD)*. MIPRO, May 2014.

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

[14] S. O. Sadjadi, M. Slaney, and L. Heck. Msr identity toolbox v1.0: A matlab toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*, November 2013.

[15] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel. A study of the cosine distance-based mean shift for telephone speech diarization. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(1):217–227, 2014.

[16] S. Shum, N. Dehak, E. Chuangsuwanich, D. A. Reynolds, and J. R. Glass. Exploiting intra-conversation variability for speaker diarization. In *INTERSPEECH*, pages 945–948, 2011.

[17] D. Sturim, P. Torres-carrasquillo, T. F. Quatieri, N. Malyska, and A. Mccree. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Proceedings of Interspeech*, 2011.

[18] E. M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal Machine Learning Research*, 1:211–244, 2001.

[19] 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 Audio/Visual Emotion Challenge and Workshop (to appear)*. SSPNET, November 2014.

[20] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998.

[21] 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*, AVEC '13, pages