Co-training based algorithm for gender detection from emotional speech

Jelena Slivka*, Aleksandar Kovačević*

* University of Novi Sad/Faculty of Technical Sciences/Computing and Control Department, Novi Sad, Serbia

slivkaje@uns.ac.rs, kocha78@uns.ac.rs

Abstract—Automatic recognition of gender from spoken data is an important prerequisite for many practical applications. In this paper we consider the problem of automatic gender recognition from emotional speech. Analyzing emotional speech adds to the complexity of the problem due to the variability of the speech signal which is driven by the speakers’ emotional state. The ever-present and dominating bottleneck in automatic analysis of spoken language is the scarcity of labeled data. Here we investigate the possibility of applying a co-training style algorithm in order to reduce the human effort needed for data labeling. We apply several co-training settings: random split of features, natural split of features, majority vote of several co-training classifiers and Random Split Statistic Algorithm (RSSalg) we have developed earlier in order to boost co-training classifiers and Random Split Statistic Algorithm features, natural split of features, majority vote of several co-training classifiers and RSSalg. These settings also proved to be the most robust ones considering the amount of the available labeled data.

Keywords: semi-supervised learning; co-training; computational paralinguistics; gender recognition; emotional speech; SVM;

I. INTRODUCTION

Automatic recognition of gender from speech plays an important role in a broad range of practical applications. Gender-dependent models achieve better performance than gender independent models in various Automatic Speech Recognition tasks as they help reduce the inter-speaker variability [1]. Gender information has been shown to aid the detection of the speakers’ emotional state [2]. It was used to leverage recommender systems [3]. Studies suggest that the acoustic analysis of pathological voices should be conveyed using gender-dependent feature subsets [4]. Gender detection module is one of the main parts of audio segmentation system presented in [5].

In [6] it was stated that one of the ever-present and dominating bottlenecks in automatic analysis of spoken language is data scarcity. Manual annotation of data instances by human experts is very tedious, time-consuming and expensive. This problem can be alleviated by application of semi-supervised learning (SSL). SSL techniques incorporate both labeled and unlabeled data in the training process, resulting with the models that have comparable or even better performance than traditional supervised models trained on much larger portions of labeled data. Successful application of SSL techniques highly reduces the human effort needed for annotation of the appropriate training set.

One of the major semi-supervised techniques is co-training [7]. Co-training assumes that the features of the dataset can be naturally separated in two disjunctive subsets called views. The two views should be based on two independent sources describing the same data, a trait not so commonly found in real-world problems.

In this paper, we investigate the possibility of applying a co-training style algorithm to the problem of automatic gender recognition from emotional speech in order to reduce the human effort needed for data labeling. The inherited difficulties of analyzing emotional speech add to the complexity of the problem [8]. For example, pitch information (commonly used for gender recognition) varies over time, driven by the speakers’ emotional state [9]. It is proven that using pitch as criterion for gender classification of emotional speech deteriorates the accuracy by 21% compared to the processing of spontaneous speech [10].

Previous work has shown that co-training can be successfully applied to the task of gender recognition from spoken data [11]. In order to apply co-training authors have identified a reasonable feature split, motivated partly by the independence of the features and partly by the size of the resulting views in order to ensure the sufficiency condition. Although these views are not completely conditionally independent [12], the experimental results demonstrate the validity and effectivity of such feature split for co-training [11][12]. In this paper we test several co-training settings: co-training run using a “natural” feature split (defined in [11] and [12]), co-training run with a random feature split, majority vote of the ensemble consisting of multiple co-training classifiers run with different random splits (MV), and Random Split Statistic algorithm (RSSalg) we have developed earlier in [13] with the goal of eliminating the need for the “natural” feature split in co-training, as well as boosting the performance of co-training.

We have tested the algorithms on interactive emotional dyadic motion capture (IEMOCAP) database [14]. We have decided to use IEMOCAP database as it is relatively large (approximately 12 hours of audiovisual data), contains emotional speech and is publicly available and free of charge.

The IEMOCAP database consists of 5 sessions, each recorded by two mixed gender actors. Sessions were
manually segmented at the dialog turn level (speaker turn), defined as the continuous segments in which one of the actors was actively speaking. Here we consider the problem of segment-level gender annotation and consider each segment independently. We evaluate our solution using a five-fold-cross evaluation scheme where in each turn one session is used as test data (1 male and 1 female actor) while the remaining 4 sessions are used as training data. In order to test the robustness of compared methods we vary the available amount of initially labeled data in our experiments.

Our results confirm that the feature split used in [11][12] is effective for co-training, as this setting has improved the initial classifier. However, in our experiments, co-training applied with random feature split has outperformed the performance of co-training applied with the defined natural split for all sizes of initial labeled set L except for the largest one. The best performing settings in our experiments were RSSalg and MV which achieved better improvement in accuracy than the competing solutions and displayed more robustness to the size of the initial training set L than the competing settings.

The rest of the paper is organized as follows. Work related to this paper is presented in Section 2. Section 3 describes our methodology. Section 4 presents the experiments conducted in this paper and discusses the achieved results. Finally, section 5 concludes the paper and gives directions for future work.

II. RELATED WORK

In this section, we present the previous effort of using semi-supervised techniques in order to leverage the problem of labeled data scarcity in gender detection from audio.

In [15] authors present an Expectation-maximization (EM) based algorithm for semi supervised learning in audio classification. In their setting, each audio class is modeled with a Gaussian mixture model, the parameters of which are estimated using EM. They apply their setting to gender and speaker identification tasks and show that adding unlabeled data in such a fashion may reduce the classification error rate by more than half.

Co-training was shown to be beneficial in paralinguistic tasks [11-12][16-18]. However, with the exception of [11], these studies focus on emotion recognition.

In [11] authors apply co-training on several representative tasks of paralinguistic phenomena. They chose to investigate three tasks officially studied in INTERSPEECH challenges from 2009-2011: recognizing emotion, sleepiness, and gender of speakers. Authors have used the standard set of acoustic features from INTERSPEECH 2010 (IS10) challenge [19] which they separated in three partitions suitable for the assumption of independence. The three partitions were rearranged into two views by agglomerating the two smaller groups of features in one view. A similar feature split was used in other co-training studies applied on computational paralinguistics task [12][16].

Authors in [11] have evaluated the task of gender recognition on Agender database, an official corpus of INTERSPEECH 2010 Paralinguistic Challenge (PC) Gender Sub-Challenge [19]. They report 2.1% improvement of unweighted average recall (UAR) when using co-training compared to the performance of the supervised classifier. However, for the gender detection task, co-training was not significantly better than self-training (improvement of 0.1% of UAR). In this paper, we follow the work presented in [11] and further evaluate the applicability of co-training style algorithms on the gender detection task. We note that we focus on the task of identifying gender form emotional speech as opposed to spontaneous speech used in [11].

III. METHODOLOGY

A. Acoustic features and the feature split for co-training

We have selected the standard set of acoustic features used in the INTERSPEECH 2010 (IS10) challenge [19]. The feature set consists of 1582 acoustic features which result from a base of 34 low-level descriptors (LLD) with 34 corresponding delta coefficients appended, and 21 functionals applied to each of these 68 LLD contours (1428 features). In addition, 19 functionals are applied to the 4 pitch-based LLD and their four delta coefficient contours (152 features). Finally, the number of pitch onsets (pseudo syllables) and the total duration of the input are appended (2 features). Features were extracted using the openSMILE framework [20].

Authors in [11] and [12] suggest the feature split for co-training which should be suitable for both sufficiency and independence assumption of co-training: the first view comprises of MFCCs (Mel-Frequency Cepstral Coefficients) while the second view comprises of LLDs (Low-Level Descriptors). In this paper, we adopt this feature split and refer to it as „natural” feature split. The numbers of features in the resulting views are 630 and 952 for the first view and the second view, respectively.

B. Base learner

As base learners in co-training, we employ Support Vector Machines (SVMs) trained with the Sequential Minimal Optimization (SMO) algorithm available in the WEKA toolkit [21]. The classifier setup was: SVMs with linear kernel and a complexity constant of 0.2. This is the same base learner that was employed in [11] and used as a set-up for INTERSPEECH 2010 (IS10) challenge [19] with the exception of the complexity constant which we empirically determined to be 0.2 as it yielded better results for IEMOCAP database. As in [12] we employed a parametric method of logistic regression in order to transform the output distances of SVM into (pseudo) probabilistic values [22].

C. The applied co-training settings

Originally, co-training was designed for the datasets that have the natural separation of the features in two feature subsets, called views. In order to guarantee the successful application of co-training, each view should be sufficient for learning (i.e. given enough labeled data, the features from each view separately should be sufficient to train an accurate model) and conditionally independent of the other view given the class label [7]. Co-training uses the feature split in order to train two different classifiers – both classifiers are trained using the same set of labeled data, but using different views of the data. Classifiers are then applied on unlabeled data and each is allowed to select a certain amount of most confidently labeled
instances to label and add to initial training set. Both classifiers are then retrained on the enlarged training set and the process iteratively continues for the predefined number of iterations.

In this paper we experiment with several different co-training settings:

- **Natural**: co-training algorithm [7] applied with “natural” feature split,
- **Random**: co-training applied with random feature split (obtained by randomly splitting features in two equal-sized feature sets)
- **Majority Vote (MV)**: an ensemble of diverse co-training classifiers is obtained by generating a number of different random feature splits. By using different feature splits, we are able to train different co-training classifiers as co-training is sensitive to the used feature split [23]. MV algorithm combines the predictions of the obtained ensemble in a simple majority vote fashion [13].
- **Random Split Statistic Algorithm (RSSalg)**: In [13] we have developed RSSalg in order to boost the performance of co-training and enable its application to single-view datasets. As in MV, in RSSalg, an ensemble of diverse co-training classifiers is obtained by training multiple co-training classifiers using different random splits of features. The enlarged training set resulting from each co-training process consists of initially labeled examples and examples labeled during co-training. Each enlarged training set is different as different views will cause the resulting classifiers to select different instances as most confident and added instances may be assigned different labels depending on the accuracy of the classifier that classified them. Enlarged training sets obtained by co-training are processed by selection of most reliably labeled instances. An instance is considered reliable if it appears in the majority of resulting training sets and if most of the resulting co-training classifiers agree on its label. The selected reliable instances form the final training set used for learning a final RSSalg model.

**RSSalgbest** will denote RSSalg optimized on the test data [13]. This is only considered as the upper bound of RSSalg, i.e. the performance it could achieve if we had additional labeled data for optimization. It is not considered a competing co-training setting but as a test of performance of the threshold optimizing technique used in RSSalg.

### IV. EXPERIMENTAL RESULTS

#### A. Experimental setup

In order to evaluate our solution we use a five-fold-cross evaluation scheme: in each iteration, a different session is used as test data, while the 4 remaining sessions are used as training data. Each session in IEMOCAP database is recorded by a different pair of mixed-gender actors. This way we ensure that the test data contains segments spoken by both genders and avoid the situation where the model would be trained and tested on segments produced by the same speaker. In this way, in each round of five-fold-cross validation we use in average 3827 instances for training and 957 instances for testing.

In order to test the robustness of the compared methods we vary the size of the initial labeled set $L$. We test the situations where we have 20, 50, 100, 200, 400 and 800 labeled instances available which make around 0.5%, 1.3%, 2.6%, 5.2% and 20.9% of the training set, respectively. All instances that don’t belong to the initial labeled set $L$ are used as unlabeled data (i.e. the information about the class label is discarded for those instances).

Initial labeled set $L$ is randomly chosen from the training data in the following way: in each round of five-fold-cross validation a different session (out of 4 used for training) is chosen to supply the labeled data; an equal number of randomly chosen ‘male’ and ‘female’ instances from that session are used as the initial labeled set $L$. Thus, the initial set $L$ contains segments recorded by one male and one female actor. This setting makes the recording and labeling of the segments much easier as we employ just one speaker in order to record several segments of speech. All of these segments are automatically labeled with the gender of that speaker, without the need for latter expert annotation. Also, employing only two professional actors is much cheaper than obtaining data from a group of different actors.

As the measure of performance, we use the accuracy measure as IEMOCAP database is relatively balanced according to the gender label (2362 and 2422 segments are spoken by female and male actors, respectively).

The rest of the co-training parameters are fixed to the following values; the number of co-training iterations is 30; the size of the unlabeled pool $U$ is 100; in each iteration of co-training 20 instances most confidently labeled as ‘male’ and 20 instances most confidently labeled as ‘female’ are added to the initial training set; the number of different random splits we use in Random, MV, and RSSalg setting is 50. All of these parameters were empirically chosen.

The performance of the supervised classifier obtained using the whole training set (labeled and unlabeled data with correct labels assigned) was evaluated to be 94.12%. This accuracy will be referred to as $All_{acc}$ in the remainder of this paper. The obtained accuracy justifies the features used for the task – it is reported that most gender classifications tested for clean speech demonstrate the accuracy below 95% [24]. We consider that IEMOCAP database contains only clean speech as it was recorded in the controlled environment.

In order to test the redundancy of the feature set as in [25], we test the performance of the supervised classifier trained using the whole training set (labeled and unlabeled data with correct labels assigned) when using different feature sets - “natural” views defined in section III.A, as well as random split of features in two equal-sized partitions. The results are shown in table I.

<p>| TABLE I. THE PERCENT ACCURACY AND STANDARD DEVIATION OBTAINED BY A SUPERVISED CLASSIFIER TRAINED USING DIFFERENT VIEWS OF THE WHOLE TRAINING SET |</p>
<table>
<thead>
<tr>
<th>View 1 (MFCCs)</th>
<th>View 2 (LLDs)</th>
<th>Random halves</th>
<th>All features</th>
</tr>
</thead>
<tbody>
<tr>
<td>90.18 ± 6.59</td>
<td>93.03 ± 3.87</td>
<td>94.11 ± 0.51</td>
<td>94.12 ± 3.26</td>
</tr>
</tbody>
</table>
We can see from the results presented in Table I that a significant amount of redundancy exists within the feature set of the constructed dataset - if we randomly select just half of the available features, the resulting performance is still very close to the performance obtained by using all available features. This suggests that co-training with a random split of features can yield an improved classifier performance, as it was shown that this setting is beneficial in the case where there is enough redundancy in the data [26].

Accuracies and standard deviations obtained by applying the compared settings using the described fivefold-cross-validation process and different sizes of initial training set \( L \) are shown in Table II. Alongside with the performance of the settings described in sub-section III.C we report the performance of the supervised classifier trained on the initial set of labeled examples \( L \), denoted as \( L_{acc} \). The highest accuracies for each setting of \( L \) are bolded.

### Table II

<table>
<thead>
<tr>
<th>( L_{acc} )</th>
<th>10/10</th>
<th>25/25</th>
<th>50/50</th>
<th>100/100</th>
<th>200/200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand.</td>
<td>73.1±4.6</td>
<td>80.8±7.1</td>
<td>81.3±7.2</td>
<td>83.6±6.6</td>
<td>83.3±6.4</td>
</tr>
<tr>
<td>Natural</td>
<td>74.0±5.0</td>
<td>85.5±5.5</td>
<td>88.3±5.3</td>
<td>88.2±2.4</td>
<td>87.2±5.3</td>
</tr>
<tr>
<td>MV</td>
<td>72.3±8.2</td>
<td>83.2±5.2</td>
<td>85.0±5.9</td>
<td>87.1±4.2</td>
<td>88.2±4.6</td>
</tr>
<tr>
<td>RSSalg</td>
<td>76.0±7.4</td>
<td>87.3±3.8</td>
<td>87.2±6.6</td>
<td>88.2±5.5</td>
<td>89.3±3.7</td>
</tr>
<tr>
<td>RSSalg best</td>
<td>81.0±6.0</td>
<td>89.3±3.3</td>
<td>88.5±6.0</td>
<td>90.2±4.4</td>
<td>90.8±3.1</td>
</tr>
</tbody>
</table>

In order to better visualize the results, we plot the percentage increase of accuracy achieved by each setting. We consider the difference \( All_{acc} - L_{acc} \) as 100% of possible increase in accuracy, i.e. we consider a 100% possible increase in accuracy if the algorithm was able to assign the correct label to all available unlabeled instances and add them to the initial training set for training a supervised classifier. The percentage increase of accuracy for each setting is calculated as \( (obtained\_accuracy/All_{acc}) \times 100\% \). The percentage increases in accuracy are plotted in figure 1.

Experimental results presented in Table II and figure 1 show that all compared co-training settings succeeded in improving the accuracy of the initial classifier for all chosen sizes of the initial training set \( L \), with the exception of Natural setting applied in the extreme case where \( L \) comprises of only 10 labeled instances from each class.

Furthermore, we observe that all settings except for Random seem to benefit from adding more examples in the initial training set \( L \). However, it should be noted that if the initial classifier is too strong (i.e. the initial training set \( L \) is large enough to apply supervised training), co-training is unable to further improve its performance [27].

From Table II and figure 1 we observe that Random yields with a higher increase in accuracy than Natural for all sizes of \( L \) except for the largest one (10.5% of the whole training set). This result could be attributed to the fact that Random feature split produced stronger views in average than MFCC and LLD views used in Natural (Table I). However, Random setting is not optimal as we cannot guarantee the success of co-training with a particular random split of features.

The Natural setting was able to improve the initial classifier for all cases except for the extreme case where there are only 0.5% labeled instances in which case it degraded the performance of the initial classifier. From these results, we can further confirm that the split suggested in [11] is effective for co-training. The Natural setting yields maximal improvement when applied on largest initial labeled set \( L \).

The best performing settings in our experiment were RSSalg and MV. In each setting, they succeeded in improving the accuracy of the initial classifier and outperformed the Natural setting. They outperformed the Random setting for each setting of \( L \) except for the size of 2.6% of the training data.

We can observe that RSSalg and MV settings are more robust to the size of the initial training set than Natural and Random. Although adding more instances does improve the performance of these settings, the increase of performance is relatively small, i.e., increasing the size of \( L \) from 50 to 400 increases the performance of RSSalg and MV by 1.4% and 1.6%, respectively, while the same increase of \( L \) results with a more significant increase of 5% for the Natural setting.

It should be noted that RSSalg and MV settings require significantly more resources than Natural and Random setting as they require multiple runs of the co-training algorithm. However, RSSalg and MV run completely offline, without any human interaction which makes the time complexity less of a problem. Moreover, these solutions are highly parallelizable as multiple runs of co-training needed in RSSalg and MV are completely independent.

Finally, we analyze how close the performance of RSSalg is to its upper bound performance \( RSSalg_{best} \). We can conclude that \( RSSalg_{best} \) does outperform RSSalg, however their performances are correlated and relatively close with the exception of the extremely small initial labeled set \( L \). This result indicates that the process of
optimizing RSSalg parameters without using any labeled data is relatively successful for this case; however there is space for further improvement: given the optimal parameters RSSalg\_opt has significantly outperformed all compared settings. It is also the most robust setting – given only 1.3% labeled instances it was able to achieve the performance of 89.3%. Further adding of data to the initial training set L yields with little improvement of RSSalg\_opt (the best accuracy for this setting was 90.8% when 10.5% of the training set is labeled).

In our experiments, we have tried using more than 400 labeled instances as the initial training set L (more than 20.9% of the training data). However, this did not result with better performance of supervised algorithm trained using L or any of the compared settings. These results are omitted here as they are very similar to the results presented in the 5th column of table II (denoted as 200/200). This is probably due to using the initial labeled data recorded by only one male and one female speaker. It might be helpful to add examples recorded by different speakers to the dataset. However, it should be noted that the results obtained by applying RSSalg on the initial data recorded by only one female and one male speaker is reasonably high (89.6% compared to the performance of 94.12% obtained by using all labeled data from different speakers).

\[
\text{V. CONCLUSION}
\]

In this paper we have considered the problem of automatic gender recognition from emotional speech. In order to alleviate the problem of scarcity of labeled data, an ever-present bottleneck in automatic analysis of spoken language, we have investigated the possibility of applying a co-training style algorithm in order to reduce the human effort needed for data labeling. With this goal in mind we have tested the performance of several co-training settings applied to the task of gender recognition from emotional speech. In our experiments we have used co-training with random split of features, natural split of features, a majority vote of several co-training classifiers (MV) and Random Split Statistic Algorithm (RSSalg) developed with the goal of boosting co-training performance and enabling its application on single-view datasets. We have tested the performance of these settings using IEMOCAP, a gender-annotated emotional speech database.

Our experiments suggested that for this task, a random feature split yields a better performance of co training than the feature split obtained by the separation of MFCC (Mel-Frequency Cepstral Coefficient) features and LLD (Low-Level Descriptor) features which might be considered as the “natural” feature split. We attributed the high performance of Random split to the fact that the significant amount of redundancy exists among the constructed features and to the fact that the random feature split produced stronger views than LLD and MFCC views. The best performing settings in our experiments proved to be MV and RSSalg settings.

We have also tested the robustness of the considered solutions to the size of the initial labeled set L. Both MV and RSSalg proved to be most robust to this aspect.

In the future we plan to conduct more experiments using other emotional speech databases in order to account for speakers of different age groups and different speech quality.

One other branch of experiments would be to track the performance of individual random splits in order to discover the feature split best suitable for this problem.

Also, in the experiments conducted in this paper we have only tested the robustness of the compared solutions to the size of the initial data set L. In the future, we plan to experiment with other co-training parameters such as the growth size and number of iterations of co-training.

Finally, we plan to parallelize RSSalg in order to reduce the problem of computational complexity of this solution.

\[
\text{ACKNOWLEDGMENT}
\]

Results presented in this paper are part of the research conducted within the Grant No. III-47003 financed by the Ministry of Education and Science of the Republic of Serbia.

\[
\text{REFERENCES}
\]


