Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines

Dong-Yan Huang[1], *Shuzhi Sam Ge*[2], *Zhengchen Zhang*[2]

[1] Department of Signal Processing
Institute for Infocomm Research/A*STAR
1 Fusionopolis Way, # 21-01 Connexis (South Tower), Singapore 138632
[2]Social Robotics Lab, Interactive Digital Media Institute,
and Department of Electrical and Computer Engineering,
National University of Singapore
(huang@i2r.a-star.edu.sg, (samge, zhangzhengchen@nus.edu.sg))

Abstract

This paper describes a Speaker State Classification System (SSCS) for the INTERSPEECH 2011 Speaker State Challenge. Our SSC system for the Intoxication and Sleepiness Sub-Challenges uses fusion of several individual sub-systems. We make use of three standard feature sets per corpus given by organizers. Modeling is based on our own developed classification method - Asymmetric simple partial least squares (ASIMPLS) and Support Vector Machines (SVMs), followed by the calibration and multiple fusion methods. The advantage of asymmetric SIMPLS is prone to protect the minority class from being misclassified and boosts the performance on the majority class. Our experimental results show that our SSC system performs better than baseline system. Our final fusion results in 1.8% absolute improvement on the unweighted accuracy value for the Alcohol Language Corpus (ALC) and about 0.7% for the Sleepy Language Corpus (SLC) on the development set over the baseline. On the test set, we obtain 1.1% and 1.4% absolute improvement, respectively.

Index Terms: Speaker State Challenge, Asymmetric SIMPLS, SVMs, Fusion, Intoxication, Sleepiness

1. Introduction

Paralinguistic analysis attracts a lot of interests of researchers and engineers due to the constantly growing demands in the fields of Multimedia Retrieval and Human-Machine Interaction. Paralinguistics include not only emotional states, speaker's gender and age, but also speaker states, mood, etc. The INTERSPEECH 2011 Speaker State Challenge addresses two new sub-Challenges: the Intoxication Sub-Challenge and the Sleepiness Sub-Challenge. The analysis of intoxication and sleepiness in speech is considered to be useful in the applications in the medical domain for some disease diagnosis and surveillance in driving, steering or controlling [1].

In both Sub-Challenges, the tasks are two-class classification problems. But the distribution of two classes in datasets - ALCOHOL LANGUAGE CORPUS (ALC) and SLEEPY LANGUAGE CORPUS (SLC) provided by organizers, are imbalance. The challenge is how to classify these two classes based on imbalance datasets with high classification accuracy for both classes.

For such kind of asymmetric class distribution, regular classifiers even like support vector machines (SVMs), may take the minority class data as noise and make the class-boundary extremely in favor of the majority class, resulting in a low accuracy in classifying the minority class [2]. A lot of work have been proposed for imbalanced classification [3]. However, these works emphasize on over-learning the minority class or under-learning the majority class. They are not able to provide a correct model based on a small training dataset, which is capable to predict the new data. Partial Least Squares is a tool which allows to build a relationship between any two datasets. Therefore, we reformulate this binary classification problem on dimension reduction method, partial least square (PLS), and propose asymmetric SIMPLS algorithm to solve the problem, which features low computational complexity, less sensitive to class distribution and more capable to generate favorable features for classification. SIMPLS is an alternative approach to PLS and widely used in practice by its fast and avoidance of matrix inverse calculation [4].

Hence, we participate in the Intoxication and Sleepiness Sub-Challenges by employing our developed speaker state classification system. The best results obtained by the calibration and fusion show an absolute improvement of 1.8% on the unweighted accuracy value for the intoxication and 0.7% for the sleepiness compared to the competition baseline system on the development set. On the test set, we obtain 1.1% and 1.4% absolute improvement, respectively.

The paper is organized as following: Section 2 describes the system developed for the Intoxication and Sleepiness Sub-Challenges. Our developed classifier - Asymmetric SIMPLS and our fusion methods will be presented in details. Section 3 shows our experimental results to illustrate that our proposed system performs better than the competition baseline system in the development and test sets. Finally, we give a conclusion and the research direction in the future.

2. Speaker State Classification System

As the problems of these two challenges can be formulated into binary classification problems, we explore several methods to improve performance upon the baseline. For acoustic evident, we tried to optimize their individual performance by different feature configuration, different classifiers (e.g., Gaussian GMM, GMM, SVM, PLS, SIMPLS, asymmetric SIMPLS) as well as their combination.

Our final system makes a prediction based on fusion of

Copyright © 2011 ISCA

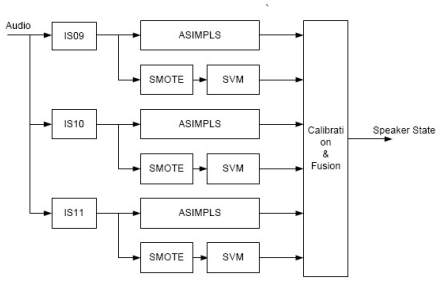

Figure 1: *Overall system structure.*

different information sources: different acoustic feature sets and different classifiers. The overall system structure is shown in Figure 1. The system is composed of six classifiers using three standard feature sets provided by organizers (e.g., IS09 → IS2009EC, IS10 → IS2010PC, IS11 → IS2011SSC) [1] and two classifiers per feature set. The base classifier (Asymmetric SIMPLS) and baseline classifier (SMO learned pairwise SVM with linear kernel, SMOTE on learning instances) are used since they achieve the best performance among several individual classifiers. The motivation of having six classifiers with different properties is that the diversity improve the combination and will ultimately lead to a better system. Finally, fusion and calibration of the six classifiers are carried out with different fusion methods.

2.1. Asymmetric SIMPLS

When we investigated different classification methods for these challenge tasks, we found that the data sets are highly unbalanced and all the existing classifiers even SVMs fail to give a high precision in classifying the minority class. The main drawback of these existing algorithms may not be able to provide a model for minority class which can capture the underlying relationships between the small amount training data and testing data. However, partial least squares (PLS) is a powerful tool to build the relationship between any two group datasets. In this paper, we address the problem of unbalanced data on PLS. As these challenge tasks are binary classification problems, we attempt to formulate these problems into PLS problems and propose an asymmetric classifier based on PLS to solve the unbalanced classification problem, especially promoting the prediction accuracy of the minority class.

Consider a set of M-dimensional features and its class label vector denoted as $\mathbf{X} \in R^{N\times M}$ and $\mathbf{y} \in R^N$, where N is the number of samples. Here $\mathbf{y}$ is one dimensional in binary classification problems. Partial least squares (PLS) is commonly used to model the relations between these two blocks by means of score vectors.

The PLS is based on the assumption that the matrix of M-dimensional $\mathbf{X}$ and its class label vector $\mathbf{y}$ are produced by a linear transformation of a speaker-independent latent variable vector $\mathbf{r}$ as

$$\mathbf{X} = \mathbf{Q}\mathbf{r} + \mathbf{e}_x \tag{1}$$

$$\mathbf{y} = \mathbf{P}\mathbf{r} + \mathbf{e}_y \tag{2}$$

where $\mathbf{Q}$ and $\mathbf{P}$ are speaker-specific transform matrices, and $\mathbf{e}_x$ and $\mathbf{e}_y$ are residual terms which cannot be modeled by the linear model.

Solving $\mathbf{Q}$ and $\mathbf{P}$ leads to the regression model leads to the regression model

$$\mathbf{y} = \mathbf{B}\mathbf{X} + \mathbf{e} \tag{3}$$

where $\mathbf{B}$ is the regression matrix which depends on $\mathbf{Q}$ and $\mathbf{P}$, and $\mathbf{e}$ is the regression residual. The difference PLS differs from the standard multivariate regression in the sense that also $\mathbf{X}$ is assumed to have a stochastic residual term. Furthermore, the rank of the regression matrix $\mathbf{B}$ is the dimensionality of the latent variable vector $\mathbf{r}$. This dimension is called the number of PLS components, and selecting it appropriately prevents overfitting effectively. The PLS model becomes equivalent to the multivariate regression if $\mathbf{B}$ has full rank, i.e., the number of latent variables equals the number of source variables.

In binary classification problem, the output, the PLS model for classification is expressed as

$$Y = \text{sign}(\mathbf{B}\mathbf{X} + \mathbf{e}) \tag{4}$$

There exists many variants for solving the PLS regression problem. In this paper, we propose an asymmetric SIMPLS algorithm which inherits the advantage of SIMPLS which has the advantages of being computationally efficient, its avoidance of matrix inverses, and operation on the original data instead of its covariances and boosts the performance on the large training data [4]. The following is a brief description of the processing steps of the algorithm.

Algorithm 1 Asymmetric SIMPLS Training

Input: Feature set $\mathbf{X}$
Label $\mathbf{y}$
Number of components l
Variables: Projection matrix $\mathbf{R}$,
score vectors $\mathbf{T}$ and $\mathbf{U}$, loading $\mathbf{P}$ and $\mathbf{Q}$
$\mathbf{R} = []; \mathbf{V} = []; \mathbf{Q} = []; \mathbf{T} = []; \mathbf{U} = [];$
$\mathbf{y} = [y_1, y_2, ..., y_N]^T; \mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^T$
$\mathbf{y}_0 = \mathbf{y} - \text{mean}(\mathbf{y}); \mathbf{X}_0 = \mathbf{X} - \text{mean}(\mathbf{X});$
$\mathbf{S} = \mathbf{X}_0\mathbf{y}_0^T$
for $i = 1$ to l **do**
  $\mathbf{q}_i =$ dominant eigenvectors of $\mathbf{S}^T\mathbf{S}$
  $\mathbf{r}_i = \mathbf{S} * \mathbf{q}_i$
  $\mathbf{t}_i = \mathbf{X}_0 * \mathbf{r}_i$
  $\text{normt}_i = \text{SQRT}(\mathbf{t}_i^T\mathbf{t}_i)$
  $\mathbf{t}_i = \mathbf{t}_i/\text{normt}_i$
  $\mathbf{r}_i = \mathbf{r}_i/\text{normt}_i$
  $\mathbf{p}_i = \mathbf{X}_0^T * \mathbf{t}_i$
  $\mathbf{q}_i = \mathbf{y}_0^T * \mathbf{t}_i$
  $\mathbf{u}_i = \mathbf{y}_0 * \mathbf{q}_i$
  $\mathbf{v}_i = \mathbf{p}_i$
  if $i > 1$ **then**
    $\mathbf{v}_i = \mathbf{v}_i - \mathbf{V} * (\mathbf{V}^T * \mathbf{p}_i)$
    $\mathbf{u}_i = \mathbf{u}_i - \mathbf{T} * (\mathbf{T}^T * \mathbf{u}_i)$
  end if
  $\mathbf{v}_i = \mathbf{v}_i/\mathbf{SQRT}(\mathbf{v}_i^T * \mathbf{v}_i)$
  $\mathbf{S} = \mathbf{S} - \mathbf{v}_i * (\mathbf{v}_i^T * \mathbf{S})$
  $\mathbf{r}_i, \mathbf{t}_i, \mathbf{p}_i, \mathbf{q}_i, \mathbf{u}_i,$ and $\mathbf{v}_i$ into
  $\mathbf{R}, \mathbf{T}, \mathbf{P}, \mathbf{Q}, \mathbf{U},$ and $\mathbf{V}$, respectively.
end for
$\mathbf{B} = \mathbf{R} * \mathbf{Q}^T$

The algorithm generally works on zero-mean training data and its class label $\mathbf{X}$ and $\mathbf{y}$, respectively, so the empirical means of the vectors are subtracted before the processing, and afterwards added to the regression results. In each iteration i, the algorithm estimates score vector $\mathbf{r}_i$ which explains most of the

Algorithm 2 Asymmetric SIMPLS Testing

Input: Projection matrix $\mathbf{R}$,
score vectors $\mathbf{T}$ and $\mathbf{U}$, loading $\mathbf{P}$ and $\mathbf{Q}$
Output: Predicted Label $\hat{Y}$ of $\mathbf{X}$,
Use Algorithm 1 to obtain projection matrix $\mathbf{R}$, loading $\mathbf{P}$ and $\mathbf{Q}$.
Calculate $\mathbf{m}, C_p, C_n, r_0$ and r_c according to their definitions in Eqs.(8) and (9)
Calculate b according to
$b = (m_1 + \Delta)(C_p - r_0(C_p - C_n)/(r_c + r_0))$
Choose Δ according to Eq.(10)
$\hat{\mathbf{S}}_0 = \mathbf{X}_t$
for $i = 1$ to l **do**
  $\hat{\mathbf{t}}_i = \hat{\mathbf{S}}_{i-1}\mathbf{r}_i$;
  $\hat{\mathbf{S}}_i = \hat{\mathbf{S}}_{i-1} - \hat{\mathbf{t}}_i\hat{\mathbf{t}}_i^T\hat{\mathbf{S}}_{i-1}/(\hat{\mathbf{t}}_i^T\hat{\mathbf{t}}_i)$;
end for
$\hat{Y} = \text{sign}(\sum_{i=1}^{A} m_i\hat{\mathbf{t}}_i - b)$

cross-covariance between $\mathbf{X}$ and $\mathbf{y}$. The i th element in vector $\mathbf{t}_i$ corresponds to a coefficient in the latent variable vector $\mathbf{r}_i$ in Eqs. (1) and (2), whereas the loading vectors $\mathbf{p}$ and $\mathbf{q}$ are the corresponding rows in matrices $\mathbf{P}$ and $\mathbf{Q}$. After each iteration, the contribution of the estimated PLS component is subtracted from the cross-covariance matrix $\mathbf{S}$. After the score vectors are extracted according to Algorithm 1, the PLS regression model is written in the matrix form as

$$\mathbf{y} = \mathbf{R}\mathbf{Q}^T\mathbf{X} + \mathbf{e} = \mathbf{B}\mathbf{X} + \mathbf{e} \tag{5}$$

where $\mathbf{B} = \mathbf{R}\mathbf{Q}^T$ and Eq. (5) is the same as Eq.(4). To make prediction on new sample data, we rewrite (5) into (6)

$$\hat{\mathbf{y}} = \sum_{i=1}^{l} b_{ii}\hat{\mathbf{t}}_i q_i \tag{6}$$

where $\hat{\mathbf{t}}_i = \hat{\mathbf{S}}_{i-1}\mathbf{r}_i$ and $\hat{\mathbf{S}}_i = \hat{\mathbf{S}}_{i-1} - \hat{\mathbf{t}}_i\hat{\mathbf{t}}_i^T\hat{\mathbf{S}}_{i-1}/\text{normt}_i, \hat{\mathbf{S}}_0 = \mathbf{X}_t$; $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_l]^T$ and $\mathbf{Q} = [q_1, q_2, \cdots, q_l]^T$. Finally, the output, the PLS model for classification is represented as

$$\hat{Y} = \text{sign}(\sum_{i=1}^{l} m_i\hat{\mathbf{t}}_i) = \text{sign}(\hat{\mathbf{t}}\mathbf{m}) \tag{7}$$

where $\hat{\mathbf{t}}_i$ is the ith score vector of $\mathbf{X}$, and $m_i = b_{i,i}q_i$.

In fact, according to our observation, $\mathbf{y} = \hat{\mathbf{t}} \cdot \mathbf{m}$ represents the boundary of two classes on a plane - a line L, which goes through the point $(0, 0)$ of the plane, away from the center of the minority class. Based on the discrimination function of PLS model, points on the left side of L are classified as majority class, while those on the right side are in minority class. Since the L passes cross the majority class, away from the minority class, the PLS Classifier favors to detect the minority class at a cost of loss of true majority class.

In practice, it is desirable to develop an algorithm possessing good sensitivity of the minority class while enhancing the performance of majority class classification. We attempt to estimate a boundary of these two classes. We can estimate the center points of positive and negative classes on the first dimension, C_p and C_n, respectively, as well as radius of two circles r_p and r_c. The relationship between two circles in position can be expressed as

$$C_p - C_n = c(r_p + r_c) \tag{8}$$

where c represents the overlapping degree between both circles, the larger value of c, the less overlapping between both circles. We attempt to find the estimated boundary b as

$$b = (m_1 + \Delta)(C_p - r_p c) \tag{9}$$

where the Δ is chosen such that the predicted label is as close as the ground truth of development set.

$$E[e_y^2] = \text{min}E[(Y - \hat{Y})^2] = \text{min}E[(Y - \text{sign}(\sum_{i=1}^{l} m_i\hat{\mathbf{t}}_i - b))^2] \tag{10}$$

2.2. Classifier Fusion

Three difference fusion methods are employed: FoCal fusion, adaboost fusion and simple fusion. Linear logistic regression fusion is done with the FoCal Multiclass Toolkit [5]. Weights are assigned to the two systems in a training step. A weighted sum of the different system scores is the combination result. An adaboost model is trained using the scores obtained by the two systems on the development data set. The combined result is the testing output on the same score set. A simple fusion method is also used to generate combination results:

$$s(t) = \sum_{i=1}^{k} \alpha_i * s_i(t) \tag{11}$$

where t is an instance of the development set and $s_i(t), i = 1, \cdots, k$ are scores generated by the subsystems on the development set and k is the number of subsystems. The parameter $\alpha_i \in \{0, 0.1, 0.2, \cdots, 1\}, i = 1, \cdots, k$, and $s(t)$ is the final score combined. Every combination of the α_i values are tried, and the value set which obtains the best score is selected as the parameters for fusing the results obtained on test set.

3. Experimental Results

3.1. Intoxication Sub-Challenge

The corpus for intoxication sub-challenge is ALC [1]. It consists of 154 speakers (77 male, 77 female) within the age range from 21-75 who are randomly partitioned into gender balanced training, development and test sets. The features provided by the baseline system are employed in our experiments. We also use the features provided by INTERSPEECH 2010 Paralinguistic Challenge [6] and INTERSPEECH 2009 Emotion Challenge [7] for training and testing.

Table 1 summarizes the results obtained on the ALC corpus development set by the individual classifier and by the fusion methods. Results are expressed in terms of accuracy of non-alcoholized (NAL) data (Acc_{NAL}), accuracy of alcoholized (AL) data (Acc_{AL}), unweighted and weighted accuracy on average per class (% UA and % WA). It can be seen that the ASIMPLS method gives better classification result for the AL data which is the minority class in the corpus than SVM method. While SVM method performs better on the NAL data which is majority class, and as a result, a better weighted accuracy is obtained. Better results can be generated by using feature sets IS11 and IS10, while the feature sets IS09 gave us worse results. One can also find that the ASIMPLS method generates comparable results with the SVM method using the IS11 feature set, and it performs better than the later using all the other three feature sets. The performances of three different fusion methods are compared in this table. According to our experiments,

the best fusion results are given by the simple fusion method which combines the ASIMPLS-IS10, ASIMPLS-IS11, SVM-IS10 and SVM-IS11 subsystems. It can be seen that the simple fusion method represents an absolute improvement of 1.8% on the development set and 1.1% on the test set over the reported challenge baseline [1].

Table 1: Experimental results obtained on the ALC development and test sets.

Classifier	Acc_{NAL}	Acc_{AL}	%UA	%WA
ASIMPLS-IS09	58.46	62.74	60.60	59.72
ASIMPLS-IS10	62.47	67.18	64.83	63.86
ASIMPLS-IS11	60.68	**69.74**	65.21	63.36
SVM-IS09	71.22	44.87	58.05	63.43
SVM-IS10	71.94	51.71	61.82	65.96
SVM-IS11	**74.73**	55.90	65.31	**69.17**
FoCal Fusion	70.68	62.14	66.41	68.16
Adaboost Fusion	66.20	66.67	66.43	66.34
Simple Fusion	67.42	66.92	**67.17**	67.27
Train+Develop vs. Test				
Simple Fusion	70.06	63.99	67.02	67.27

3.2. Sleepiness Sub-Challenge

The corpus and features provided for the sleepiness sub-challenge are used for testing our classification system. In the SLC, 99 participants within the age range 20-52 years took part in the studies. The sampling rate of speech is 44.1 kHz and down-sampled to 16 kHz.

Table 2 shows the system performance on the SLC development and test sets. The ASIMPLS method gives higher accuracy on the minority spleepy data again of the SLC, while SVM performs better on non-sleepy (NSL) data. The ASIMPLS method performs worse than the SVM method using IS11 feature set, and it generates comparable results with the later using the other feature sets on this corpus. The best fusion result is produced by the adaboost fusion method which combines ASIMPLS-IS09, ASIMPLS-IS10, ASIMPLS-IS11, SVM-IS09, SVM-IS10, SVM-IS11 subsystems. After fusion, we obtained 1.5% improvement to the SVM-IS11 method which is the best individual subsystem. It is worth noting that the performance given by SVM-IS11 method 66.58% UA here is lower than the reported challenge baseline 67.3% UA. It is acceptable because the SMOTE technology uses generated random parameters to balance the training data. As a result, we obtain different results for different testing platform. Finally we obtain 68% UA which is 0.7% absolute improvement on the development set and an absolute improvement of 1.4% over the reported baseline system.

4. Conclusion

In this paper, we presented our Speaker State Classification System for INTERSPEECH 2011 Intoxication and Sleepiness Sub-Challenges. The classification system is composed by the fusion of six individual sub-systems trained on the provided corpora. An asymmetric simple partial least squares method was proposed aiming at promoting the predication accuracy of the minority class data in unbalance data set. Linear logistic regression fusion, adaboost fusion and simple fusion methods are employed, and their performances are compared. The best results obtained by the whole system show an absolute improvement of 1.8% on the unweighted accuracy value on the development set and 1.1% on the test set for the Intoxication Sub-Challenge compared to the baseline system. For the Sleepiness Sub-Challenge, our system achieves performance of an absolute improvement of 0.7% on the development set and an absolute improvement of 1.4% on the test set over the baseline system. The predication accuracy of the minority class data is more than 65% which is 10% absolutely higher than the SVM method.

Table 2: Experimental results obtained on the SLC development and test sets.

Classifier	Acc_{NSL}	Acc_{SL}	% UA	% WA
ASIMPLS-IS09	58.55	**71.55**	65.05	63.36
ASIMPLS-IS10	58.55	69.69	64.12	62.68
ASIMPLS-IS11	62.04	68.67	65.36	64.49
SVM-IS09	63.94	65.43	64.69	64.49
SVM-IS10	75.65	53.66	64.66	67.51
SVM-IS11	**78.76**	54.40	66.58	69.74
FoCal Fusion	72.77	62.28	67.52	68.89
Adaboost Fusion	72.39	61.54	66.96	68.37
Simple Fusion	75.71	60.33	**68.02**	**70.02**
Train+Develop vs. Test				
Simple Fusion	79.10	64.28	71.69	74.61

The experimental results showed that our proposed asymmetric SIMPLS is prone to protect the minority class from being misclassified and boosts the performance on the majority class. The theoretical performance of ASIMPLS on the classification accuracy of majority and minority classes will be further studied in the future.

5. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech (2011)*, (Florence, Italy), ISCA, 2011.

[2] G. Wu and E. Y. Chang, "Kba: Kernel boundary alignment considering imbalanced data distribution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 786–795, 2005.

[3] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263 – 1284, 2009.

[4] S. de Jong, "SIMPLS: an alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993.

[5] N. Brummer, "Focal multiclass toolkit." http://niko.brummer.googlepages.com/focalmulticlass.

[6] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. M "uller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[7] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.