Speech-based emotion recognition: Application of collective decision making concepts

Christina Brester, Eugene Semenkin,
Siberian State Aerospace University named after academician M. F. Reshetnev,
Krasnoyarsk, Russian Federation

Maxim Sidorov
Ulm University, Ulm, Germany

Wuhan, 2014

Outline

- Background and Motivation
  - Some Examples
  - Problem Definition
  - Corpora Description
- Conventional models
  - Experiment Conducted
  - Results Obtained
  - Inferences #1
- Collective decision making in emotion recognition
  - Main Concepts
  - Results Obtained
  - Inferences #2
- Conclusions and Future work

Example #1

Human-Human Communication

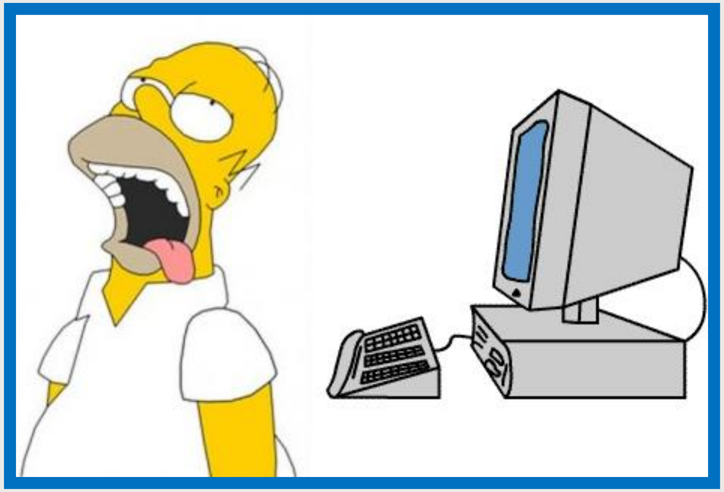
First 30 min

After a while

Human-Machine Communication

First 30 min

After a while



To show regret

or

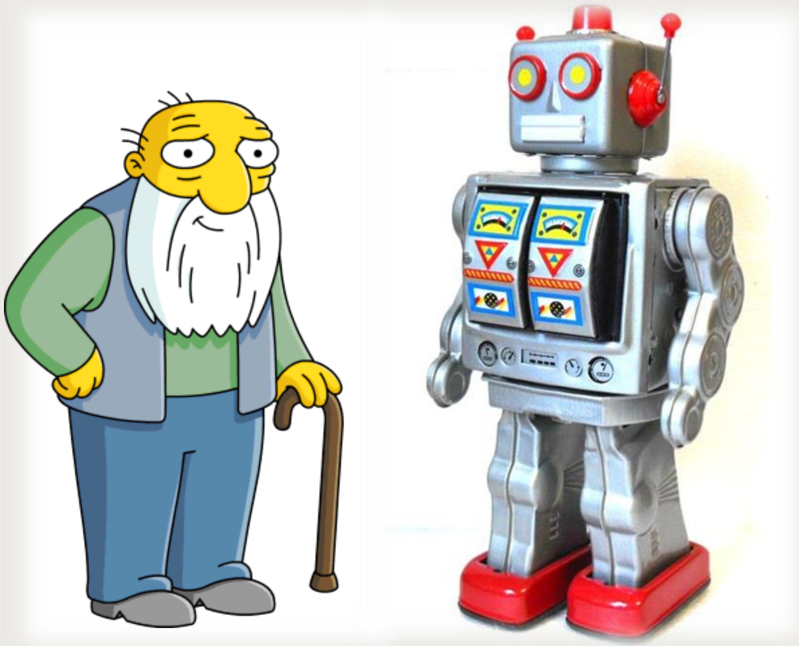
To express happiness

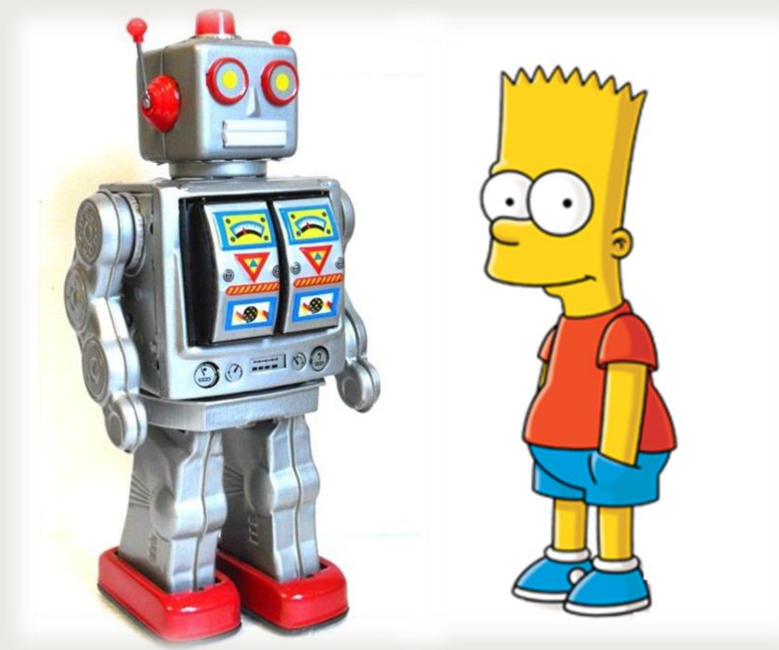
Example #2

To personalize a response

Good morning,
Mister!
Can I help you?

Hey, guy!
What's up?





Example #3

Quality monitoring of call centres

Are you kidding?
I've been waiting for two hours!

Please, wait a minute, Sir!

An agent

..okay

Speech-based Emotion Recognition Problem

List of extracted features

- General features: Power, Mean, Root mean square, Jitter, Shimmer
- Mel-frequency cepstral coefficients (MFCCs):12 MFCCs
- Formants: 5 Formants
- Pitch, Intensity and harmonicity based features: Mean, Minimum, Maximum, Range, Deviation
- Etc.

Voice

Voice conversion into the digital form

Extraction of numerical characteristics

Classification of sound signals

The **emotion** is detected

Sample

$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	y_1
$x_{2,1}$	$x_{2,2}$	…	$x_{2,m}$	y_2
$x_{3,1}$	$x_{3,2}$	…	$x_{3,m}$	y_3
…	…	…	…	…
$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	y_n

$\bar{x}_i$ – independent variable,
y_i – dependent variable, $i=\overline{1,n}$,
$y_i \in C$, where $C=\{c_1,c_2,\ldots,c_r\}$ – finite set,
r – the number of classes.

New examples

$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	?
…	…	…	…	…
$x_{l,1}$	$x_{l,2}$	…	$x_{l,m}$	?

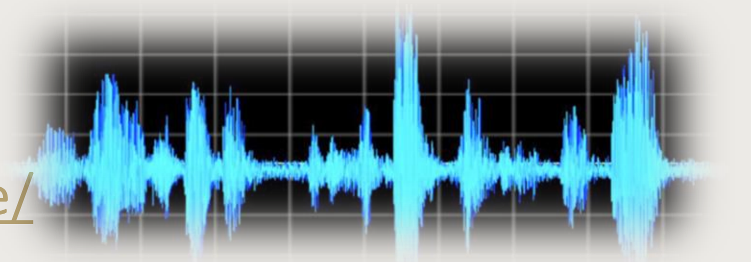
Goal:
To classify new objects based on the sample (supervised learning).

*To get the conventional feature set introduced at **INTERSPEECH 2009**, the following systems might be used*

- **Praat**

http://www.fon.hum.uva.nl/praat/

University of Amsterdam



- **OpenSMILE**

http://sourceforge.net/projects/opensmile/

Technical University of Munich

Speech-based Emotion Recognition Problem

List of extracted features

- General features: Power, Mean, Root mean square, Jitter, Shimmer
- Mel-frequency cepstral coefficients (MFCCs):12 MFCCs
- Formants: 5 Formants
- Pitch, Intensity and harmonicity based features: Mean, Minimum, Maximum, Range, Deviation
- Etc.

Voice

Voice conversion into the digital form

Extraction of numerical characteristics

Classification of sound signals

The **emotion** is detected

Sample

$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	y_1
$x_{2,1}$	$x_{2,2}$	…	$x_{2,m}$	y_2
$x_{3,1}$	$x_{3,2}$	…	$x_{3,m}$	y_3
…	…	…	…	…
$x_{n,1}$	$x_{n,2}$	…	$x_{n,m}$	y_n

$\bar{x}_i$ – independent variable,
y_i – dependent variable, $i=\overline{1,n}$,
$y_i \in C$, where $C=\{c_1,c_2,\dots,c_r\}$ – finite set,
r – the number of classes.

New examples

$x_{1,1}$	$x_{1,2}$	…	$x_{1,m}$	?
…	…	…	…	…
$x_{l,1}$	$x_{l,2}$	…	$x_{l,m}$	?

Goal:
To classify new objects based on the sample (supervised learning).

Corpora description

Database	Language	Full length (min.)	Number of emotions	File level duration		Notes
				Mean (sec.)	Std. (sec.)	
EMO-DB	German	24.7	7	2.7	1.02	Acted
SAVEE	English	30.7	7	3.8	1.07	Acted
LEGO	English	118.2	5	1.6	1.4	Non-acted
VAM	German	47.8	4	3.02	2.1	Non-acted
RadioS	German	278.5	4	6.26	5.17	Non-acted
UUDB	Japanese	113.4	4	1.4	1.7	Non-acted

Conventional classification models used

* Multilayer Perceptron (MLP)
* Support Vector Machine (SVM)
* Linear Logistic Regression (Logit)
* Radial Basis Function network (RBF)
* Naive Bayes
* Decision trees (J48)
* Random Forest
* Bagging
* Additive Logistic Regression (LogitBoost)
* One Rule (OneR)

Experiment conducted

For each classifier the *F-score* metric was evaluated to estimate the results of the 6-fold cross-validation procedure:

the more effective the classifier that we used, the higher F-score value we obtained.

$$\boxed{F\_score = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}}$$

F-score definition

		True_class			
		Class$_1$	Class$_2$	…	Class$_N$
Predicted_class	Class$_1$	a_{11}	a_{12}	…	a_{1N}
	Class$_2$	a_{21}	a_{22}	…	a_{2N}
	…	…	…	…	…
	Class$_N$	a_{1N}	a_{2N}	…	a_{NN}

$$precision_l = \frac{a_{ll}}{\sum_j a_{lj}},$$

$$recall_l = \frac{a_{ll}}{\sum_i a_{il}},$$

$$F\_score = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision}, \quad Precision = \sum_l precision_l, \quad Recall = \sum_l recall_l.$$

Experimental results for conventional classifiers, *F-score*, %

	Emo-DB	SAVEE	LEGO	VAM	RadioS	UUDB
MLP	**82.87**	**61.72**	67.53	41.08	**34.81**	25.48
SVM	81.71	59.22	**70.81**	43.57	27.26	35.59
Logit	80.04	57.20	70.75	36.88	31.91	36.72
RBF	68.93	43.27	52.61	37.87	23.14	26.75
Naive Bayes	66.91	43.64	57.00	40.86	34.02	36.52
J48	50.15	42.46	57.55	36.20	29.81	38.70
Random Forest	54.69	38.60	65.47	**45.66**	30.31	40.11
Bagging	60.60	42.99	67.53	37.24	26.37	40.94
Logit Boost	66.66	49.08	67.66	40.06	31.24	41.28
OneR	29.20	30.41	59.01	33.34	23.94	**41.92**

Inferences #1

- There is no particular model that is equally effective for all of the databases.

- The random choice of the classifier may lead to significant performance deterioration.

- For the used corpora Multilayer Perceptron (MLP), Support Vector Machine (SVM) and Linear Logistic Regression (Logit) demonstrated rather high performance.

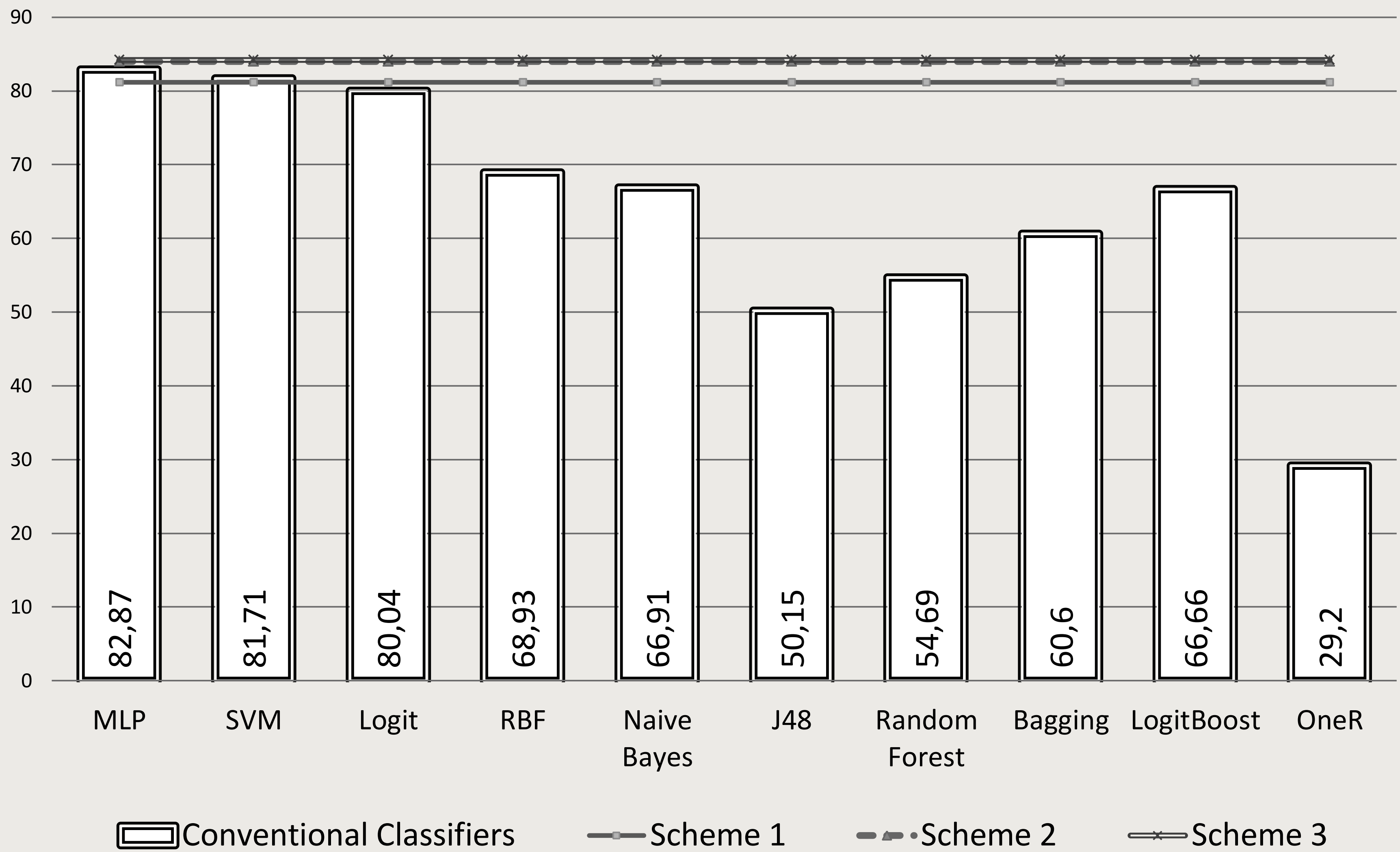
Collective decision making

Concept	Detailed information
Scheme 1. For each test example: *Choose a model that classifies correctly k-nearest neighbours from the training data set.*	1. For each test example it is necessary to determine k-nearest neighbours from the training data set. 2. The prediction of the model that classifies these k-nearest neighbours correctly is used as the final decision. (If several models demonstrate equal effectiveness, choose one of them randomly).
Scheme 2. *Voting procedure is realized with the usage of the majority rule.*	1. For each test example the engaged models vote for different classes according to their own predictions. 2. The final decision is defined as a collective choice based on the majority rule.
Scheme 3. *Combination of Scheme 1 and Scheme 2.*	Combine Schemes 1 and 2 in the following way: - fulfil the voting procedure as it is described in Scheme 2; - if several classes have the maximum number of votes, apply Scheme 1.

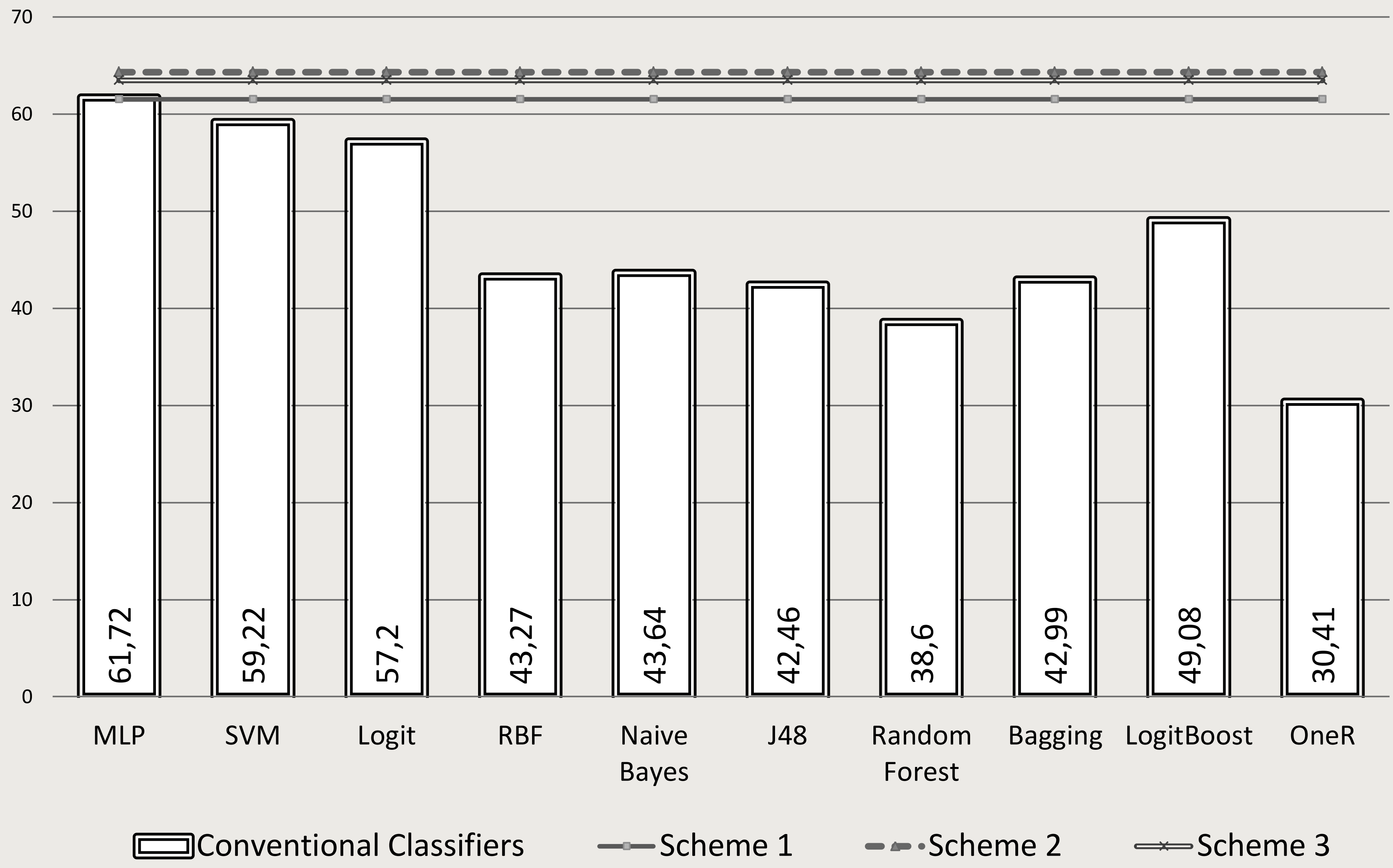
Experimental results for collective decision making schemes

	Scheme 1	Scheme 2	Scheme 3
Berlin	81.18	84.01	**84.23**
SAVEE	61.52	**64.33**	63.50
LEGO	70.52	**71.19**	71.13
VAM	42.29	**50.19**	43.69
RadioS	**30.68**	26.39	26.39
UUDB	37.96	36.41	**39.78**

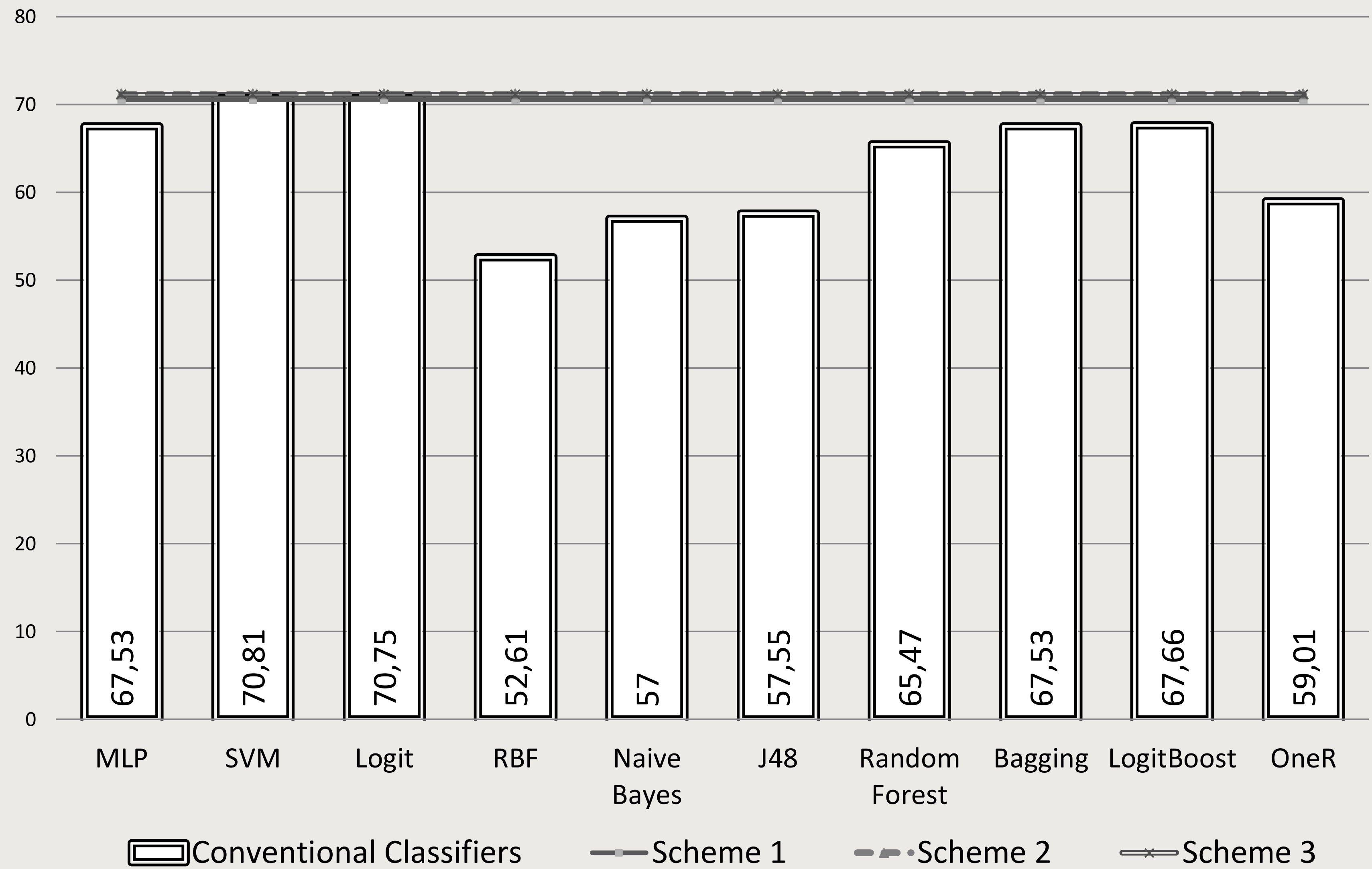
Classification results for Emo-DB



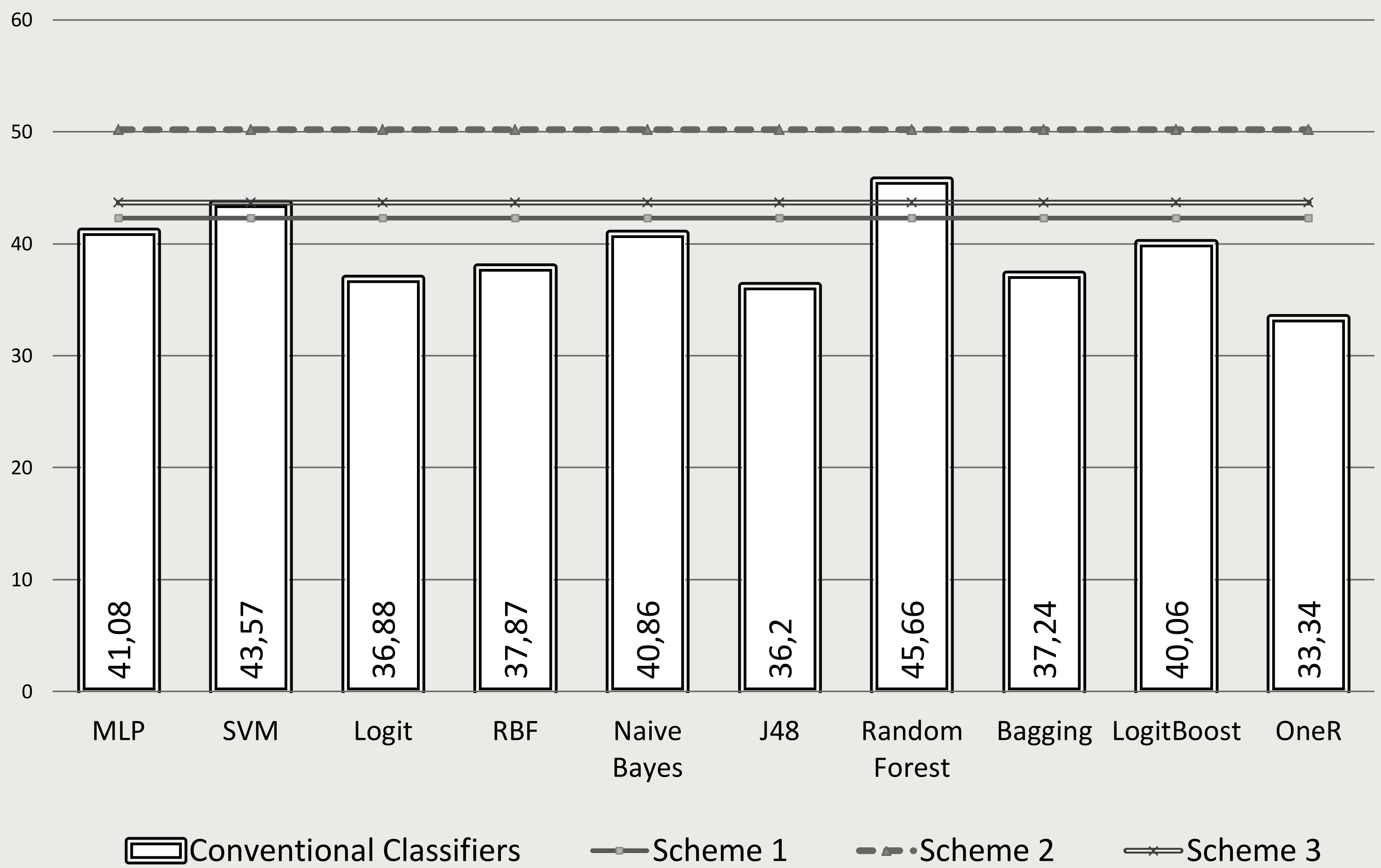
Classification results for SAVEE



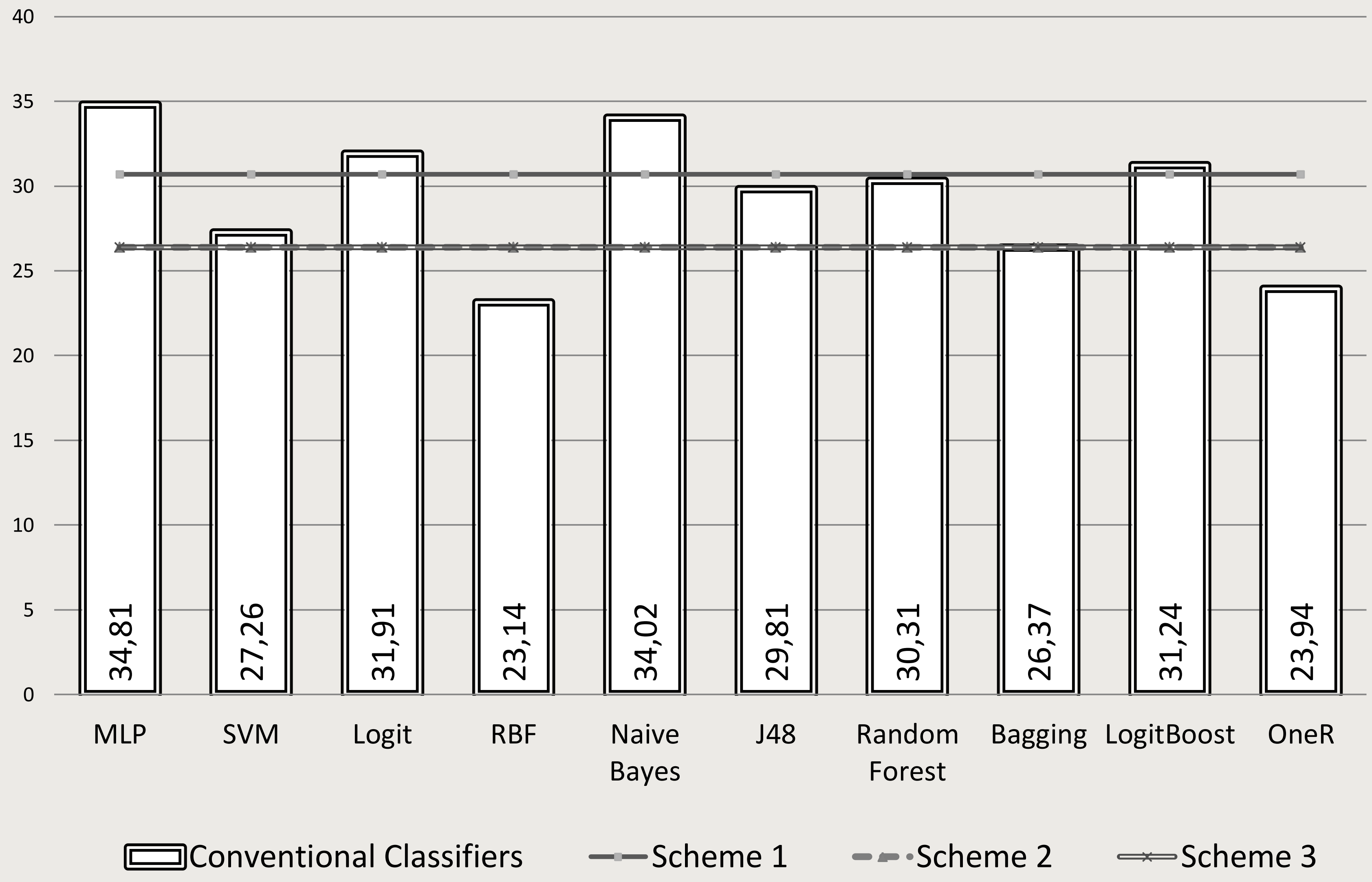
Classification results for LEGO



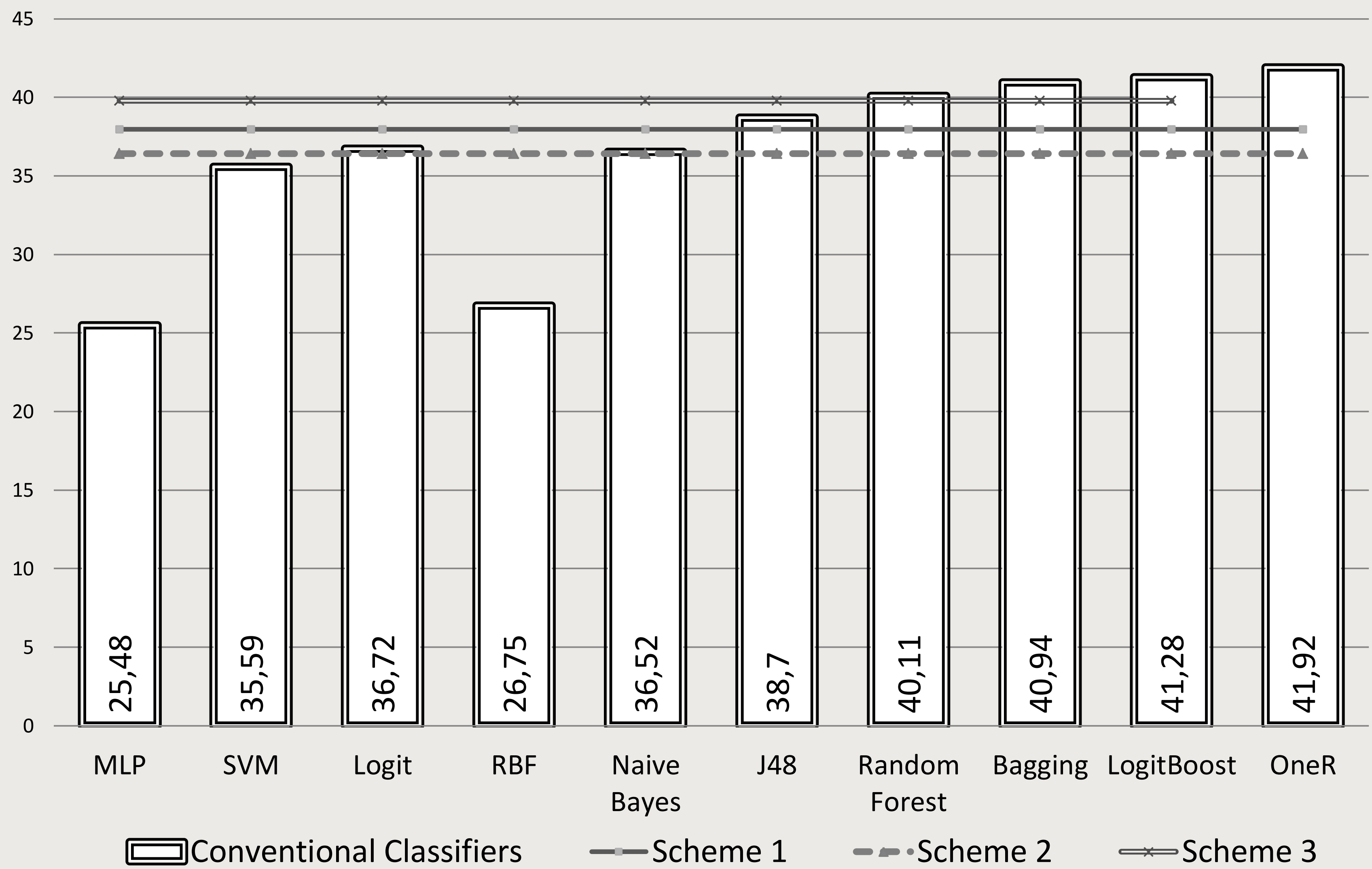
Classification results for VAM



Classification results for RadioS



Classification results for UUDB



Inferences #2

- Due to the usage of the proposed techniques it became possible **to improve** the classification results for most of the corpora.

  (In some cases even by up to **9.93%** relative improvement)

- On the set of the presented databases Scheme 2 was the most effective for the collective classification process.

Conclusions and Future work

1. Although we managed to achieve some good results, there are a number of **questions**:

- *How many classifiers should we use to provide the most reliable scheme? What kind of models should it be compulsory to include in the ensemble of classifiers?*

2. There are some other aspects related to recognition of qualities of the user such as **gender** and **speaker identification**. Consequently, the proposed schemes might be applied to solve these problems.

Thanks a lot