MixedEmotions

Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets

D6.5 Call Center Operations Pilot Implementation, initial Version

Project ref. no	H2020 644632
Project acronym	MixedEmotions
Start date of project (dur.)	01 April 2015 (24 Months)
Document due Date	31 March 2016 (Month 12)
Responsible for deliverable	Phonexia
Reply to	matejka@phonexia.com
Document status	Final

Project reference no.	H2020 644632

Project working name	MixedEmotions
Project full name	Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets
Document name	MixedEmotions_D6.5_31_03_16_Pilot3_Callcenter
Security (distribution level)	PU
Contractual delivery date	31 March 2016
Deliverable number	D6.5
Deliverable name	Call Center Operations Pilot Implementation, initial version
Type	Demonstrator
Version	Final
WP / Task responsible	WP6 / Phonexia
Contributors	Phonexia(Pavel Matejka, Filip Povolny, Aneta Cerna) University of Passau(Hesam Sagha)
EC Project Officer	Susan Fraser

Table of Contents

Executive Summary

Pilot 3 is one of the examples of the output of MixedEmotions recognition and deals with Call Center operation. This document, Deliverable 6.5, describes Phonexia Platform for Call Centers - Speech Analytic Platform - SPAS and implementation of emotion recognition modules. It also addresses other related issues like REST server for calling the specific modules, gender and age recognition for better analysis and conditioning the output and Speech To Text (STT) advances in Phonexia.

1 Introduction

The first part of the document describes Phonexia Speech Analytic Platform for Call Centers monitoring (SPAS). It starts with the Call Centers operation description, use case definition and technical details about the platform. Next there are screenshots of the application itself together with the description of how to use it and set it up. Next part describes the implementations and visualizations of the emotion recognition done with project partners within this project. We created 3 modules of acoustic emotion recognition from the beginning of the project and all three were implemented and tested with SPAS. We also added emotion recognition from predefined keywords which are spotted in audio signal.

In the Call Centers (CC), and also in this document, there are used to following designation:

- **client** - is a person who contacts / is contacted by Call Centers because of some services or products
- **agent** - is a staff attending to clients' contacts / requirements
- **supervisor** - is a direct superior of agents, having to worry about their motivation and also about the responsibility for the results of his team

2 SPAS - Speech Analytics Platform

SPAS is a solution for Call Centers allowing automated analysis of all calls using speech technologies. With this solution, we can dramatically improve the quality and efficiency while reducing cost of supervision of the calls. Additionally, the calls contain other interesting information, such as demographic information about client, that was not used before, because it was too expensive to acquire.

2.1 Introduction

There are huge amount of calls per day in Call Centers (see the table below). The supervisor of a Call Center manages to manually inspect about 1-3% of the calls on random bases. The checking is performed manually by supervisors, at a cost of approximately 4 EUR per call for the CC. With SPAS, we developed solutions where 100% of the calls is automatically processed and ranked.

For the CC, this means that the supervision costs can be either reduced, or much more efficiently used for problematic calls that would have otherwise remained undetected.

CC size	S	M	L	XL
Amount [min] / per month	20 000	100 000	300 000	700 000

size of Call Centers in Czech Republic

Part of the success of CC marketing campaigns lies in targeting the right people at the right time. For instance, the CC has to ensure that they are calling people which are targets from the campaign's guidelines. With SPAS, the CC will be able to automatically verify the identity of the called person (provided the client was contacted before) and to gain key data for the marketing campaign: age range segment, gender, reaction speed, etc.

SPAS Solution automatically analyzes 100% of calls - identifies and analyzes critical places in calls, evaluate agents' efficiency in a team scope, gets detailed statistics from all calls and key information about client (e.g. estimate the age and gender of the speaker). We added emotion recognition module from MixedEmotions project to the SPAS. This helps to detect problematic calls which were not spotted before. In this document there is a special section describing more about this use-case (*3 Emotions in SPAS*).

2.2 Use-cases

SPAS helps users in many different ways, and each new client usually comes up with a new way that can be used. But mostly it is used in the following use-cases:

CLIENT'S speech:

- outbound - marketing research - opinion poll
- inbound - reason of the call - divided into categories
- client response / satisfaction - positive, negative, neutral
- tracking trends
- getting information about demographic elements (age / gender of a client)

AGENT'S speech – control of:

- compliance call script
- communication standards - process of handling objections / argumentation / forbidden phrases

- suitable offer according to metadata (e.g. offer another product, if the client has met the conditions)

DIALOGUE:

- providing approval on legislative issues (card activation, money transfer, monitoring, etc.)

2.3 Technical details about platform

SPAS is web based application written in Java EE and Javascript. Web server runs on Tomcat server and uses MySql database. Many frameworks are used, for example - Spring, MyBatis, Wicket, Dozer, Birt, etc.

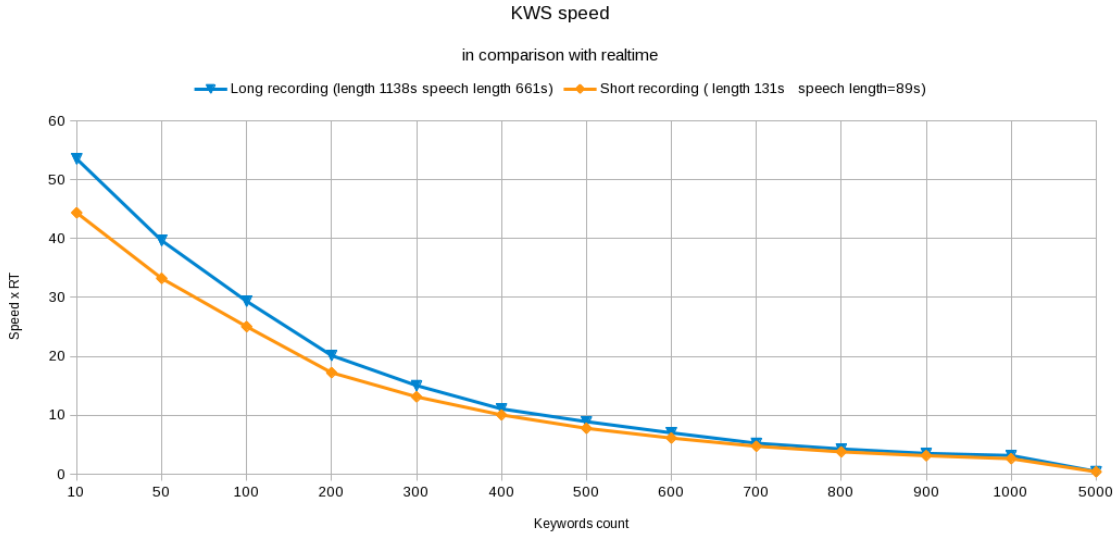
SPAS is all in one solution which encapsulates speech technologies. Outputs from these technologies are used for further speech analytics - Workflow, Visual Analytics, Dashboards, etc.

SPAS includes the following technologies:

- Acoustic Keyword Spotting (KWS)
- Transcription (STT = speech to text)
- Age Estimator (AE)
- Gender Identification (GI)
- Language Identification (LI)
- Speaker Identification (SI)
- Time Analysis Extractor (TAE) - basic time statistics - who spoke, when spoke, how long, how quickly, detects pauses / cross-talks in a speech

Most of the processing and checking of the calls is done by KWS. The software works by analyzing the content of clients' and agents' speech - the words and phrases they use, its ordering, and the questions they ask - as well as the context of the call and their intonation.

Hardware requirements depend on types and quantity of the used speech technologies. The amount of instances of the individual technologies is counted from client's requirement on number of recordings processed per a day. The length of the recording and number of analyzed keywords are very important too. Next figure shows the influence of the KWS engine on the number of processed keywords and its real-time processing (50 means 50 times faster than real-time)



As an example, the following table shows real traffic in two Call Centers Phonexia is working with.

Call Center	# Recordings per day [pcs]	# Recordings length [s]	# Recordings processed by KWS [%]
A	10 000	140	54
B	850	95	99

2.4 First steps with SPAS

SPAS is designed to have a user friendly and easy to use interface. A lot of interesting data in the recordings can be obtained only by uploading them to the campaign. The first step is to process them by Time Analysis Extractor. Users can see many interesting results without any settings (cross-talks, detection of silent, speed of speech, reaction times, etc.).

Other technologies are not possible to blindly run and have to be adjusted for particular campaign. There are only three main and simple key steps for analysis settings in the SPAS:

1. **to define keywords / group of keywords** - sources for this list can be different, it depends on the use-case (control of the script / forbidden phrases / negation / client reaction / mention of competition ...)

2. **to create a workflow** = define the conditions under which words from step one have / do not have to be spoken
   - definition nodes of workflow = specification when / where / who / what / whether should be said including timing and logical sequence
   - specification of call categories / metadata for using workflow
   - in this document there is a special section describing more about this step (2.5.2 Workflow management)

3. **to set up jobs and sources** = in this step a user defines which recordings have to be analyzed
   - selection and settings campaign (inbound / outbound, mono / stereo etc.)
   - creating and setting a job (period / which technologies and workflows have to be used)

When the job is saved and run, all workflows settings and groups of keywords are fixed. No other changes to the settings will affect the ongoing analysis. These will be used in newly created jobs.

2.5 Visualization

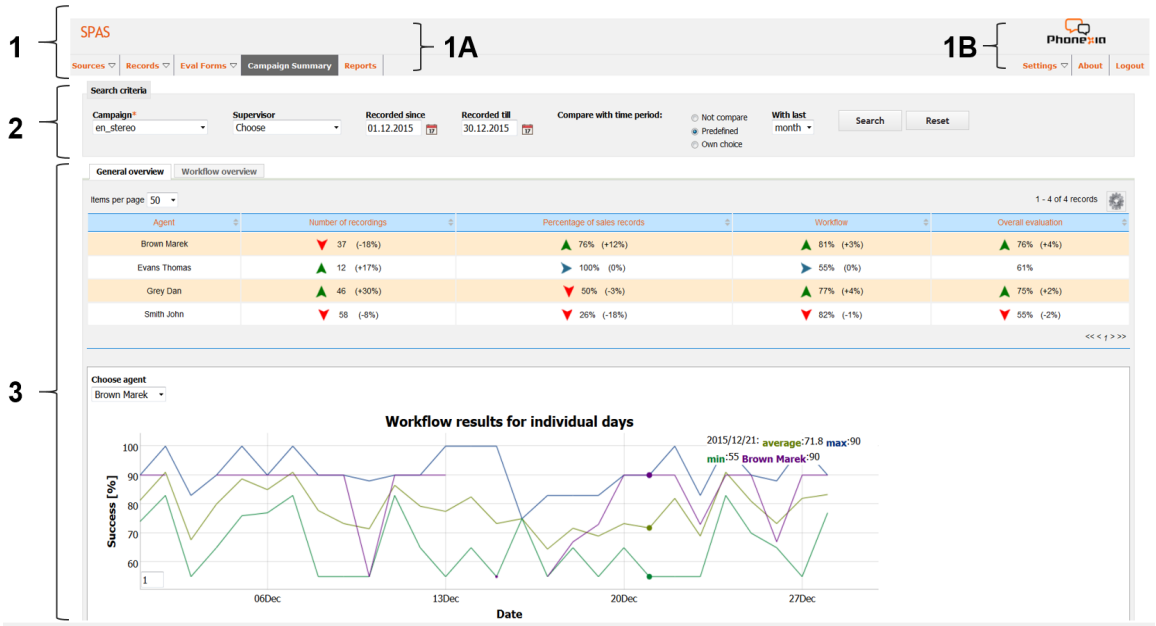
2.5.1 GUI

There are three basic part in GUI of SPAS on many pages:

1. MAIN MENU contains two parts:
   - 1A (see picture) is the layout for analysis
   - 1B (see picture) is the layout application and users

2. FILTERING MENU
   - there are different possibilities for choosing filtering criteria on every page, but often the choice is f Campaign / Period / Agent / Supervisor / Job / Workflow

3. DASHBOARD



campaign summary

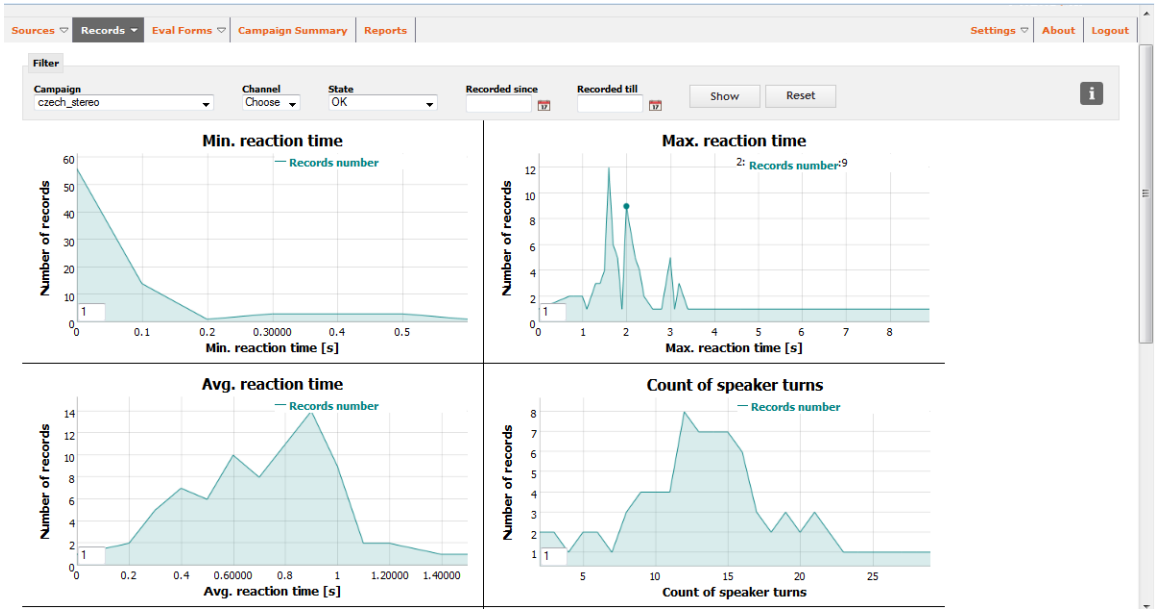
1A: Sources | Records | Eval Forms | Campaign Summary | Reports

is part for analysis (settings + getting results) - user have these elections:

- **Sources** - page presents complete view of record sources that are prepared in your file system to be processed by SPAS and there are also all settings for analyzes (page for defining keywords / workflows / campaigns)
- **Records** - the menu provides mainly filtering of individual records including results, also there are statistics criteria settings for campaign and results for this in graphs - some of the screens:

ID	Recorded	Name	Selling call	Record length [s]	Max. silence length [s]	Speech length [s]	Speech speed [ph/s]	Speaker turn count	Successful nodes	Workflow [%]	Detail	Analytics
10011	04.12.2015 00:00:00	Smith John	Yes	57	9.2	27	12.21	5	[6/9]	61	Detail	Add
10012	12.12.2015 00:00:00	Brown Marek	Yes	79	1.1	46	12.55	8	[7/9]	86	Detail	Add
10014	26.01.2016 00:00:00	Brown Marek	No	61	8.2	27	12.32	6	[5/9]	54	Detail	Add
10015	25.12.2015 00:00:00	Smith John	Yes	79	1.5	50	11.86	8	[8/9]	75	Detail	Add
1001	06.12.2015 00:00:00	Brown Marek	Yes	79	1.1	46	12.55	8	[7/9]	86	Detail	Add
1002	01.12.2015 00:00:00	Smith John	No	57	9.2	27	12.21	5	[6/9]	61	Detail	Add
1005	17.01.2016 00:00:00	Evans Thomas	Yes	75	0.9	39	13.07	6	[6/9]	61	Detail	Add
1006	05.01.2016 00:00:00	Grey Dan	No	72	2.2	39	12.72	5	[9/9]	100 ✔	Detail	Add
1009	17.12.2015 00:00:00	Smith John	No	70	0.1	36	13.04	5	[9/9]	100 ✔	Detail	Add
101010	08.12.2015 00:00:00	Grey Dan	Yes	75	1.3	39	13.47	6	[6/9]	61	Detail	Add

records page



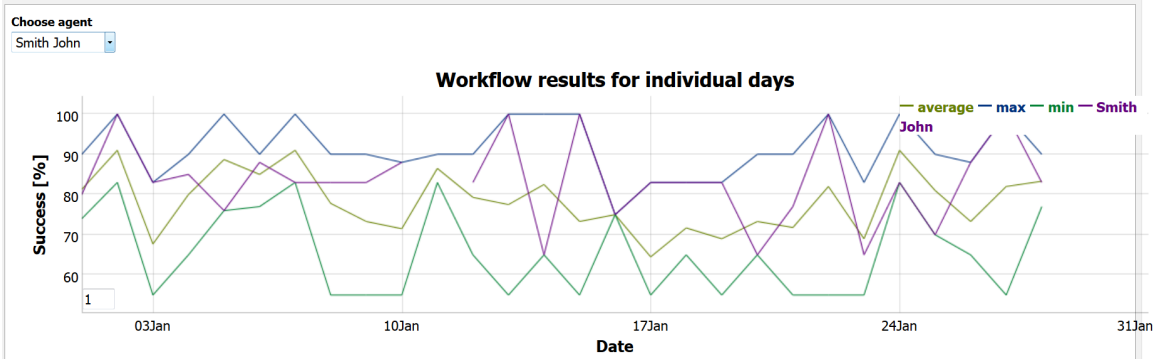
statistics page

- **Eval Forms** - the menu provides system support for manual evaluation: create forms, planners, and sites for the activity of evaluation. In the CC they still use manual evaluation for small part of the controls (e.g. soft skills - proactivity, client access) - they can use this part of application and create forms of the controls and planner = definition of some parameters (which campaign, period, who is response for control). Supervisors only open appropriate page and start evaluation - the system selects the appropriate recordings, displays form for evaluation and audio player with the results of automatic analysis.

- **Campaign Summary** - this page presents summarized results for campaigns and agents from specific period - it can also quickly compare outcomes from two periods and calculates the percentage of improvement / deterioration. On this page, the user can compare results of automatic analysis and manual evaluation (= last two columns).

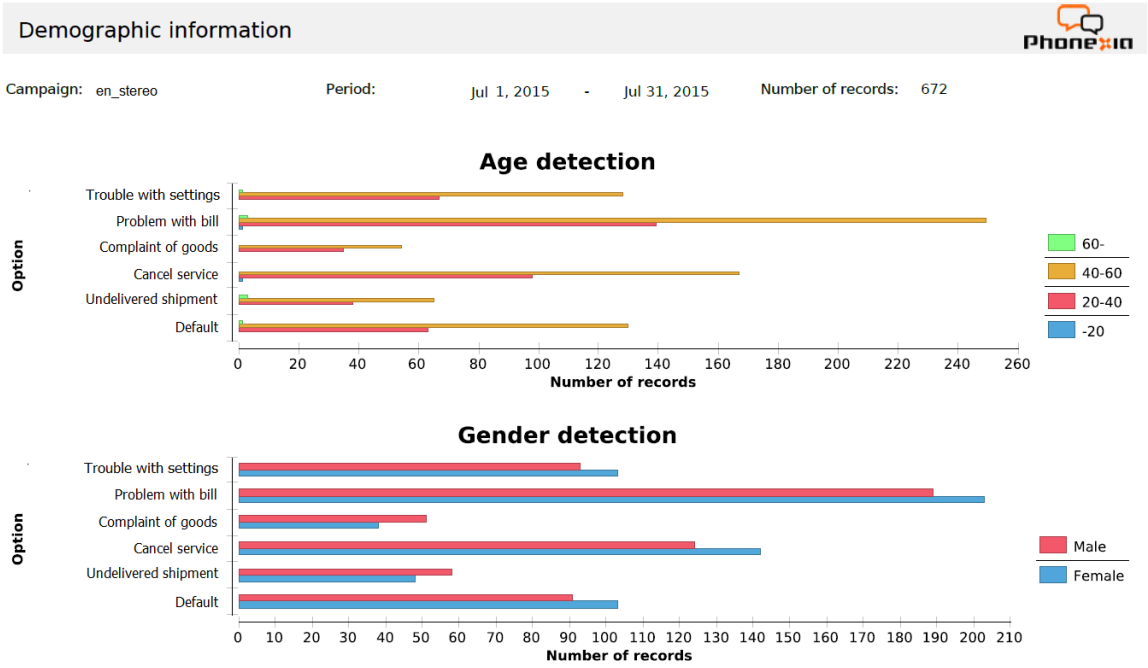
Agent	Number of recordings	Percentage of sales records	Workflow	Overall evaluation
Brown Marek	▼ 37 (-18%)	▲ 76% (+12%)	▲ 81% (+3%)	▲ 76% (+4%)
Evans Thomas	▲ 12 (+17%)	▶ 100% (0%)	▶ 55% (0%)	61%
Grey Dan	▲ 46 (+30%)	▼ 50% (-3%)	▲ 77% (+4%)	▲ 75% (+2%)
Smith John	▼ 58 (-8%)	▼ 26% (-18%)	▼ 82% (-1%)	▼ 55% (-2%)

<< < 1 > >>



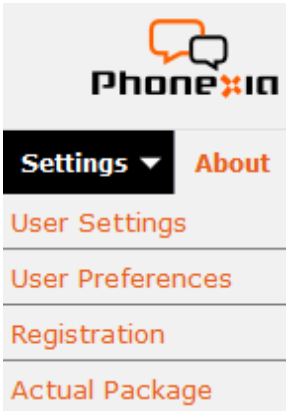
campaign summary - agents' results

, **Reports** - the page contains a filter for report generation and has a section to set the automatic emails with this reports. There are many reports (different views of the results for speech statistic, keyword spotting, manual evaluation, demographic information). The user can set up the frequency of emails together with report type and target email address. An example of such report can be seen on the following picture. It shows the analysis of the calls given the topic and demographic information (age and gender).



example of report

1B: is the section for managing user accounts (options are listed below) and also there is info about application (version, contact for technical support)



Setting of the users and their **roles**. There is also possibility to register a new user, but this is accessible only for quality manager.

- set password, language, email, campaigns, role and change name
- specify prefer options which will be automatically filled in search
- create a new account of user (accessible for QM)
- check actually package (accessible for QM)

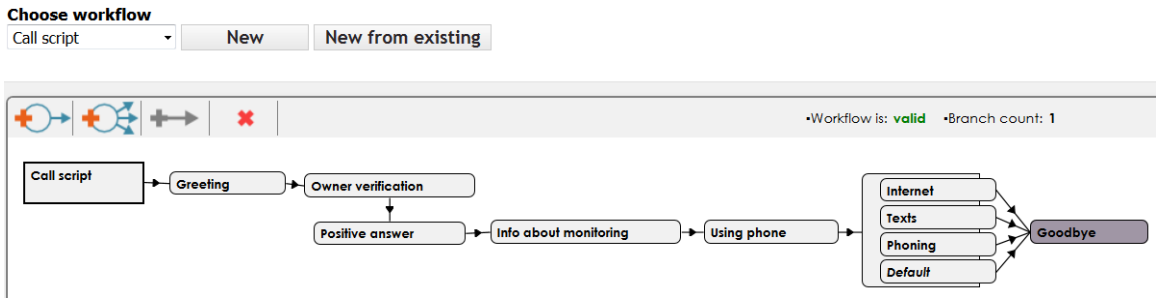
There are two basic roles in SPAS:

- SUPERVISOR (SV) is default set to every user - this role has access to the results (also create reports) and evaluate the recordings
- QUALITY MANAGER (QM) has an access to the entire application - defines all of the analysis settings / manual evaluation / creates new users / rights of supervisor

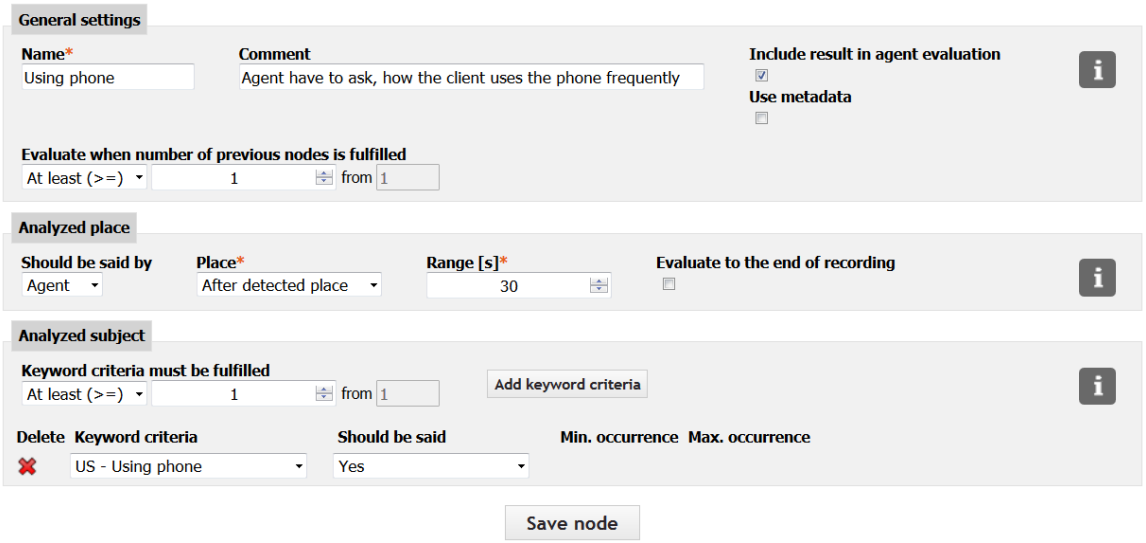
2.5.2 Workflow management

This section extends in more detail section 2.4 with details about settings of workflow management. The core function for it is Keyword spotting. Every workflow consists of **nodes** and **connections**.

node - represents one analysis in the recording (keywords in certain place of recording) - for every node there is specification when / where / who / what / whether should be said including timing and logical sequence.
connection - is oriented connectors between two nodes, which presents logical relation, ie. the node, where connection is entered, depends on node where connection starts

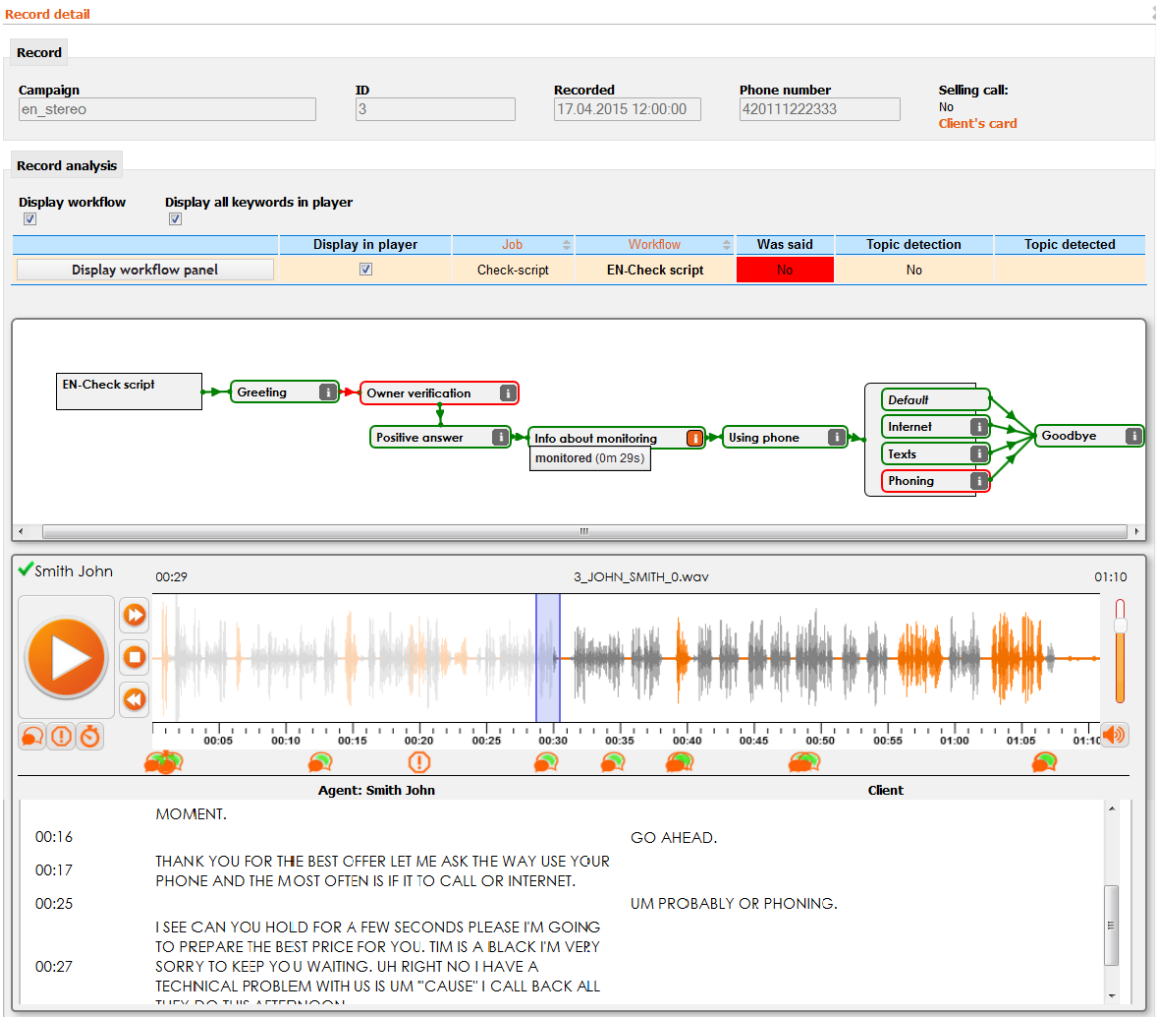


workflow creator



specification for node

For every record it is possible to use many differents workflows. It depends what the user needs to control in the recordings (it can be used for checking of the call categories, control scripts ... - more about it at 2.5 Use-cases)
Record detail page - shows the results of the workflows, and outputs of other technologies like



transcription. Results are divided into two channels for stereo recordings.

records detail page

3 Emotions in SPAS

3.1 Introduction

The objective of Pilot 3 is to utilize acoustic emotion recognition from the MixedEmotions platform. The obtained data will help with rating of recordings in Call Centers. Together with other parameters that are already used (especially cross-talks, speed of speech, speaker turn count, key-word spotting for emotion words) it will help to detect problematic parts in recordings and scripts.

Functionalities, which can help identify emotions in the speech (recordings) and analyze these parts, are important for Call Centers because these are the key moments of unsuccessful / successful calls. Supervisors / quality managers could use the information to improve results and increase success (weaker agents in this area; verbal expressions and phrases that cause these emotions; also the phrases which are useful; in a real-time, during a call, warning for agents which will help them handle the situation ...).

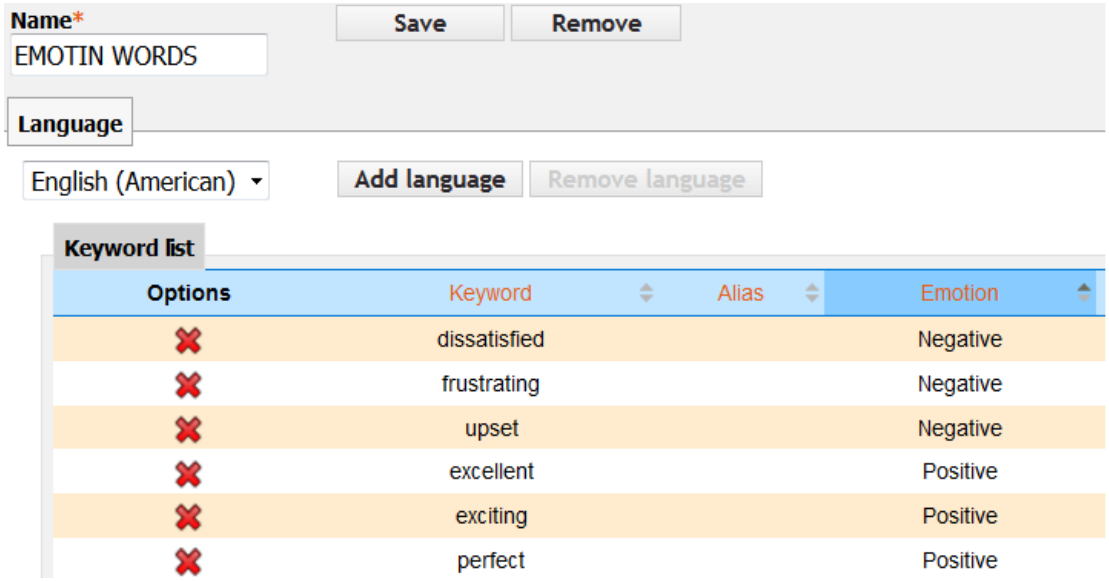
Two new functionalities have been introduced in SPAS by the MixedEmotions project. First is emotion recognition using pre-identified emotion specific keywords, and second is emotion recognition from audio signal.

These new functionalities will define three types of emotions in the speech: positive, neutral and negative and mark them in a call. This results can be used by different ways – for analysis and statistics, training communication skills, inform supervisors about problem in a speech in a real-time etc .

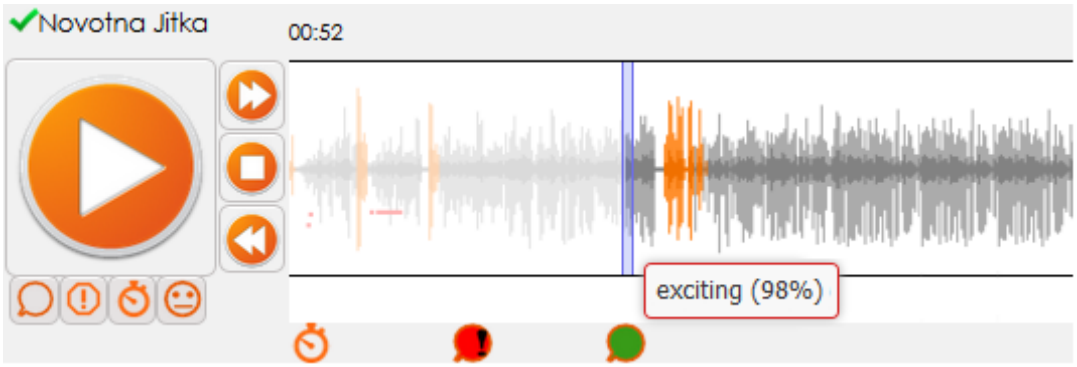
3.2 Visualization

3.2.1 Keywords

The application now let to set the emotional index for each keyword. If the system detects the keyword it show it with corresponding color (red is negative / green for positive) in the player.



keyword definition



SPAS player

3.2.2 Acoustic emotion recognition

Next visualisation is based on the module which recognize emotions from audio signal. We use two continuous output between -1 to +1 which correspond to negative and positive emotion respectively. User can define thresholds for the positive and negative emotions (for example positive emotion is if the score is greater than 0.3). These thresholds may be different for individual campaigns and emotions.

Criterion	Unit	Description	Minimum	Maximum	Weight [%]
Min. reaction time	[s]	Minimal reaction time for all call			
Max. reaction time	[s]	Maximal reaction time for all call	0	2	
Avg. reaction time	[s]	Average reaction time for all call			
Record length	[s]	Length of record			
Max. silence length	[s]	Max. silence length during all call	0	4	100
Silnce length at the end of record	[s]	Silence length at the end of record			
Speech length	[s]	Length of speech			
Speech speed	[phoneme/s]	Speed of speech			
Speaker turns count	[count]	Number of speaker turns per call			
Cross-talks count	[count]	Number of cross-talks per call			
Positive emotion threshold	[%]	Positive emotion threshold setting	20	100	0
Negative emotion threshold	[%]	Negative emotion threshold setting	50	100	0

statistics criteria

Next figure shows the player which has a button for displaying the emotion recognition output. If there is an emotion, it is shown as the curve at particular time. Red curve for negative emotions, and green one for positive. It also displays the confidence of the system - higher curve represents more positive, while lower means more negative.

SPAS player

Other places where user can check emotions results - columns in the scoreboards for recording / agent:

Agent	Number of recordings	Percentage of sales records	Workflow	Overall evaluation	Pozitive emotion	Negative emotion
Brown Marek	2 (0%)	75% (-25%)	85% (-8%)	N/A	0 (0%)	0 (0%)
Evans Thomas	1	100%	53%	N/A	0	0
Grey Dan	3 (+67%)	44% (-56%)	64% (+4%)	N/A	0 (0%)	8 (-4%)
Smith John	4 (+75%)	37% (+37%)	88% (+8%)	N/A	0 (0%)	6 (-6%)

Name	Record length [s]	Crosstalk count	Max. silence length [s]	Successful nodes	Workflow [%]	Pozitive emotion	Negative emotion	Detail
Modra Hana	338	0	2.7	[6/10]	56	0	21	Detail
Novak Tomas	104	0	2.6	[2/10]	13	0	6	Detail
Cerna Jitka	223	0	2.4	[8/10]	88	1	2	Detail
Modra Hana	51	0	0.9	[6/10]	56	0	3	Detail
Cerna Aneta	167	0	5.6	[7/10]	44	0	2	Detail
Sefr Richard	200	0	2.2	[8/10]	69	0	0	Detail
Sefr Richard	257	0	1.8	[8/10]	69	0	17	Detail

campaign summary - agents' results records page

3.3 Emotion recognition module

This section describes the evolution of the "production" modules. The main idea was to build the language independent emotion recognizer which would be easier to deploy for any language. But results are in favor for the language dependent variant. In this regard, Phonexia and University of Passau built 3 systems for the integration.

3.3.1 Module v.1

This module works on the utterance level. It outputs a score for emotion at the end of the utterance. It is trained on several databases in different languages.

Datasets:

We have used four emotional speech databases in different languages (German [1], English [2], Roman [3], and Italian [4]). Some specifications of these datasets are given in table below:

Corpora information and the mapping of class labels onto Negative/Positive valence. (#m): number of male speaker, (#f): number of female speakers, (Rate): Sampling rate. Labels are: (A)nger, (Sa)dness, (F)ear, (D)isgust, (B)oredome, (N)eutral, (H)appy, (Su)rprise, (J)oy

Corpus	Language	#m	#f	Rate	Negative Valence (#)		Positive Valence (#)	
EMODB	German	5	5	16	A,Sa,F,D,B	(385)	N,H	(150)
SAVEE	English	4	0	44	A,Sa,F,D	(240)	N,H,Su	(240)
EMOVO	Italian	3	3	44	A,Sa,F,D	(336)	N,J,Su	(252)

Polish	Polish	4	4	44	A,Sa,F,B	(160)	N,J	(80)

Feature Extraction:
From each utterance we have extracted 6,373 features using openSMILE [5]. This feature set has been used in Interspeech 2013 Computational Paralinguistics challenge [6].
The ComParE feature set contains functionals of acoustic low-level descriptors (LLDs). The LLDs include prosodic features (signal energy, perceptual loudness, fundamental frequency), voice quality features (jitter and shimmer of the fundamental frequency, voicing probability, and the harmonics-to-noise ratio), spectral features (spectrum statistics such as variance and entropy and energies in relevant frequency bands), and cepstral features (Mel-Frequency cepstral coefficients – MFCC). From these LLDs, the first order delta coefficients are computed and both LLDs and delta coefficients are smoothed with a 3 tap moving average filter over time. Then, functionals are applied to the LLDs and their delta coefficients over a complete speech segment. The functionals include statistical measures such as moments (means, variances, etc.), statistics of peaks (mean amplitude of peaks, mean distance between peaks, etc.), distribution statistics such as percentiles (especially quartiles and inter-quartile ranges), regression coefficients obtained by approximating the LLD over time as linear or quadratic function and the errors between the approximation and the actual LLD, temporal characteristics such as positions of maxima and the percentage of values above a certain threshold, and modulation characteristics expressed as linear predictor (autoregressive) coefficients of a predictor of five frames length.

Feature Selection and classification:
A random forest with 30 trees has been used to rank the features. We selected 250 top ranked features and trained a Support Vector Machine (SVM) classifier with linear kernel. This classifier is trained on the concatenation of the four datasets. A 10-fold cross-validation yields
about 80% of emotion recognition accuracy.

3.3.2 Module v.2

This version of the module outputs the score for arousal and valence for each frame - every 40ms. It is trained on several databases in different languages.

Datasets:
Emotional databases in different languages and environments with both acted and spontaneous emotions were used. Emotional labels were mapped into 2-dimensional model: arousal (i. e. passive vs. active) and valence (i. e. negative vs. positive) [7]. Final emotion is represented as point with real coordinates <-1, 1>. Details of each database are listed in the table below:

Distribution of arousal and valence values in each database, its language and whether emotions are acted or spontaneous.

name	language	duration [h:mm:ss]	arousal			valence		
			low	medium	high	negative	neutral	positive
ABC	GER	1:14:51	0:03:35	0:40:25	0:30:51	0:18:51	0:40:25	0:15:34
AVEC2015	FRA	1:30:00	0:43:19	0:00:00	0:46:41	0:18:09	0:00:00	1:11:51
DES	DAN	0:17:16	0:03:43	0:03:27	0:10:05	0:06:56	0:03:27	0:06:53
EMODB	GER	0:24:47	0:06:46	0:06:51	0:11:10	0:14:55	0:06:51	0:03:01
EMOVO	ITA	0:33:16	0:05:13	0:04:22	0:23:41	0:19:42	0:04:22	0:09:13
eNTERFACE	ENG	1:07:04	0:12:02	0:00:00	0:55:02	0:46:43	0:00:00	0:20:21
Polish	POL	0:09:14	0:01:38	0:03:16	0:04:20	0:04:32	0:03:16	0:01:25
SAVEE	ENG	0:30:43	0:04:29	0:07:13	0:19:01	0:15:54	0:07:13	0:07:36
SRoL	RUM	0:14:02	0:03:20	0:02:10	0:08:33	0:07:14	0:02:10	0:04:39
VAM	GER	0:47:48	0:24:14	0:02:25	0:21:09	0:44:40	0:02:25	0:00:43
All		**6:49:02**	**1:48:18**	**1:10:10**	**3:50:34**	**3:17:37**	**1:10:10**	**2:21:16**

Feature Extraction, Selection and classification:

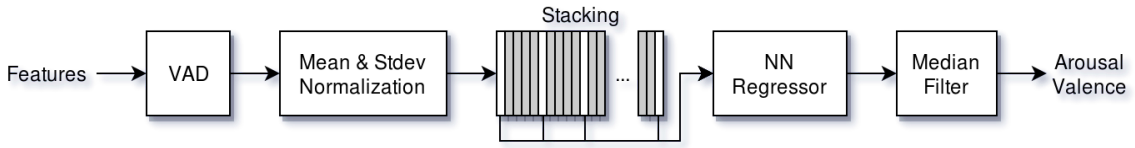
Two sets of features were used: eGeMAPS and BN features. The former, extended version of Geneva Minimalistic Acoustic Parameter Set was used as baseline feature set for AVEC 2015 emotion recognition challenge [11]. These features were obtained using openSMILE toolkit [12]. First, 42 low-level acoustic descriptors (LLDs) containing spectrum, energy and voicing related information are extracted from short speech frames. Second, functionals over 3 s segments with 40 ms shift are computed. Functionals consist of mean, variance, percentiles, mean and standard deviation of slope and temporal features, leading to total vector size of 102.

The latter, Stacked Bottleneck Features were originally used for language identification [13], obtained using two neural network architecture. First, 24 MFCCs and 2 F0 related features are concatenated into 26-dimensional vector. Second, 11 frames are stacked and Hamming window followed by DCT are applied, leading to 156-dimensional input feature vector. Configuration of both NNs is 156x500x500x80. Output from the first NN is stacked and forms context-dependent 400-dimensional input feature vector for the second NN. Output from this NN are the final 80 BN features used in our emotion recognition system.

System:

A single system is trained for each dimension (arousal and valence) separately and its scheme is shown in figure. After feature extraction, voice activity detection (VAD) and global mean and

standard deviation normalization are applied. 41 frames are stacked, downsampled (every 5th frame is taken) and used as input for NN with one hidden layer trained as regressor. Output of the NN is post-processed with median filter of size 71 and every 40ms gives a real value from -1 to 1 indicating either arousal or valence of current emotion.



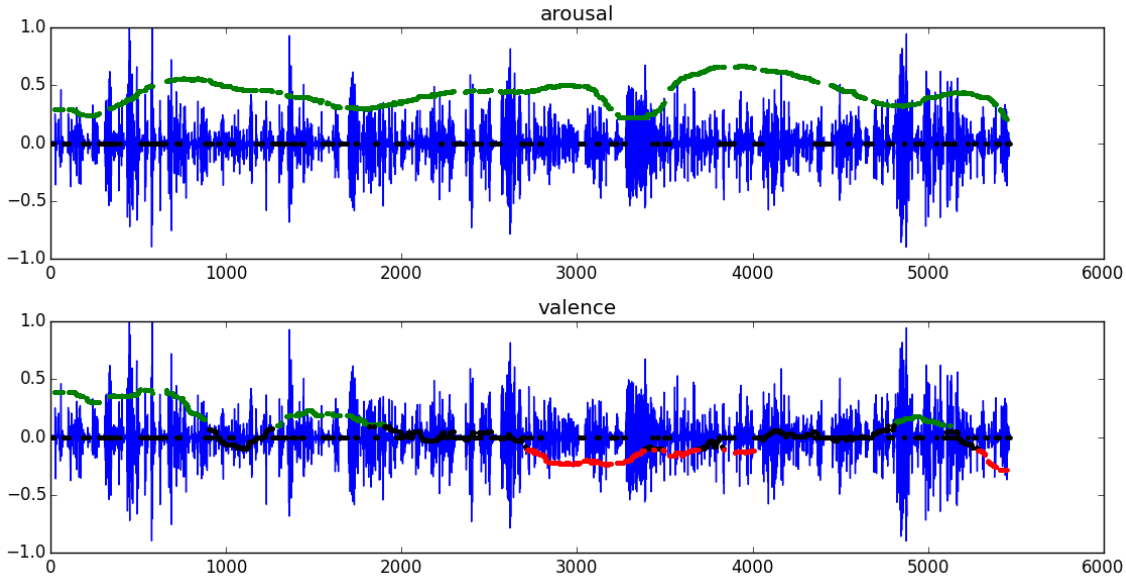
3.3.3 Module v.3

This version of the system focuses on Czech real Call Center data and therefore it was trained only on target data, described in the table below. The data was manually segmented on the speaker level and annotated on emotion level. The system scheme is identical to the one in section 3.3.2 with minor change in configuration. The number of stacked frames has increased to 71 and median filter size to 121. Outputs for both arousal and valence are shown in Figure.

Distribution of arousal and valence values in used Czech Call Center data.

name	duration [h:mm:ss]	arousal			valence		
		low	medium	high	negative	neutral	positive
Call Center1	2:09:16	0:05:42	1:18:42	0:44:53	0:25:49	1:18:42	0:24:45
Call Center2	1:21:41	0:07:10	0:39:33	0:34:58	0:33:13	0:39:33	0:08:55
All	**3:30:57**	**0:12:51**	**1:58:15**	**1:19:51**	**0:59:02**	**1:58:15**	**0:33:40**

Output of the system for one record for both arousal (top) and valence (bottom). Horizontal axis represents sequence of frames with step 40ms. Values in green and red color represent positive and negative emotion respectively. Thresholds were experimentally set to ±0.1.



3.4 Results

This section shows the comparison of the results between the systems. Table below compares and shows the improvement on several databases between multilingual system v1 versus v2.

Comparison of unweighted average recall (UAR) of systems v1 and v2 for both arousal and valence, evaluated on utterance level. Leave-one-database-out cross validation was used in case of system v2.

arousal

test db	system	
	v1	v2
DES	**71.00%**	54.49%
SRoL	69.00%	**69.58%**
VAM	70.00%	**73.35%**

valence

test db	system	
	v1	v2
DES	50.00%	**53.37%**
SRoL	**55.00%**	50.99%
VAM	50.00%	**65.80%**

Then we show the results of multilingual system v2 and monolingual v3 on target data from Czech Call Center in next Table. More detailed analysis will be in the deliverable from Workpackage 4. The results on the target data are the best for the monolingual system, but our goal is still to build the good working multilingual system, because of the deployment for several languages.

Comparison of utterance level unweighted average recall (UAR) of systems v2 and v3 for both arousal and valence.

arousal

test db	system	
	v2	v3
Call Center1	60.44%	**70.07%**
Call Center2	61.63%	**66.49%**

valence

test db	system	
	v2	v3
Call Center1	53.61%	**72.27%**
Call Center2	50.00%	**61.19%**

4 Conclusion

SPAS solution is designed to help Call Centers to monitor their traffic, agents and also clients.
This document describes basic operation of SPAS and implementation of emotion recognition from audio signal. Two new functionalities for emotion recognition were added to SPAS. The first one is based on pure audio signal and the second one is based on detecting keywords which have assigned positive or negative emotion label. The supervisor of the Call Center is able to see the ranking of the "suspicious" calls and listen only the bad ones. Emotion recognition greatly helps to detect such calls.
SPAS generates also reports for a given period of time and there is possible to see the different statistics for the agents and its trend in time.
Future work will focus on improving accuracy of acoustic emotion recognition module, better visualization of emotions; mainly on visualization in reports. Next we will focus also on the speech to text engine and analysis of the output given the detected emotions which helps in defining appropriate keywords for detecting emotions states or words preceding or leading to emotional state of the client.

5 REFERENCES

[1] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech,"in Proc. of Interspeech, 2005, vol. 5, pp. 1517–1520.
[2] S. Haq, P. JB. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in Proc. of the International Conference on Auditory-Visual Speech Processing, Tangalooma, Australia, 2008, pp. 185–190.
[3] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "Emovo corpus: an italian emotional speech database," in Proc. of the 9th International Conference on Language Resources and Evaluation, Iceland, 2014, pp. 3501–3504.

[4] P. Staroniewicz and W. Majewski, "Polish emotional speech database — recording and preliminary validation," in Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions, vol. 5641 of Lecture Notes in Computer Science, pp. 42–49. Springer Berlin Heidelberg, 2009.
[5] F. Eyben, F.Weninger, F. Groß, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In Proc. of ACM MM, pages 835–838, Barcelona, Spain, 2013.
[6] Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., … Kim, S. (2013). The INTERSPEECH 2013 Computational Paralinguistics Challenge : Social Signals, Conflict, Emotion, Autism. Interspeech2013, 148–152.
[7] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, Merano, 2009, pp. 552-557.
[8] S. E. Shepston, Z.-H. Tan, and S. H. Jensen, "Audio-based age and gender identification to enhance the recommendation of tv content," IEEE Transactions on Consumer Electronics, vol. 59, pp. 721–729, 2013.
[9] M. Feld, F. Burkhardt, and C. Muller, "Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.
[10] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," Computer Speech and Language, vol. 27, pp. 151–167, 2013.
[11] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, … M. Pantic, "Av Ec 2015," *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, 2015.
[12] F. Eyben, F. Weninger, F. Groß, and B. Schuller. "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," In Proc. of ACM MM, pp. 835–838, Barcelona, Spain, 2013.
[13] P. Matějka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, and B. Zhang, "Neural network bottleneck features for language identification," in Proceedings of Odyssey 2014, pp. 299–304, 2014.