Could Speaker, Gender or Age Awareness be beneficial in Speech-based Emotion Recognition?

Maxim Sidorov[1], **Alexander Schmitt**[1], **Eugene Semenkin**[2] **and Wolfgang Minker**[1]
[1]Ulm University, [2]Siberian State Aerospace University
Ulm Germany, Krasnoyarsk Russia
maxim.sidorov@uni-ulm.de, alexander.schmitt@uni-ulm.de, eugene.semenkin@sibsau.ru, wolfgang.minker@uni-ulm.de

Abstract

Emotion Recognition (ER) is an important part of dialogue analysis which can be used in order to improve the quality of Spoken Dialogue Systems (SDSs). The emotional hypothesis of the current response of an end-user might be utilised by the dialogue manager component in order to change the SDS strategy which could result in a quality enhancement. In this study additional speaker-related information is used to improve the performance of the speech-based ER process. The analysed information is the speaker identity, gender and age of a user. Two schemes are described here, namely, using additional information as an independent variable within the feature vector and creating separate emotional models for each speaker, gender or age-cluster independently. The performances of the proposed approaches were compared against the baseline ER system, where no additional information has been used, on a number of emotional speech corpora of German, English, Japanese and Russian. The study revealed that for some of the corpora the proposed approach significantly outperforms the baseline methods with a relative difference of up to 11.9%.

Keywords: Adaptive speech-based emotion recognition, classification performances, corpora evaluation.

1. Introduction

By deploying the ER component within the SDS, its quality could be significantly increased. It might be beneficial during human-robot or even human-human interaction. Whereas the majority of studies concentrate on speaker-independent ER experiments, in some cases speaker-awareness can bring an additional advantage. Despite the fact that the basic emotions are shared between cultures and nationalities (Scherer, 2002) obviously, each person expresses his emotions individually. This thesis lies behind the idea of building different emotional models for each speaker independently or incorporating the speaker-specific information within the single ER model in a different way. On the one hand it results in a problem-decomposition, similar to the cluster-then-classify approach, but on the other hand by deploying different models for each speaker, the individual features of the corresponding speaker can be caught and utilised properly.

Furthermore, as has been mentioned in many studies (Brody, 1985), (Hall et al., 2000) the gender difference in emotional expression has been detected during several psychological investigations. In contrast to the very specific nature of speaker-adaptive ER, gender-adaptive ER might be more general. A similar idea is behind the age-adaptive ER models, where each user has one of the age-specific labels (for example *youth* or *adult*).

The global aim of the study is to figure out whether the speaker-, gender- or even age-related information of an end-user might be utilised in order to improve the quality of the ER models. We proposed here a two-stage approach, where firstly the speaker or other additional information (gender or age) is identified and secondly, an adaptive ER procedure is performed. We intend to study both cases: the theoretically possible improvement when the known speaker-related information is taken into account, and the actual difference, which can be observed by deploying speaker-state recognition models, i.e. Speaker Identification (SI), Gender (GR), or Age Recognition (AR). Thus, in the first case we took the ground-truth information about the speaker, gender and age (the *G* experiments, for *Ground truth*), whereas in the second series of experiments we deployed the actual SI, GR, and AR models to estimate the corresponding hypothesis (the *E* experiments, for *Estimated*).

Since the emotions themselves have a subjective nature and generally may vary depending on what language one speaks, we carried out the experiments based on 8 different emotional corpora of English, German, Russian and Japanese in order to gain generalizability of the results obtained.

The rest of the paper is organised as follows: Significant related work is presented in 2. Section, whereas 3. Section describes the applied corpora and outlines their differences. Our approach to incorporating the additional speaker-related information within the ER system is presented in 4. Section. The results of numerical experiments are demonstrated in 5. Section. Finally, the conclusion and future work are described in 6. Section.

2. Significant related work

The authors in (Lopez-Otero et al., 2015) researched dependencies between speaker-dependent and -independent approaches when the depression level of the speaker is under examination. It has been concluded that the system performance is much better when the test speaker is in both the training and testing sets. Intuitively, the results could be extrapolated in the case of other speaker traits such as emotions, in a similar way to how it was implemented in the case of the speaker identification approach (Kockmann et al., 2011).

The authors in (Vogt and André, 2006) improved the performance of emotion classification by automatic gender detection. The authors have used two different classifiers in

Database	Language	# speakers	Paralinguistic Labels	Naturalness
AVEC-2014	German	58	Happy-exciting, angry-anxious, sad-bored, relaxed-serene	Non-acted
Emo-DB	German	10	Anger, boredom, disgust, anxiety/fear, happiness, sadness, neutral	Acted
RadioS	German	69	Neutral, happy, sad, angry	Non-acted
VAM	German	47	Happy-exciting, angry-anxious, sad-bored, relaxed-serene	Non-acted
LEGO	English	279	Anger, neutral, non-speech	Non-acted
SAVEE	English	4	Anger, disgust, fear, happiness, sadness, surprise, neutral	Acted
Ruslana	Russian	61	Neutral, surprise, happiness, anger,sadness, fear	Acted
UUDB	Japanese	14	Happy-exciting, angry-anxious, sad-bored, relaxed-serene	Non-acted

Table 1: Databases description.

order to classify male and female voices from the Emo-DB (Burkhardt et al., 2005) and the SmartKom (Steininger et al., 2002) corpora. They concluded that the combined gender and emotion recognition system improved the recognition rate of a gender-independent emotion recognition system by 2–4% relatively by applying the Naive Bayes classifier for building the emotion models.

3. Corpora description

All evaluations were conducted using several audio emotional databases. Here is a brief description of them and their statistical characteristics.

The **AVEC-2014** database was used for the fourth Audio-Visual Emotion Challenge and Workshop 2014 (Valstar et al., 2014). This corpus is a subset of the AVEC'13 database (Valstar et al., 2013) consisting of 150 videos. Only two tasks in a human-computer interaction scenario have been selected to be included in the dataset. During the *Northwind* scheme participants read aloud an extract of the story 'The North Wind and the Sun' in German. The *Freeform* task is participants' answers to several general questions such as 'What was your best gift, and why?', again in German. Each affect dimension (Arousal, Dominance, and Valence) has been annotated separately by a minimum of three and a maximum of five human raters. We averaged the valence and arousal values over the whole recording's duration to obtain only one pair of continuous labels.

The **Emo-DB** emotional database (Burkhardt et al., 2005) was recorded at the Technical University of Berlin and consists of labelled emotional German utterances which were spoken by 10 actors (5 females). 10 German sentences of non-emotional content have been acted by professional actors so that every utterance has one of the following emotional labels: anger, boredom, disgust, anxiety/fear, happiness, sadness and neutral. The total number of utterances in the corpus is 535.

The **RadioS** database consists of recordings from a popular German radio talk-show. Within this corpus, 69 native German speakers talked about their personal troubles. The labelling was performed by a human rater so that each utterance has one of the following emotional labels: neutral, happy, sad and angry.

The **VAM** (Grimm et al., 2008) dataset was created at Karlsruhe University and consists of utterances extracted from the popular German talk-show 'Vera am Mittag' (Vera in the afternoon). For this database 12 broadcasts of the talk-show have been recorded. Each broadcast consists of several dialogues of between two and five people each. Continuous emotional labels have been set by evaluators using the valence, activation and dominance basis.

The **LEGO** emotional database (Schmitt et al., 2012) comprises non-acted English (American) utterances which were extracted from the SDS-based bus-stop navigational system (Eskenazi et al., 2008). The utterances are requests to the system spoken by real users with real concern. Each utterance has one of the following emotional labels: anger, slight anger, much anger, neutral, friendliness and non-speech - critical noisy recordings or just silence. The corpus was manually annotated by a human rater who chooses one of the labels. We combined all the utterances with different anger levels into a single class with anger labels. Moreover, since there are very few friendly recordings we removed them from the database. As a result we operated only with recordings of 3 labels, namely anger, neutral and non-speech.

The **SAVEE** (Surrey Audio-Visual Expressed Emotion) corpus (Haq and Jackson, 2010) was recorded as a part of research into the field of audio-visual emotion classification, from four native English male speakers aged from 27 to 31. The emotional label for each utterance is one of the standard set of emotions (anger, disgust, fear, happiness, sadness, surprise and neutral).

The corpus of Russian emotional speech (Makarova and Petrushin, 2002) **Ruslana** includes records of utterances from 61 subjects (49 females). Each native Russian speaker (aged from 16 to 28 with the average equalling 18.7) read aloud 10 sentences of different content conveying the following six emotional states: neutral, surprise, happiness, anger, sadness and fear. Altogether the database contains 3,660 emotional utterances (61 speakers x 10 sentences x 6 emotional primitives).

The **UUDB** (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database

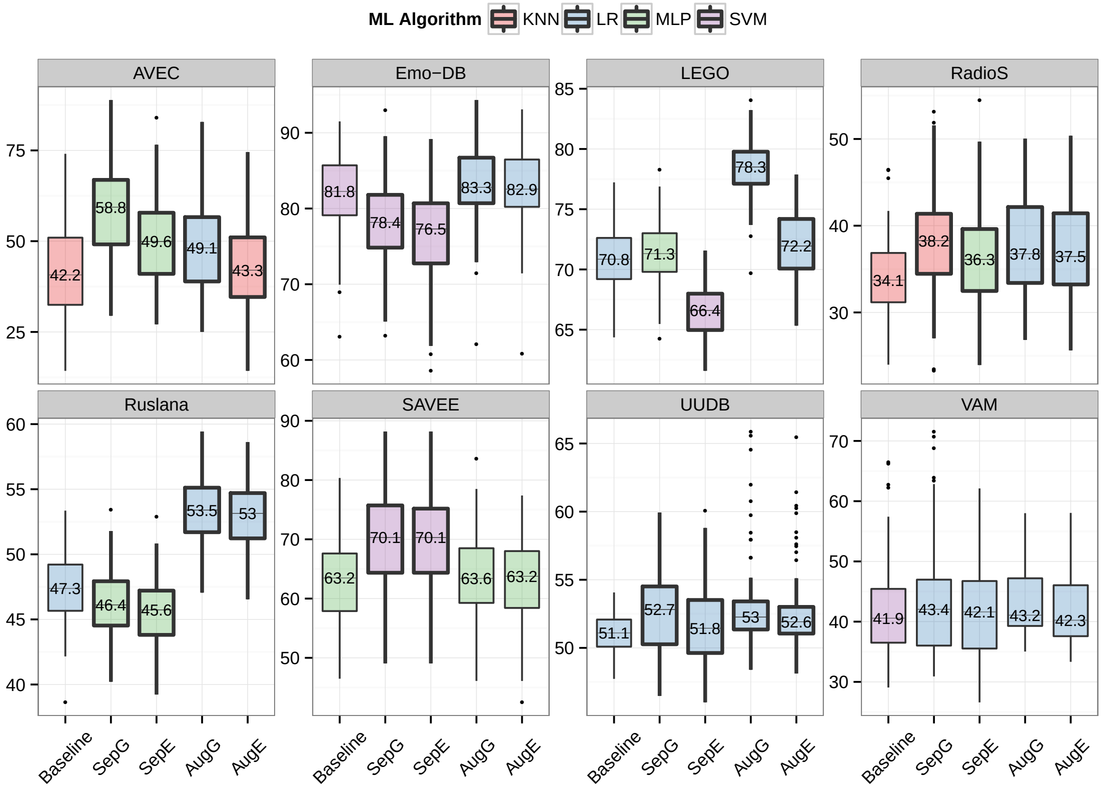

Figure 1: F_1 measure of speech-based emotion recognition with ground-truth speaker-related information (*AugG* and *SepG*), with the estimated speaker-related hypothesis (*AugE* and *SepE*), and without any additional information (baseline). Colours show the optimal classifiers. All the experiments are 10 repetitions of 10-fold cross-validation emotion-stratified. Box-plots with bold frames indicate T-test-based significant differences against the baseline results (at least with $p = 0.05$).

(Mori et al., 2011) consists of spontaneous Japanese speech through task-oriented dialogue which was produced by 7 pairs of speakers (12 females), 4,737 utterances in total. Emotional labels for each utterance were created by 3 annotators on a 5-dimensional emotional basis: interest (interested–indifferent), credibility (credible–doubtful), dominance (dominant–submissive), arousal (aroused–sleepy) and pleasantness (pleasant–unpleasant). The human raters evaluated the perceived emotional state of the speakers for each utterance on a 7-point scale. Thus, on the pleasantness scale, 1 corresponds to extremely unpleasant, 4 to neutral, and 7 to extremely pleasant.

Since a classification task is under consideration, we have used just pleasantness (a synonym for evaluation) and arousal axes from the AVEC-2014, VAM, and UUDB corpora. The corresponding quadrant (anticlockwise, starting in the positive quadrant, assuming arousal as abscissa) can also be assigned emotional labels: happy-exciting, angry-anxious, sad-bored and relaxed-serene (Schuller et al., 2009b).

There is a description of the used corpora in Table 1.

4. The two-stage adaptive emotion recognition

Incorporating speaker-specific information into the emotion recognition process may be done in several ways. A very straightforward way is to add this information to the set of features as an additional variable; we will refer to this approach as System *Aug* for *augmented feature vector* (Sidorov et al., 2014a). Another way is to create speaker-dependent models: While, for conventional emotion recognition, one statistical model is created independently of the speaker, one may create a separate emotion model for each speaker, we will refer to this approach as System *Sep* for *separate model* (Sidorov et al., 2014b). Both approaches result in a two-stage recognition procedure: First, the speaker is identified and then this information is included into the feature set directly (for the System *Aug*), or the corresponding emotion model is used for estimating the emotions (for the System *Sep*). Both emotion recognition-speaker identification hybrid systems have been investigated and evaluated in this study.

To investigate the theoretical improvement of using speaker-specific information for ER, the ground truth information about the speaker has been used (*AugG* and *SepG* approaches). Then, in order to perform experiments in real-world conditions, an actual speaker identification component has been applied (*AugE* and *SepE* systems). We used a number of classification algorithms, namely k-Nearest Neighbours (KNN) (Cover and Hart, 1967), Multi-Layer Perceptron (MLP), Support Vector Machine (SVM) (Xuegong, 2000) trained by the sequential minimal optimisation

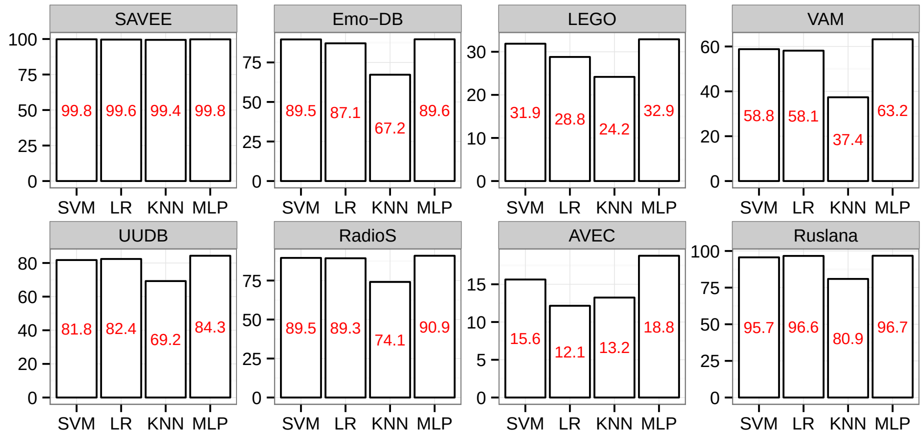

Figure 2: F_1 measure of speech-based speaker recognition. All the experiments are 10 repetitions of 10-fold cross-validation emotion-stratified. Values within the graphs are average F_1 measures.

algorithm (Platt and others, 1999), and boosted Logistic Regression (LR) (Menard, 2002), in order to provide statistically reliable and algorithm-independent results.

In the first experiment, the focus was on investigating the theoretical improvement, which may be achieved using speaker-based adaptiveness. For this, known speaker information (true labels) was used for both approaches. In System *Aug*, the speaker information was simply added to the feature vector as an additional variable. Hence, all utterances with the corresponding speaker information were used to create and evaluate an emotion model through the augmented feature vector. For the System *Sep*, individual emotion models were built for each speaker. During the training phase all speaker utterances were used for creating the emotion models. During testing, all speaker utterances were evaluated with the corresponding emotion model, based on known speaker-related information.

Additionally, a second experiment was conducted including an actual speaker identification module instead of using known speaker information. First, a speaker identifier was created during the training phase. Furthermore, for System *Aug*, the known speaker information was included into the feature vector for the training of the emotion classifier. The testing phase starts with the SI procedure. Then, the speaker hypothesis was included into the feature set which was in turn fed into the emotion recogniser. For System *Sep*, an emotion recogniser was created for each speaker separately. For testing, the speaker hypothesis of the speaker recognition is used to select the emotion model which corresponds to the recognised speaker to create an emotion hypothesis. In contrast to the first experiment, these experiments are not free of speaker identification errors. Therefore, relatively worse results were expected here.

It should be noted that similar experiments have been performed in the case of gender- and age-adaptive studies, where instead of using speaker ID directly (*AugG* and *SepG* experiments) and the speaker-identification procedure (*AugE* and *SepE* experiments), both gender- and age-related information, as well as gender- and age-recognition systems have been used correspondingly.

Within the *Sep* systems, one may perform normalisation only once for the whole set of utterances, or speaker-wise (similarly for gender and age groups). We used Z-transformation, as it was found to perform best for the problem of ER previously (Zhang et al., 2011), using both strategies described.

Regarding the *Aug* system, one may consider an augmented feature vector with speaker ID as a unique integer or as a dummy variable (one-hot encoding). When the dummy coding is applied, for all values of the speaker ID attribute a new attribute is created. Next, in every utterance, the new attribute which corresponds to the actual nominal value of the example gets the value 1 and all other new attributes get the value 0. It means that each utterance gets N additional binary variables where N is equal to the number of speakers in the training set, where all the values except for a single one are equal to 0. In such cases when an utterance of an unknown speaker is in the testing set (which could be a case when the number of utterances of this particular speaker is not high enough, provided random emotion-stratified cross-validation splitting) all new attributes are set to 0. Another aspect is whether these additional speaker-related attributes (either unique integer or dummy variable) should be normalised.

5. Numerical evaluations

As a baseline for acoustic features we consider the 384-dimensional feature vector which was used within the InterSpeech 2009 Emotion Challenge (Schuller et al., 2009a), (Eyben et al., 2010).

We used 10 repetitions of the 10-fold Cross-Validation (CV) emotion-stratified experiment and F_1 measure as a main performance metric. We deployed four machine learning algorithms of different nature to avoid algorithm-dependent results.

Thus, in the case of speaker identity for the system *Sep*, we performed 4 (classification algorithms) x 8 (corpora) x 2 (known speaker-related information - *SepG* vs. estimated one - *SepE*) x 2 (normalisation once vs. speaker-wise) =

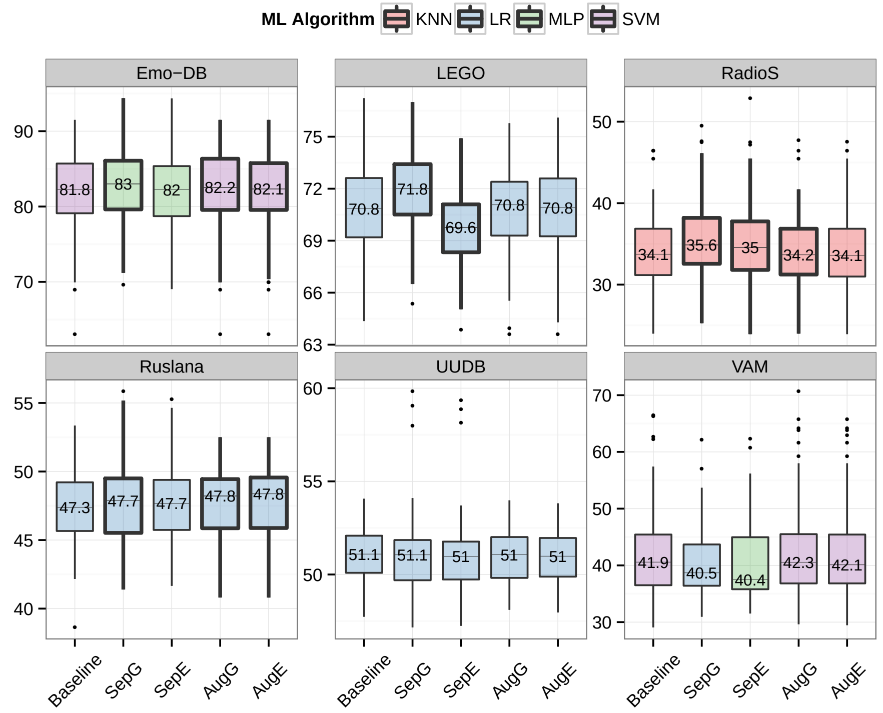

Figure 3: F_1 measure of speech-based emotion recognition with ground-truth gender-related information (*AugG* and *SepG*), with the estimated gender-related hypothesis (*AugE* and *SepE*), and without any additional information (baseline). Colours show the optimal classifiers. All the experiments are 10 repetitions of 10-fold cross-validation emotion-stratified. Box-plots with bold frames indicate T-test-based significant differences against the baseline results (at least with $p = 0.05$).

SepG		SepE	
Each	Once	Each	Once
1.19	1.81	**1.31**	1.69

Table 2: Ranks of *Sep* while speaker-adaptiveness is under examination.

AugG			
Dummy		Unique	
Non-norm	Norm	Non-norm	Norm
1.5	1.63	3.19	3.69
AugE			
Dummy		Unique	
Non-norm	Norm	Non-norm	Norm
2.06	**1.81**	2.56	3.56

Table 3: Ranks of *Aug* while speaker-adaptiveness is under examination.

128 experiments, each of them is 10 repetitions of 10-fold cross-validation. For the system *Aug*, we performed 4 (classification algorithms) x 8 (corpora) x 2 (known speaker-related information - *AugG* vs. estimated one - *AugE*) x 2 (speaker incorporating method - unique integer vs. dummy coding) x 2 (speaker ID normalised vs. non-normalised) = 256 experiments, each of them is 10 repetitions of 10-fold cross-validation. It should be noted that gender-related information was available only for 6 corpora, and age-related information was found only within the LEGO corpus, therefore the total number of experiments for gender- and age-adaptive ER systems was less than for the speaker-adaptive experiments.

For each experiment we calculated the mean of F_1 measure (over 100 runs, that is - 10 repetitions of 10-fold CV) and among 4 classifiers we selected the algorithm with the highest mean. After that, for each combination of normalisation and speaker ID incorporation methods we calculated average ranks, that is - the nominal value depending on the performance of the system on a particular data set, similar to Friedman's statistic (Theodorsson-Norheim, 1987), (Demšar, 2006). Thus, the best approach will be assigned rank 1, while the runner-up, 2, etc. In the case of identical highest average F_1 measures, we set 1.5 to both approaches. We chose this ranking method due to its simplicity and since it has been observed that the average ranking outperformed more advanced ones, when the performance of classification algorithms was analysed (Brazdil and Soares, 2000).

We calculated the ranks separately for the systems which used known speaker-related information and estimated one, as well as for two groups of settings related to *Sep* and *Aug* systems, since they have different nature and potentially may result in very different levels of performances. Next, we calculated these ranks for all the corpora considered and used average ranks for the eventual assessment of approaches. Hence, the lower the average rank, the better the system performed on average on all the emotional corpora.

For speaker-adaptive systems, the average ranks for the systems *Sep* and *Aug* are depicted in Table 2 and in Table 3, respectively.

Subsequently, we selected the systems with the highest ranks to include the corresponding results in graphs. As a visualisation tool we selected a box-plot graph (Williamson et al., 1989) for its high descriptive ability. We used a rather standard declaration of box-plots: the upper hinge is the first quartile (the 25th percentile), the lower hinge is the third quartile (the 75th percentile), upper (lower) whisker - to the highest (lowest) value within $1.5 * IQR$ (Inter-Quartile Range), points are outliers, lines within boxes depict medians, numbers within boxes are means.

Figure 1 depicts the following systems' results for each database: baseline approach - without any additional speaker-related information, *SepG* and *SepE* systems performing Z-transformation speaker-wise, *AugG* system with non-normalised speaker-related attributes within the dummy variable, and *AugE* approach with normalised speaker-related attributes. We chose these settings due to their highest average ranks (see Table 2 and Table 3).

Since we did not pay any attention to a significance test while performing the rank calculation, now we performed the paired Student's T-Test, comparing the proposed systems with the baseline approach for each corpus independently. Thus, the box-plots with bold outlines indicate significant difference against the baseline approach with at least $p = 0.05$. The speaker identification procedure has been performed in such a way, that we used the same algorithm as for the ER task. We used the SI procedure in a corpus-based manner, which means that for each corpus on each iteration of the cross-validation experiments we used exactly the same speech data and features to train both the ER and SI models. Since the results of speaker recognition have changed dramatically depending on the corpus and the algorithms used, we also presented the results of speaker recognition in Figure 2. Next, we repeated the same experiments for gender-adaptive settings. The average ranks of the systems proposed are depicted in Table 4 and Table 5.

SepG		SepE	
Each	Once	Each	Once
1.33	1.66	1.58	**1.42**

Table 4: Ranks of *Sep* for gender-adaptive ER.

AugG			
Dummy		Unique	
Non-norm	Norm	Non-norm	Norm
2.66	**1.92**	3.08	2.33
AugE			
Dummy		Unique	
Non-norm	Norm	Non-norm	Norm
2.42	**2**	3.08	2.5

Table 5: Ranks of *Aug* for gender-adaptive ER.

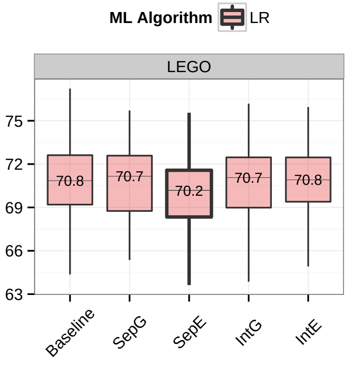

Figure 4: F_1 measure of speech-based emotion recognition with ground truth age-related information (*AugG* and *SepG*), with the estimated age-related hypothesis (*AugE* and *SepE*), and without any additional information (baseline). All the experiments are 10 repetitions of 10-fold cross-validation emotion-stratified. Box-plots with bold frames indicate T-test-based significant differences against the baseline results (at least with $p = 0.05$).

The results, which correspond to the systems with the highest ranks, are depicted in Figure 3.

Finally, we performed age-adaptive experiments on the LEGO corpus since it has age-related information including the following 3 classes: *youth*, *adult* and *elder*. Again, we selected the highest average F_1 measure among all the algorithms considered and depicted the results obtained in Figure 4. Since only one emotional corpus has been analysed within the age-adaptivity, we did not calculate average ranks for this system.

5.1. Speaker identity

It turned out that in both cases - using actual and estimated speaker identity, the system *Sep* performed the best with speaker-wise normalisation (see corresponding cells in Table 2). Similar results were previously obtained for speech recognition, where speaker normalisation improved the performance of speech recognisers (Giuliani et al., 2006).

Regarding the *Aug* systems, the one-hot codding performed better than using a unique integer. This was expected due to fact that speaker ID is not numerical but a nominal value and dummy-coding allows this fact to be handled in a more proper way than with a unique integer. Moreover, non-normalised speaker-related attributes resulted in the best performance within the *AugG* systems, whereas the normalised version achieved a higher F_1 measure within the *AugE* system (see corresponding cells in Table 3).

The proposed system using an actual SI module resulted in a significant improvement on most of the corpora with a remarkable enhancement of the F_1 measure on the AVEC corpus (49.6 *SepE* vs. 42.2 baseline), Ruslana (53 *AugE* vs. 47.3 baseline), and SAVEE (70.1 *SepE* vs. 63.2 baseline). The results may be even better if the SI component performs more accurately (compare *E* and *G* systems in Figure 1).

However, the performances of the *Sep* system dropped on

Emo-DB, LEGO, and Ruslana.

In the case of the Emo-DB corpus this can be explained by highly unbalanced coverage of emotions by the speakers. Thus, the speakers ID03 and ID10 have only one single utterance with the disgust label. By using 10-fold CV we ensured that this particular utterance will appear in the testing data exactly once. Let us consider this case and suppose that in a particular iteration of the CV we have the disgust recording in the testing set. When we train the model speaker-wise, then the emotional model for the speaker ID03 and ID10 has no chance to recognise it properly, since during the training phase there were not any recordings with the disgust label. Alternatively, during the baseline approach we operate with the whole training data in order to build only one single emotional model for all the speakers from the training set. It means that the algorithm is trained not only on disgust samples of ID03 or ID10 but all speakers from the training set. Therefore, on each iteration the model could operate with enough samples of all possible labels, enhancing the probability of proper recognition of a particular sample from the testing set.

Regarding the LEGO corpus, it was collected from the bus-navigation system (see 3. Section) containing real-user requests. Each dialogue consists of from 5 to 9 system-user turns in which the speaker tried to determine an optimal bus-route from the current to the desired location within the city. We supposed that each dialogue had been initiated by a new user and therefore each speaker in the database has very few utterances. As a result, in each iteration of CV we do not have enough data to build a reasonable speaker identification model (see rather poor SI performance on the LEGO in Figure 2). Therefore, the performance of the *SepE* system is much lower than that of the *SepG* system.

The Ruslana corpus contains 10 recordings for each emotional tag for each of 61 speakers. It means that if we perform speaker-dependent modelling for the *Sep* system, then on each iteration of the CV a modelling algorithm could operate at most with 10 recordings for each emotional label (in this case all the recordings of a particular label should be placed by chance in all the CV folds but not in the one which is currently used for testing) – obviously it is not enough to obtain a reasonable model which would show good generalisation ability. On another hand, the baseline approach operates with the whole set of recordings from all speakers which are in the training set. Therefore, a modelling algorithm within the baseline approach operates with more data of a particular emotional label which in turn lead to higher generalisation ability and recognition performance.

5.2. Gender-awareness

The results of gender recognition itself were rather high and quite similar for the 4 algorithms used, having an F_1 measure on average (over 4 algorithms) of 97.2 on Emo-DB, 85.7 on LEGO, 93.1 on VAM, 97.1 on UUDB, 94.9 on RadioS, and 98.5 on the Ruslana corpus. It turned out that for the approach which used actual gender information speaker normalisation performed best, whereas for the system with the actual GR component normalisation should be performed only once for all the utterances (see the corresponding ranks in Table 4). Regarding the system *Aug*, in both cases normalised dummy-based speaker ID encoding resulted in the highest average ranks (see the corresponding ranks in Table 5). The results of gender-adaptive ER are more regular, without large variability, and in the case of most corpora resulted in improvement (see Figure 3).

5.3. Age-awareness

The result of age recognition itself was equal to 67.7, 70.9, 64.9, and 68.6 using SVM, LR, KNN, and MLP, respectively. However, no improvement on LEGO has been achieved by performing age-adaptiveness (see Figure 4). We state that more sophisticated experiments with several corpora are needed.

6. Conclusion and future work

We concluded that the speaker-adaptive ER can significantly improve the performance using both approaches proposed. However, the *Sep* system requires balanced data and enough training material of all the target users of the ER system. Moreover, the *Sep* systems tend to be more sensitive to both the speaker identification error and statistical characteristics of the databases. Indeed, when the *Aug* systems are applied all of the utterances from the training set are used in order to train the model, whereas only the utterances of the corresponding speaker are used to build the *Sep* models.

In terms of future work, applying multi-agent emotional models can be considered by performing a simple vote or by building a meta-classifier based on individual single classifiers. In this paper we took into account only audio signals, however a dialogue might consist of visual representation, and by analysing visual cues, ER might be more successful. An additional use of advanced machine learning algorithms and contemporary feature selection methods may further improve the ER performance. Specifically, we consider using the deep learning concept to perform ER (Kim et al., 2013), and the multi-objective genetic algorithm-based feature selection (Sidorov et al., 2015) and state-of-the-art iVector-based SI procedure to further enhance the performance of the ER systems.

7. Bibliographical References

Brazdil, P. B. and Soares, C. (2000). A comparison of ranking methods for classification algorithm selection. In *Machine Learning: ECML 2000*, pages 63–75. Springer.

Brody, L. R. (1985). Gender differences in emotional development: A review of theories and research. *Journal of Personality*, 53(2):102–149.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Interspeech*, pages 1517–1520.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

Eskenazi, M., Black, A. W., Raux, A., and Langner, B. (2008). Let's go lab: a platform for evaluation of spoken dialog systems with real world users. In *Ninth Annual Conference of the International Speech Communication Association*.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM.

Giuliani, D., Gerosa, M., and Brugnara, F. (2006). Improved automatic speech recognition through speaker normalization. *Computer Speech & Language*, 20(1):107–123.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865–868. IEEE.

Hall, J. A., Carter, J. D., and Horgan, T. G. (2000). Gender differences in nonverbal communication of emotion. *Gender and emotion: Social psychological perspectives*, pages 97–117.

Haq, S. and Jackson, P., (2010). *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pages 398–423. IGI Global, Hershey PA, Aug.

Kim, Y., Lee, H., and Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3687–3691. IEEE.

Kockmann, M., Burget, L., et al. (2011). Application of speaker-and language identification state-of-the-art techniques for emotion recognition. *Speech Communication*, 53(9):1172–1185.

Lopez-Otero, P., Docio-Fernandez, L., and Garcia-Mateo, C. (2015). Assessing speaker independence on a speech-based depression level estimation system. *Pattern Recognition Letters*.

Makarova, V. and Petrushin, V. (2002). Ruslana: A database of russian emotional utterances. In *Proc. Int. Conf. Spoken Language Processing (ICSLP 2002)*.

Menard, S. (2002). *Applied logistic regression analysis*, volume 106. Sage.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50.

Platt, J. et al. (1999). Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methodssupport vector learning*, 3.

Scherer, K. (2002). Emotion. In *Sozialpsychologie*, pages 165–213. Springer.

Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated spoken dialog corpus of the cmu let's go bus information system. In *LREC*, pages 3369–3373.

Schuller, B., Steidl, S., and Batliner, A. (2009a). The interspeech 2009 emotion challenge. In *INTERSPEECH*, volume 2009, pages 312–315.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009b). Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 552–557. IEEE.

Sidorov, M., Ultes, S., and Schmitt, A. (2014a). Comparison of gender-and speaker-adaptive emotion recognition. *International Conference on Language Resources and Evaluation (LREC)*, pages 3476–3480.

Sidorov, M., Ultes, S., and Schmitt, A. (2014b). Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4803–4807. IEEE.

Sidorov, M., Brester, C., and Schmitt, A. (2015). Contemporary stochastic feature selection algorithms for speech-based emotion recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany, September.

Steininger, S., Schiel, F., Dioubina, O., and Raubold, S. (2002). Development of user-state conventions for the multimodal corpus in smartkom. In *LREC Workshop on "Multimodal Resources", Las Palmas, Spain*.

Theodorsson-Norheim, E. (1987). Friedman and quade tests: Basic computer program to perform nonparametric two-way analysis of variance and multiple comparisons on ranks of several related samples. *Computers in biology and medicine*, 17(2):85–99.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM.

Vogt, T. and André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*. Citeseer.

Williamson, D. F., Parker, R. A., and Kendrick, J. S. (1989). The box plot: a simple visual method to interpret data. *Annals of internal medicine*, 110(11):916–921.

Xuegong, Z. (2000). Introduction to statistical learning theory and support vector machines. *Acta Automatica Sinica*, 26(1):32–42.

Zhang, Z., Weninger, F., Wöllmer, M., and Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 523–528. IEEE.