Journal of Siberian Federal University. Mathematics & Physics 2015, 8(2), 208–216

УДК 519.87

Automated Recognition of Paralinguistic Signals in Spoken Dialogue Systems: Ways of Improvement

Maxim Sidorov[*]
Alexander Schmitt[†]
Institute of Communications Engineering
Ulm University
Albert Einstein-Allee, 43, Ulm, 89081
Germany
Eugene S. Semenkin[‡]
Institute of Computer Science and Telecommunications
Siberian State Aerospace University
Krasnoyarskiy Rabochiy, 31, Krasnoyarsk, 660014
Russia

Received 19.01.2015, received in revised form 25.02.2015, accepted 20.03.2015

The ability of artificial systems to recognize paralinguistic signals, such as emotions, depression, or openness, is useful in various applications. However, the performance of such recognizers is not yet perfect. In this study we consider several directions which can significantly improve the performance of such systems. Firstly, we propose building speaker- or gender-specific emotion models. Thus, an emotion recognition (ER) procedure is followed by a gender- or speaker-identifier. Speaker- or gender-specific information is used either for including into the feature vector directly, or for creating separate emotion recognition models for each gender or speaker. Secondly, a feature selection procedure is an important part of any classification problem; therefore, we proposed using a feature selection technique, based on a genetic algorithm or an information gain approach. Both methods result in higher performance than baseline methods without any feature selection algorithms. Finally, we suggest analysing not only audio signals, but also combined audio-visual cues. The early fusion method (or feature-based fusion) has been used in our investigations to combine different modalities into a multimodal approach. The results obtained show that the multimodal approach outperforms single modalities on the considered corpora. The suggested methods have been evaluated on a number of emotional databases of three languages (English, German and Japanese), in both acted and non-acted settings. The results of numerical experiments are also shown in the study.

Keywords: recognition of paralinguistic signals, machine learning algorithms, speaker-adaptive emotion recognition, multimodal approach.

Introduction

A system which is able to recognize paralinguistic signals from human-machine and human-human dialogues is useful in various applications. Thus, emotion-related information is used for the assessing of the user's satisfaction while using a Spoken Dialogue System (SDS) or for the automated monitoring of call-centres. A negative emotion-based signal can be used for changing the dialogue strategy. Depression-based information can be helpful for physicians in

[*] maxim.sidorov@uniulm.de
[†] alexander.schmitt@uniulm.de
[‡] eugenesemenkin@yandex.ru

order to support their decisions and to avoid critical mistakes in practice. Some of the user's characteristics (such as openness, agreeableness or engagement) can be utilized to create speaker-adaptive systems.

We proposed a number of ways to improve the recognition performance. Firstly, it is obvious that the dialogue consists not only of audio-signals, but also of visual-cues. Therefore, the recognition performance might be increased by also analysing visual-signals. Secondly, such subjective concepts as emotions or personal traits differ from user to another one. Hence, speaker-specific information might be beneficial for building speaker-adaptive models. Lastly, more than 6,000 numerical features can be extracted out of each audio-signal. Moreover, nearly 16,000 visual-based features can be extracted out of visual-cues per time frame. However, not all of them are relevant for the recognition task. Furthermore, highly correlated features might even result in a decrease in performance. Nevertheless, through the advanced feature selection techniques a number of features can be decreased and the performance of the recognition procedure can be increased, simultaneously. In this study we propose a number of ways of improving performance.

This paper is organized as follows: the used corpora are described in section 2; section 3 briefly describes the proposed methods and the machine learning algorithms used, as well as evaluation results and their analysis; finally, there are some conclusions in the section 4.

1. Corpora Description

All evaluations were conducted using several audio and audio-visual databases. Here are their brief description and statistical characteristics.

Berlin emotional database [1] was recorded at the Technical University of Berlin and consists of labelled emotional German utterances which were spoken by 10 actors (5 females).

LEGO emotional database [2] comprises non-acted English (American) utterances which were extracted from the SDS-based bus-stop navigational system.

SAVEE (Surrey Audio-Visual Expressed Emotion) corpus [3] was recorded as a part of an investigation into audio-visual emotion classification, from four native English male speakers.

UUDB (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database [4] consists of spontaneous Japanese speech through task-oriented dialogue which was produced by 7 pairs of speakers (12 females), 4,737 utterances in total.

VAM [5] dataset was created at Karlsruhe University and consists of utterances extracted from the popular German talk-show "Vera am Mittag" (Vera in the afternoon).

RadioS database consists of recordings from a popular German radio talk-show. Within this corpus, 69 native German speakers talked about their personal troubles.

AFEW audio-visual emotion corpus [6] was used for the first and the second Emotion Recognition in the Wild Challenges.

AVEC'14 database was used for the fourth Audio-Visual Emotion Challenge and Workshop 2014 [7]. In order to obtain the level of depression, participants have been asked to fill in a standard self-assessed depression questionnaire (the Beck Depression Inventory-II) consisting of 21 questions. Each affect dimension (Arousal, Dominance, and Valence) has been annotated separately by a minimum of three and maximum of five raters.

MATPRAITS database [8] is a subset of the SEMAINE corpus consisting of 44 videos, which are in four situational contexts of 11 subjects. The following personality and social aspects have been chosen as labels in the dataset: the Big Five personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and openness) and four additional social dimensions (engagement, facial attractiveness, vocal attractiveness and likability).

It should be noted, that only in the case of the AVEC'14 dataset has the regression-based procedure been applied. All continuous labels of the MAPTRAITS database have been converted to nominal values in range of [1, 9]. There is a statistical description of the corpora used in Tab. 1.

Table 1. Databases description

Database	Language	Full length(min.)	File level duration(sec.)		Paralinguistic Labels (Type)
			Mean	Std.	
Berlin	German	24.7	2.7	1.02	Anger, boredom, disgust, anxiety, happiness, sadness, neutral (Categories)
SAVEE	English	30.7	3.8	1.07	Anger, disgust, fear, happiness, sadness, surprise, neutral (Categories)
VAM	German	47.8	3.02	2.1	Valence, Activation, Dominance (Dimensions)
RadioS	German	278.5	6.26	5.17	Neutral, happy, sad, angry (Categories)
UUDB	Japanese	113.4	1.4	1.7	Pleasantness, arousal, dominance, credibility, positivity (Dimensions)
LEGO	English	118.2	1.6	1.4	Angry, slightly angry, very angry, neutral, friendly, non-speech (Categories)
AFEW	English	55.38	2.43	1.03	Angry, disgust, fear, happy, neutral, sad, surprise (Categories)
MAPTRAITS	English	11.0	15.0	0.0	Extraversion, agreeableness, conscientiousness, neuroticism, openness, engagement, facial attractiveness, vocal attractiveness, likability (Dimensions)
AVEC'14	German	164.08	65.63	46.22	Valence, arousal, dominance (Dimensions)

2. The proposed approaches

2.1. Speaker- and gender-adaptive emotion recognition

It is evident that the expressed emotions differ from one person to another Therefore we propose using speaker-specific information with emotion recognition models. Incorporating speaker-specific information into the emotion recognition process may be done in many ways. A very straightforward way is to add this information to the set of features (System A). Another way is to create speaker-dependent models: while, for conventional emotion recognition, one statistical model is created independently of the speaker, a separate emotion model may be created for each speaker (System B). Both approaches result in a two-stage recognition procedure: Firstly, the speaker is identified and then this information is included in the feature set directly (System A), or the corresponding emotion model is used for estimating the emotions (System B). Both emotion recognition speaker identification (ER-SI) hybrid systems have been investigated and evaluated [10].

As a baseline, an emotion recognition process without speaker-specific information has been conducted. The training set was used to create and train an artificial neural network (ANN) based emotion model. The test set was used to evaluate the model. Hence, one single neural network has been created addressing the emotions of every speaker in the database.

In the first experiment, the focus was on investigating the theoretical improvement, which may be achieved using speaker-based adaptivity. For this, known speaker information (true labels) was used for both approaches (see True SI in Fig. 1). In System A, the speaker information was simply added to the feature vector. Hence, all utterances with the corresponding speaker information were used to create and evaluate an ANN-based emotion model. For System B, individual emotion models were built for each speaker. During the training phase, for each speaker, all speaker utterances were used for creating the emotion models. During testing, all speaker utterances were evaluated with the corresponding emotion model.

Additionally, a second experiment was conducted including a real speaker identification mod-

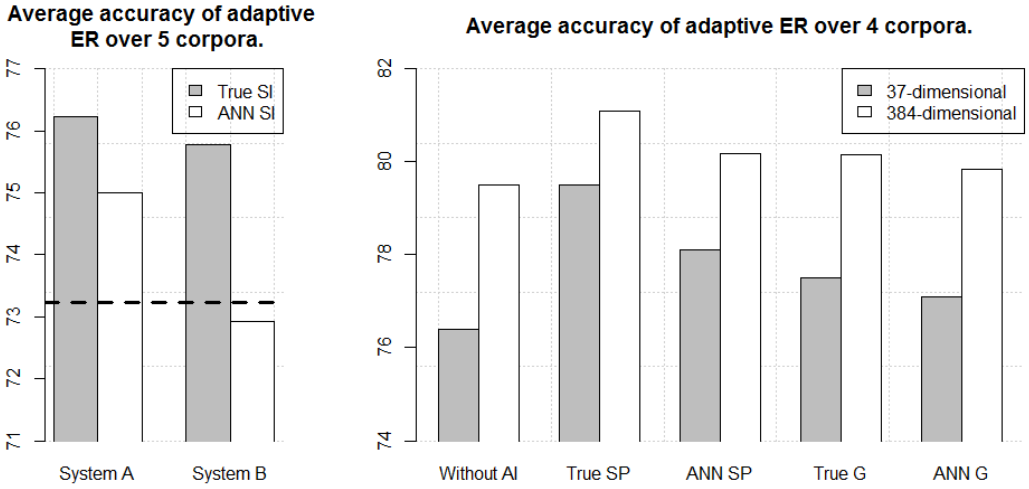

Fig. 1. The results of speaker- and gender-adaptive emotion recognition. Speaker-related information is included in the feature vector (System A), separate emotion model for each speaker (System B). The dashed line is ER accuracy without speaker-specific information. Without AI denotes performance of the system without any additional information, ground true information is used (True SP, True G), and ANN-based hypotheses of speaker- and gender-specific information is used (ANN SP, ANN G). All results on the left-side of the figure have been achieved using System A

ule instead of using known speaker information. Firstly, an ANN-based speaker identifier was created during the training phase (See ANN SI in Fig. 1). Furthermore, for System A, the known speaker information was included into the feature vector for the training of the emotion classifier. The testing phase starts with the speaker identification procedure. Then, the speaker hypothesis was included in the feature set which was in turn fed into the emotion recognizer. For System B, an ANN-based emotion recognizer 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 recognized speaker to create an emotion hypothesis. In contrast to the first experiment, these experiments are not free of speaker identification errors.

During these experiments, all audio-signals were characterised by 37-dimensional feature vectors in static mode (i.e. one feature vector per recording). A multi-layer perception has been chosen as a modelling algorithm for both emotion recognition and speaker-identification. 5 emotional databases, namely LEGO, Berlin, SAVEE, UUDB, and VAM, have been used in order to evaluate the proposed methods.

By analysing the results obtained it could be concluded that System A outperforms System B. Obviously, the recording which was incorrectly classified by the MLP-based speaker model cannot be properly analysed by the emotional model of another speaker. Moreover, the balanced data must be used in order to build separate emotional models for each speaker. What is more, it is obvious that using the speaker-specific information significantly improves the performance of the ER procedure, even in the case of the automated MLP-based speaker recognition component.

Further, the used feature vector has been extended to a 384-dimensional one; moreover, the proposed speaker-adaptive ER has been compared against gender-adaptive ER using System A only (since System A outperformed System B, see Fig. 1). The following corpora have been used in this experiment: VAM, LEGO, Berlin, and UUDB. As a baseline, the experiments without any Additional Information (see Without AI bars in Fig. 1) have been conducted. Then, analogically, ground true speaker- and gender-related information has been included in the feature vector directly (see True SI and True G bars in Fig. 1). Lastly, in order to evaluate ANN-based speaker and gender recognizers, MLP-based models have been incorporated to obtain speaker- and gender-related hypotheses. After this, during the testing phase, the obtained hypotheses have been included in the feature vector directly, then the extended feature vector has been used by the MLP-based emotional model to produce a final emotional-based hypothesis (see ANN SI and ANN G bars in Fig. 1). It should be noted that during the training phase, the ground true information has been included in the extended feature vector.

From the results obtained, it can be concluded that the speaker-identification procedure results in much higher ER performance than the gender-recognition component. Nonetheless, speaker-based adaptive ER required the training recordings from all speakers who are intended to use the system. It results in a closed-set of end-users of the developed system, but such limitations are not always appropriate in practice. However, both speaker- and gender-adaptive systems outperform the system which does not use any additional information about the user. It is true for the systems which use not only the ground-true information, but also the estimated hypotheses.

Further, the 384-dimensional feature vector seems to be an optimal choice to describe the audio-based recordings (see Fig. 1).

2.2. Speaker State Recognition with Feature Selection Techniques

To perform the feature selection-based approach two techniques have been chosen, Information Gain Ratio (IGR) and a Genetic Algorithm (GA). Since, the speaker- and gender-related information is useful for the emotion recognition, the following problems have been examined with feature selection techniques: ER, Speaker Recognition (SR), and Gender Recognition (GR). The results of the numerical experiments, using the same 384-dimensional feature vector and emotional corpora are in Fig. 2. In order to obtain more statistically significant results the optimization procedure has been implemented using the cross-validation method with the number of folds equal to 6 (stratified sampling).

Information Gain Ratio is a state-of-the-art approach for feature weighting and selection. This algorithm is used in a filter approach i.e. the weighted procedure is applied only once before the modelling technique and it uses only the labels and statistical characteristic of the dataset, independently from the classification method. This is in contrast to the GA-based feature selection which uses the output of the classification algorithm in every population of the evolution process in order to assess the fitness of the individual. A Boolean true corresponds to the relevant feature and a Boolean false denotes an unessential one. All the evolution procedures (selection, recombination, mutation) are used to find a quasi-optimal set of features.

By analysing the results in Fig. 1, it can be concluded that using the feature selection techniques results in a significant increase in ER performance. Both techniques outperform the results of ER without any feature selection method. However, the IGR-based approach results in higher performance with much more features than the GA-based approach. It turns out that the IGR-based feature selection technique uses twice as many features (on average) than the GA-based one. Thus, the choice of the feature selection method in a particular case is a matter of a trade-off between the accuracy and the number of features used [9].

2.3. Multimodal recognition of paralinguistic features

The last three databases from Tab. 1 (AVEC, MAPTRAITS, and AFEW) comprise not only audio-signals, but also visual ones. Therefore, these corpora can be used in order to investigate

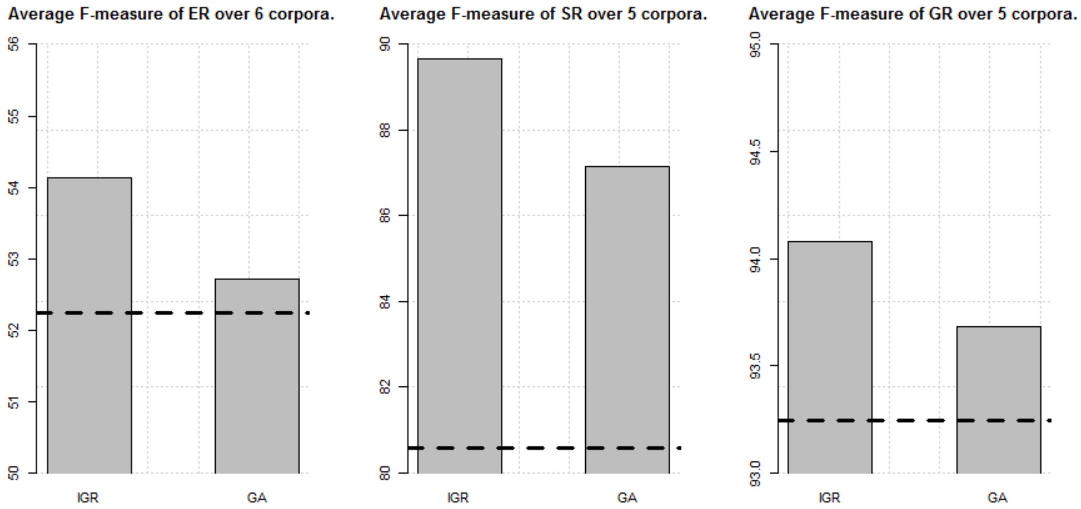

Fig. 2. The results of speaker-state recognition (ER, SR, GR denote emotion, speaker and gender recognition, respectively) with feature selection techniques. Dashed lines are the average F-measure without any feature selection techniques. IGR is Information Gain Ratio-based feature selection, and GA is genetic algorithm-based feature selection when the wrapper approach is applied. MLP is used as a modelling algorithm

multimodal approaches. We suggest using the feature-level fusion, when audio- and visual-based features are concatenated in order to form the multimodal feature vector.

The audio part of the AFEW recordings has been characterised by a 1,582-dimensional feature vector similar to the features employed in Audio Video Emotion Recognition Challenge 2011 motivated from the INTERSPEECH 2010 Paralinguistic challenge [11].

For the visual part of the AFEW corpus, the face localisation using the Mixture of Parts framework and tracking using the IntraFace library have been performed. First, the fiducial points generated by IntraFace have been utilized in order to align the faces. Second, the local appearance descriptor LBP-TOP [12] has been extracted from non-overlapping spatial 4x4 blocks. Finally, the LBP-TOP features from each block are concatenated to form one feature vector.

The audio part of the recordings from AVEC'14 has been characterised by a 2,268-dimensional feature vector.

Regarding the visual part of the AVEC'14 dataset, the local dynamic appearance descriptor LGBP-TOP from the eMax face analysis toolbox has been utilised as the visual-based features. The original video recordings have been resampled to uniform 30 frames per second at 640 x 480 pixels. Further, face localisation and segmentation have been performed with the publicly avail- able Viola & Jones face detector. The following parameters set of LGBP-TOP [13] feature extraction procedure has been used: 18 filters with variable orientation and frequencies, but constant amplitude and phase; a fixed window of 5 overlapping frames. It should be noted that only features extracted from XY image planes have been included into the feature set.

For the MAPTRAITS dataset, the audio part of the recordings has been characterised by a 6,376-dimensional feature vector.

As visual-features for the MAPTRAITS, Quantised Local Zernike Moments (QLZM) [14] have been utilised. Face localisation by detecting 49 landmark points per frame has been performed with the publicly available Xiong and De la Torre face detector [15]. The face is divided into subregions by applying a 5x5 outer grid and a 4x4 inner grid which yielded a 656-length feature vector. Another video-based feature set has been created by determining the two eye areas and the mouth area, and extracting QLZM features on these parts. For part-based representation, the eye and mouth areas have not been divided into subregions. This resulted in feature vectors having a length of 48.

The results of audio-only-, visual-only- and combined audio-visual-based recognition are shown in Fig. 3. The test results have been achieved with the corresponding portions of the datasets, which have been provided by the authors of the corpuses. As a modelling algorithm, the SVM trained by Sequential Minimal Optimization (SMO) algorithm has been chosen for all databases and modalities. It should be noted, that in the cases of the MAPTRAITS and AVEC'14 databases, the average values over all segmentation methods have been calculated (see the corresponding papers).

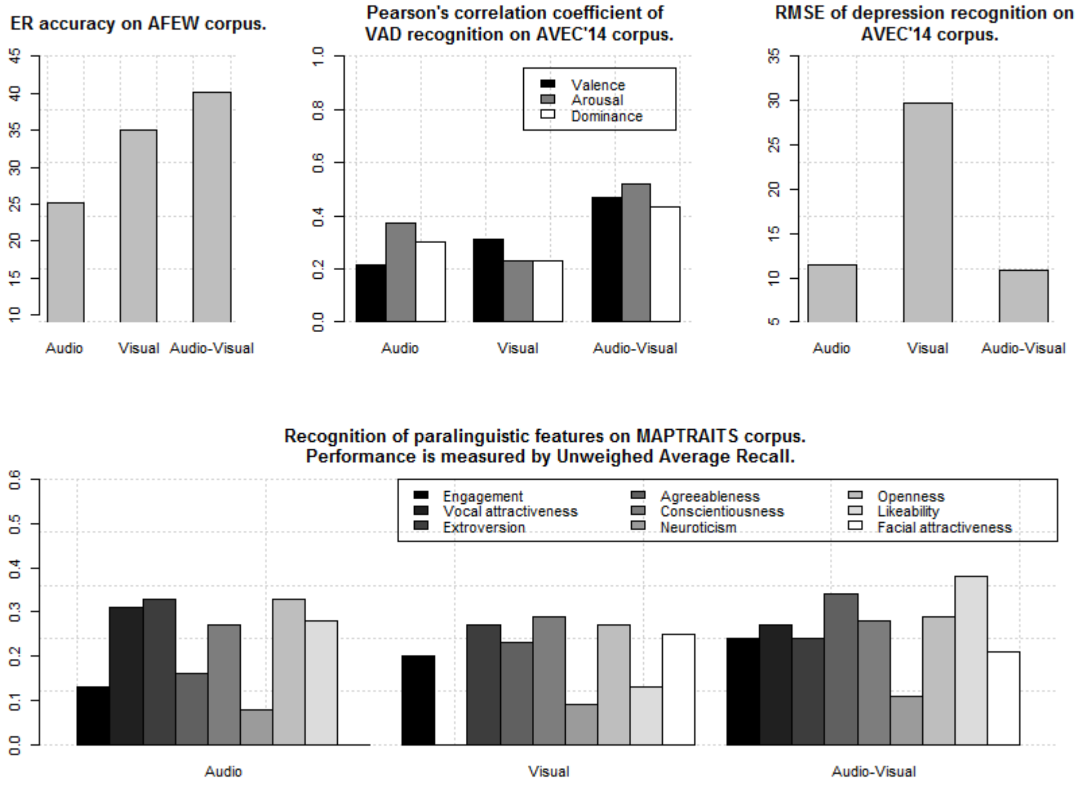

Fig. 3. Recognition of paralinguistic signals under three modalities (audio-, visual-only, and audio-visual cues). Emotion recognition on AFEW dataset. Recognition of emotion-related continuous measurements (valence, arousal, and dominance) on AVEC'14 database. Depression recognition on AVEC'14 dataset. Recognition of paralinguistic measurements (Big Five personality traits plus engagement, facial and vocal attractiveness, and likeability) on MAPTRAITS corpus.

The results obtained show that for all of the databases a multimodal approach significantly outperforms the single modalities-based approaches.

3. Conclusion and future work

We have proposed a number of approaches resulting in improvement of ER, as well as recognition of personality traits. The suggested approaches have been evaluated on a number of databases.

It has already been shown that such a very simple method as extending the feature vector with additional speaker-specific information can improve the recognition performance. This improvement is more significant when using true speaker-specific information. These results are very encouraging leading to further more sophisticated approaches on speaker-dependent emotion recognition, e.g., applying speaker adaptation methods already known from speaker-adaptive speech recognition.

Moreover, an application of the proposed feature selection systems in order to select the most representative features and maximize the accuracy of particular tasks could decrease the number of features and increase the accuracy of the system simultaneously. In most of the cases, the IGR-based technique outperforms baseline results. It should be noted that the number of selected features using the IGR method is quite high. It means that in some cases the number of features was equal to 384, i.e. an optimal modelling procedure has been conducted without feature selection at all. That is why a GA-based approach might be used in the cases where the decreasing of the feature set is the important task.

Finally, it has been shown that the usage of not only audio-cues, but also of visual-ones in order to recognize emotions or personal traits might be beneficial in most of the cases.

References

[1] F.Burkhardt et al., A database of German emotional speech, Proceedings of the Interspeech 2005 Conference, 2005, 1517–1520.

[2] A.Schmitt, B.Schatz, W.Minker, Modeling and predicting quality in spoken human-computer interaction, Proceedings of the SIGDIAL 2011 Conference, 2011, 173–184.

[3] S.Haq, P.J.B.Jackson, Multimodal Emotion Recognition, In W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, IGI Global Press, 2010, 173–184.

[4] H.Mori, T.Satake, M.Nakamura, Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics, *Speech Communication*, **53**(2011), no. 1, 36–50.

[5] M.Grimm, K.Kroschel, S.Narayanan, The Vera am Mittag German audio-visual emotional speech database, in IEEE International Conference Multimedia and Expo, Hannover, 2008, 865–868.

[6] A.Dhall et al., Collecting large, richly annotated facial-expression databases from movies, *IEEE MultiMedia*, **19**(2012), no. 3, 34–41.

[7] M.Valstar et al., AVEC 2013: the continuous audio/visual emotion and depression recognition challenge, Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, New York, USA, 2013, 3–10.

[8] O.Celiktutan et al., MAPTRAITS 2014: The First Audio/Visual Mapping Personality Traits Challenge, Proceedings of the Personality Mapping Challenge & Workshop, Istanbul, Turkey, 2014.

[9] M.Sidorov, Ch.Brester, E.Semenkin, W.Minker, Speaker State Recognition with Neural Network-based Classification and Self-adaptive Heuristic Feature Selection, Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics, Vienna, Austria, Vol. 1, 2014, 699–703.

[10] M.Sidorov, S.Ultes, A.Schmitt, Emotions are a personal thing: Towards speaker-adaptive emotion recognition, IEEE International Conference on Acoustics, Speech and Signal Processing, Florence, Italy, 2014, 4803–4807.

[11] B.Schuller et al., The INTERSPEECH 2010 paralinguistic challenge, Proc. of the Interspeech, 2010, 2794–2797.

[12] G.Zhao, M.Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions., *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **29**(2007), no. 6, 915–928.

[13] T.R.Almaev, M.F.Valstar, Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition., In: IEEE Conference on Affective Computing and Intelligent Interaction, 2013, 356–361.

[14] N.Singhal et al., Robust image watermarking using local Zernike moments, *Journal of Visual Communication and Image Representation*, **20**(2009), no. 6, 408–419.

[15] Xiong X., De la Torre F. Supervised descent method and its applications to face alignment, In: IEEE Conference on Computer Vision and Pattern Recognition, 2013, 532–539.

Автоматическое распознавание паралингвистических характеристик говорящего: способы улучшения качества классификации

Максим Сидоров
Александр Шмитт
Евгений С. Семенкин

Способность искусственных систем распознавать паралингвистические характеристики говорящего, такие как эмоциональное состояние, наличие и степень депрессии, открытость человека, является полезной для широкого круга приложений. Однако производительность таких систем далека от идеальных значений. В этой статье мы предлагаем подходы, применение которых позволяет существенно улучшить производительность систем распознавания. В работе описывается метод построения адаптивных эмоциональных моделей, позволяющих использовать характеристики конкретного человека для построения точных моделей. В статье представлены алгоритмы выявления наиболее значимых характеристик речевых сигналов, позволяющие одновременно максимизировать точность решения поставленной задачи и минимизировать количество используемых характеристик сигнала. Наконец, предлагается использовать комбинированные аудио визуальные сигналы в качестве входов для алгоритма машинного обучения. Указанные подходы были реализованы и проверены на 9 эмоциональных речевых корпусах. Результаты проведенных экспериментов позволяют утверждать, что предложенные в статье подходы улучшают качество решения поставленных задач с точки зрения выбранных критериев.

Ключевые слова: распознавание паралингвистических характеристик, алгоритмы машинного обучения, адаптивная процедура распознавания эмоций, мультимодальность.