Interspeech 2022
18-22 September 2022, Incheon, Korea

Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework

Alena Velichko[1], *Maxim Markitantov*[1], *Heysem Kaya*[2], *Alexey Karpov*[3]

[1]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), Russia
[2]Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
[3]ITMO University, St. Petersburg, Russia

alena.n.velichko@gmail.com, m.markitantov@yandex.ru, h.kaya@uu.nl, karpov@iias.spb.su

Abstract

In this paper, we present a hierarchical framework for complex paralinguistic analysis of speech including gender, emotions and deception recognition. The main idea of the framework is built upon the research on interrelation between various paralinguistic phenomena. It uses gender information to predict emotional states, and the outcome of the emotion recognition to predict the truthfulness of the speech. We use multiple datasets (aGender, Ruslana, EmoDB and DSD) to perform within-corpus and cross-corpus experiments using various performance measures. The experimental results reveal that gender-specific models improve the effectiveness of automatic speech emotion recognition in terms of Unweighted Average Recall up to an absolute 5.7%, and the integration of emotion predictions improves the F-score of automatic deception detection compared to our baseline by an absolute 4.7%. The obtained cross-validation results of 88.4±1.5% for deception detection beat the existing state-of-the-art by an absolute 2.8%.

Index Terms: Computational Paralinguistics, Gender Recognition, Emotion Recognition, Deception Detection

1. Introduction and background

Paralinguistic analysis deals with various speech phenomena beyond words and meaning: semi-permanent speaker characteristics (e.g. gender, ethnicity), semi-permanent speaker traits (e.g., age, openness, likability) and transitory speaker states (e.g., emotions, stress, health conditions). The major driving force in the field of computational paralinguistics is the annual ComParE challenges [1] that provide data, baselines and evaluation standards for recognition of various paralinguistic speech phenomena, for example age and gender, dialects, emotions, speaker traits, sincerity and so on. However, most of the research deals with these classification tasks in isolation, without considering the fact that all the paralinguistic phenomena co-exist in the same speech signal and are likely to have high correlations or even influence each other.

The benefit of multitasking, mostly incorporating speaker gender into other applications, was explored earlier in several works. Examples of holistic assessment of paralinguistic speech phenomena include simultaneous analysis of speaker characteristics such as age, gender, race and height [2]; age, height, weight and smoking habits [3]; emotions, likability and personality assessment [4]; deception and sincerity [5]. Some works focused on co-learning of several related paralinguistic aspects for speaker verification and diarization [6]. There also were attempts at transferring age and gender attributes for dimensional emotion prediction [7].

Using gender for emotion recognition task has been shown effective both by directly including this information to the set of features [8] and by adaptively training separate systems on different sets of data [9]. Neural network architectures were also used for speech emotion recognition with auxiliary learning of gender recognition [10] and allow extracting new types of knowledge that represents gender as a distributed feature [11].

The general idea of using emotions for deception detection is based on psychological and paralinguistic studies. According to four-factor theory [12], deception includes different psychological processes and conditions that affect human behaviour. Key factors that can reveal a deceiver in this model are: emotional responses, activation, cognitive efforts and attempts to control the behaviour. Emotional states associated with deception evoke uncontrolled behavioural changes that can be detected in different nonverbal channels [13]. A renowned psychologist, Paul Ekman [14], suggests three main emotions that refer to deception: fear, shame and duping delight. The use of high-level feature generation (activation, valence, regulation and emotional classes) was explored by Amiriparian et al. [15]. Mendels et al. [16] proposed another model that uses emotional and lexical features for deception detection task.

Nevertheless, until now there has been no research on analysis and prediction of gender, emotions and deception in a holistic manner. Our work presents the first complex paralinguistic system that improves the speech deception detection by incorporating information about gender and emotion, emphasizing the connection between these phenomena.

2. Proposed framework

The overall architecture of the proposed hierarchical framework is depicted in Figure 1. It consists of 3 stages. In the first stage, we perform gender classification as the base for further analysis. In the second stage, the results of gender recognition are used to build adaptive gender-specific models. Finally, we use the results of both gender and emotion recognition for deception detection, expanding the feature set with probabilities of different emotional states.

In each stage, we use a different performance evaluation measure to compare the results to existing literature. In the first and second stages, we use Unweighted Average Recall (UAR) because this performance measure is continuously used in ComParE challenges since 2010 as a standard measure for imbalanced classes [17]. UAR is also more important than precision for emotion recognition: we are more interested in identifying all instances of emotional expressions even though some of them might not be relevant. In the third stage, F-measure is used

10.21437/Interspeech.2022-11294

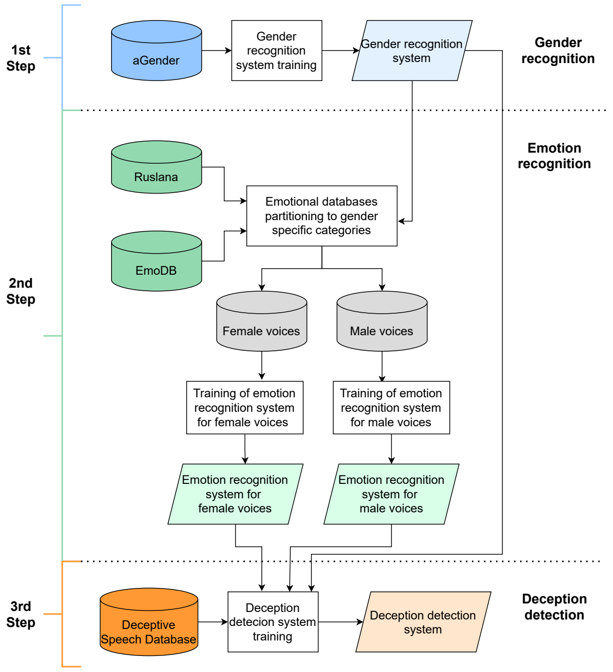
1st Step
aGender
Gender recognition system training
Gender recognition system
Gender recognition
Emotion recognition
Ruslana
Emotional databases partitioning to gender specific categories
EmoDB
2nd Step
Female voices
Male voices
Training of emotion recognition system for female voices
Training of emotion recognition system for male voices
Emotion recognition system for female voices
Emotion recognition system for male voices
3rd Step
Deceptive Speech Database
Deception detecion system training
Deception detection system
Deception detection

Figure 1: *Hierarchical framework of the complex paralinguistic system.*

to evaluate the performance of the deception detection because it is the most commonly used measure by other authors [18] and it balances the precision and recall.

2.1. Databases

2.1.1. Gender corpus

For training and testing the gender recognition system we used aGender corpus [19] because of its highly popular use since ComParE 2010 [20]. It consists of 49 hours of telephone speech, recorded from 945 German speakers in 6 sessions. The total number of utterances is 65364. The length of the utterances varies: for command words, months names, dates, time, telephone numbers, names and surnames, the duration falls in the range of 1 to 6 seconds. Every utterance is annotated in accordance with the speaker's age and gender. This corpus has seven groups (classes) of speakers: children, youth (male and female), adults (male and female), and seniors (male and female).

2.1.2. Emotional corpora

EmoDB [21] is a well-known German-language database that was recorded by 10 actors (5 males and 5 females) with ages ranging from 21 to 35, according to predefined scenarios. It contains a total of 535 speech utterances with a duration ranging from 1.2 to 9.0 seconds and a median of 2.6 seconds. The utterances were pronounced with 7 emotions: neutral, anger, fear, joy, sadness, disgust and boredom. The recordings were taken in an anechoic chamber with high-quality recording equipment. Five short and five longer phrases were chosen with neutral semantic content that could be used in day-to-day conversations. The naturalness of the emotions was evaluated in a perceptual test to assure the quality of the portrayed emotions. In this work the original 535 instance set was used for training and testing.

Ruslana [22] is a Russian-language acted database with 3661 phonetically representative utterances pronounced by 61 native speakers (12 males and 49 females) with ages ranging from 16 to 28. The structure of the corpus is similar to EmoDB but Ruslana features much more variety of speakers. The recordings were made in a sound proof recording studio and comprise the following six emotional classes: neutral, surprise, happiness, anger, sadness and fear. The duration of the recordings ranges from 1.2 to 7.8 seconds with a median at 2.3 seconds. Human evaluators monitored how well the subjects portrayed the intended emotions to make sure the expressions are natural.

2.1.3. Deception corpus

We used Deceptive Speech Database (DSD) [17] for both training and testing of the deception detection model as it is one of the few publicly available corpora for this task. It was collected at the University of Arizona (USA) and used in the INTERSPEECH Computational Paralinguistics Challenge in 2016 [17] as well as in the other research on the deception detection task. The total length of the corpus is approximately 162 minutes of speech; it was recorded by 72 English-speaking university students, which were randomly split into two groups. They played a role of either deceivers who stole papers from teacher's office, or honest students. The total amount of audio files is 1058. The corpus has two classes – deception and truth. The duration of audio files ranges between 0.62 and 161.42 seconds with a median of 2.93 seconds.

2.2. Experiments

The experiments were designed in three steps: (1) gender recognition, (2) emotion classification, and (3) deception detection. At each step, the baseline performance was established by evaluating the proposed systems in an isolated manner. Next, the systems were re-evaluated by incorporating the knowledge from the previous steps. The data and experimental procedures are described in detail below.

2.2.1. Gender recognition

The pipeline of the gender recognition system is presented in Figure 2. Since the aGender corpus files were stored as raw audio, first, we converted these files to wav format and applied a voice activity detector on each audio file. Unlike previous research [20], we used pretrained enterprise-grade Voice Activity Detector (VAD) by Silero[1]. After that, MelSpectrograms, log scale (dB) MelSpectrograms with 64 Mel filterbanks and MFCCs with their Δ and $\Delta\Delta$ were extracted from the speech signal using Python library librosa [23]. The window width was 22 ms. and the step was 5 ms. In addition, we have split these features into chunks of the length of the shortest audio file. In our case, it was 188. We also padded obtained features with zeros when necessary.

We used a large-scale pretrained audio neural network (PANN) for Audio Pattern Recognition [24]. Our studies presented recently have shown the effectiveness of using pretrained models [25, 26]. We compared pretrained models and models trained from scratch such as CNN-6, CNN-10, CNN-14. The last layer was replaced with a new one with the number of neurons equal to the number of classes.

The aGender corpus provides a predefined data split: train-

[1] https://github.com/snakers4/silero-vad

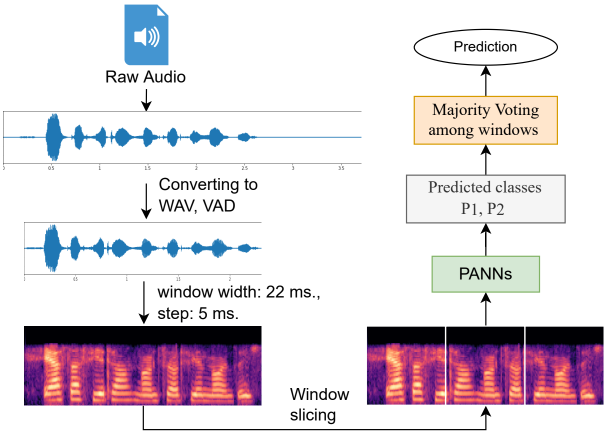

Figure 2: *Speaker gender recognition system. VAD: Voice Activity Detector. PANNs: pre-trained audio neural networks.*

ing (23 hours, 471 speakers), validation (14 hours, 299 speakers), and testing (12 hours, 175 speakers). However, because the test set labels are unavailable, we only used the train and validation sets to perform the experiments. In addition, we did not use children class in this experiments. We were limited to only females and males classes.

In the binary classification (females and males), the proposed approach resulted in UAR of 97.72% for the validation dataset. Table 1 shows the results of speaker's gender recognition on the cross-corpus setup. Ruslana corpus shows 86.09% UAR, EmoDB - 66.67% UAR, and DSD - 57.30% UAR. Low classification rate on DSD corpus can be explained by the nature of the speech utterances: a lot of DSD speech samples contain either too short (monosyllabic answers "yes" or "no") or too long (up to 160 sec.) utterances, which greatly differ from the original aGender utterances that range from 1 to 6 seconds.

Table 1: *Gender recognition UAR (%) of the proposed system trained on aGender corpus in single and cross-corpus setups*

Test DB	Male/Female ratio (%)	Gender UAR (%)
aGender	47/53	97.72
Ruslana	20/80	86.09
EmoDB	44/56	66.67
DSD	53/47	57.30

2.2.2. Emotion recognition

The small size of emotional speech corpora is often a restriction for training complex deep learning systems. Therefore, we used Support Vector Machine (SVM) classifier as it has been proved effective for many speech paralinguistic tasks [17]. We used de-facto standard openSmile suprasegmental features by summarizing the LLDs over the whole utterance as described in [27], which gives a 6373-dimensional feature vector. Principle Component Analysis (PCA) was applied to reduce the negative effects of high dimensionality of the feature space. The data was z-normalized before fitting the classifiers. Because of the small dataset size, all experiments were performed in a k-fold cross-validation fashion with the value of k equal to 5.

To investigate the effect of gender information on emotion recognition we built three types of emotion classifiers: (1) using no gender information (gender-agnostic), (2) using true gender information obtained from the data provided by the donors of the corpus, and (3) using gender predictions obtained from the proposed gender recognition system. To incorporate the gender information, we split the available data into male and female voices and trained gender-specific models on each subset. In this way, the systems can adapt to the corresponding gender categories and produce better results.

Table 2 shows within-corpus emotion classification performance of the proposed systems on two datasets, Ruslana and EmoDB. The baseline performance of gender-agnostic models in terms of UAR is 48.34±4.59% for Ruslana and 81.41±5.05% for EmoDB, which is on par with other systems proposed recently [28, 29]. With gender specific models it is possible to achieve up to 5.7% absolute improvement for female voices and 3.54% for male voices. Moreover, the standard deviation drops when considering gender-specific models. Generally, the performance is lowered when using gender predictions instead of true gender categories; however, it still remains higher than the baseline models that do not take gender information into account.

Table 2: *Within-corpus emotion recognition UAR (%) of the proposed gender-agnostic and gender-specific systems*

Emotion DB	Gender info	Emotion UAR (%)	
		Male	Female
Ruslana	-	48.34±4.59	
	Predicted	48.41±1.94	53.07±0.67
	True	51.88±2.42	54.05±1.82
EmoDB	-	81.41±5.05	
	Predicted	84.21±4.18	82.55±1.94
	True	77.66±3.84	86.32±4.03

2.2.3. Deception detection

For deception detection experiments, we annotated the DSD dataset manually to get the true gender label for every speech utterance. Three acoustic feature sets from INTERSPEECH ComParE challenges were extracted using openSMILE toolkit: ComParE 2013 [27], ComParE 2016 [17] (revised version of 2013 set) and ComParE 2011 [30] (consists of acoustic features that were used in the detection of speaker state challenge), thus a union set with total dimensionality of about 8000 features was computed.

To augment the data and cope with class imbalance we used the SMOTE method [31] with nearest neighbours parameter set to 3. The SMOTE method was used to augment the data for deception detection system due to the fact that the DSD dataset is the smallest and the most imbalanced dataset in the hierarchical system; two other datasets (for gender and emotion recognition) are more representative and balanced, thus, we have augmented only the DSD dataset. We also used PCA with 99% of preserved variance to reduce the dimensionality of the feature space. The resulting dataset has 1494 instances for training and testing with 507 features per each instance, so the final feature vector does not include any redundant features.

We used 3 different implementations of the gradient boosting algorithm in an ensemble similar to [18]: Catboost [32], XGBoost [33], LightGBM [34]. Due to the differences in implementations of the gradient boosting algorithm, the ensemble is expected to improve the final quality of each model. To combine predictions of these models we used a two-level method of stacking with three boosting algorithms on the first level and logistic regression algorithm on the second level of the model. SMOTE, PCA and stacking method were performed using Scikit-learn Python library [35]. The overall process is shown in Figure 3. All models were trained using a 5-fold cross-validation approach. Our baseline model trained without information about emotions and gender achieved an F-score of 83.7±0.2%.

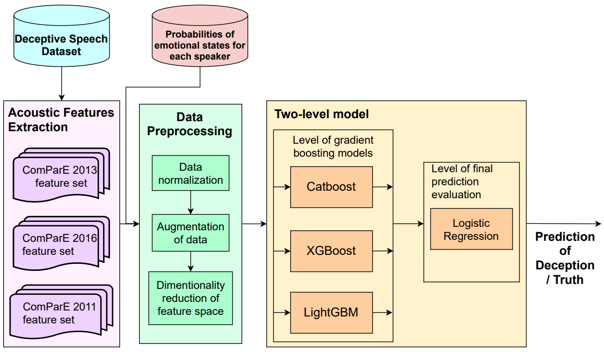
Deceptive Speech Dataset
Probabilities of emotional states for each speaker
Acoustic Features Extraction
ComParE 2013 feature set
ComParE 2016 feature set
ComParE 2011 feature set
Data Preprocessing
Data normalization
Augmentation of data
Dimentionality reduction of feature space
Two-level model
Level of gradient boosting models
Catboost
XGBoost
LightGBM
Level of final prediction evaluation
Logistic Regression
Prediction of Deception / Truth

Figure 3: *Deception detection system with two-level model.*

During the next set of experiments, we expanded the feature vector by adding probabilities of emotional states obtained at the previous stage. A comparison of the results achieved with the baseline model (no emotions, no gender), the emotion-aware model (using emotions only), and the complex model (using both emotions and gender) is shown in Table 3. Here, we compared the performance based on true gender information and gender predictions from the first processing stage. Dashes mean that information about gender was not used during training process. The best F-score of 88.4%±1.5 was achieved using emotional features from Ruslana corpus, and a predicted gender for both Ruslana and DSD. This result outperforms the current DSD state-of-the-art [18] by 2.8% in terms of F-score.

Table 3: *Deception recognition F-score (%) of the proposed system on DSD dataset subject to gender and emotion information*

Emotion DB	Gender info		Deception
	Emotion	Deception	F-score (%)
Ruslana	-	-	85.5±0.2
	Predicted	Predicted	**88.4±1.5**
	Predicted	True	88.0±1.2
	True	Predicted	88.2±1.7
	True	True	85.4±0.3
EmoDB	-	-	85.4±0.6
	Predicted	Predicted	87.5±1.0
	Predicted	True	87.7±1.3
	True	Predicted	87.9±1.6
	True	True	85.4±0.1
-	-	-	83.7±0.2

3. Discussion and conclusions

The effectiveness of the developed gender-specific and emotion-aware deception detection models was empirically shown using multiple sets of experiments with various corpora: aGender, Ruslana, EmoDB and DSD. Using gender-specific models allowed improving the emotion recognition rate as compared to gender-agnostic models by an absolute 5.7% and 3.54% for female and male voices, respectively. The recognition rate of the systems trained on female voices is notably higher than models trained on male voices on all datasets. This may happen due to females voices having wider range of voice characteristics and being more emotionally expressive. In addition, the performance on Ruslana dataset was much lower than the EmoDB; the difference may be attributed to the noise in labels - Ruslana was recorded by amateur (student) actors, while EmoDB was created by professionals. Additionally, Ruslana contains a much higher variety of speakers compared to EmoDB, which also contributes to more challenges during modelling. However, this fact does not affect the positive influence of inclusion of emotional features for deception detection from both datasets: Ruslana, despite having lower emotion recognition performance, produces even better results than EmoDB in the deception detection task.

Experiments on integration of emotional features in the system for deception detection also show positive impact. The proposed hierarchical framework achieved an absolute improvement in terms of F-score on DSD corpus of 4.7% compared to the baseline, and 2.8% compared to the known state-of-the-art [18]. The achieved results produced with Ruslana features are slightly better than the results produced with EmoDB, which can be attributed to greater speaker variability in Ruslana and perhaps similar recording conditions. The most informative emotional states were anger, sadness and neutral condition - these results correlate with most of known psychological and practical literature (e.g. [14] and [15]). Moreover, it turns out that the presence of such emotional state as happiness also can be a marker of truthfulness/deceptiveness in speech. It can be explained by the idea that some people can use cheerful attitude to cover their real feelings, or feel elevated from adrenaline caused by their deception.

An interesting observation can be made on using estimated gender instead of the true labels: its effect seems to be the opposite for emotion recognition and deception detection. In emotion recognition, the predicted gender labels generally worsened the performance; however, in deception detection, estimated gender labels worked better than the true labels. This can be explained by the fact that in DSD corpus, true gender labels may be misleading: there are several speech instances with female voices heavily resembling male characteristics and vice versa. The gender recognition system groups voices with similar characteristics and therefore allows for more accurate modelling in the following steps. These findings open an opportunity for future research on deception detection differences between male and female speakers.

The proposed system can be used in spoken dialog systems and voice assistants that require an adaptation to individual voice characteristics and requests of every user/speaker. For example, the system can be applied to automatic addressee recognition that can be a part of a smart home and intelligent spaces that analyze spoken conversations between several people [36].

4. Acknowledgements

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730321P5Q0002), agreement No. 70-2021-00141.

5. References

[1] B. Schuller, A. Batliner, C. Bergler *et al.*, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *Proc. of INTERSPEECH*, 2020, pp. 2042–2046.

[2] B. Schuller, M. Wöllmer, F. Eyben *et al.*, "Semantic speech tagging: Towards combined analysis of speaker traits," in *Proc. of AES International Conference Semantic Audio*. Audio Engineering Society, 2011, p. 9.

[3] A. Poorjam, M. Bahari, and H. Van hamme, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *Proc. of International Conference on Computer and Knowledge Engineering (ICCKE)*, 2014, pp. 7–12.

[4] Y. Zhang, Y. Zhou, J. Shen *et al.*, "Semi-autonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6090–6094.

[5] Y. Zhang, F. Weninger, Z. Ren *et al.*, "Sincerity and Deception in Speech: Two Sides of the Same Coin? A Transfer- and Multi-Task Learning Perspective," in *Proc. of INTERSPEECH*, 2016, pp. 2041–2045.

[6] Y. Zhang, F. Weninger, B. Liu *et al.*, "A paralinguistic approach to speaker diarisation: using age, gender, voice likability and personality traits," in *Proc. of the 25th ACM international conference on Multimedia*, 2017, pp. 387–392.

[7] H. Zhao, N. Ye, and R. Wang, "Transferring age and gender attributes for dimensional emotion prediction from big speech data using hierarchical deep learning," in *Proc. of the 4th International Conference on Big Data Security on Cloud (BigDataSecurity), International Conference on High Performance and Smart Computing (HPSC), and International Conference on Intelligent Data and Security (IDS)*, 2018, pp. 20–24.

[8] O. Verkholyak, D. Fedotov, H. Kaya *et al.*, "Hierarchical two-level modelling of emotional states in spoken dialog systems," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6700–6704.

[9] M. Sidorov, A. Schmitt, E. Semenkin *et al.*, "Could speaker, gender or age awareness be beneficial in speech-based emotion recognition?" in *proceedings of Language Resources and Evaluation (LREC)*, 2016, pp. 61–68.

[10] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7179–7183.

[11] L. Zhang, L. Wang, J. Dang *et al.*, "Gender-aware cnn-blstm for speech emotion recognition," in *Proc. of International Conference on Artificial Neural Networks*. Springer, 2018, pp. 782–790.

[12] B. D. G. and S. S. E., "Predicting organizational effectiveness with a four-factor theory of leadership," *Administrative Science Quarterly*, vol. 11(2), p. 238–263, 1966.

[13] M. Zloteanu, "Emotions and deception detection," Ph.D. dissertation, University College London, 2017.

[14] P. Ekman, *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. W W Norton & Co., 2009.

[15] S. Amiriparian, J. Pohjalainen, E. Marchi *et al.*, "Is deception emotional? an emotion-driven predictive approach," in *Proc. of INTERSPEECH*, 2016, pp. 2011–2015.

[16] G. Mendels, S. Levitan, K. Lee *et al.*, "Hybrid acoustic-lexical deep learning approach for deception detection," in *Proc. of INTERSPEECH*, 2017, pp. 1472–1476.

[17] B. Schuller, S. Steidl, A. Batliner *et al.*, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity and native language," in *Proc. of INTERSPEECH*, 2016, pp. 2001–2005.

[18] A. Velichko and A. Karpov, "Automatic detection of deceptive and truthful paralinguistic information in speech using two-level machine learning model. computational linguistics and intellectual technologies," in *Proc. of the International Conference "Dialogue 2021"*, 2021, pp. 698–704.

[19] F. Burkhardt, M. Eckert, W. Johannsen *et al.*, "A database of age and gender annotated telephone speech," in *Proc. of Language Resources and Evaluation (LREC)*. Malta, 2010, pp. 1562–1565.

[20] M. Markitantov and O. Verkholyak, "Automatic recognition of speaker age and gender based on deep neural networks," in *Proc. of International Conference on Speech and Computer (SPECOM)*. Springer, 2019, pp. 327–336.

[21] F. Burkhardt, A. Paeschke, M. Rolfes *et al.*, "A database of german emotional speech," in *Proc. of INTERSPEECH*, 2005, pp. 1517–1520.

[22] V. Makarova and V. Petrushin, "Ruslana: a database of russian emotional utterances," in *Proc. of the 7th International Conference on Spoken Language Processing*, vol. 1, 2002, pp. 2041–2044.

[23] B. McFee, C. Raffel, D. Liang *et al.*, "librosa: Audio and music signal analysis in python," in *Proc. of the 14th python in science conference*, 01 2015, pp. 18–24.

[24] Q. Kong, Y. Cao, T. Iqbal *et al.*, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[25] M. Markitantov, "Transfer learning in speaker's age and gender recognition," in *Proc. of International Conference on Speech and Computer (SPECOM)*. Springer, 2020, pp. 326–335.

[26] M. Markitantov, D. Dresvyanskiy, D. Mamontov *et al.*, "Ensembling end-to-end deep models for computational paralinguistics tasks: Compare 2020 mask and breathing sub-challenges," *Proc. of INTERSPEECH*, pp. 2072–2076, 2020.

[27] B. Schuller, S. Steidl, A. Batliner *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH*, 2013, pp. 148–152.

[28] A. Bhavan, P. Chauhan, R. Shah *et al.*, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.

[29] Ž. Nedeljković, M. Milošević, and Ž. ĐJurović, "Analysis of features and classifiers in emotion recognition systems: Case study of slavic languages," *Archives of Acoustics*, vol. 45, no. 1, pp. 129–140, 2020.

[30] B. Schuller, A. Batliner, S. Steidl *et al.*, "The interspeech 2011 speaker state challenge," in *Proc. of INTERSPEECH*, 2011, pp. 3201–3204.

[31] N. Chawla, K. Bowyer, L. Hall *et al.*, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research (JAIR)*, vol. 16, pp. 321–357, 2002.

[32] L. Prokhorenkova, G. Gusev, A. Vorobev *et al.*, "Catboost: Unbiased boosting with categorical features," in *Proc. of the 32nd International Conference on Neural Information Processing Systems (NIPS)*, 2018, p. 6639–6649.

[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[34] D. Wang, Y. Zhang, and Y. Zhao, "Lightgbm: An effective mirna classification method in breast cancer patients," in *Proc. of the 2017 International Conference on Computational Biology and Bioinformatics*, 2017, p. 7–11.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] M. Nagano, Y. Ijima, and S. Hiroya, "Impact of Emotional State on Estimation of Willingness to Buy from Advertising Speech," in *Proc. of INTERSPEECH*, 2021, pp. 2486–2490.