Speech emotion, gender and age classification and speaker verification

Albina Klepach [1] **Assel Yermekova** [1] **Nina Konovalova** [1] **Dmitrii Korzh** [1]

Abstract

Accurately recognizing gender, age and emotion from speech is one of the most challenging problems as well as speaker verification task. A lot of studies related to the classification task have been carried out focusing on feature extraction and improving classifier design. In this research two different approaches were used: extracting features from speech melspectrogram and from audio using "wav2vec" embeddings. In this research we tested embeddings from classification task for verification task, but finally came up with the solution, provided by SpeechBrain toolkit, based on Emphasized Channel Attention, Propagation and Aggregation-Time Delay Neural Network (ECAPA-TDNN) and conducted several experiments. We conducted experiments on Common voice dataset (CV) of only Russian CV and combination of Slavic languages (Russian, Ukrain, Czech and Slovak) and with addition of TIMIT dataset.

Classification task

1. Introduction

With the fast development of automatic speech recognition and natural language processing, a number of spoken dialogue systems are being investigated to conduct specific tasks. Smart bots and smartphone assistants are widely used in different area of people lives. Correct identification of person's characteristics, such as gender, age or emotion, can improve the quality of its work.

A main stage in identifying speakers age, gender and emotion is to extract effective features that represent the speaker's characteristics uniquely. Another important stage is to build good classifier. A classifier uses the extracted features to predict the speakers' age and gender.

In this research two main approaches are represented:

1. Using mel-spectrogam and building the model, based on the article (Li et al., 2019);
2. Using Speeachbrain embeddings (Ravanelli et al., 2021), based on Wav2Vec approach (Schneider et al., 2019) and building the model, based on the article (Leonardo Pepino, 2021).

Both of these approaches will be discussed in details in the next section.

2. Algorithms and Models

2.1. Emotion parsing

As Common Voice and Timit datasets don't provide information about speakers' emotions (however, these are most helpful datasets for our tasks, as there are not a lot of free datasets with Russian language), we conducted emotion parsing using Selenium Python library on `https://speech-recog-demo.vsrobotics.ru/` web-site. This web-site provides a time-dependent probabilities of several classes of emotions for the audio. To simplify the task we took only one emotion, with highest probability.

We should admit, that results of this parsing are not stable.

2.2. Mel-spectrogram feature extraction

2.2.1. Spectrogram Extraction

Firstly, all audio were converted to mel-spectrogram with sample rate 16000 Hz, $hann$ window type, window time set at 22. Minimum frequency 1 Hz and 8162 Hz - maximum frequency were set in order to mimic the non-linear human ear perception of sound. Then all mel-spectrogram were padded (or cut) to the same length (in our case all mel-spectrogram were padded to the max length of all audio in dataset).

The spectrogram of a speech segment were denoted as $X = \{x_1, ...x_L\}$, where $x_i \in \mathbb{R}^{d_{spec}}$, L - is the tempo-

*Equal contribution [1]Skolkovo Institute of Science and Technology, Moscow, Russia. Correspondence to: Albina Klepach <Albina.Klepach@skoltech.ru>, Assel Yermekova <Assel.Yermekova@skoltech.ru>, Nina Konovalova <Nina.Konovalova@skoltech.ru>, Dmitrii Korzh <Dmitrii.Korzh@skoltech.ru>.

Final Report of the Industrial Immersion(Nanosemantics), Skoltech, Moscow, Russian Federation, 2021.

ral length of the spectrogram, and d_{spec} is the dimension of a spectrogram feature vector. We encode the spectrogram X as a fixed-length vector z, and conduct classifications on z.

2.2.2. CNN and Pooling Layer

This part consists of three Convolution layers and 3 Max-Pooling layers with stride 1. More details about parameters of CNN and MaxPooling layers can be founded in Table 1.

2.2.3. BLSTM Layer

The bidirectional LSTM network encodes global contexts by updating its hidden states recurrently. More details about parameters of BLSTM Layer can be founded in Table 1.

2.2.4. Self Attention Layer

After the BLSTM layer, a structured self attention network aggregates information from the BLSTM hidden states and produces a fixed-length vector as the encoding of the speech segment. More details about parameters of Attention Layer can be founded in Table 1.

Table 1. Dimension details.

Notation	**Meaning**	**Value**
d_{spec}	Number of spectrogram features	128
d_{cnn1}	Number of convolution filters for 1-st conv layer	64
d_{cnn2}	Number of convolution filters for 2-d conv layer	128
d_{cnn3}	Number of convolution filters for 3-d conv layer	256
d_{lstm}	Number of LSTM hidden units	64
o_{lstm}	Dropout for LSTM	0.5
d_{attn}	Hidden size of attention head	512
n_{cw}	Size of convolutional window	3
n_{pw}	Size of pooling window	2
n_{attn}	Number of attention head	16

2.2.5. Linear layers

As a final stage for classification three linear fully-connected output layers generate the probability distributions over emotions genders and age, respectively.

This approach was based on the article (Li et al., 2019). The full picture of the model presented on the Figure 1.

2.3. Wav2Vec feature extraction

Another approach based not on spectrogram but on Speechbrain embeddings (Ravanelli et al., 2021).

First of all Speechbrain classification embedding were used.

Table 2. Classes details.

Notation	Meaning	Value
num_g_class	Gender classes	2
num_e_class	Emotion classes	6
num_a_class	Age classes	5

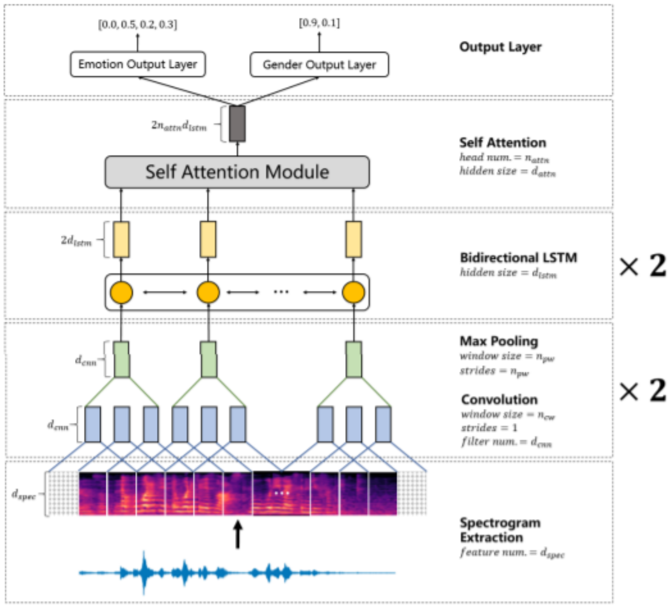


Figure 1. CNN BLSTM ATTENTION model

The remaining part of the model was constructed in two different ways:

1. Model with linear layer (Algorithm 1);
2. Model with lstm layer (Algorithm 2);

Algorithm 1 Model with linear layer

```
1. Linear layer ($n_l = 128$ - number of filters); ReLU layer;
2. DropOut layer ($p = 0.5$);
3. lstm layer (hidden size: 64,number of layers: 2 and bidirectional); ReLU layer;
4. DropOut layer ($p = 0.5$).
```

This approach was based on the following article: (Leonardo Pepino, 2021)

Algorithm 2 Model with lstm layer

```
1. Linear layer ($n_l = 128$ - number of filters); ReLU layer;
2. DropOut layer ($p = 0.5$);
3. Linear layer ($n_l = 128$ - number of filters); ReLU layer;
4. DropOut layer ($p = 0.5$).
```

2.4. Datasets

We verify our approach on the publicly available datasets and their combinations and our own voices. The diagrams for each dataset is presented in section of **??** Supplementary Material.

We prepossessed datasets, simply dropped all the entries with NaNs.

2.4.1. Common Voice

The first dataset which was taken into consideration was Common Voice (Common Voice). This is an open source multi-language dataset of voices for different speech-enabled applications. We were using the recordings of Russian language.

2.4.2. Combination of Slavic Datasets of Common Voice (International CV)

Russian CV was unbalanced with respect to the gender. Initially dataset contained 22159 males 3609 females. Therefore we decided to add data from other Slavic languages to balance it.

2.4.3. OurDataset

Despite the fact, that accuracy for gender and age was ≈ 0.9 on CV, the results on our voices were not so good as it was expected.Therefore we decided to build our own custom dataset, that contains recordings of our relatives and our voices, recordings of several famous people's voices from the Internet and CV utterances. OurDataset was used only for tests and not for training.

2.4.4. TIMIT

To increase diversity of voices, we also added TIMIT dataset to Slavic dataset. TIMIT contains recordings of 630 individuals/speakers with 8 different American English dialects, with each individual reading upto 10 phonetically rich sentences.

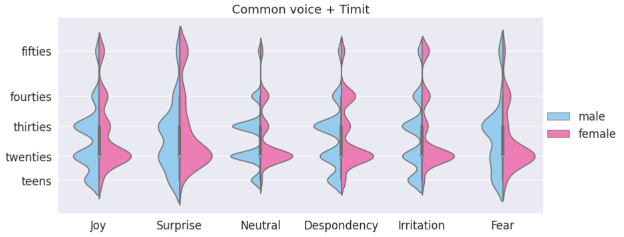


Figure 2. Diagram of final Dataset (Common voice + Timit)

2.5. Experiments

Experiments were conducted for three main types of models:

1. CNN-BLTM-Attention model;
2. SpeechBrain embeddings with linear layer model;
3. SpeechBrain embeddings with lstm layer model.

For each type of models we conduct experiment with different scaling:

1. without any normalization;
2. mean-std normalization $\frac{\text{array}-\text{mean(array)}}{\text{std(array)}}$;
3. Standardize $\frac{\text{array}-\text{min(array)}}{\text{max(array)}-\text{min(array)}}$.

.

CNN-BLTM-Attention model was tested only on Russian commonvoice dataset, SpeechBrain embeddings with lstm layer model was tested on Russian and Russian + another languages Common Voice datasets and SpeechBrain embeddings with linear layer model was tested on Russian, Russian + another languages commonvoice and combination of these datasets with TIMIT dataset.

The experiments were conducted in Python via PyTorch framework.

All models were trained with Adam optimizer for **30 epochs** and the batch size of 64. The initial learning rate was 10^{-3} with scheduler = 0.1 every 8 epochs.

2.6. Experiment results

The tables below represents the main results for our experiments. Also additional graphics of losses and accuracy, depending on epochs are represented in supplementary materials.

Table 3. Gender accuracy for different models for train, val and test datasets (Russian language)

	train	val	our test
CNN-LSTM-Attention	0.987	0.985	-
SpeechBrain+linear layer	0.994	0.996	0.87
SpeechBrain+lstm layer	0.995	0.990	0.89

Table 4. Age accuracy for different models for train, val and test datasets (russian language)

	train	val	our test
CNN-LSTM-Attention	0.868	0.861	-
SpeechBrain+linear layer	0.929	0.926	0.355
SpeechBrain+lstm layer	0.873	0.886	0.355

Verification task

3. Verification

3.1. Intro

Speaker verification task is 1-1 check for the specific enrolled voice and the new voice. This task needs higher accuracy than speaker identification which is N-1 check for N enrolled voices and a new voice.

There are two types of speaker verification: text dependent speaker verification (TD-SV) and text independent speaker verification (TI-SV). The former uses text specific utterances for enrollment and verification, whereas the latter uses text independent utterances. In our investigation we came up with TI-SV solution.

3.2. Pipeline description

The first idea was to use the model for age/gender/emotion classification task with an addition of a new linear layer and adjusting the model's last layers using cosine similarity between vectors as loss for speaker verification (or at least cross-entropy loss).

However, in 2021, an open-source and special to speech related tasks toolkit based on PyTorch, SpeechBrain was released (Ravanelli et al., 2021). SpeechBrain provides different models for speaker recognition and identification. In the research to perform speaker verification a pretrained ECAPA-TDNN model was used (Desplanques et al., 2020). It is a combination of convolutional and residual blocks. The embeddings are extracted using attentive statistical pooling. The system is trained with Additive Margin Softmax Loss. Speaker Verification is performed using cosine distance between speaker embeddings. It is trained on Voxceleb 1 + Voxceleb 2 training data.

Table 5. Emotion accuracy for different models for train, val and test datasets (russian language)

	train	val	our test
CNN-LSTM-Attention	0.52	0.518	-
SpeechBrain+linear layer	0.554	0.574	0.22
SpeechBrain+lstm layer	0.448	0.463	0.27

Table 6. Gender accuracy for different models for train, val, test and our test datasets for big datasets

	train	val	test	our test
SpeechBrain+linear layer for Common Voice	0.991	0.994	1.0	0.903
SpeechBrain+lstm layer for Common Voice	0.990	0.992	1.0	0.919
SpeechBrain+linear layer for Common Voice+Timit	0.991	0.992	0.917	0.935

3.3. Datasets

For experiments and hyper-parameters tuning Mozilla Common Voice dataset of Russian language was used as well as our small custom dataset described in section 2.4.3. (OurDataset).

3.4. Speaker verification task's experiments

3.4.1. Finding equal error rate (EER)

The model receives two paths to audios as inputs and answers whether there is the same speaker presented on two audios or not and provide confidence (a number from 0 to 1, cosine similarity). As we started our experiments with pretrained model on (probably) English voices with threshold for cosine similarity around 0.25 we decided to test this model on Ru CV and find suitable threshold for us.

For this purpose, we were looking for False Accept - False Reject rates trade-off:

False Accept rate – percentage of situations when the model incorrectly predicts audios to be spoken by the same person, comparing audios of different persons. False Reject rate – percentage of situations when the model incorrectly predicts them to be spoken by the same person, comparing audios of the same persons.

We obtained such result 3. However, later we found out that for threshold around 0.4 we obtained quite high FRR ($\approx$ 0.2) and decreased it to 0.3, having the suitable almost suitable FAR $\approx$ 0.03 and FRR $\approx$ 0.08. All experiments we conducted on rather small random subgroups of dataset (Ru CV has $\sim$17000 utterances, totally we can test model on 17000 $\times$ 17000 comparisons, but we used only several

Table 7. Age accuracy for different models for train, val, test and our test datasets for big datasets

	train	val	test	our test
SpeechBrain+linear layer for Common Voice	0.894	0.900	0.88	0.420
SpeechBrain+lstm layer for Common Voice	0.838	0.857	0.92	0.354
SpeechBrain+linear layer for Common Voice+Timit	0.843	0.858	0.834	0.420

Table 8. Emotion accuracy for different models for train, val, test and our test datasets for big datasets

	train	val	test	our test
SpeechBrain+linear layer for Common Voice	0.494	0.495	0.52	0.242
SpeechBrain+lstm layer for Common Voice	0.456	0.460	0.37	0.227
SpeechBrain+linear layer for Common Voice+Timit	0.590	0.600	0.66	0.220

thousands comparisons to decrease time).

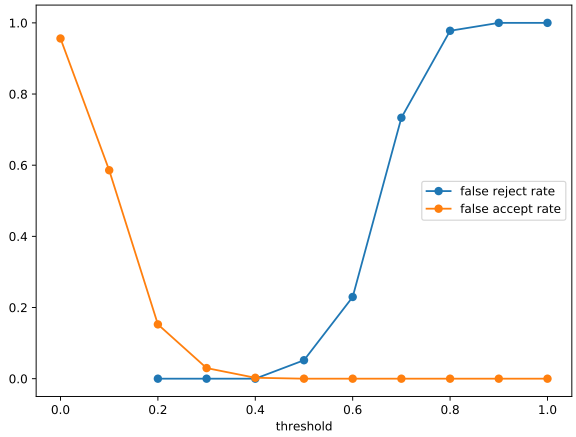


Figure 3. Estimation of FAR-FRR trade off on Russian Common Voice dataset

3.4.2. Verification's quality dependency on audio time

The shorter audios we can use for a reliable verification, the better. Average length of Ru CV utterance ≈ 5.7 s, std ≈ 1.7.

We conducted several experiments and observed, that for 5s results are pretty satisfactory.

Verification results on OurDataset with threshold 0.3: FRR = 0%, FAR = 6%.

Table 9. Verification dependency on audio length

Time	Threshold	FAR	FRR
1s	0.3	0.02	0.33
3s	0.3	0.03	0.16
5s	0.3	0.04	0.03

The experiments were conducted in Python via PyTorch framework on GPUs and Intel(R) Xeon(R) Platinum 8168 CPU @ 2.70GHz and one Tesla V100 GPU.

4. Docker and web-site

We created a local web-site, that can be launched with Dockerfile. Figures of this site you can see in supplementary materials.

5. Conclusion

Dealing with classification task we faced a problem of dependency of our model on the dataset. Our models shows quite good results of age and gender recognition on the dataset, the models were trained on. But on the dataset, made of audio from another source of information the quality decrease a lot. That means that it is better to made bigger dataset from different sources.

Emotion prediction shows not really good results and one of the problem is the initial data markup.

Regarding the verification task, SpeechBrain works well and language independently. To achieve even better results one can fine-tune the model on his/her own dataset or simply find the most suitable threshold. On the other hand, one can change parameters of the model and train it from scratch.

The future work may focus on more qualified emotion marking, increasing the size and diversity of datasets, setting up experiments of new architectures. Also one can try to estimate the robustness of verification, e.g. to the spoofing attacks and to voice-cloning technologies.

References

Desplanques, B., Thienpondt, J., and Demuynck, K. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech 2020*, Oct 2020. doi: 10.21437/interspeech.2020-2650. URL `http://dx.doi.org/10.21437/Interspeech.2020-2650`.

Leonardo Pepino, Pablo Riera, L. F. Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. 2021.

Li, Y., Zhao, T., and Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In *Proc. Interspeech 2019*, pp. 2803–2807, 2019. doi: 10.21437/Interspeech.2019-2594. URL `http://dx.doi.org/10.21437/Interspeech.2019-2594`.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. Speechbrain: A general-purpose speech toolkit, 2021.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. wav2vec: Unsupervised pre-training for speech recognition. Apr 2019. doi: http://doi.org/10.21437/Interspeech.2019-1873. URL `https://arxiv.org/abs/1904.05862`.

A. Supplementary materials

In this section we want to provide some experiments, that weren't considered in the main part of the report.

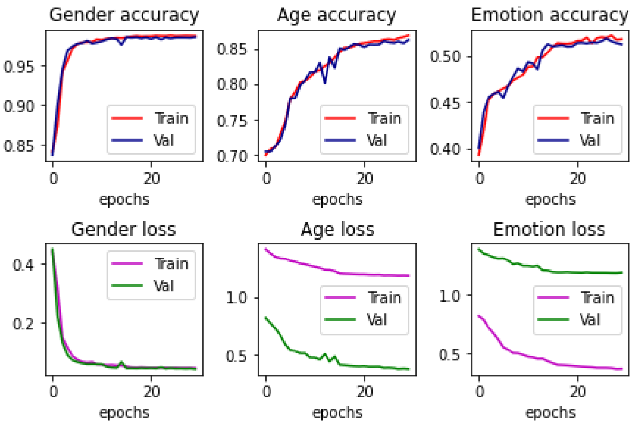


Figure 4. Age, gender and emotion losses and accuracy for train and val datasets for CNN-BLTM-Attention model

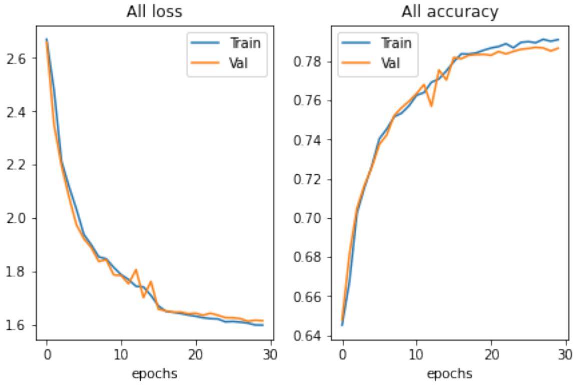


Figure 5. All loss and accuracy depending on epochs for CNN-BLTM-Attention model

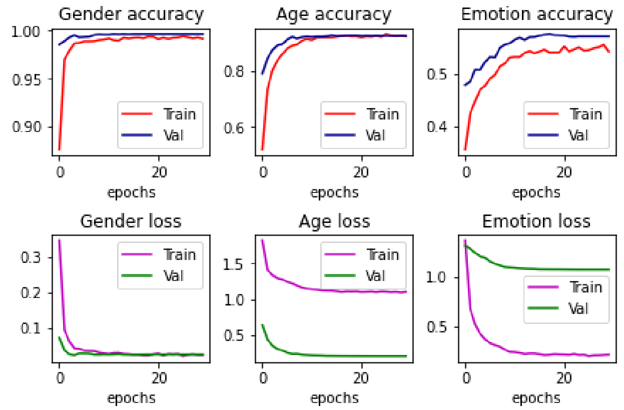


Figure 6. Age, gender and emotion losses and accuracy for train and val datasets for SpeechBrain embeddings model with linear layer for Russian common voice dataset

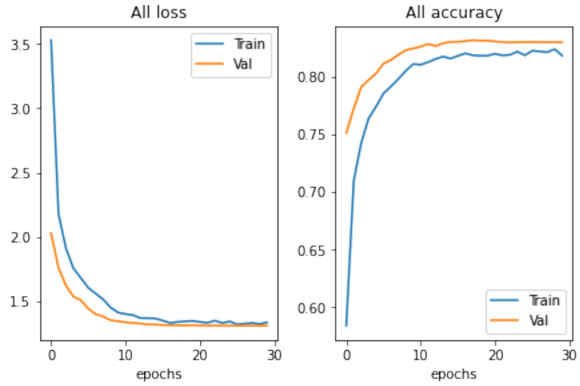


Figure 7. All loss and accuracy depending on epochs for SpeechBrain embeddings model with linear layer for Russian common voice dataset

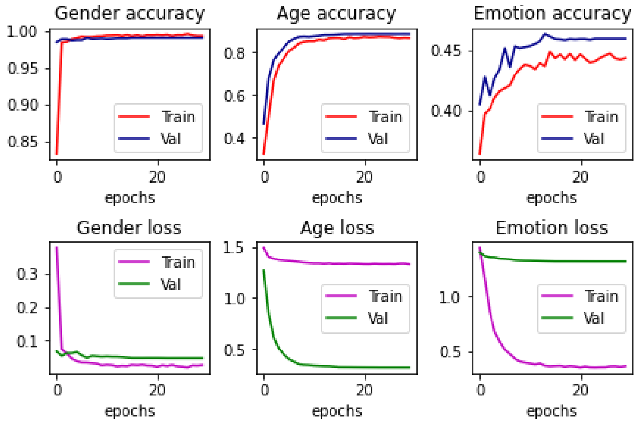


Figure 8. Age, gender and emotion losses and accuracy for train and val datasets for SpeechBrain embeddings with lstm layer model for Russian common voice dataset

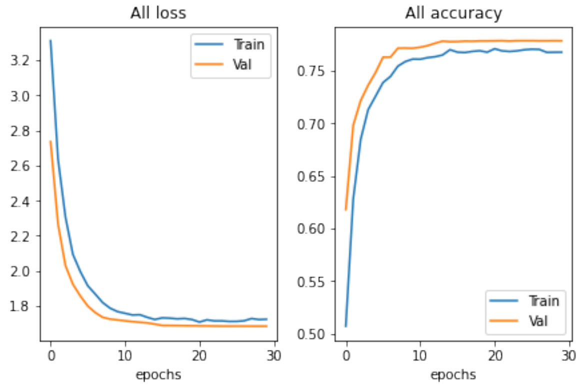


Figure 9. All loss and accuracy depending on epochs for SpeechBrain embeddings with lstm layer model for Russian common voice dataset

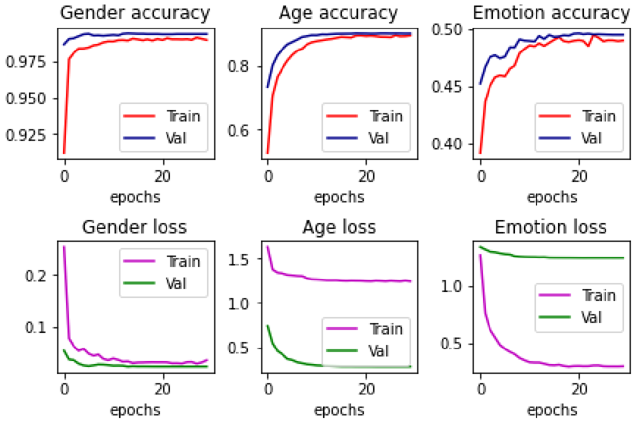


Figure 10. Age, gender and emotion losses and accuracy for train and val datasets for SpeechBrain embeddings with linear layer model for different language common voice dataset

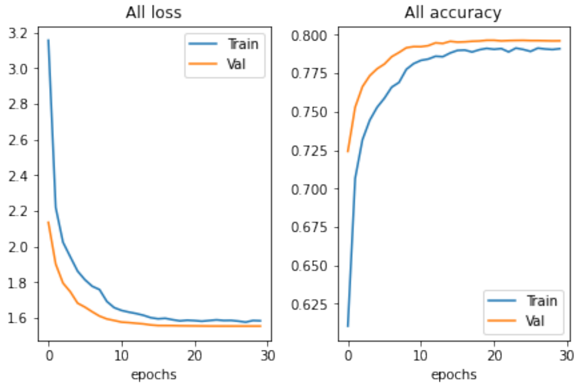


Figure 11. All loss and accuracy depending on epochs for SpeechBrain embeddings with linear layer model for different language common voice dataset

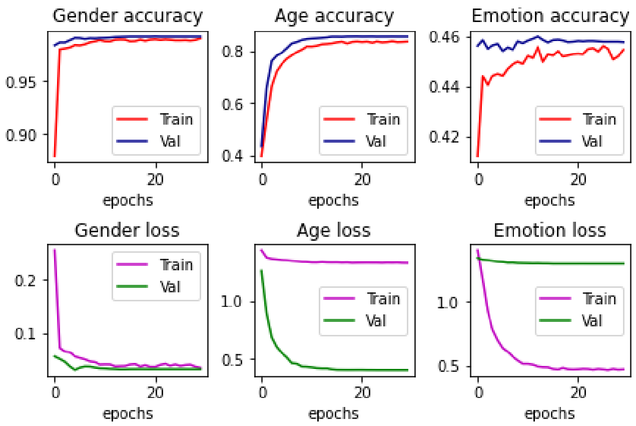


Figure 12. Age, gender and emotion losses and accuracy for train and val datasets for SpeechBrain embeddings with lstm layer model for different language common voice dataset

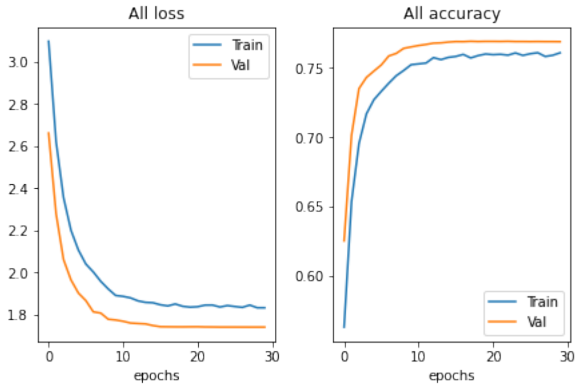


Figure 13. All loss and accuracy depending on epochs for SpeechBrain embeddings with lstm layer model for different language common voice dataset

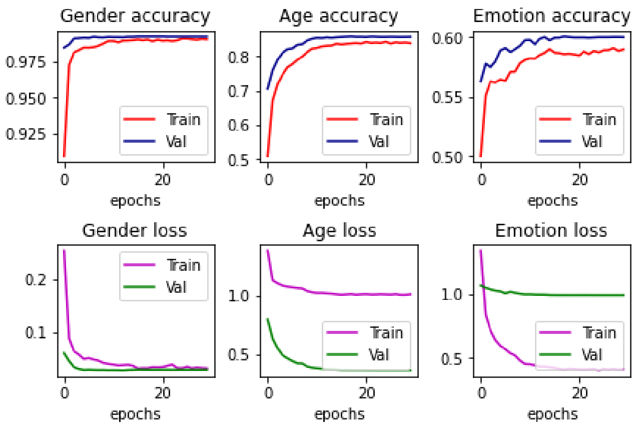


Figure 14. Age, gender and emotion losses and accuracy for train and val datasets for SpeechBrain embeddings with linear layer model for different language common voice and Timit datasets

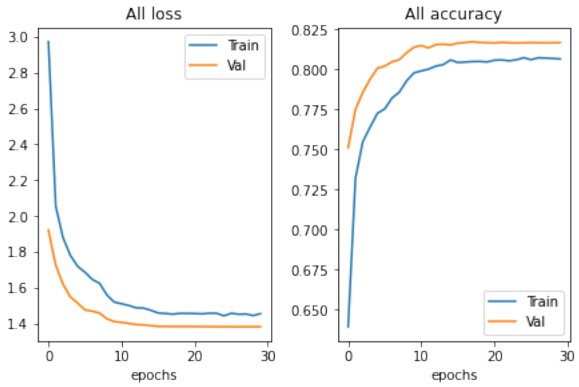


Figure 15. All loss and accuracy depending on epochs for SpeechBrain embeddings with linear layer model for different language common voice and Timit datasets

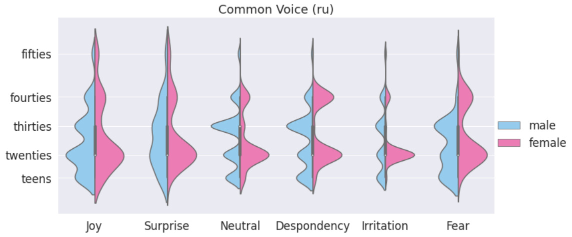


Figure 16. Diagram of Common Voice Dataset(Russian language)

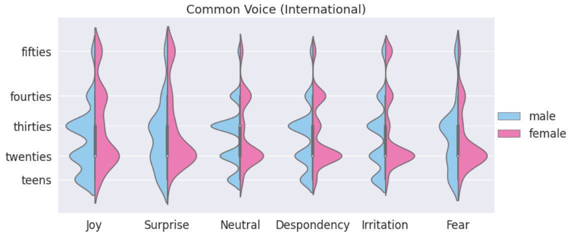


Figure 17. Common Voice

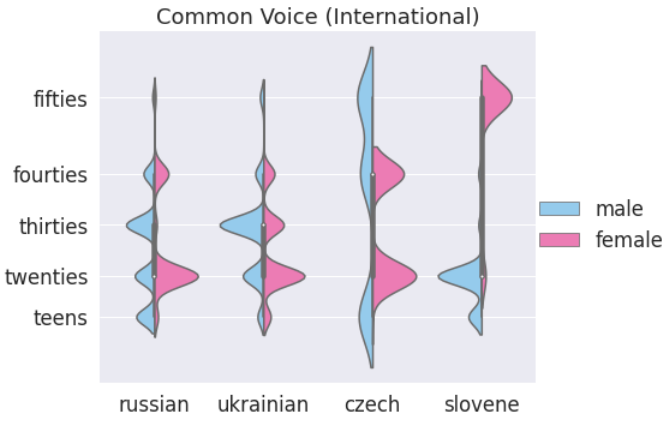


Figure 18. Diagram of Common Voice Dataset (languages

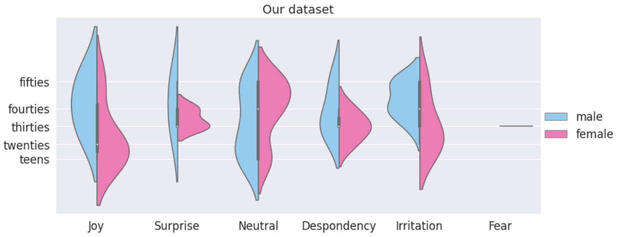


Figure 19. Diagram of the dataset from our voices and voices of celebrities

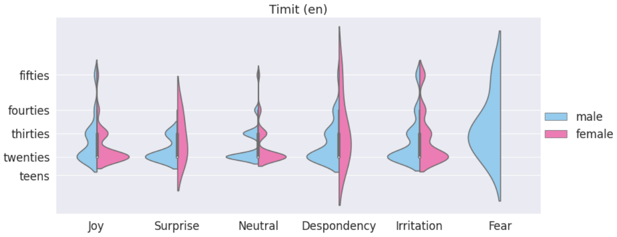
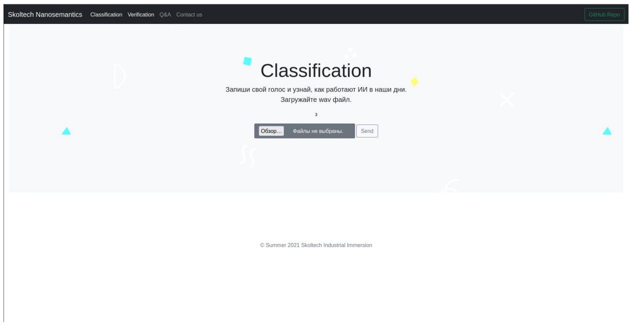


Figure 20. Diagram of TIMIT Dataset (en)



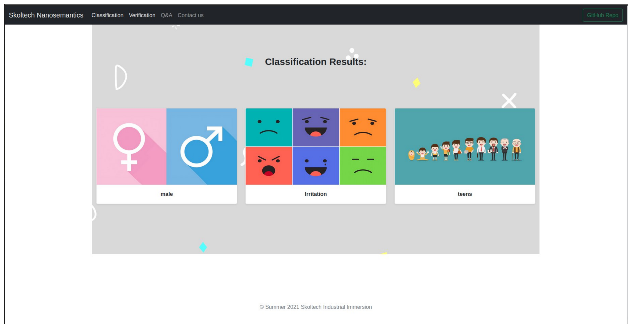
Classification

Запиши свой голос и узнай, как работают ИИ в наши дни. Загружайте wav файл.

з

© Summer 2021 Skoltech Industrial Immersion

Figure 21. Page for classification on the site



Classification Results:

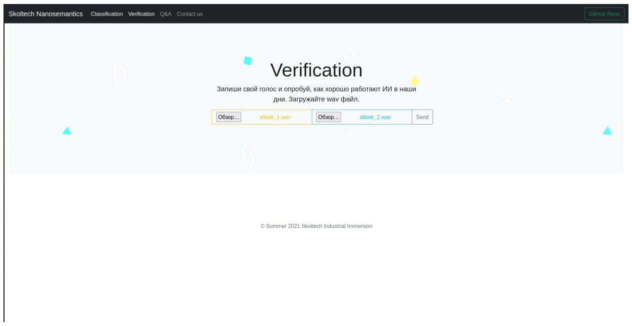
male

Irritation

teens

© Summer 2021 Skoltech Industrial Immersion

Figure 22. Results of classification on the site

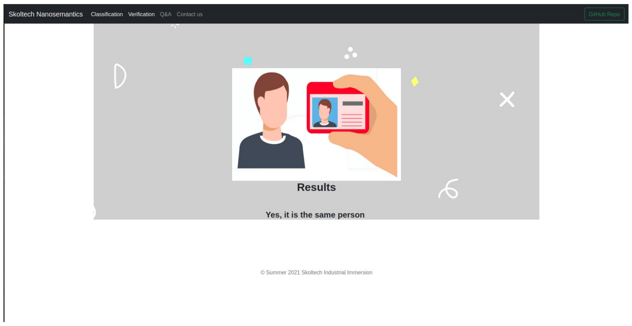


Verification

Запиши свой голос и опробуй, как хорошо работают ИИ в наши дни. Загружайте wav файл.

© Summer 2021 Skoltech Industrial Immersion

Figure 23. Verification page on the site



Results

Yes, it is the same person

© Summer 2021 Skoltech Industrial Immersion

Figure 24. Results of verification on the site