Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/140108

Speech Emotion Recognition For Kazakh And Russian Languages

*Kanat Kozhakhmet**, *Rakhima Zhumaliyeva, Aisultan Shoiynbek and Nazerke Sultanova*

Suleyman Demirel University, Kaskelen, Kazakhstan

Received: 2 Jun. 2019, Revised: 2 Aug. 2019, Accepted: 6 Aug. 2019
Published online: 1 Jan. 2020

Abstract: In the age of information and automation, where the robotics sphere is increasing each day, and people interact with automation systems, emotion recognition mechanisms will play an important role for better interactions between humans and machines. The emotion recognition is very important in AI spheres since it will make human-computer interface (HCI) more user-friendly and similar to the real-man behavior. The audio emotion corpus of Kazakh and Russian languages which has more than 16 000 records on 8 type of emotions is collected during the research. One hundred and one participants participated in the assembly of the corps. Extensive amount of work has been done related to sorting and human recognizing of emotion using the majority voting method. The data set was divided into TRAIN (80%), VALIDATION (10%) and TESTING (10%) sets. For this problem, Deep Neural Network has been applied with Stochastic Gradient Descent optimizer with batch normalization. MFCC feature supported by the LIBROSSA library was used as an input.

Keywords: MFCC, spectral audio features, speech emotion recognition, Deep Neural Networks, Kazakh language, Russian language

1 Introduction

Emotion recognition is one of the progressive tasks in the field of AI, as people are emotional beings and it is important that machines react and act taking into account person's emotions. Emotion can be recognized from various sources such as face analysis, skin temperature, galvanic resistance, and gesture recognition [2]. SER is important part of ER and can be useful in interactive systems such as in call center applications, virtual reality, tutoring systems, psychiatric aids, and interactive games, among others. Apart from this, emotion recognition could also form an essential step towards personal wellness studies such as automatic detection of fatigue, stress, depression, and above all, in identifying personality. Considering the progress in speech synthesis, inclusion of emotions can increase the naturalness of the synthesized speech.

The aim of this paper is to create the emotional corpus in Kazakh and Russian. After creating this emotions dataset, it is necessary to recognize emotions using deep neural networks (DNN) models.

The organization of this work is as follows: section 1 describes the introduction, section 2 describes related works and literature review. In sections 3 and 4, preprocessing and the architecture using deep learning is described. Results are shown in section 5, and section 6 concludes the work.

2 Literature Review

Speech Emotion Recognition (SER) aims to automatically identify the current emotional state of a person from his or her speech [3]. In speech, discrete emotion expressions are associated with characteristic variations in the acoustic structure of the speech signals and the relative perturbation of specific acoustic cues over the course of an utterance [4]. During speech emotion analysis, these vocal cues are extracted from speech as a marker for the emotional state by assuming that there are objectively measurable cues that can be used for emotion recognition [5]. Vast amount of work has been done to determine the factors that influence emotion identification from expressive speech such as gender and age [6]. While analyzing the influences of language in SER, Pell [4] has highlighted the importance of acoustic data such as fundamental frequency and speaking rate for indicating vocal emotion in languages. In research [2] the KNN

* Corresponding author e-mail: nazerke.sultanova@sdu.edu.kz

model was trained on the basis of Berlin Emotional Database (EMO-DB) [7] to predict some pool of emotional sentences in English, Malay, Mandarin languages. In a result accuracy levels were 78.5%, 71.0% and 72.5% respectively, but they used four emotions (sad, happiness, anger, neutral). Emotion recognition system based on audio (which can also be seen as voice-based) has very low requirement for hardware, even though multi modal speech processing can improve speech related system performance [8], [9]. Therefore, the audio-based emotion recognition system is easier to be employed on AI products [10], [11] than other means. However, current voice-based AI products, such as Siri, Google voice search and Cortona, lack of emotion recognition capability, which make people feel them as "machine". This shows the importance of exploring the emotion recognition system.

Many researches have been done for decades [10], [12], [13], [14]. So far, most of the work has been done on the data collected in the studio environment. The data collection was well controlled, therefore the data is clean and well segmented. But SER strong dependent from language. Emotional corpuses in Russian in the public domain are very few and in the Kazakh language there are almost none.

3 Preprocessing

The dataset is collected from one hundred and one people. 48 of them are women (average age 20 year), 53 are men (average age 21 year). On a national basis, 99.1% of them are Kazakhs, and 0.9% are Russian people, where all participants are not professional actors.

20 sentences (10 in Kazakh, 10 in Russian) were compiled. All sentences are shown in Table 1. Each sentence is uttered in eight emotions (anger, boredom, disgust, happiness, sadness, neutral, fear). As a result, 160 sentences in Kazakh and Russian language from one person was obtained. Final dataset contains 16160 audio files in wav extension and more than 14 hours of duration. Emotions are recorded on Dictaphone of mobile phone with different number of sample and later down sampled do 16kHz (mono).

After getting the dataset, several audio files are randomly listened. As expected, many emotions do not match their labels. This issue is dealt with the fact that the participants are not professional actors, and their emotions are artificially caused.

Some files also have a background noise. To solve this issue, the majority voting method is applied. We have collected five focus groups (nine people in the first two groups, in the remaining groups by ten people). The dataset is divided into five equal parts and distributed between groups. In each group, each person had to listen to the emotion and assign the label to it manually, based on his feelings and ability to recognize the emotion. Thus, listening to the same file in the group, each person votes

Table 1: Kazakh and Russian sentences for FOREIGN dataset.

N	Kazakh language	Russian Language	Meaning in English
1	Оларға сен мұны қалай істедін?	Как ты мог так поступить с ними?	How could you do this to them?
2	Ол маған қарап тұрды, бірақ байқамады.	Он смотрел на меня в упор, и не замечал.	He looked at me point-blank, and did not notice.
3	Бүгін менде емтихан болады, сондықтан мені алаңдатпаңыз, мен дайындап отырмын	Сегодня у меня экзамен, поэтому не беспокойте меня, я готовлюсь	Today I have an exam, so do not touch me, I am preparing
4	Мен кеше кешке хат жібердім	Я отправил письмо еще вчера вечером	I sent the letter last night
5	Олар оны жоғары көтеріп, енді қайтадан түсіп кетті	Они просто подняли его наверх, и теперь они снова спускаются	They just lifted him upstairs, and now they descend again.
6	Демалыс күндері мен әрқашан үйге оралып, Арсенді көрдім	В выходные дни я всегда возвращался домой и видел Арсена	On weekends I always came back home and saw Arsen
7	Ол әрқашан оны сақтайтын жерде болады	Он будет в том месте, где мы всегда храним его.	He will be in the place where we always keep him
8	Ол сәрсенбі күні келеді.	Она придет в среду.	She will come on Wednesday.
9	Төсек үстел тоңазытқышта орналасқан.	Скатерть лежит на холодильнике.	The tablecloth is on the fridge.
10	Бүгін кешке мен оған айта аламын.	Сегодня я мог сказать ему.	Today I could tell him.

Table 2: Comparative analysis of spectral features.

N	Accuracy(%) Epoch 5	Accuracy(%) Epoch 10	Accuracy(%) Epoch 20	Accuracy(%) Epoch 40
1	70.7	67.92	76.41	82.07

for assigning a label. The label for each file is determined by the majority vote. Due to this approach, we got a completely new dataset. Finally, we have listened all happiness and anger records of each participant and select only 20 participants based on emotion anger and happiness, because they are extremely high emotional coloring.
Data set is divided into training set 80%, developing set 10%, test set 10%. The training set has 3 emotions: happiness with 308 records, anger with 208 records, neutral with 332 records. The developing set has: happiness 38 records, anger 27 records, neutral 41 records. The test set has: happiness 38 records, anger 27 records, neutral 41 records.

4 Feature extraction and DNN architecture

For feature extraction we have used LibROSA [15] python package and converted all to MFCC feature. The chosen DNN architecture contains six fully-connected layers with activation function relu [16], and last layer is also fully-connected but with activation function softmax. The structure is as follows: 1 – 320 neurons, 2 – 160 neurons, 3- 80 neurons, 4 – 40 neurons, 5- 20 neurons, 6 – 10 neurons. The last layer with activation function softmax contains 3 neurons. For the regularization of the DNN, we have used a 0.2 dropout [17] between the third and the fourth layers and batch normalization before the first layer. All layers are initialized using Glorot uniform initialization [18]. The detailed information about the architecture is shown in Figure 1.

For training the proposed model we utilized Stochastic Gradient Descent algorithm with the fixed learning rate of 0.11 to optimize a Binary cross entropy loss function also known as logloss [19]. Metrics of model is the accuracy. The input data are presented to the DNN in batches of size 16 in multiple epochs (iterations).

5 Results

For each test set, a model with a different number of epochs is trained. The results of the accuracy in the context of 5, 10, 20, 40 epochs are displayed in the Table 2. Based on these results, the best accuracy of emotion recognition is 82.07%. The confusion matrix is given in the Table 3.

In Test set, the anger emotion was confused 7 times with hapiness, and once with the neutral emotion. Neutral emotion was confused 3 times with happiness and 3 times with anger. Hapiness emotion was confused 3 times with anger and 3 times with neutral.

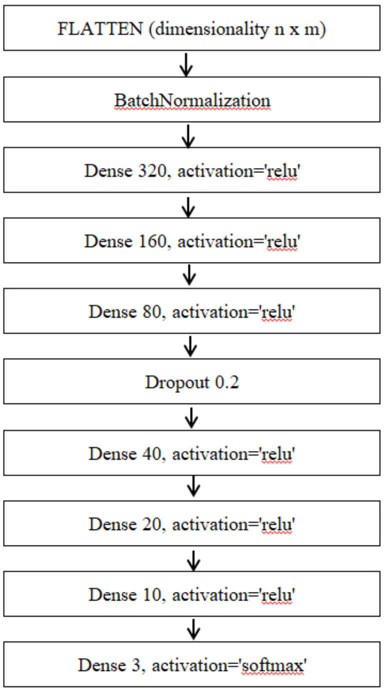
FLATTEN (dimensionality n x m)
BatchNormalization
Dense 320, activation='relu'
Dense 160, activation='relu'
Dense 80, activation='relu'
Dropout 0.2
Dense 40, activation='relu'
Dense 20, activation='relu'
Dense 10, activation='relu'
Dense 3, activation='softmax'

Figure 1: Detailed architecture of proposed DNN

Table 3: Confusion matrix of prediction emotion on German language

Test set	anger	happiness	neutral
anger	21	3	3
happiness	7	28	3
neutral	1	2	38

6 Conclusion

The aim of this work is to create an emotional corps in Kazakh and Russian and to recognize emotions using the DNN model. The dataset with emotion audio records in Kazakh and Russian languages is created and used in this study. The dataset consists of more than 16.000 records. Based on completed work, we have got new emotion corpus and have recognized the emotions with accuracy 82.07%. The model often confuses anger emotion with happiness emotion, it is related with extremely high emotional coloring, and its prosody could be easy to confuse. It is necessary to highlight that the 82.07% of accuracy is the strong result for speech emotion recognition in Kazakh and Russian languages.

References

[1] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman. *Emotion Recognition in Speech using Cross-Modal Transfer in the Wild*. in Proc. 26th ACM international conference on Multimedia.Association for Computing Machinery, New York, NY, USA, 292–301, (2018).

[2] R. Rajoo and C. C. Aun. *Influences of languages in speech emotion recognition: A comparative study using Malay, English and Mandarin languages*. in Proc. 2016 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE), Batu Feringghi, 35-39, (2016).

[3] R.W. Picard. *Affective computing. Technical Report 321*. MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, (1995).

[4] M.D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz. Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, 37, 417-435, (2009).

[5] M. Gjoreski, H. Gjoreski, and A. Kulakov. Machine Learning Approach for Emotion Recognition in Speech. *Informatica*, 38, 377–384, (2014).

[6] S. Paulmann, M.D. Pell, and S.A. Kotz. How aging affects the recognition of emotional speech. *Brain and Language*, 104, 262–269, (2008).

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. *A database of German emotional speech*. in Proc. Interspeech, 5, 1517-1520, (2005).

[8] F. Tao and C. Busso. *Lipreading approach for isolated digits recognition under whisper and neutral speech*. in Proc. Interspeech 2014, 1154–1158, (2014).

[9] F. Tao and C. Busso. *Bimodal recurrent neural network for audiovisual voice activity detection*. in Proc. Interspeech 2017, 1938–1942, (2017).

[10] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. *Analysis of emotion recognition using facial expressions, speech and multimodal information*. in Proc. Sixth International Conference on Multimodal Interfaces ICMI 2004, 205–211, (2004).

[11] B. Schuller, G. Rigoll, and M. Lang. *Hidden Markov model-based speech emotion recognition*. in Proc. ICASSP 2003, 2, 1–4, (2003).

[12] F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech, in Proc. ICSLP 1996, 3, 1970–1973, (1996).

[13] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan. *Emotion recognition based on phoneme classes*. in Proc. ICSLP 2004, 889–892, (2004).

[14] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, D. Grandjean, and B. Schuller. *Fisher kernels on phase-based features for speech emotion recognition*. in Proc. Dialogues with Social Robots, 195–203, (2017).

[15] McFee, Brian, C. Raffel, D. Liang, D. PW Ellis, M. McVicar, E. Battenberg, and O. Nieto. *librosa: Audio and music signal analysis in python*. in Proc. 14th python in science conference, 18-25, (2015).

[16] V. Nair and G.E. Hinton. *Rectified linear units improve restricted boltzmann machines*. in Proc. 27th International Conference on Machine Learning (ICML-10), 807-814, (2010).

[17] N. Srivastava, G.E. Hinton, A. Krizhevsky,I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929-1958, (2014).

[18] X. Glorot, and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks*. in Proc. Aistats, 9, 249- 256, (2010).

[19] T. Zhang. *Solving large scale linear prediction problems using stochastic gradient descent algorithms*. in Proc. Twenty-first international conference on Machine learning, 116, (2004)

Kanat Kozhakhmet PhD, Assoc. Professor at Astana IT University, Nur-Sultan, KAZAKHSTAN

Rakhima Zhumaliyeva PhD, Assoc. Professor at Faculty of Education and Humanities, Abylaikhan Street, No:1/1, Karasai district, Kaskelen, Almaty, KAZAKHSTAN

Aisultan Shoiynbek PhD student at Faculty of Engineering & Natural Sciences, Abylaikhan Street, No:1/1, Karasai district, Kaskelen, Almaty, KAZAKHSTAN

Nazerke Sultanova PhD student at Faculty of Engineering & Natural Sciences, Abylaikhan Street, No:1/1, Karasai district, Kaskelen, Almaty, KAZAKHSTAN