**УДК 004.85**

**Sarsembayev A.A., Tolganbayeva G.A., Janybekova S.T.**

International Information Technology University, Almaty, Kazakhstan

# SOLVING EMOTION CLASSIFICATION PROBLEM USING DEEP LEARNING

***Abstract**: Speech emotion classification is one of the most interesting and complicated problems in today's world. One of the main obstacles to this task is that emotions are subjective and difficult to capture. In this paper, we proposed deep learning methods that solve emotion classification problems based on audio streams. Three methods are propagated and compared throughout the paper. Within the first method a Multilayer Perceptron model was built. A second method shows decreased accuracy building Long Short Term Memory models. Finally, the third method that reached the best accuracy among others is convolutional neural network models. A speech corpus consisting of acted and spontaneous emotion samples in English language is described in detail. This dataset was tested and trained using these proposed methods. The CNN model for our emotion classification problem achieved a validation accuracy of 70%.*

***Key words:** Speech emotion recognition, convolutional neural network, deep neural network, long short-term memory, multilayer perceptron.*

## Introduction

Deep neural networks are quickly becoming a fundamental component of high performance speech recognition systems. Deep neural network (DNN) acoustic models perform substantially better than the Gaussian mixture models (GMMs) typically used in large vocabulary continuous speech recognition (LVCSR)[1].

Emotions are a tool for expressing an opinion, an existing situation or one's psychological state. Some of the emotions include negative, positive, neutral, static, dynamic and can be used as input to the human-computer interaction system for accurate recognition. The importance of automatic recognition of emotions in human speech has increased with the growing role of oral language interfaces in this area to make them more effective.

The most commonly used acoustic features in the literature are LPC features, prosody features like pitch, intensity and speaking rate. Although it seems easy for a human to detect the emotional classes of an audio signal, researchers have shown an average score of identifying different emotional classes such as neutral, surprise, happiness, sadness and anger. Emotion recognition is one of the fundamental aspects to build a man–machine environment that provides the theoretical and experimental basis of the right choice of emotional signal for understanding and expression of emotion. Emotional expressions are continuous because the expression varies smoothly as the expression is changed. The variability of expression can be represented as amplitude, frequency and other parameters. But the emotional state is important in communication between humans and has to be recognised properly [2].

In this paper we explore various deep learning based architectures to get the best individual detection accuracy from each of the different modes. Our work consists of Long Short Term Memory networks, Convolution Neural Networks, fully connected Multi-Layer Perceptrons and we complement them using techniques such as Dropout, adaptive optimizers such as Adam and pre trained word-embedding models.

The paper is structured as follows. Section 1 introduces the importance of this work. Section 2 represents the related literature. We discussed the work background and related work of the deep learning and conventional methods. The proposed methods have been explained in Section 3. Section 4 discusses the evaluation criteria and the results. Finally, Section 5 concludes the work.

**Relation to prior work**

Speech emotion recognition is a challenging task due to its complexity of emotional expressions. Researchers have developed various methods for speech emotion recognition based on traditional methods and deep learning techniques. Recent studies have shown that deep neural networks (DNNs) perform significantly better than shallow networks and Gaussian mixture models (GMMs) on large vocabulary speech recognition tasks [3].

In particular, a global statistics framework of an utterance is classified by Gaussian mixture models using derived features of the raw pitch and energy contour of the speech signal and hidden Markov model-based speech emotion recognition considering several states using low-level instantaneous features instead of global statistics were proposed as traditional methods [4].

Recently, deep learning methods have been evolving rapidly. The SER system using RNNs with an efficient learning approach was proposed in [5], which takes into account the long-range context effect and the uncertainty of emotional label expressions.

Convolutional Neural Networks have been performed for speech emotion recognition in many researches [6-8]. From [8] , we learnt a CNN-based SER method that learns salient features of SER using semi-CNNs.

In this paper, we propose the comparison of models like MLP, LSTM, CNN models for emotion classification problems. Then we applied the given methods using the RAVDESS speech dataset, and the classification results were tested to be better than traditional methods.

**Algorithm details**

a. Algorithm overview - MLP

We built a Multilayer Perceptron model, LSTM model and CNN models. The MLP and LSTM were not suitable as it gave us low accuracy. As our project is a classification problem where we categorize the different emotions, CNN worked best for us. The MLP model formulation:

1. The relu, sigmoid and softmax are used as activation functions in the layer network.
2. As a loss function we used the categorical cross entropy loss between the labels and predictions.
3. As an optimization algorithm we used Adam or Adaptive Moment Optimization algorithms that combined the heuristics of both Momentum and RMSProp optimizations.

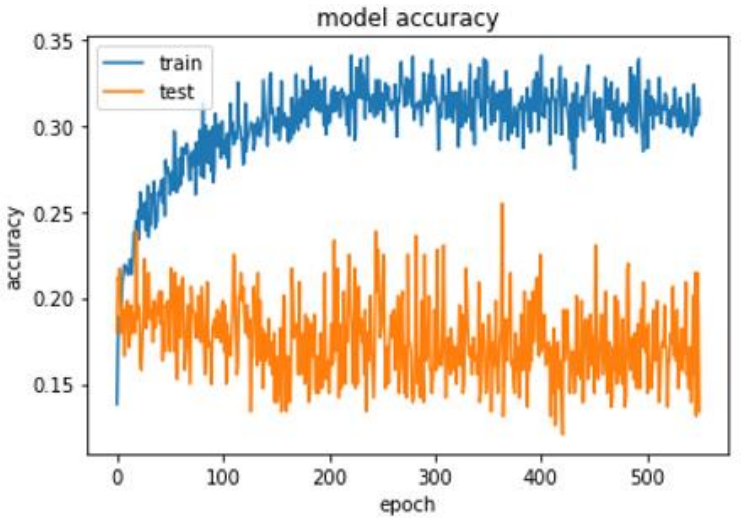


*Figure 1: Representation of loss function for MLP model.*

b. Algorithm overview - LSTM

The LSTM model had the lowest training accuracy of around 15% with 5 layers. The LSTM model formulation:

1. The softmax and tanh are used as activation functions in the layer network.
2. Stochastic gradient descent algorithm is used to adjust the weights of corresponding layers.
3. As a loss function we used the categorical cross entropy loss between the labels and predictions.

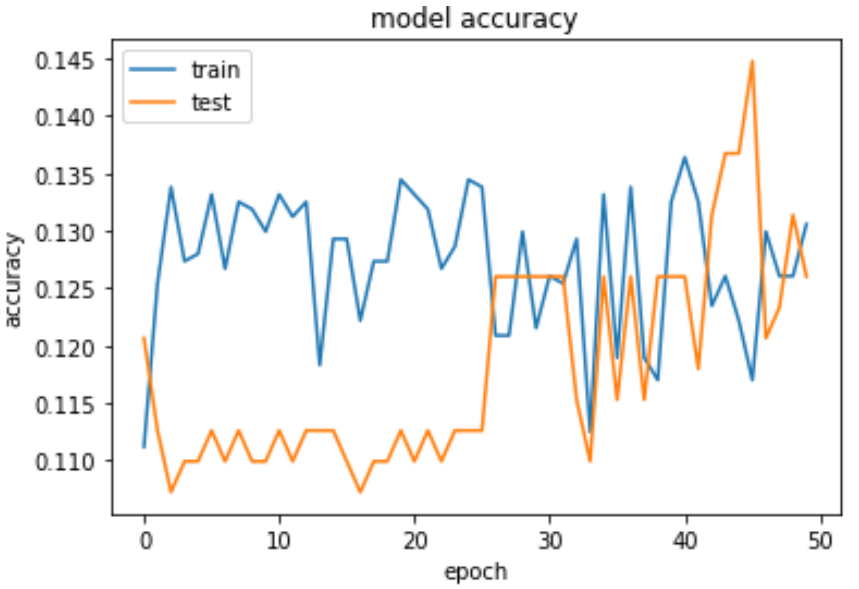


*Figure 2: Representation of loss function for LSTM model.*

c. Algorithm overview - CNN

CNN model was the best for our classification problem. After training numerous models we got the best validation accuracy of 70% with 18 layers, relu and softmax activation function. The CNN model formulation:

1. The softmax and relu are used as activation functions in the layer network.
2. Stochastic gradient descent algorithm is used to adjust the weights of corresponding layers.
3. Root Mean Square Propagation (RMSProp) algorithm is used in the hidden layer to compute the derivatives of weights.

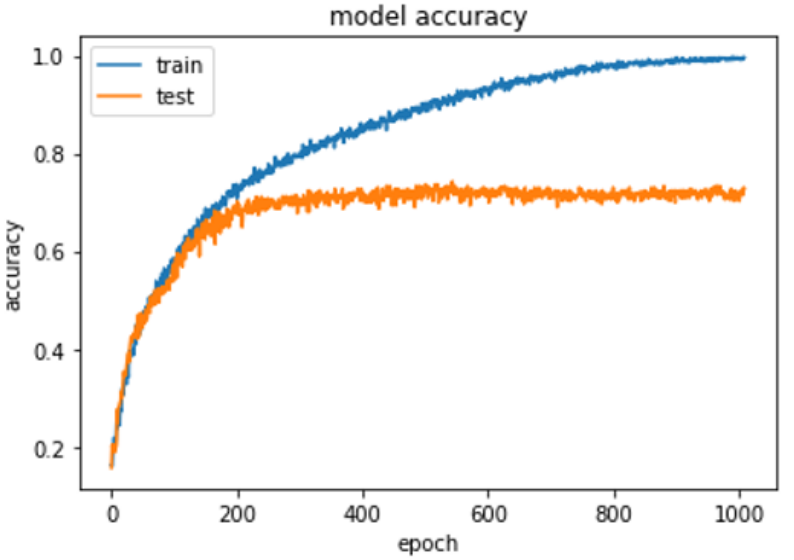


*Figure 3: Representation of loss function for CNN model.*

d. Feature extraction

The first step in speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the Mel Frequency Cepstral Coefficient accurately represents this envelope.

Firstly, we frame the input signal into short frames, then the periodogram estimate of the power spectrum for each frame was calculated. Next step is to apply the mel filterbank to the power spectra and sum the energy in each filter. Once we have the filterbank energies, we take the logarithm of them. This is also motivated by human hearing: we don't hear loudness on a linear scale. Generally to double the perceived volume of a sound we need to put 8 times as much energy into it. This means that large variations in energy may not sound all that different if the sound is loud to begin with. This compression operation makes our features match more closely what humans actually hear. The final step is to compute the DCT of the log filterbank energies. There are 2 main reasons this is performed. Because our filterbanks are all overlapping, the filterbank energies are quite correlated with each other. The DCT decorrelates the energies which means diagonal covariance matrices can be used to model the features in e.g. a HMM classifier.

We have 5 different emotions in our dataset (Calm, Happy, Sad, Angry, Fearful).We also separated out the females and males' voices by using the identifiers provided in the website. This was because as an experiment we found out that separating male and female voices increased by 15%. It could be because the pitch of the voice was affecting the results.

**Experimental results**

a. Algorithm evaluation

As the main evaluation metric we used accuracy score, which was around 25% for multilayer perceptron. To further enhance the recognition accuracy of the proposed solution, we tried to build an LSTM layer, which in turn showed the lowest training accuracy of around 15%. Finally, the convolutional neural network model was the best for our classification problem. After training numerous models we got the best validation accuracy of 70%. As a result, the CNN model was the best for our classification problem.

b. Experimental setting

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and songs contain calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound).

c. Results

After building different models, we have found our best CNN model for our emotion classification problem. We achieved a validation accuracy of 70% with our existing model. Our model could perform better if we have more data to work on. What’s more surprising is that the model performed excellent when distinguishing between a males and female voice. We can also see above how the model predicted against the actual values.

**Conclusion**

This audio-based emotion recognition task is difficult due to the fact that emotions are a subjective concept and are difficult to classify even for a human. As a result, the field of speech emotion recognition is still a challenging problem. In this paper, we proposed the CNN based network with-

out using any traditional hand-crafted features to classify emotional speech. For SER, we performed CNNs feature extraction architecture. Moreover, we investigated the classification result by comparing with the basic CNN, LSTM and MLP based emotion recognition results. We verified that CNN-based networks show better results. This comparison of results provides a baseline for future research, and we expect that it can give a better result when using more concatenated CNNs. In future, we are planning to study the audio based multimodal emotion recognition task.

REFERENCES


1. Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In International Conference on Machine Learning (ICML).
2. H.K. Palo, Mihir Narayan Mohanty, Mahesh Chandra. “Use of Different Features for Emotion Recognition Using MLP Network”. Advances in Intelligent systems and computing. vol. 332, Springer India, Jan. 2015.
3. Yu, D., Seltzer, M.L., Li, J., Huang, J., and Seide, F. “Feature Learning in Deep Neural Networks Studies on Speech Recognition Tasks”. In ICLR, 2013.
4. Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition." Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on. Vol. 2. IEEE, 2003.
5. Jinkyu Lee and Ivan Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” 2015.
6. Michael Neumann and Ngoc Thang Vu, “Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech,” arXiv preprint arXiv:1706.00612, 2017.
7. Yoon Kim, “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882, 2014.
8. Huang, Zhengwei, et al. "Speech emotion recognition using CNN." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.
9. H.K. Palo, Mihir Narayan Mohanty, Mahesh Chandra. “Use of Different Features for Emotion Recognition Using MLP Network”. Advances in Intelligent systems and computing. vol. 332, Springer India, Jan. 2015.


**Сарсембаев А.А., Толғанбаева Г.А., Джаныбекова С.Т.**

**Решение задачи классификации эмоций с помощью глубокого обучения**

**Аннотация.** Классификация речевых эмоций - одна из самых интересных и сложных задач в современном мире. Одним из основных препятствий на пути к этой задаче является то, что эмоции субъективны и их трудно уловить. В этой статье предложены методы глубокого обучения, которые решают задачи классификации эмоций на основе аудиопотоков. В статье распространяются и сравниваются три метода. В рамках первого метода была построена модель многослойного персептрона. Второй метод показывает снижение точности построения моделей долгосрочной краткосрочной памяти. Наконец, третий метод, достигший лучшей точности среди других - это модели сверточных нейронных сетей. Подробно описан речевой корпус, состоящий из образцов разыгрываемых и спонтанных эмоций на английском языке. Этот набор данных был протестирован и обучен с использованием предложенных методов. Модель CNN для нашей проблемы классификации эмоций достигла точности подтверждения 70%.

**Ключевые слова:** распознавание речевых эмоций, сверточная нейронная сеть, глубокая нейронная сеть, долговременная краткосрочная память, многослойный перцептрон.

**Сарсембаев А.А., Толғанбаева Г.А., Джаныбекова С.Т.**

**Эмоциялық классификация мәселелерін терең оқыту арқылы шешу**

**Аңдатпа.** Сөйлеу эмоциясын жіктеу – қазіргі әлемдегі ең қызықты және күрделі мәселелердің бірі. Бұл тапсырманың басты кедергілерінің бірі – эмоциялар субъективті және оларды тану қиын. Осы жұмыста біз аудио негізінде эмоцияны жіктеу мәселелерін шешетін терең оқыту әдістерін ұсындық. Ал енді жұмыста үш әдіс қарастырылады және салыстырылады. Бірінші әдіс шеңберінде көп қабатты Перцептрон моделі құрылды. Екінші әдіс ұзақ мерзімді жад модельдерінің дәлдігі төмендеуін көрсетеді. Сонымен басқалардың арасында ең жақсы дәлдікке жеткен үшінші әдіс – бұл жүйкелік жүйенің конволюциялық модельдері. Ағылшын тіліндегі әрекет етуші және спонтанды эмоциялар үлгілерінен тұратын сөйлеу корпусы егжей-тегжейлі сипатталған. Аталған деректер базасы осы ұсынылған әдістердің көмегімен тексеріліп, оқытылды. Біздің эмоцияны жіктеу мәселесі үшін CNN моделі 70% дәлдікке қол жеткізді.

**Түйінді сөздер:** сөйлеу эмоциясын тану, конволюциялық жүйке жүйесі, терең жүйке жүйесі, ұзақ мерзімді есте сақтау, көп қабатты перцептрон

**Сведения об авторах:**

Сарсембаев Айдос Айдарович, PhD, ассистент-профессор кафедры «Компьютерной инженерии и информационной безопасности» Международного университета информационных технологий.

Джаныбекова Салтанат Талгатбековна, докторант кафедры «Компьютерной инженерии и информационной безопасности» Международного университета информационных технологий.

Толғанбаева Гауһартас Алғабасқызы, докторант кафедры «Компьютерной инженерии и информационной безопасности» Международного университета информационных технологий.

**УДК 004.65.004**

**Bektemyssova G.U., Ainabek Zh.B.**

International Information Technologies University, Almaty, Kazakhstan

## OBJECT TRACKING

***Abstract.*** *Moving object tracking is very useful in many computer vision applications. The most famous examples are surveillance systems in crowded public places, traffic control systems, motion capture systems for electronic games, applications for human-computer interaction, and many others. Recently, a large number of approaches have been proposed for tracking objects. However, no algorithm has yet been developed that would cope with all the existing problems of object tracking. This article aims to analyze the existing problems, as well as consider ways to solve them.*

***Key words****: object detection, object tracking, background subtraction, image subtraction, optical flow, speeded-up robust features.*

### Introduction

Object tracking is one of the most researchable topics in computer vision today, with interest increasing dramatically over the last few decades. This demand has been due to the rapid development of information technologies, the availability of high-quality, low-cost cameras, and the increased need for tracking applications in various fields such as traffic monitoring, human- computer interaction, surveillance and medical imaging. Reliable detection and tracking of an object in a video remains an open research problem even after several years of study in this field. In spite of sig-