Mathematical Problems of Computer Science 56, 35–47, 2021.
doi: 10.51408/1963-0077

UDC 004.934

Deep Learning Approaches for Voice Emotion Recognition Using Sentiment-Arousal Space

Narek T. Tumanyan

Weizmann Institute of Science, Israel
e-mail: narek.tumanyan@weizmann.ac.il

Abstract

In this paper, we present deep learning-based approaches for the task of emotion recognition in voice recordings. A key component of the methods is the representation of emotion categories in a sentiment-arousal space and the usage of this space representation in the supervision signal. Our methods use wavelet and cepstral features as efficient data representations of audio signals. Convolutional Neural Network (CNN) and Long Short Term Memory Network (LSTM) architectures were used in recognition tasks, depending on whether the audio representation was treated as a spatial signal or as a temporal signal. Various recognition approaches were used, and the results were analyzed.
Keywords: Voice emotion recognition, Sentiment-arousal space, Spectral features, Speech sentiment classification.
Article Info: Received 19 July 2021; accepted 26 October 2021.

1. Introduction

In this work, we address the problem of emotion recognition from voice recordings. Recognizing emotion from voice can have various real-world applications, such as in recommendation systems, security systems, customer services, etc. Defining the recognition task formally, we want to come up with a model F, such that given a voice recording X in some representation, the model will give us a mapping $F(X) = y$, where y is some descriptor of the recognized emotion from the audio signal. Now, the question is, what space does y belong to? Is it discrete or continuous, and how are emotion values organized in this space? To answer these questions, we utilize a sentiment-arousal space described in the paper, which allows us to tackle the recognition task in different approaches, depending on how we use this space for defining the set of y values.

Previous methods for the voice emotion recognition problem include SVM-based classification algorithms [1], which also consider visual data of the facial expression of the speaker as an additional signal, as well as Deep Neural Network Extreme Learning method with an efficient performance on small datasets [2].

We use Mel Frequency Cepstral Coefficients (MFCC) and Continuous Wavelet Transforms (CWT) for representing audio signals in spectral features. Convolutional Neural Networks (CNN) and Long Short Term Memory Networks (LSTM) were used as deep learning model architectures.

2. Datasets

In our work, we used 3 databases of labeled voice recordings: Surrey Audio-Visual Expressed Emotion (SAVEE) [3], Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [4], and Toronto Emotional Speech Set (TESS) [5]. The databases are comprised of voice recordings of individuals who pronounce a statement with an exerted emotion, which is the label of the given voice recording. The emotion labels in the RAVDESS database are: "neutral", "calm", "happy", "sad", "angry", "fearful", "disgust", "surprised". TESS and SAVEE datasets have the same emotion labels except the "calm" one. The distribution of samples and labels of the databases are summarized in Table 2 and in Table 1.

Table 1: Number of voice recordings per emotion label across all databases.

Neutral	Calm	Sad	Fear	Anger	Surprises	Happiness	Disgust
616	192	652	652	652	652	652	652

Table 2: Summary of datasets used.

Database	Num of Recordings	Num of Actors	Emotion Labels
RAVDESS	1440	24	8
SAVEE	480	4	7
TESS	2880	2	7

3. Method

3.1 Feature Extraction

The first step in data preparation is resampling the voice recording signal in a certain sampling rate. As the signal in interest is human voice, which is known to lie in frequency ranges 4-10 Khz, we chose 22.05 Khz sampling rate. The resampled signal includes the human voice signal along with some possible frequency variations, which can be caused by possible pronunciation of high frequency sounds, such as fricatives. As a result, we obtain a temporal signal representation of the voice recording, which at a given time point shows the amplitude of air pressure oscillations from 0 frequency.

3.1.1 Fourier Representation

A temporal signal $x(t)$ can be represented as a combination of periodic functions of varying frequencies [6]

$$x(t) = \int_{-\infty}^{\infty} X(w) e^{jwt} dw,$$

where w denotes the frequency of the periodic function. Having the coefficients $X(w)$ is equivalent to having the original signal $x(t)$, and these coefficients are used as a representation of the signal in frequency domain. Such a representation is obtained by the Fourier Transform operation [6]. Discrete Fourier Transform (DFT) is the discrete equivalent of Fourier Transform, which we leverage for representing our discrete resampled signal $x[n]$ of length k in frequency space through coefficients / intensities $X[k]$ for each frequency k [7]:

$$X[k] = \sum_{n=1}^{K} x[n] e^{-i2\pi kn/N};\ 1 \leq k \leq K.$$

Usually, representing the entire discrete signal $x(t)$ with Fourier coefficients can result in loss of temporal resolution, since having a Fourier representation for the entire signal does not include changes of the signal in small temporal windows. For obtaining higher resolution in temporal domain, Short-Time Fourier Transform (STFT) [6] of a signal is used in some of the approaches, which basically calculated Fourier coefficients of the signal in temporal windows.

3.1.2 Continuous Wavelet Transform

The continuous wavelet transform is a method of analyzing the frequency components of a signal at specific time intervals. The advantage that CWT has over STFT is that it solves the problem of trade-off between frequency resolution and time resolution. When performing an STFT on a signal, one has to choose the window length for dividing the signal into sub-signals and performing DFT on each window, meaning that the larger the window size is set, a higher frequency resolution (the frequency components are better explained for the signal as a whole) and a lower time resolution (the changes of frequencies across time are not explained well) is obtained. The opposite holds as well: STFT with a smaller window size has higher time resolution but lower frequency resolution. CWT solves this problem of trade-off by representing the signal at different frequency scales, larger scale corresponding to lower frequencies, and lower scales to higher ones. At smaller scales, the signal is divided into smaller time windows, and lower frequency information is extracted, resulting in higher temporal resolution but lower frequency resolution. At larger scales, the signal is divided into larger time intervals, and higher frequency information is processed, resulting in higher frequency resolution but lower temporal resolution.

CWT makes use of wave-like functions called wavelets, and, at each step of the algorithm, the original signal is convolved by the wavelet function for deriving the corresponding frequency-domain value. The requirements for a function $f(t)$ to be considered a wavelet function as follows (complex wavelets are not considered in this paper, the following conditions relate to the real-valued wavelet qualifications only) [8]:

$$E = \int_{-\infty}^{\infty} |f(t)|^2 dt < \infty, \text{where } E \text{ is termed as the energy of } f,$$

$$\int_{0}^{\infty} \frac{|F(k)|^2}{k} dk < \infty, \text{where } F(k) \text{ is the Fourier transform of } f.$$

The most commonly used wavelet functions are Gaussian wave, Mexican hat, Haar and Morlet [8], the latter of which we utilized in speech signal processing (visualized in Fig. 1.)

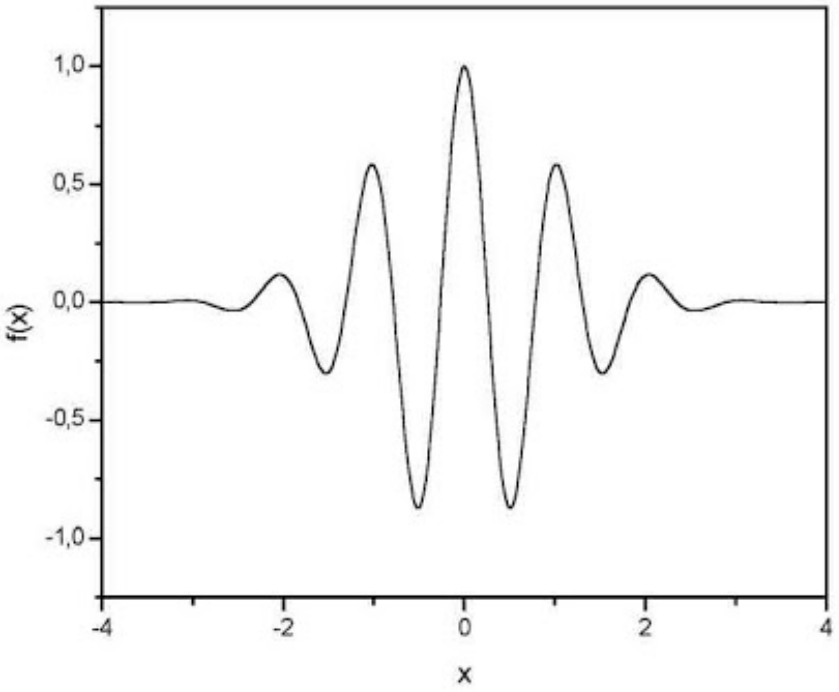

Fig. 1. Morlet wavelet function.

After choosing the wavelet function $\Phi(t)$, the CWT of the signal $x(t)$, denoted as $T(a, b)$, is computed as follows:

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \Phi \left(\frac{t-b}{a} \right) dt,$$

where a is the scale at which the signal is processed, and b indicates the time interval at which the signal is convoluted with the wavelet function. An example of a heatmap resulting from CWT is visualized in Fig. 2.

3.1.3 Mel Frequency Cepstral Coefficients

Another representation of audio signals that our methods use are Mel-frequency ceptstral coefficients (MFCC). MFCCs represent a temporal signal by cepstral energy coefficients at specific time intervals. The motivation of using MFCCs is to represent a signal by features that replicate the perception of audio signal by a human ear. Such representation is obtained by processing the signal with cepstral filters across frequency scales, the length of which is directly proportional to the scale of the frequency [9].

The resulting MFCC representation of a signal is given as a function $P_i(k)$, the output of which is the value of k-th cepstral coefficient at i-th temporal frame index. An example of an extracted MFCC feature is demonstrated in Fig. 3.

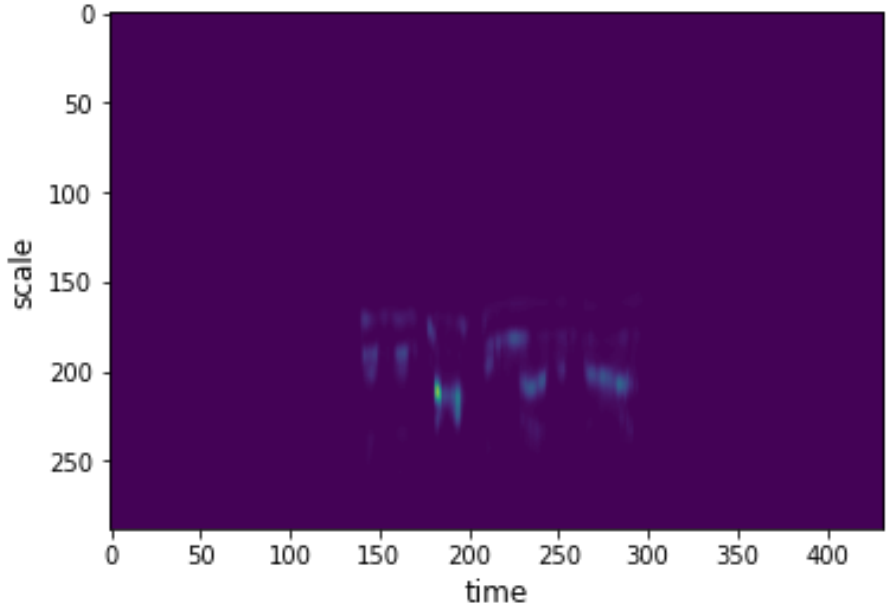

Fig. 2. Sample CWT heatmap of an audio signal.

3.2 Technical Details

For the extraction of audio signal features described in section 3.1, we use the "librosa" library for python [10]. "Pytorch" was used as a deep learning library for training our models [11]. The architectural details of each model are described in their respective sections.

3.3 Recognition Approaches

Having the labeled audio signals and their feature representations from CWT and MFCC, the next step is designing a method for emotion recognition from those signals. Following [12], the approach that this work relies on is using a sentiment-arousal space of emotions, which is depicted in Fig. 4. The idea is to come up with an intuitive 2-dimensional organization space of emotions by defining 2 axes: the arousal axis, and the positivity axis. By assigning these 2 values to every emotion label, we come up with an intuitive organization of emotion values in this space, as demonstrated in Fig. 4.

Having the sentiment-arousal space allows us to come up with different emotion recognition approaches, such as defining each quadrant of the 2D space as a classification label (i.e., whether the emotion is active-positive, active-negative, passive-positive, or passive-negative), or viewing the sentiment-arousal space as a continuous one, and solving the recognition task as a regression problem. In the upcoming sections, we show each of such approaches used along with the corresponding extracted features and the neural network architecture.

To the best of our knowledge, our proposed methods are the first try on tackling the problem in the specified setups. An exception is the setup of classification in sentiment-arousal space using CWT features and CNNs, where we compare to a method that has some of its aspects of setup shared with ours.

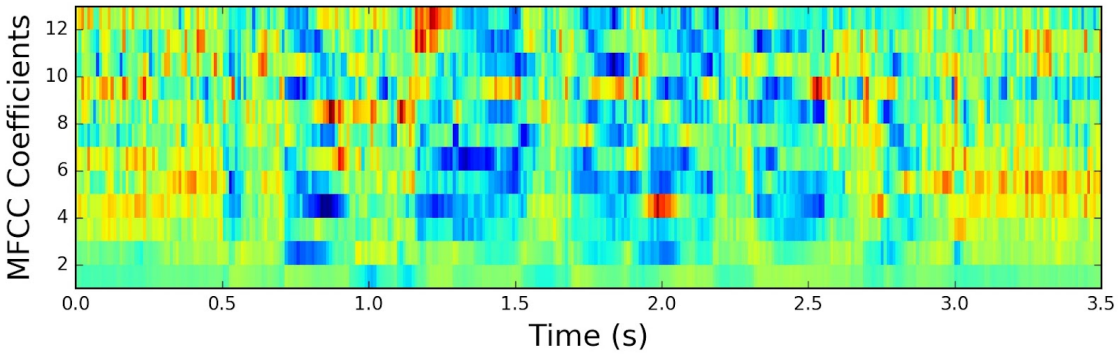

Fig. 3. Sample MFCC representation of a voice recording signal.

3.4 Architectures and Results

3.4.1 Mapping Emotion to Continuous Sentiment-Arousal Space

Table 3: Mapping of emotion values in sentiment-arousal continuous space

Emotion	Sentiment-Arousal Coordinates
Neutral	[0,0]
Calm	[0.25,-1]
Sad	[-0.75,-0.5]
Happy	[1,0.75]
Angry	[-0.75,1]
Fear	[-1,0.25]
Disgust	[-0.25,0.25]
Surprise	[0.25,1]
U.	1

An interesting approach that we can take towards the voice emotion recognition task is using the sentiment-arousal dimensions for defining a continuous space of emotion values, and solving a regression problem of emotion prediction. Specifically, for each emotion label coming from the datasets, we define sentiment and arousal values, as described in Tab. 3, which results in the organization of emotion values in a continuous sentiment-arousal space. Thus, the objective of the problem can be defined as:

$$L(\hat{y}) = \frac{1}{2}(\hat{y} - y)^T(\hat{y} - y) + \lambda \sum_{w \in W} w^2,$$

where $\hat{y}$ is the predicted point in the continuous space, y is the point in the 2D space corresponding to the ground-truth emotion label. W is the set of the trainable parameters, and thus the last term serves as a regularization to the optimization problem.

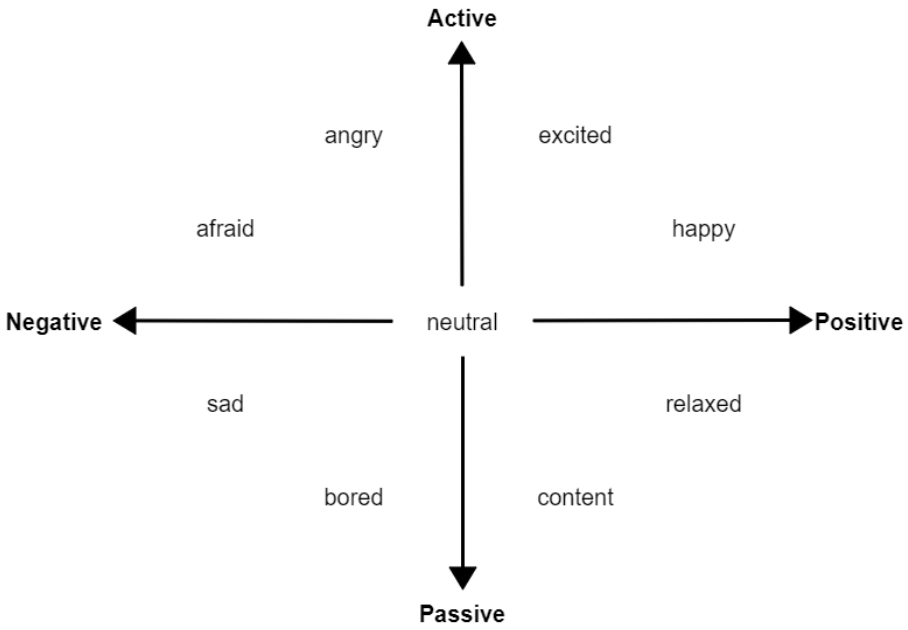

Fig. 4. The dimensions of the sentiment-arousal space, and how different emotions are organized in the space.

For solving the resulting regression task, we utilize MFCC features of audio recordings as inputs. We use CNN architecture for the model, which is shown in Fig. 5. Average pooling of size (2x2) is used for downsampling between the layers. The last layer is a fully connected layer that maps the flattened output of convolutional layers to the 2-dimensional sentiment-arousal space. Each layer has 32 output channels. The first layer has kernels of size (10x3), which is followed by a layer with (5x5) and a layer with (3x3) kernel sizes. Between layers, leaky relu activation function was used.

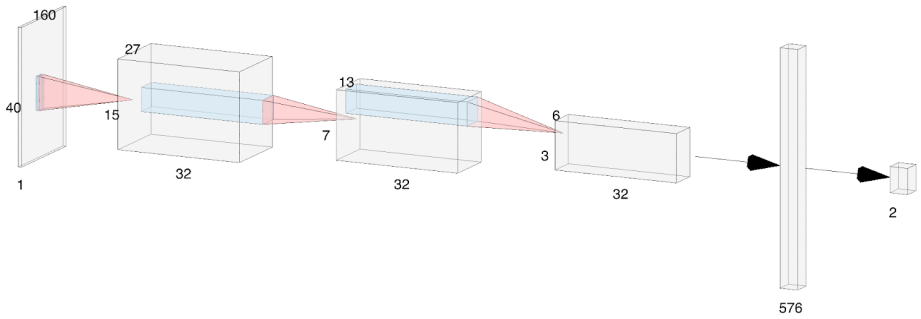

Fig. 5. CNN architecture used for the continuous emotion recognition task.

Fig. 6 shows the output of the model on voice recordings with the corresponding emotion labels. In the majority of cases, the network correctly identifies both the sentiment and

arousal of speech. It rarely fails to identify both components and it can at least identify the arousal of the speech. One of the shortcomings we see is that the significant proportion of the recordings with a "happy" label were identified as negative by the network. On the contrary, fear, disgust, anger and sadness were correctly positioned in the space. This, as also pointed out in the previous sections, shows us that the network is struggling to determine the positivity, but is good at differentiating between active and passive emotions.

Fig. 6. Performance of the CNN model on the continuous emotion recognition task.

3.4.2 Classification Using Sentiment-Arousal Space: LSTM with MFCCs

First, we solve a classification problem defined by the quadrants of the sentiment-arousal space. We use the extracted MFCC features as our input, and, viewing MFCC's as temporal signals, we use LSTMs [13] as our neural network architecture. Only the first 40 cepstral coefficients were considered. The datasets used were RAVDESS and TESS datasets (in some scenarios, only RAVDESS was considered.) For all classification models, for a single audio recording, given its ground-truth label values $\{y_1, y_2, ..., y_n\}$ and the estimated label values $\{\hat{y_1}, ..., \hat{y_n}\}$, the objective function is:

$$L = -\sum_i y_i \log(\hat{y_i}) + \lambda \sum_{w \in W} w^2,$$

where W is the set of all trainable weights.

There were 4 scenarios of splitting the dataset into train and test subsets: 1. 10% testing and 90% training (standard), 2. all the recordings of the first 2 actors as the test dataset and the rest as the train 3. all the recordings of the first 3 actors as the test dataset and the rest as the train, 4. all the recordings of the first 4 actor as the test dataset and the rest as the train. The architecture of LSTM model depicted in Fig. 7 was used for all scenarios. A dropout layer with probability p=0.3 was used.

The results of the experiment are summarized in Tab.4.

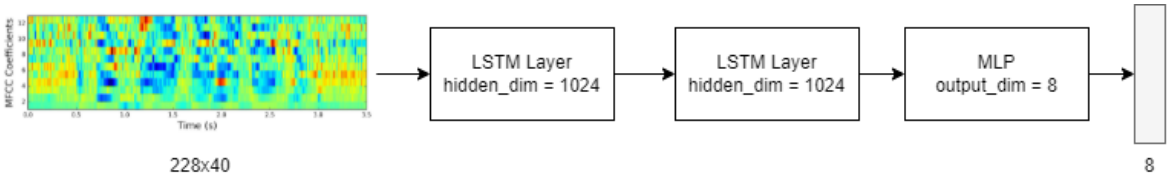

Fig. 7. The architecture of the trained LSTM model.

Table 4: Classification Results of LSTM model on different scenarios

Zones	Data separation	Datasets	Train acc.	Test acc.	AUC
4 zones	standard	RAVDESS	96.30%	67.36%	–
4 zones	2 new actors	RAVDESS	96.13%	74.16%	–
Arousal zones	standard	RAVDESS	97.76%	87.14%	0.91
Arousal zones	2 new actors	RAVDESS	99.54%	90.83%	0.94
Arousal zones	3 new actors	RAVDESS+TESS	94.23%	86.11%	0.91
Arousal zones	4 new actors	RAVDESS+TESS	98.40%	83.33%	0.87
Arousal zones	2 new actors	RAVDESS+TESS	95.63%	93.30%	0.97
Sentiment zones	standard	RAVDESS	98.30%	80.00%	0.81
Sentiment zones	2 new actors	RAVDESS+TESS	96.89%	84.14%	0.91
Sentiment zones	3 new actors	RAVDESS+TESS	93.69%	79.44%	0.84

From the results, we can see that the model managed to learn meaningful representations from the supervision signal. Since the same LSTM architecture gave performance for all classification scenarios, it indicates that the architecture is a good fit considering the datasets available. Also, the results indicate that the performance was good in classifying the arousal level of the speech, but classifying positivity is a bigger challenge for the model. This can be explained by the fact that MFCCs represent the energy amount in the signal in specific frequency or cepstral ranges, and, intuitively, larger amounts of energies correspond to higher arousal level. However, both negative and positive emotions can correspond to a high arousal level (i.e., surprised and angry), but it is harder to say how energy features can distinguish the positivity of a given speech.

3.4.3 Classification Using Sentiment-Arousal Space: CNN with CWT

The next experiment that we conducted is solving the problems of arousal level classification and positivity classification with CWT as inputs, and using CNN as the neural net architecture. Only RAVDESS dataset was considered in this experiment, and it was divided into a 10% test and 90% train datasets in both classification problems. Fig. 8. illustrates the architecture of CNN used for the classification tasks. Dropout with p=0.4 was used between each convolutional layer to prevent overfitting. Leaky ReLU was used as an activation function between layers and for preventing vanishing gradients. The results of the experiment

are summarized in Fig. 9 and in Tab. 5.

Table 5: Classification results of CNN model trained on CWT data

Zones	Train accuracy (%)	Test accuracy (%)	AUC
Arousal zones	98.70	83.76	0.84
Sentiment zones	87.76	75.71	0.77

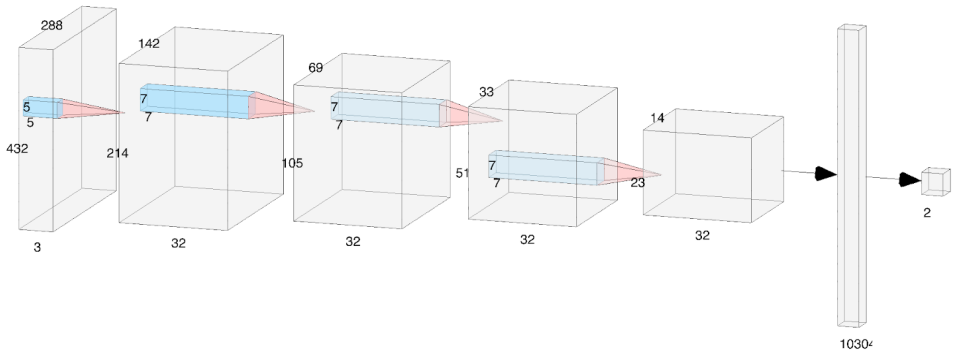

Fig. 8. CNN model architecture used for the CWT-based classification tasks.

As noted in the previous experiment, the models encounter the difficulty of classifying the positivity of the speeches.

[14] proposes methods for classifying arousal and sentiment in speech. They use the DEAP database [15], and their setup considers only "happiness", "sadness" and "cheerfulness" emotional labels. In their 2-label classification setting (high/low arousal; positive/negative sentiment), the arousal classification and sentiment classification accuracies are 61.23% and 92.19%, respectively, which are comparable results to our method.

4. Conclusion

This work proposes methods for solving voice emotion recognition tasks based on deep learning models. Audio signals were represented by features resulting from MFCC and CWT transforms. A pivotal component in the approaches is defining a 2D sentiment-arousal space, where the emotion values are organized in an intuitive way, allowing to define the recognition problem within this space either as a classification or a regression. The main challenge identified in all the proposed methods was the difficulty of recognizing the positivity aspect of the recordings, a possible explanation to which is the absence of such information in the features used, which mainly encode energies corresponding to frequency ranges. Overall, the results indicate that the models manage to learn features meaningful for the emotion recognition task. As one of the main challenges was the scarcity of labeled data, possible

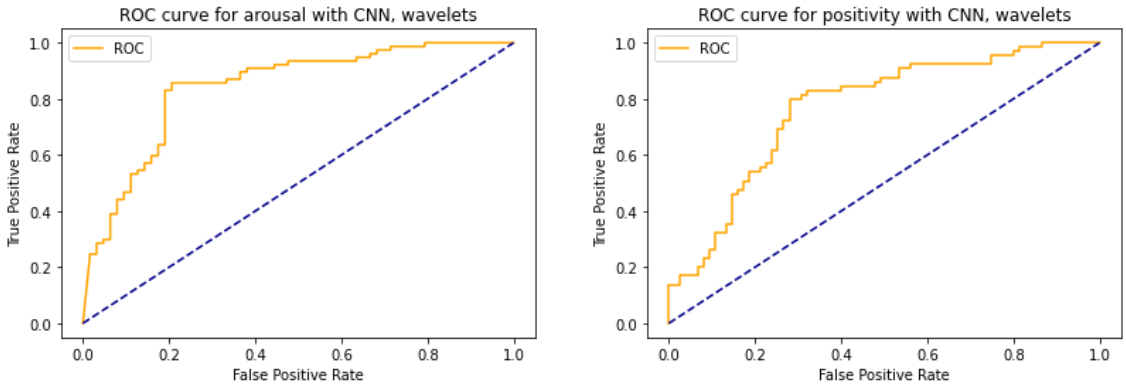

Fig. 9. ROC curves of the CNN classifier.

future directions include the use of data augmentations on voice recordings, as well as self-supervised approaches for learning semantic representations of the audio signals and fine-tuning those features for emotion recognition task, which doesn't require any labeled data.

References

[1] E. Mower, M. J.Mataric and S.Narayanan, "A framework for automatic human emotion classification using emotion profiles", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2010.

[2] S. Glge, R. Bck and T. Ott,"Emotion recognition from speech using representation learning in extreme learning machines", *Proceedings of the 9th International Joint Conference on Computational Intelligence*, Funchal, Portugal, pp. 179–185, 2017.

[3] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database", University of Surrey: Guildford, UK. 2014.

[4] S.R.Livingstone, and F.A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English", *PLoS ONE* , vol. 13, no. 5, 2018.

[5] Pichora-Fuller, M. Kathleen and K. Dupuis, "Toronto emotional speech set (TESS)", Scholars Portal Dataverse, 2020. https://doi.org/10.5683/SP2/E8H2MF

[6] K. Grchenig, *Foundations of Time-Frequency Analysis*, First Edition. Birkhuser, Boston, MA, 2001.

[7] A. Kulkarni, M.F. Qureshi and M. JHA, "Discrete fourier transform: Approach to signal processing", *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 03, pp. 12341–12348, 2014.

[8] P. S. Addison, *The Illustrated Wavelet Transform Handbook* , Second Edition. CRC Press, 2017.

[9] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition", *Speech Communication* , vol. 54, no. 4, pp. 543–565, 2012.

[10] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thom, C. Raffel, F. Zalkow, A. Malek, D. Kyungyun Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth. (2022). librosa/librosa: 0.9.0 (0.9.0). Zenodo. https://doi.org/10.5281/zenodo.5996429

[11] A. Paszke, S.Gross, F.Massa, A. Lerer, J. Bradbury,G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library", *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019.

[12] J. Posner, J.A. Russell and B.S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology", *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[13] S. Hochreiter and J. Schmidhuber,"Long Short-Term Memory", *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] G. Garg and G. K. Verma, "Emotion recognition in valence-arousal space from multichannel EEG data and wavelet based deep learning framework", *Procedia Computer Science*, vol. 171, pp. 857–867, 2020.

[15] S. Koelstra, C. Muhl, M. Soleymani, J.S. Lee, A. Yazdani, T. Ebrahimi and I. Patras, "Deap: A database for emotion analysis; using physiological signals", *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

Խորը ուսուցման մեթոդներ ձայնագրությունների էմոցիայի գնահատման համար օգտագործելով տրամադրական կոորդինատային համակարգ

Նարեկ Տ. Թումանյան

Վեյցմանի գիտությունների համալսարան
e-mail: narek.tumanyan@weizmann.ac.il

Ամփոփում

Այս հոդվածում ներկայացվում են խորը ուսուցման վրա հիմնված մոտեցումներ՝ ձայնագրությունների էմոցիայի գնահատման խնդրի համար: Առաջադրված մոտեցումների բանալի բաղադրիչ է հանդիսանում էմոցիաների դասերի ներկայացումը տրամադրական երկչափ կոորդինատային համակարգում, որտեղ աքսիսների չափման միավոր են հանդիսանում էմոցիայի դրական/բացասական լինելը և ակտիվ/պասիվ լինելը, ինչպես նաև այդ ներկայացման օգտագործումը ուսուցման վերահսկման մեջ: Աուդիո ազդանշանները մշակելու համար օգտագործվել են ձայնագրությունների հաճախական տվյալներ: Որպես խորը ուսուցման մոդելներ, առաջադրված մեթոդներում օգտագործվում են լրիվ փաթույթային նեյրոնային ցանցեր (CNN) և երկար կարճաժամկետ հիշողություն (LSTM): Ներկայացվում են տարբեր էմոցիայի գնահատման մոտեցումներ և վերլուծվում են արդյունքները:

Բանալի բառեր՝ ձայնագրության էմոցիայի գնահատում, տրամադրական կոորդինատային համակարգ, հաճախական հատկանիշներ, խոսքի տրամադրության դասակարգում:

Глубокое обучение для распознавания эмоций в записях голоса с использованием валентно-возбужденного пространства

Нарек Т. Туманян

Институт Вейцмана, Израиль
e-mail: narek.tumanyan@weizmann.ac.il

Аннотация

В этой статье представлены основанные на глубоком обучении подходы к задаче распознавания эмоций в записях голоса. Ключевым компонентом этих методов является представление категорий эмоций в валентно-возбужденном пространстве, и использование этого пространства в качестве обучающего сигнала. Наш метод использует вейвлетные и кепстральные признаки для эффективного представления аудиосигнала. Для задачи распознавания были использованы сверточные нейронные сети (CNN) и сети долгой краткосрочной памяти (LSTM). Архитектура выбиралась в зависимости от того, каким образом был представлен сигнал - в пространственном или временном виде. Были использованы различные подходы к задаче распознавания, и были проанализированы результаты.

Ключевые слова: распознавание эмоций в голосе, валентно-возбужденное пространство, кепстральные признаки, классификация настроения голоса.