Mathematical Problems of Computer Science 57, 7–17, 2022.
doi:10.51408/1963-0082

UDC 004.934

Emotion Classification of Voice Recordings Using Deep Learning

Narek T. Tumanyan

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

Abstract

In this work, we present methods for voice emotion classification using deep learning techniques. To processing audio signals, our method leverages spectral features of voice recordings, which are known to serve as powerful representations of temporal signals. To tackling the classification task, we consider two approaches to processing spectral features: as temporal signals and as spatial/2D signals. For each processing method, we use different neural network architectures that fit the approach. Classification results are analyzed and insights are presented.
Keywords: Voice sentiment detection, Mood recognition, Speech emotion recognition, Cepstral features.
Article info: Received 10 February 2022; received in revised form 17 April 2022; accepted 25 April 2022.

1. Introduction

The problem that is addressed in this work is the emotion classification from voice recording. Formally, given some representation X of voice recording data and a set of n emotion labels/classes $\{y_1, y_2, ..., y_n\}$, the aim is to come up with a classifier $F(X) = y_i$ that maps X to a label $y_i \in \{y_1, ..., y_n\}$. Practically, having such a classifier F can have a wide range of applications, such as recommendation systems of movies or music driven by users' mood, systems for tracking the emotional state and satisfaction of clients through time, security systems for preventing harmful actions based on emotion, and so on.

Previous attempts to tackle the voice emotion classification problem include SVM-based algorithms of classifying voice into 5 categories - angry, happy, neutral, sad, or excited [1], which also considers the facial expression of the speaker during speech as an additional signal. Glüge et al. [2] propose a Deep Neural Network Extreme Learning method with efficient performance on small datasets. Eskimez et al. [3] tackle the speech emotion recognition problem through an unsupervised approach, by which they come up with meaningful speech representations by learning the underlying structure of the data, which aids in solving the main task. Bertero et al. [4] introduce a Convolutional Neural Network (CNN)-based approach of 3-label ("angry", "happy", "sad") emotion recognition of speech, where they use

the standard pulse-code modulation (PCM) temporal representation of the audio signal as input. Mirsamadi et al. [5] propose a 4-label ("angry", "happy", "sad", "neutral") speech emotion recognition model based on Long Short Term Memory Network (LSTM) architecture and local attention, and base their model on Mel-Frequency Cepstral Coefficients (MFCC), Fast Fourier Transform (FFT), fundamental frequency and zero-crossing rate features of the audio. In our setups, we experiment with both CNN-based and LSTM-based architectures and consider 8 emotional labels for classification, which are described in Section 2.

In this paper, we use cepstral features as representations of voice data, particularly, we utilize Mel-Frequency Cepstral Coefficients (MFCC) for representing the audio signal. We experiment with two views for processing MFCCs: processing them as sequential data in the time domain, and processing them as spatial data. For each of the approaches, we use the appropriate neural network architecture. Specifically, for processing MFCCs as temporal data, we utilize Long Short Term Memory Networks (LSTM), and for processing MFCC as spatial/2D data, we make use of Convolutional Neural Networks (CNN).

2. Datasets

In our setup, we consider 8 emotion labels for classification. The databases used in the paper are as follows: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6], Surrey Audio-Visual Expressed Emotion (SAVEE) [7] and Toronto Emotional Speech Set (TESS) [8]. Each item in the datasets is a recording of an actor that pronounces some statement with a certain expressed emotion. Voice recordings in the databases come in a .wav format, which describes the amplitude of air pressure oscillations in the temporal domain. Each voice recording has an emotion label attached to it. The RAVDESS database has 24 actors that pronounce 2 phrases: "Kids are talking by the door" and "Dogs are sitting by the door" with 2 intensities: Normal and High each repeated twice. Neutral emotion has no high intensity so it is only repeated twice. The emotion labels are: "neutral", "calm", "happy", "sad", "angry", "fearful", "disgust", "surprised". TESS dataset has 2 actors, young and old, and both of them are female. There are 2800 voices in total with each phrase being of the form "Say the word x, where x stands for some word. Recordings in the TESS dataset have the same labeled emotions as in RAVDESS, except for the calm label, which is absent in this dataset. SAVEE dataset has 4 English male actors with 480 voice recordings. 7 emotions are present, with the calm emotion missing. In total, there are 4720 samples. The distribution of samples and classes is summarised in Table 1 and in Table 2.

Table 1: 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

Table 2: 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

3. Method

3.1 Feature Extraction

To extract audio features from voice recordings, we use librosa library for python [9]. It handles most of the transformations done to voice recordings to get final features used for classification. The first step before extracting features is to resample voice recording files to obtain their time domain and amplitude representation. Voice recordings from our databases have different original sampling rates, which range from 22Khz to 48Khz. However, the content that we are trying to analyze from those recordings are the human voices themselves. Normally, the human voice ranges from low range frequencies 300Hz to higher ranges of 4 - 10Khz. This means that we can use lower sampling rates to resample our voice recording. We chose 22.05Khz sampling rate, which preserves all human voices in original audio recordings and also preserves some possible frequency deviations from the normal range, which can be caused by pronouncing high-frequency tones, e.g. fricatives. The result is a floating-point time series describing the amplitude of air pressure oscillations from a mean frequency of 0 at each time point. Thus, we obtain a time-domain representation of the signal. An example is illustrated below in Fig. 1.

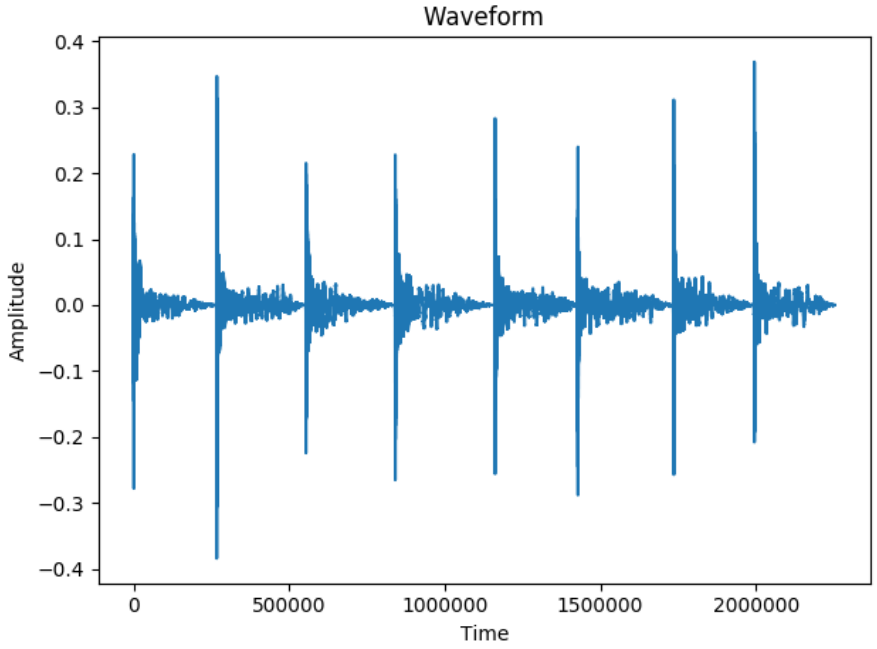

Fig. 1. Sample waveform representation of a voice recording signal.

Having the temporal signal representation of the voice signal, we then process it to obtain its spectral features, which serves as the main data representation for our models.

3.2 Spectral Features Extraction

Conceptually, given a temporal signal $x(t)$, we can represent it as a combination of periodic functions of varying frequencies [10]:

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

where w is the frequency of the corresponding periodic function. Thus, having the coefficients $X(w)$ is equivalent to having the original signal $x(t)$, and we can use these coefficients as a representation of the temporal signal in the frequency space. To achieving such a representation, the Fourier Transform operation is used [10]. Since we are dealing with discrete data, the equivalent operation used is Discrete Fourier Transform (DFT), which converts discrete temporal signal $x[n]$ of length K to a representation of this signal in frequency space by obtaining the coefficients / intensities $X[k]$ for each frequency k [11]:

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

In signal processing, frequency decomposition is often performed by dividing the signals into time intervals of specified window size and performing DFT on each windowed signal, thus coming up with frequency components in multiple time intervals. Such representation of a signal is called the Short-Time Fourier Transform (STFT) of a signal [10].

For audio signals, in some cases, more sophisticated representations of the signal based on STFT are necessary for higher efficiency. Mel-frequency cepstral coefficients, a.k.a. MFCCs, are features, which represent a given signal by cepstral energy coefficients at specific short intervals of time. The advantage of MFCC features is that they represent the signal in a way that is close to the signal perception by the human ear, which, is intuitively achieved by applying smaller window-sized cepstral filters on low frequencies on a signal and increasing the window size of the filters as the considered frequency increases. The reason behind such intuition is that the human ear perceives frequencies in lower ranges much better than in higher ones. Hence, higher resolution at lower ranged frequencies is used while computing MFCCs [12].

In its final form, the MFCC of a signal can be represented simply as a function $P_i(k)$, where the outputted value is the intensity of k-th cepstral coefficient in i-th temporal frame index.

An example of an extracted MFCC feature is demonstrated in Fig. 2

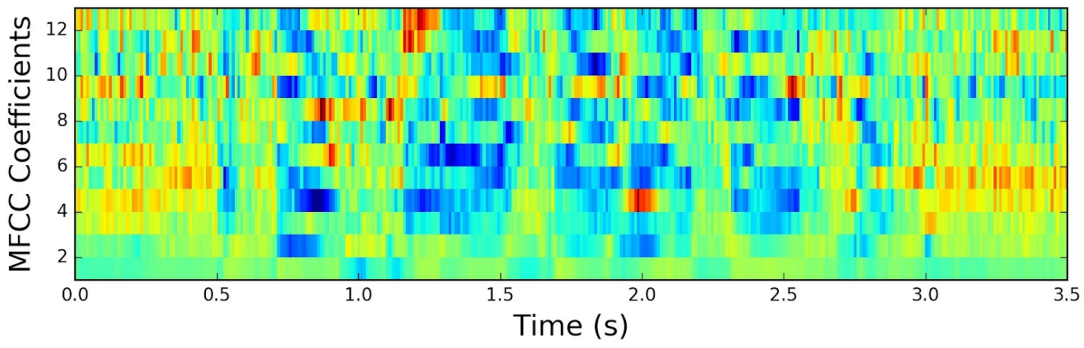

Fig. 2. Sample MFCC representation of a voice recording signal.

3.3 Architectures and Results

3.3.1 Long Short Term Memory Networks (LSTMs)

Considering the temporal nature of the data in hand, i.e., the voice recordings that are represented as magnitudes of air pressure (amplitude) across time, and the computed MFCC's that are a time series of energy coefficient values, it is sensible to use architectures that are by design intended for processing sequential data and have the appropriate inductive bias. One example of such architectures are Long Short Term Memory Networks (LSTM) [13], which are a variant of Recursive Neural Networks (RNN). The main idea behind LSTM is the usage of feedback connections for preventing the vanishing gradient problem. The architecture of LSTM used is summarized in Fig. 3.

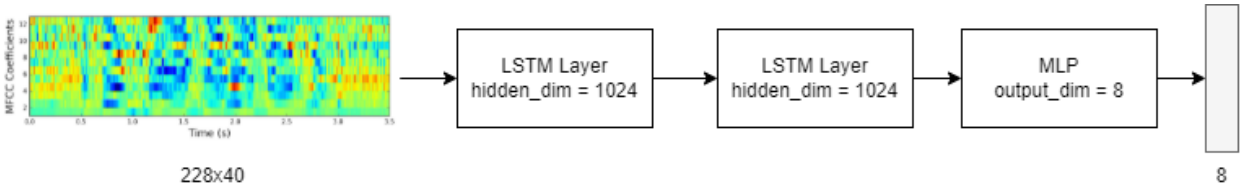

Fig. 3. The architecture of the trained LSTM model.

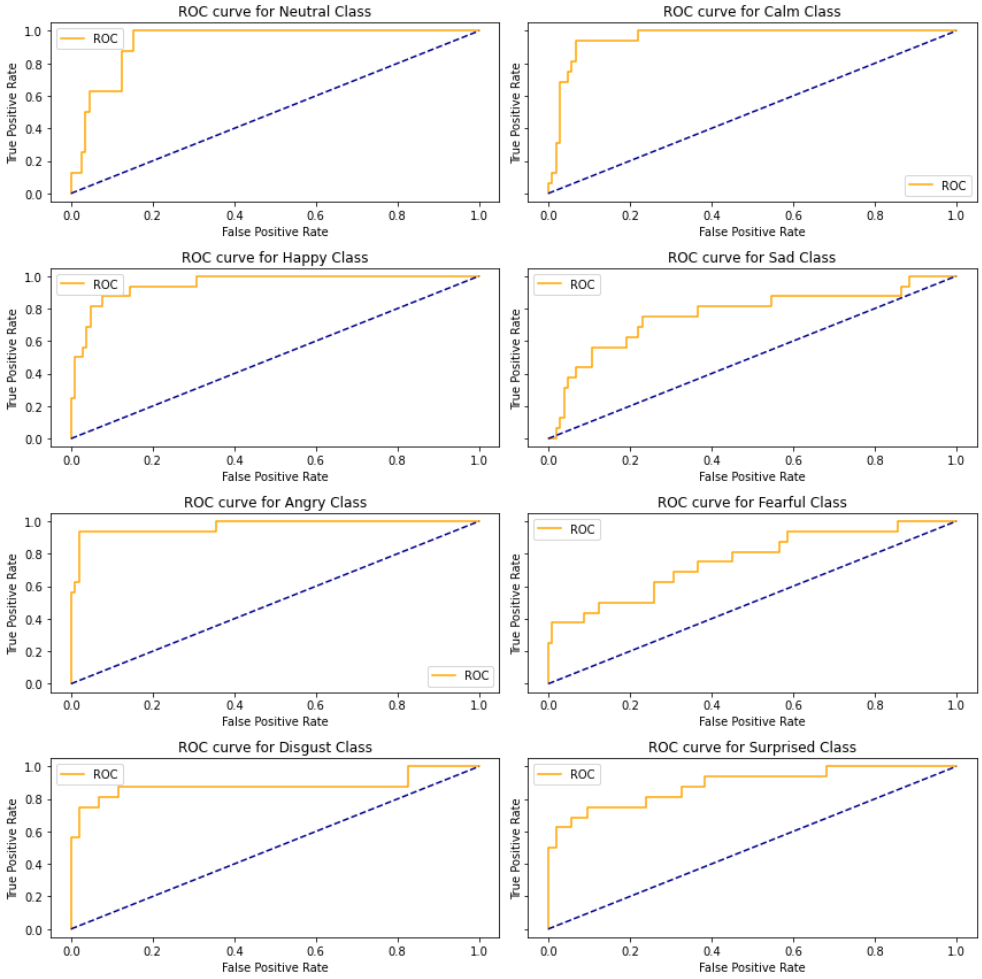

Fig. 4. ROC curves of the trained LSTM model. Each curve corresponds to an emotion label.

MFCC sequences are fed into an LSTM recurrent layer with a hidden dimension of size 1024. There are 2 LSTM layers stacked on top of each other, meaning that the outputs of the first layer are processed by the second one. This increases the perceptiveness of the network towards the features present in the sequence. Due to the dataset being small, we used dropout with high probability ($p = 0.5$) on the outputs of the first LSTM unit to prevent overfitting. The output of the last LSTM layer is then passed to a Multilayer perceptron (MLP), which outputs an 8-dimensional vector representing the logits of each emotion label.

The network was trained using only the RAVDESS dataset. The recordings of the 1st and 2nd actors (one male and one female) were used as a testing set, the rest of the recordings were used for training the network. Adam optimizer with learning rate of 0.0005 was used and the loss function to minimize was cross entropy loss given by:

$$l\left(\hat{y}_i\right) = \log\left(\frac{\exp\left(\hat{y}_i\right)}{\sum_j \exp\left(\hat{y}_j\right)}\right),$$

$$L\left(\widehat{y}_i\right) = -\sum_i y_i l\left(\widehat{y}_i\right),$$

where $\{\hat{y}_i\}$ are the estimated class labels, and $\{y_i\}$ are the ground-truth labels.

The classification results and comparison to the existing relevant method are demonstrated in Table 3. The Receiver Operating Characteristic curves (ROC curves) of the results are shown in Fig. 4.

3.3.2 Convolutional Neural Networks (CNNs)

As stated in subsection 3.2, the MFCC of a recording can be observed as a 2D feature map of a signal, with one dimension being the temporal dimension and the other being the cepstral coefficient dimension. Thus, a possible approach to working with MFCC's is processing them as spatial signals. Convolutional Neural Networks (CNN) are one of the most prominent architectures used for processing spatial data due to their shift equivariance, their inductive bias in searching for local patterns, and many other inherent benefits.

Thus, we consider solving the voice emotion classification task by training a CNN on extracted MFCC data. For MFCC calculation, the window size of 4096 and the overlap of between subsequent windows were chosen. Decreasing the window size by half degrades the performance of the network. On average, these settings produced better results. 4096 for a window size is good because it allows computing the FFT of length 4096 on that window to capture frequency spectrum of up to 4Khz. This means that the majority of human speech in those recordings is captured in each window. After calculating MFCCs for every recording and padding sequences with less length than the longest sequence, we obtain input matrices to our network of size (40 x 160) where at each sequence point we have 40 MFCCs.

The architecture of the CNN used is depicted in Fig. 5, and the method is summarized as follows:

There are 3 convolutional layers in the network followed by average pooling layers of size (2x2). The last layer is a fully connected layer that maps output of convolutional layers to an 8 length vector. Log softmax activation is applied to use cross entropy loss. Each layer has 32 kernels of parameters. The first layer has kernels of size (10x3), and it is deliberately chosen to be narrow and heighty to capture features from change of MFCCs through the sequence. Between layers, leaky rectified linear unit (ReLU) activation function given as h(x)=max(x, 0)+0.01*min(0, x)is used both to enable fast training and to prevent neurons

from dying. Leaky ReLU adds a small slope to non activated neurons thus preventing them from becoming 0 and not contributing to backpropagation in later epochs [14]. Since our dataset is very small, we used dropout with high probability (p = 0.5) as well as L2 regularization to prevent overfitting, which penalizes the sum of squares of the weights of the model.

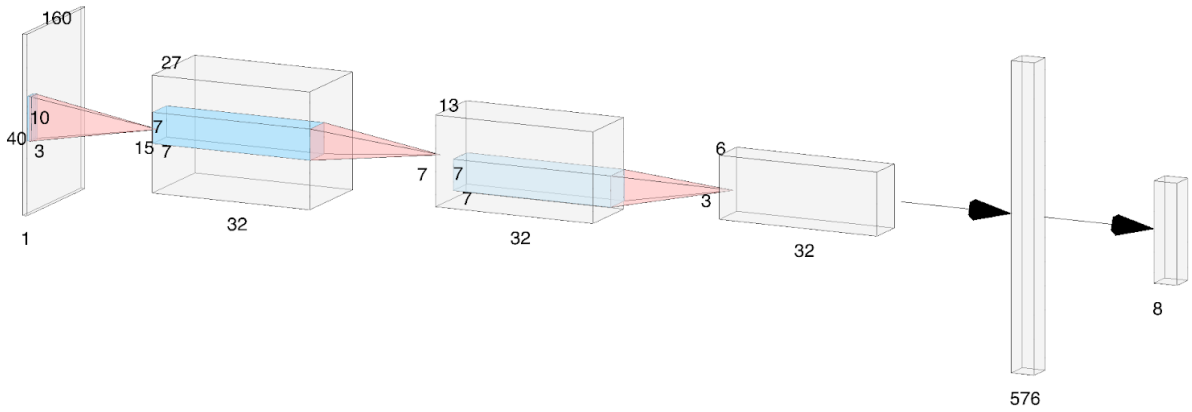

Fig. 5. The architecture of the trained CNN model.

All recordings of the 1st and 2nd actors, one male and one female from the RAVDESS database were used for testing, which the neural network was not trained on. All remaining recordings were used for training. We used Adam optimizer with a learning rate of 0.00005 and L2 regularization with decay of 10^{-4}. The final loss function becomes:

$$l(\widehat{y}_i) = \log\left(\frac{\exp(\hat{y}_i)}{\sum_j \exp(\hat{y}_j)}\right),$$

$$L(\widehat{y}_i) = -\sum_i y_i l(\widehat{y}_i) + \lambda \sum_{w \in W} w^2,$$

where W is the set of all trainable weights of the CNN.

The classification results and comparison to the existing related method are summarized in Table 3. Average ROC Area Under Curve (AUC) for all classes was 0.927. ROC curves for all classes are demonstrated in Fig. 6.

Table 3: Classification results.

Architecture	Train Accuracy	Test Accuracy	Mirsamadi et al. [5] Test Accuracy	Bertero et al. [4] Test Accuracy
LSTM	93.58%	65%	63.5%	-
CNN	96%	67.5%	-	66.1%

As it can be observed, the network captures some emotions more easily than others. For instance, Neutral, Calm, Angry and Surprise were captured better than the rest. ROC-AUC metric also suggests that the model learned meaningful representations for the task.

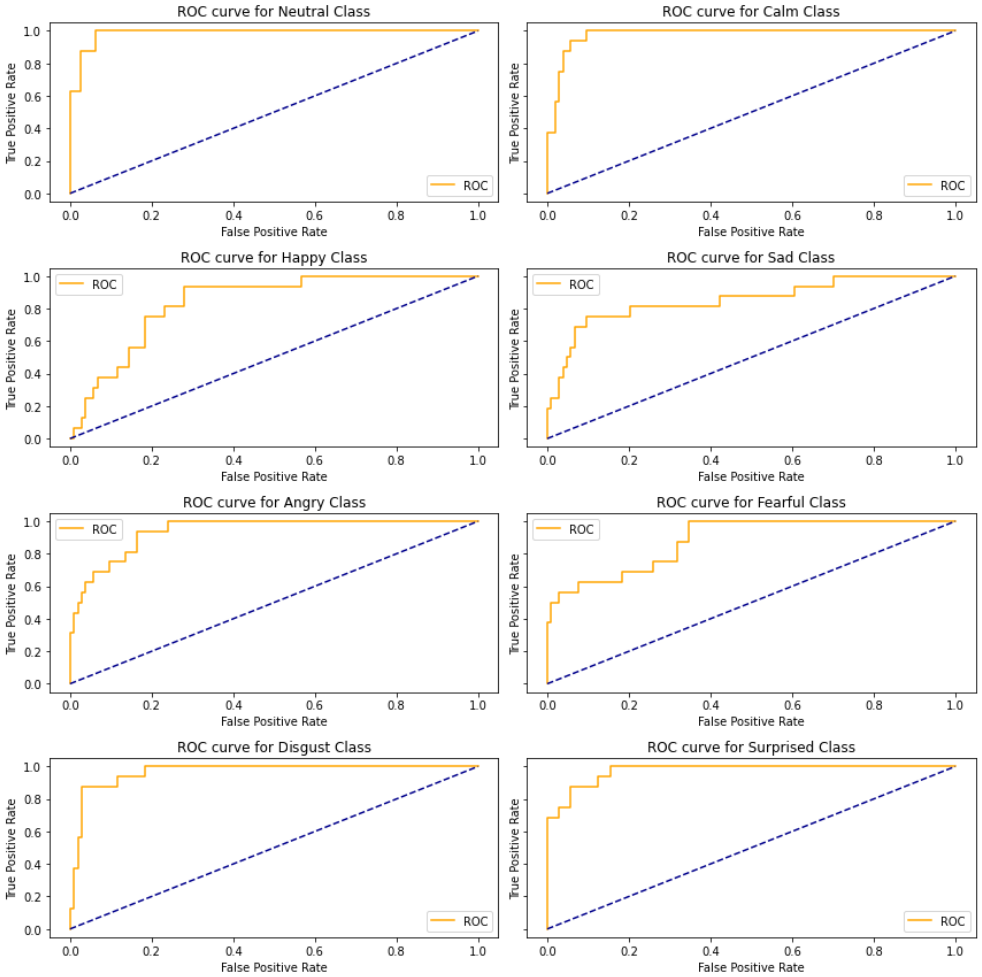

Fig. 6. ROC curves of the trained CNN model. Each curve corresponds to an emotion label.

Overall, the results show that the models managed to learn meaningful representations from the training procedure. In Table 3, we compare our results to the LSTM-based method of Mirsamadi et al. [5], which was trained and tested on the IEMOCAP benchmark [16] with a 4-label ("angry", "happy", "sad", "neutral") classification setting, as well as to the CNN-based method of Bertero et al. [4], which was trained and tested on the TED-LIUM benchmark [15] with a 3-label ("angry", "happy", "sad") classification setting. As it can be observed, our method gains superior results on our 8-label classification setting. In contrast to the 2 methods, we leverage only the MFCC representation of the signal, which highlights the efficiency of the MFCC representation and its usage with deep learning methods for the task.

4. Discussion and Conclusion

This paper proposes deep learning approaches for the voice emotion classification problem. Particularly, CNN and LSTM architectures were trained on MFCC features of voice recordings, depending on processing MFCCs either as a spatial signal or as a sequential signal. The results indicate that the networks have learned meaningful representations from the training data. A possible future direction for improving the classification performance of the pro-

posed models could be adding augmentations to the audio data. The recent advancements in using transformers [17] for multi-modal representation learning [18] and the expressiveness of the resulting feature space can also be a promising direction for solving the speech emotion recognition task.

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. Glüge, R. Böck 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] S.E. Eskimez, Z. Duan and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada., pp. 5099–5103, 2018.

[4] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA., pp. 5115–5119, 2017.

[5] S.Mirsamadi, E. Barsoum and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA., pp. 2227–2231, 2017.

[6] 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.

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

[8] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)", Scholars Portal Dataverse, 2020.

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

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

[11] 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.

[12] 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.

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

[14] B. Xu, N. Wang, T. Chen and M. Li, "Empirical evaluation of rectified activations in convolutional network", *CoRR* , vol. abs/1505.00853, 2015.

[15] A. Rousseau and P. Deleglise, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks", *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, pp. 3935-3939, 2014.

[16] C. Busso, M. Bulut, Chi-Chun Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database", *Language Resources and Evaluation* , vol. 42, no. 4, pp. 335–359, 2008.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need", *CoRR* , vol. abs/1505.00853, 2017.

[18] H. Akbari, L. Yuan, R. Qian, W.H. Chuang, S.-Fu Chang, Y. Cui and B. Gong, "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in Neural Information Processing Systems*, 2021.

Խորը ուսուցման վրա հիմնված ձայնագրությունների էմոցիաների դասակարգման մեթոդներ

Նարեկ Տ. Թումանյան

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

Ամփոփում

Տվյալ հոդվածում ներկայացվում են խորը ուսուցման վրա հիմնված ձայնագրությունների դասակարգման մեթոդներ: Առաջիո ազդանշանները մշակելու համար օգտագործվում են ձայնագրությունների հաճախական տվյալներ, որոնք հայտնի են ժամանակային ազդանշանների արդյունավետ ներկայացմամբ: Դասակարգման խնդիրը լուծելու համար հոդվածում հաշվի են առնվում հաճախական հատկանիշների մշակման երկու մոտեցում՝ որպես ժամանակային ազդանշանների մշակման մոտեցում և որպես տարածական ազդանշանների մշակման մոտեցում: Յուրաքանչյուր մոտեցման համար կիրառվում են համապատասխան արհեստական ցանցերի մոդելներ: Ներկայացվում է դասակարգման արդյունքների վերլուծություն, կատարվում են եզրակացություններ:

Բանալի բառեր՝ ձայնի տրամադրության ճանաչում, խոսքի էմոցիայի դասակարգում, հաճախական հատկանիշներ:

Классификация эмоций в голосе с использованием глубокого обучения

Нарек Т. Туманян

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

Аннотация

В этой статье мы представляем методы классификации эмоций в голосе с использованием методов глубокого обучения. Для обработки аудиосигналов, данный метод использует частотные признаки извлеченные из голосовых записей, которые, как известно, служат мощным представлением временных сигналов. Для решения задачи классификации, в данной работе рассматриваются два подхода обработки частотных признаков: как временные сигналы и как пространственные/2D-сигналы. Для каждого из подходов мы используем подходящие архитектуры нейронных сетей. Были проанализированы результаты классификации и представлены выводы.

Ключевые слова: определение настроения по голосу, распознавание настроения, классификации эмоций в голосе, частотные признаки.