An application of Artificial Intelligence in emotion detection by speech only

¹Arvind Sharma, ²R K Gupta

¹RJIT Tekanpur, arvinddevansh@rediffmail.com, ²MITS Gwalior, iitmrkg@gmail.com

Abstract:

Emotions of human are very important in our daily communications and recent years have witnessed a lot of research works to develop reliable emotion recognition systems based on various types data sources such as audio and video. Since there is no apparently visual information of human faces, emotion analysis based on only audio data is a very challenging task. In this work, a novel emotion recognition is proposed based on robust features and machine learning from only audio speech. For a person independent emotion recognition system, audio data is used as input to the system from which, Mel Frequency Cepstral Coefficients (MFCC) are calculated as features. The MFCC features are then followed by critical analysis to minimize the inner-class scatterings and maintain max value of the inter-class scatterings. The robust indistinguishable features are then applied with an efficient and fast deep learning approach Neural Structured Learning (NSL) for emotion training and recognition. The proposed approach of combining MFCC, discriminate analysis and NSL generated superior recognition rates compared to other traditional approaches such as MFCC-DBN, MFCC-CNN, and MFCC-RNN during the experiments on an emotion dataset of audio speeches. The system can be adopted in smart environments such as homes or clinics to provide affective healthcare. Since NSL is fast and easy to implement, it can be tried on edge devices with limited datasets collected from edge sensors. Hence, we can push the decision-making step towards where data resides rather than conventionally processing of data and making decisions from faraway of the data sources. The proposed approach can be applied in different practical applications such as understanding peoples' emotions in their daily life and stress from the voice of the pilots or air traffic controllers in air traffic management systems, and identification of leader's emotions or intentions behind their speech.

Keywords: Deep learning, MFCC, Adaptive system, Audio speech recognition etc.

1. Introduction

Speech is a natural way of humans to communicate each other in daily life. In affective computing research, speech has a vital role in promoting harmonious HCI systems and emotion recognition from speech is the first step. However, due to the lack of an exact definition of emotion, robust emotion recognition from audio speech seems to be quite complex. Hence, it demands a lot of research to solve the challenging problems beneath the audio-based emotion recognition [4].

Sound is a wave of pressure arises out of the vibration of molecules in a substance. The sound waves from a single source usually spread in all directions, eventually creates physical pressure on the ears. These waves are interpreted as electrical signals once transmitted to neurons. When a sound wave is generated from its source, it also vibrates the particles such as solid, liquid, and gas in its environment due to its energy. Sound waves always need a material form (i.e., solid, liquid, or gas) in an environment to travel. Sound waves can be divided into three categories based on their frequencies: below 20, between 20 and 20000 Hz, and more than 20000 Hz. Among which, sound waves in the middle category are human audible sound waves. These waves can be generated in various ways (e.g., musical instruments and vocal cords). Sound waves less than 20 Hz are called infrasonic sound waves. For example, earthquake waves. Sound waves of more than 20000 Hz are called ultrasonic sound waves [8]. Sound waves are used in different ways in science and technology such as ultrasound devices are used for imaging internal organs. High frequency ultrasonic waves are also used to break up stones in the kidneys.

Speech signal carries feelings and intentions of the speaker [7]. Speech signal analysis can be done in both time and frequency domains to obtain features to model underlying events (e.g., speaker, meaning of the speech, and emotion recognition) in the signals. Hence, original speech signal and corresponding spectrum diagram can be explored for robust emotion recognition for both the domains. In the speech signal in time domain was used as input and combined with machine learning model for emotion recognition. While studying the affective identification of speech information, the typical approach is to first use the raw audio signal processing and then followed by learning the extracted features with some machine learning models, for comprehensive pattern recognition or event prediction. Spectrogram analysis of the speech signal is also very common for speech pattern recognition [9]. In that case, the speech signal is windowed to small chunks and then divided into narrowband and broadband spectrum. Deep learning algorithms have been getting huge attention by pattern recognition and artificial intelligence researchers these days [10].

Deep neural network is typically better than the conventional neural networks. However, they often result in over fitting problem and take much time during training. Deep Belief Network (DBN) was a pioneer deep learning approach which utilizes Restricted Boltzmann Machines (RBMs) for training. Use of RBM makes DBN faster than typical neural network. Later, Convolutional Neural Networks (CNN) became very popular because of its improved discriminative power compared to DBN [11]. A typical CNN algorithm consists of convolution, pooling, tangent squashing, rectifier, and normalization. CNN consists of

feature extractions and some convolutional stacks to create a progressive hierarchy of useful features, especially effective for image processing tasks. Basically, it follows a hierarchical neural network structure where convolutional layers are followed by sub sampling layers. Finally, they are followed by fully connected layers that are identical to typical multilayer perceptron-based neural network. CNN-based deep learning approaches are very much used in visual scenery-based applications e.g., object detection in a large image achieve. Though CNN is used for many applications such as computer vision, most of the analysis of temporal events in time-sequential applications is adopted as Long Short-Term Memory (LSTM)-based Recurrent Neural Networks (RNNs). Hence, RNNs have become very popular for time-sequential event analysis. Besides, it can provide better discriminative power over DBN and CNN so far, for sequence-based pattern analysis. Amongst the available deep learning tools, Tensor Flow is one of the famous tools for deep learning-based event modeling tasks such as classification, prediction, and perception. Very recently, Google has introduced Neural Structured Learning (NSL), an open-source framework to learn neural networks. NSL can be utilized to construct robust models for in a wide range of research fields such as vision, natural language processing, and prediction in general [12]. It can be applied for training deep neural networks by leveraging structured signals with feature inputs. An NSL algorithm basically implements neural graph learning to train neural networks with the help of graphs and structured data.

Figure 1: A schematic picture of audio-based emotion recognition for human machine interaction in a room.

Structured data usually contains good relational information among the data samples. During the training process of a deep learning model, leveraging the structured signals provide supports to obtain better model accuracy. The structured signals are primarily applied to regularize the training of a neural network by driving the model towards learning as much as accurate predictions along with maintaining the input structural similarity. Thus, training with structured signals can lead to more robust deep learning models. Hence, NSL is adopted in this work from audio speech for better emotion recognition than the traditional approaches. Besides, the approach seems to be fast and accurate enough to be applied on edge devices (e.g., Raspberry Pi) for smart audio-based emotion recognition system with a limited number

of training samples. Figure 1 shows a schematic setup of rim audio-based emotion recognition in a smart room.

In this work, a novel emotion recognition method is proposed combining NSL with the Mel Frequency Cepstrum Coefficients (MFCC) based robust features obtained from audio speech. The MFCC features are first extracted from the raw speech data and then followed by discriminant analysis to represent a feature space highlighting to minimize the inner-class discrimination while maximizing the inter-class discrimination of different emotions. For the person independent emotion recognition modeling, the robust features are followed by a combination of neural graph learning of structured data as well as adversarial learning (i.e., NSL). The proposed emotion recognition method combining MFCC-based robust features discriminant analysis, and deep learning by NSL was compared with the traditional deep learning approaches such as DBN, CNN, or RNN where it showed the superior recognition performance over all the conventional methods. Since the approach is fast and robust on small datasets, it can be also adopted in various smartly controlled environments such as edge devices in smart homes or clinics to provide better affective healthcare. The proposed system can be adopted in various practical applications. For instance, such systems can be applied in a smart home for understanding the emotions in the daily life of people (e.g., elderly in smart old homes or clinics) or for predicting the mental stress of the pilots and air traffic controllers in the air traffic management systems.

2. Study of method

In the proposed emotion recognition method, the audio sensor data of all emotions is acquired for feature extraction and then applied for training an NSL model. For recognition process, the edge device obtains the features from a small chunk of audio speech and apply on the trained model for emotion recognition. Fig. 2 shows the basic architecture of the proposed method from signal data collection to the recognition process via feature extraction and training.

A sound signal is an electrical form of sound. An analog sound signal is the some copy of the sound whereas a digital sound signal is a numerical form derived from the analog sound followed by sampling and transforming them into digital ranging between 1 and 0, as shown in Fig. 3. Speech signals are basically complex signals in different frequencies where the spectrum analysis is done using a spectrogram. The Fast Fourier Transform (FFT) is most popular for spectrogram analysis in this kind of signals. FFT is applied to a window of signals to convert a sound signal from the time domain to the frequency domain. FFT is often used to measure of the frequencies of a signal using spectrogram. Spectrogram shows the intensity of vibrations based on frequencies. FFT can adopt a fast algorithm to reduce the time complication in the measurement of Discrete Fourier Transform (DFT).

Figure 2: Training and testing flowcharts of the proposed system.

Figure 3: A sample anger audio file from the experimental dataset.

Feature extraction with Mel Frequency Cepstrum Coefficients MFCC9

Mel Frequency Cepstrum coefficients (MFCC) are coefficients of a short-time windowed signal that is obtained through FFT. MFCC provides better results than the operations in time domain. MFCC utilizes the Mel scale based on the sensitivity of the human ear. MFCC is very popular and often used for feature extraction in the frequency domain for different sound-based applications.

Human ears usually pick up sound frequencies until 1000 Hz in Mel scale. A triangular filter is applied in the Mel spectrum where the bandwidth varies according to the Mel scale. A normal frequency E is converted to Mel frequency F as

$$F = 1127 \log_{10}(1 + f/700)$$

Furthermore, DFT coefficients are explored according to the amplitude frequency response of the Mel filter bank. The amplitude spectrum of the signal is distributed along the Mel scale. The spectrum contains equal intervals and multiplied by the triangular filter. Then, the logarithm of the remaining energy is calculated. As the logarithm of the Mel spectrum coefficients are real numbers, the time domain values can be returned using the discrete cosine transform. The coefficients obtained via this process are called MFCC. One set of 12 MFCC coefficients is extracted for each frame.

All 12 coefficients are shown in Figure -4

Figure 4: All 12 coefficients of sample of audio speech

The road map for MFCC is given in Figure-5

Figure-5: A road map or steps of MFCC.

Emotion modeling is based on Neural Structured Learning (NSL) and it is a deep learning approach that focuses on training the neural networks by leveraging structured signals along with the input features. The structured signals are used to regularize the training of a neural network that forces to learn accurate predictions with the help of minimizing supervised loss. At the same time, it tries to maintain the input structural similarity with the help of minimizing the neighbor loss. The approach is very generic and can be utilized on any arbitrary neural architectures such as typical Feed-forward neural networks, CNN and RNN. Hence steps for emotion detection by speech can be summarized in following steps-

1. Frame the signal into short frames.
2. For each frame calculate the period gram estimate of the power spectrum.
3. Apply the Mel filter bank to the power spectra; sum the energy in each filter.
4. Take the logarithm of all filter bank energies.
5. Take the DCT (Discrete Cosine Transform) of the log filter bank energies.
6. Keep DCT coefficients 2-13, discard the rest

Various method of emotion detection studied well during this article and found that basic method is ANN (Artificial Neural Network) and after refinement we get CNN (Convolutional Neural Networks) and then LSTM (Long Short-Term Memory) to get epochs and fivefold confusion matrices [13]. A sample is shown here-

Figure: (a) Loss and (b) accuracy of the LSTM model for 100 epochs where Err varies from 0 to 1 and Epochs from 0 to 100.

Figure: Confusion matrices of five folds from (a) to (e) using LSTM.

3. Experiments and results

An audio speech database was collected for this work containing four expressions: namely neutral, angry, happy, and sad [3]. There were 339 audio clips where the duration of each clip was around 2 s. For the experiments, five-fold cross validation was applied. The summary of the results is shown in Table 1. To achieve a good audio quality, the recordings were done in an anechoic chamber of the Technical University Berlin, Technical Acoustics Department. Each audio consists of a sampling frequency of 48 kHz, which later was made to 16 kHz by down sampling. The actors stood before the

microphone and spoke in the direction of the microphone from about the distance of 30 cm. Three phoneticians supervised the recording of each session of the users where two of them were giving instructions and one was monitoring the functions of the recording equipment. For some emotions, there were several variants of the same emotions. The actors were also instructed about not to unnecessarily shout to express anger or to avoid whispering while expressing their anxiety during the audio recording [14]. This was done due to maintain and analyze the voice quality. In addition, the recording levels were adjusted between very looted speech and very quite speech. Finally, twenty subjects evaluated the speech data with satisfactory results, ensures the emotional quality and naturalness of the utterances of the audio data.

Table-1

The mean recall rates of the emotions for all the folds using different approaches.

Emotion	odel	ANN	CNN	LSTM	NSL
Neutral		0.75	0.94	0.89	0.92
Angry		0.82	0.90	0.92	0.95
Happy		0.37	0.71	0.84	0.85
Sad		0.82	0.94	0.8	0.97
Mean		0.69	0.87	0.88	0.93

The proposed NSL-based experiments for emotion modeling and recognition is shown in Figure 7

Figure 7: Confusion Matrices using NSL

4. Conclusion

A basic audio-based emotion recognition system consists of three major parts: acquisition of audio signals, feature processing that tries to obtain distinguishable robust features for each emotion so that each expression can be represented as much different from each other, and emotion recognition that recognizes emotion by applying robust features on a strong pre-trained expression model. In this work, we have proposed a novel approach for emotion recognition from audio speech signals where MFCC features are tried with discriminate analysis and a state-of-the-art deep learning approach i.e., neural structured learning based on neural graph learning and adversarial learning. The proposed emotion recognition approach was compared with traditional approaches where it showed its superiority over others. System could be adopted to contribute in any smartly controlled environment for audio-based emotional healthcare

5. References

1. Aghdam, H.H., Heravi, E.J., Puig, D., 2017. *Explaining adversarial examples by local properties of convolutional neural networks*. In: *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.
2. Anon, 2020. *Neural structured learning: Training with structured signals*, Tensor-flow. [Online]. Available: https://www.tensorflow.org/neural_structured_learning. (Accessed: 06-June-2020).
3. Asl, B.M., Setarehdan, S.K., Mohebbi, M., 2008. *Support vector machine-based arithmetic classification using reduced features of heart rate variability signal*. *Artif. Intell. Med.* 44, 51–64.
4. But, T.D., Ravi, S., Ramavajjala, V., 2018. *Neural graph learning*. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*.
5. Burkhardt, F., Paeschke, A., Rolfes, S., Sendlmeier, W.F., Weiss, B. *A database of German emotional speech in Interspeech ZOOS- Euro special*, 9th European Conference on Speech and Technology (Lisbon). pp. 1517–1520, 2005
6. Fischer, D.C., 2014. *Training restricted Boltzmann machines: An introduction*. *Pattern Recognition* 47 (1), 25–39.
7. Fleury, A., Noury N., Yacher M., Glasson, H., Seri, I.-F., 2008. *Sound and speech detection and classification in a Smart Home*. In: *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. pp. 4644–4647.
8. Fujimoto, N., 2017. *Factored deep convolutional neural networks for noise robust speech recognition*. In: *Interspeech 2017*.
9. Gravina, R., Fortino, G., 2016. *Automatic methods for the detection of accelerative cardiac defense response*. *IEEE Transaction* 7 (3), 280–298.
10. Hassan, M.M., Alam, M.G.R., Uddin, M.Z., Huda, S., Almogren, A., Forting, G., 2019.
11. *Human emotion recognition using deep belief network architecture*. *Fusion* 51, 100–110, 2006.
12. *A fast learning algorithm for deep belief nets*. *Neural Computer* 18 (7), 1527–1554.

13. Le, D., Provost, E.M., 2013. *Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks*. In: *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*.
14. Li Yun, K.C. Popescu, 2012. *A microphone array system for automatic fall detection* *IEEE Trans. Biomed. Eng.* 59 (5), 1291—1301.