Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/130521

The Robust Spectral Audio Features for Speech Emotion Recognition

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

Suleyman Demirel University, Kaskelen, Kazakhstan

Received: 1 Mar. 2019, Revised: 2 Aug. 2019, Accepted: 6 Aug. 2019
Published online: 1 Sep. 2019

Abstract: This paper describes a revealing robust spectral feature for speech emotion recognition using Deep Neural Network (DNN) architecture with six fully-connected layers. We have used 3 class subset (angry, neutral, sad) of German corpus (Berlin database of emotional speech) containing 271 labeled recordings with a total length of 783 seconds. All data was divided into TRAIN (80 %) VALIDATION (10 %) and TESTING (10 %) sets. DNN is optimized using Stochastic Gradient Descent. And we have used batch normalization. As input, fourteen features were used and supported by the LIBROSSA library. Features are compared between each other. In accordance with the experiment we have discovered that MFCC with 100 percent accuracy is a reliable function for the task of recognizing emotions.

Keywords: MFCC, spectral audio features, speech emotion recognition, Deep Neural Network, feature extractions.

1 Introduction

Humans are emotional creatures. People aspire to understand each other and want that other people react to their emotion. In the age of information and automation when the robotics sphere is improving each day and people interact with automation systems while it is necessary to recognize emotions for better relationships between humans and machines. Undoubtedly emotion recognition is very important in AI spheres since it will make Human-Computer Interface (HCI) more friendly and similar to the real man behavior.

In the process of analysis of Speech Emotion Recognition (SER) the first issue after getting a dataset is how to present audio voice to machine data for working with building the ML models. There are many options for how to represent the sound in machine data, from the simplest ones like Fourier transform to spectral and prosodic features. In the work[1] the spectral and prosodic features were compared and it was proved that the spectral features are stronger for emotion recognition. Most of the famous studies (mentioned in the second part of this article) related to SER used Mel-Frequency Cepstral Coefficients (MFCC) feature or mentioned the use of spectral features without details.

The aim of this research work is to discover existing types of spectral features and define the robust type of feature through comparative analysis based on open existed Berlin Dataset (BD) and model of Deep Neural Networks (DNN).

2 Literature Review

The paper[2] written by Fei Tao, Gang Liu, Qingen Zhao imparts us that they used openSMILE[3] for the extract feature set. But nothing is said which features they used.The official site of openSMILE[3] has the next extraction of speech-related features: Signal energy, Loudness, Mel-/Bark-/Octave-spectra, MFCC, PLP-CC, Pitch, Voice quality (Jitter, Shimmer), Formants, LPC, Line Spectral Pairs (LSP), Spectral Shape Descriptors.

Pavol Harár, Radim Burget and Malay Kishore Dutta in their thesis[4] mentioned the next features: Mel energy spectrum dynamic coefficient (MEDC), Linear prediction cepstrum coefficient (LPCC), Perceptual linear prediction cepstrum coefficient (PLP), pitch, formants, and energy. But nothing is said whose features they used in this main experiment which gave a 96.97% of accuracy.

The MFCC feature was used in the scientific work[5] and had the 71.33% accuracy.

* Corresponding author e-mail: aisultan.shoiynbek@sdu.edu.kz

Based on the last actual articles in SER and comparative analysis shown in[1], we can make a consequence that MFCC is the most popular feature for SER. In the next sections, we compare MFCC with other existing spectral features.

3 Description of Dataset

The data set that was used in this work is German corpus (Berlin database of emotional speech) which contains about 800 sentences (7 emotion classes * 5 female and 5 male actors * 10 different sentences + some second versions). All sentences are recorded in the anechoic chamber using high-quality equipment with sampling frequency of 48kHz and later down sampled do 16kHz (mono)[6].

This preliminary experiment was conducted on a smaller subset of this corpus containing 271 labeled recordings with total length of 783 seconds. Because of non-equality between classes and in order to get comparable results with[7], we used all sentences from all actors but only from 3 emotional states: angry (127 recordings, 334 seconds), neutral (79 recordings, 186 seconds), sad (65 recordings, 263 seconds).

The dataset is splitted into TRAINING 80,81% (219 files) VALIDATION 9.594% (26 files: 12-angry; neutral-8, sad - 6) and TESTING 9.594% (26 files: 12-angry; neutral-8, sad - 6). TESTING set used for testing is taken from files that have not been seen by DNN during training or validation.

4 Features

For feature extraction, LibROSA[8] used python package for music and audio analysis. Librosa have a 14 spectral features. They are as follows:

1. Chroma_stft. Compute a chromagram from a waveform or power spectrogram.This implementation is derived from chromagram_E .
2. Chroma_cqt. Constant-Q chromagram
3. Chroma_cens. Computes the chroma variant “Chroma Energy Normalized” (CENS), following[9]
4. Melspectogram.
5. MFCC
6. RMSE. Compute root-mean-square (RMS) energy for each frame, either from the audio samples y or from a spectrogram S. S is spectrogram magnitude. It is required if y is not input. Computing the energy from audio samples is faster as it doesn’t require a STFT calculation. However, using a spectrogram gives a more accurate representation of energy over time because its frames can be windowed, it is preferable to use S if it’s already available.
7. Spectral_centroid. Compute the spectral centroid. Each frame of a magnitude spectrogram is normalized and treated as a distribution over frequency bins, from which the mean (centroid) is extracted per frame.
8. Spectral_bandwidth
9. Spectral_contrast. Compute spectral contrast[10]
10. Spectral_flatness. Compute spectral flatness. Spectral flatness (or tonality coefficient) is a measure to quantify how much noise-like a sound is, as opposed to being tone-like[11] A high spectral flatness (closer to 1.0) indicates that the spectrum is similar to white noise. It is often converted to decibel.
11. Spectral_rolloff. Compute roll-off frequency.The roll-off frequency is defined for each frame as the center frequency for a spectrogram bin such that at least roll_percent (0.85 by default) of the energy of the spectrum in this frame is contained in this bin and the bins below. This can be used to, e.g., approximate the maximum (or minimum) frequency by setting roll_percent to a value close to 1 (or 0).
12. Poly_features. Get coefficients of fitting an nth-order polynomial to the columns of a spectrogram.
13. Tonnetz. Computes the tonal centroid features (tonnetz), following the method o[12].
14. Zero-crossing rate. Compute the zero-crossing rate of an audio time series.

5 DNN architecture

DNN architecture contains six fully-connected layers with activation function relu[13], and last layer is also fully-connected but with activation function softmax. The first layer of neural network is a layer with 320 neurons; the second layer contains 160 neurons; The third layer has 80 neurons; the fourth layer contains 40 neurons; The fifth layer has 20 neurons; the sixth layer contains 10 neurons. The last layer with activation function softmax contains 3 neurons. For regularization of the DNN, we have used a 0.2 dropout[14] between third and fourth layers and batch normalization before first layer. All layers were initialized using Glorot uniform initialization[15]. Detailed information about the architecture is shown in Figure 1.

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

From audio data we alternately extract into the analyzed features in section 4 and get an input data to the above DNN model.

6 Results

For each function, a model with a different number of epochs is trained. In the experiment, we have considered

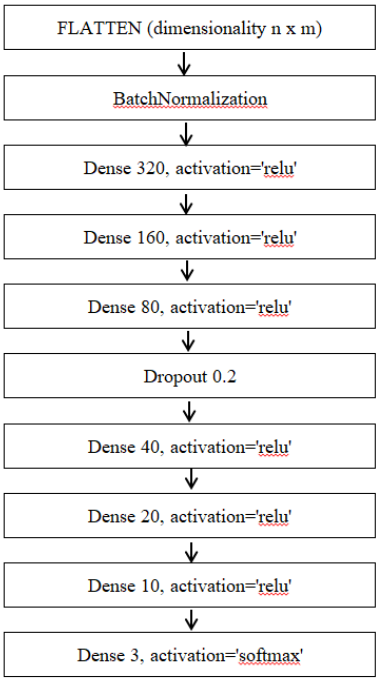

Figure 1. Detailed architecture of proposed DNN

loss and accuracy in the context of 5,10,20,40 epochs. The results are displayed in the Table 1.

Table 1: Comparative analysis of spectral features.

Feature	Dim	A5	A10	A20	A40
Chroma_stft	12x320	57,69	34.61	61.53	61.53
Chroma_cqt	12x320	30.76	65.38	53.85	53.85
Chroma_cens	12x320	73.08	73.08	84.61	76.92
Melspectogram	128x320	80.77	76.92	80.77	84.61
MFCC	20x320	100	96.15	96.15	96.15
RMSE	1x320	69.23	76.92	76.92	76.92
Spec_centroid	1x320	69.23	69.23	76.92	76.92
Spec_bandwidth	1x320	50	53.84	53.84	65.38
Spec_contrast	7x320	38.46	65.38	65.38	69.23
Spec_flatness	1x320	61.53	73.07	76.92	73.07
Spec_rolloff	1x320	61.53	73.07	73.07	69.23
Poly_features	2x320	76.92	57.69	73.07	69.23
Tonnetz	6x320	53.85	50	57.69	50
Zero-crossing rate	1x320	84.61	88.46	88.46	73.07

A – is accuracy of Test set Dim – is Dimensionality In the Table 1 are presented accuracy and losses of TEST set. According to the data of Table 1, features with an accuracy greater than 80% are defined, such as: Chroma_cens feature and Melspectogram feature achieved 84.61% accuracy on 20 epochs in training and 40 epochs in training respectively. Zero-crossing rate feature got the 88.46% accuracy on 10 and 20 epochs in training. MFCC has reached the 100% accuracy on 5 epochs in training. MFCC feature with growing number of epoch accuracy decreases to 96.66%. It means that training set with MFCC feature learned faster than other features.

7 Conclusion

As it is mentioned above the aim of this paper is to discover existing types of the spectral features and define the robust type of feature through comparative analysis based on open existed Berlin Dataset (BD) and model of Deep Neural Networks (DNN). We have compared 14 spectral features described in the second part of the research paper. We have designed the DNN model with six fully-connected layers set the same parameters and conditions for comparing each feature.

Based on Table 1 we define 4 features with an accuracy greater than 80%. MFCC has a best accuracy - 100%. Second feature is zero-crossing rate with 88.46% accuracy. The difference between Zero-crossing rate and MFCC is about 11.54% of accuracy. That difference is significant; also MFCC learning is faster than Zero-crossing rate. Evidently that MFCC reached its best accuracy on fifth epochs in training set while Zero-crossing rate reached only tenth epochs.

Due to the experiment we have discovered that MFCC is a reliable function for the task of recognizing emotions.

References

[1] Dmitri Bitouk, Ragini Verma,and Ani Nenkovab *Class-Level Spectral Features for Emotion Recognition*, Speech Commun, 52(7-8): 613–625, (2010)

[2] Fei Tao1, Gang Liu2 , Qingen Zhao2 , *An ensemble framework of voice-based emotion recognition system for films and tv programs*, IEEE International Conference on Acoustics, Speech and Signal Processing, (2018)

[3] Florian Eyben, Martin Wöllmer, Björller, *openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor*, In Proc. ACM Multimedia (MM), 1459-1462, (2010).

[4] Pavol Harár, Radim Burget and Malay Kishore Dutta , *Speech Emotion Recognition with Deep Learning*, 4th International Conference on Signal Processing and Integrated Networks (SPIN) , 137, (2017)

[5] Eduard FRANT, Ioan ISPAS, Voichita DRAGOMIR, Monica DASCALU Elteto ZOLTAN, and Ioan Cristian STOICA, Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots, *ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY* , Volume 20, Number 3, 222–240, (2017)

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

[7] Brian McFee, Colin Raffel§ , Dawen Liang, Daniel P.W. Ellis , Matt McVicar, Eric Battenberg, Oriol Nietok. *librosa: Audio and Music Signal Analysis in Python*. PROC. OF THE 14th PYTHON IN SCIENCE CONF(SCIPY)., 18, (2015).

[8] Ellis, Daniel P.W. *Chroma feature analysis and synthesis* In Proceedings of the International Conference on Music Information Retrieval, 04-21, (2007)

[9] Meinard Müller and Sebastian Ewert, *Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features* , In Proceedings of the International Conference on Music Information Retrieval (ISMIR), (2011).

[10] Jiang, Dan-Ning, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, *Music type classification by spectral contrast feature.*, IEEE International Conference, vol. 1, 113-116. IEEE, (2002)

[11] Dubnov, Shlomo, Generalization of spectral flatness measure for non-gaussian linear processes, *IEEE Signal Processing Letters*, Vol. 11, (2004).

[12] Harte, C., Sandler, M., & Gasser, M., *Detecting Harmonic Change in Musical Audio*. In Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia , 21-26, (2006).

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

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

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

[16] Zhang, T., *Solving large scale linear prediction problems using stochastic gradient descent algorithms*. In Proceedings of the twenty-first international conference on Machine learning , 116, (2004).

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

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

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

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