Emotion Recognition in Raw Speech Recordings Using Convolutional Neural Networks and Spectrograms

Mateusz Półtorak
Pearson IOKI Sp. z o. o.
Dąbrowskiego 77A, 60-529 Poznań, Poland
e-mail: *mateusz.poltorak@pearson.com*

Mikołaj Morzy
Poznań University of Technology
Piotrowo 2, 60-965 Poznań, Poland
e-mail: *mikolaj.morzy@put.poznan.pl*

Abstract

Emotion recognition from raw speech is an important and challenging task which has not found a satisfactory solution as of today. Over the last years, many SER (speech emotion recognition) methods have been proposed. The majority of proposals depend on hand-crafted features based on the continuous and spectral characteristics of the audio signal. Sometimes, these methods are augmented with affective evaluation of spoken contents, which requires an automatic speech recognition system (ASR) output. The common weakness of current methods is the fact that their effectiveness depends directly on the set of selected features. In this work we present a method of automatic feature extraction for speech emotion recognition using convolutional neural networks (CNNs). We use spectrograms to perform convolutions and extract useful features from the raw audio signal. Experiments conducted on the IEMOCAP benchmark dataset prove the efficiency of our approach.

1 Introduction

Expression of emotions is one of the most interesting forms of communication developed by humans. Emotions are ensembles of non-verbal and sound signals. As such, they constitute an important part of the communication process. In order to enable robust and natural human-computer interfaces, machines have to be able to recognize human emotions precisely. Unfortunately, the task of emotion recognition from speech has not yet found a satisfactory solution and current SER systems fail to predict human emotions with sufficient accuracy.

Most of the state-of-the-art SER systems employ either Support Vector Machine (SVM) classifiers or dense neural networks. Features used to train these classifiers usually include a mixture of characteristics extracted from the energy of the speech system, e.g., pitch, formants, or spectral features such as mel-frequency cepstral coefficients (MFCCs) and linear predictive cepstral coefficients (LPCCs) [1]. It is very questionable, however, whether this set of features corresponds to the features used by human brains when recognizing emotions in speech. Another approach is to use affective information contained in the semantics of spoken language, but this requires a robust automatic speech recognition as an upstream task and introduces an inherent bias of the selected sentiment analysis method applied to the speech transcript (not to mention the systematic word recognition errors introduced by ASRs). One might also argue that sentiment is different from emotion. Besides, emotion is often carried by non-word utterances, which are seldom correctly identified by ASRs which are limited to the vocabulary of the training set.

In this work we argue that it is possible to perform emotion recognition in raw speech using automatic feature extraction, without the need to recognize spoken words. We use a pre-trained convolutional neural network VGG-19 [4] and we fine-tune it to detect spectral features carrying affective information. In order to be able to use CNNs on raw speech recordings, we have to transform input signals into data representation format suitable for convolution, where physical proximity of pixels can be harvested to discover useful affective features. We use spectrograms generated using the Short Time Fourier Transform [3], which contain all temporal, as well as spectral information from speech signal. In addition, the use of spectrograms allows us to process not only information related to speech energy (pich, volume, frequency), but to exploit prosodic features as well (intonation, stress, pause, tone).

2 Architecture

We use a deep, convolutional neural network model with additional fully connected layer at the end of the network. The architecture of the CNN is identical to the VGG-19, but we replace the last layer with a 4 neuron softmax for final emotion classification. All hidden layers use ReLU as their activation function,. Image downsampling is performed with max pooling with the stride of 2. We align the shape of input spectrograms to the input of the VGG-19, which has the size of 224x224x3 pixels.

3 Experiment

The efficiency of our proposal is evaluated on the IEMOCAP benchmark [2]. This database contains one of the most extensive collection of labeled speech data, thus it is widely used in SER systems evaluation. The training dataset is highly imbalanced as the number of examples within each emotion class varies significantly. Usually the benchmarks of SER systems are performed with only four emotions (sad, happy, angry, neutral) because only these classes are

represented in the training set by a sufficient number of examples. We use overall (weighted) accuracy as our main evaluation metrics.

Input spectrograms must have a constant size, but some speech samples are much longer than others. As the countermeasure, we have to segment these long samples into pieces of equal length. Segmentation is done by moving a two-second-long window along audio samples with a 0.5-second step. This mechanism creates multiple spectrograms annotated with the same label as the audio file from which the spectrograms are generated. The whole process is depicted in Figure 1. During inference, CNN analyses all spectrograms which belong to a given audio file and the final prediction is the maximum summary response over these spectrograms.

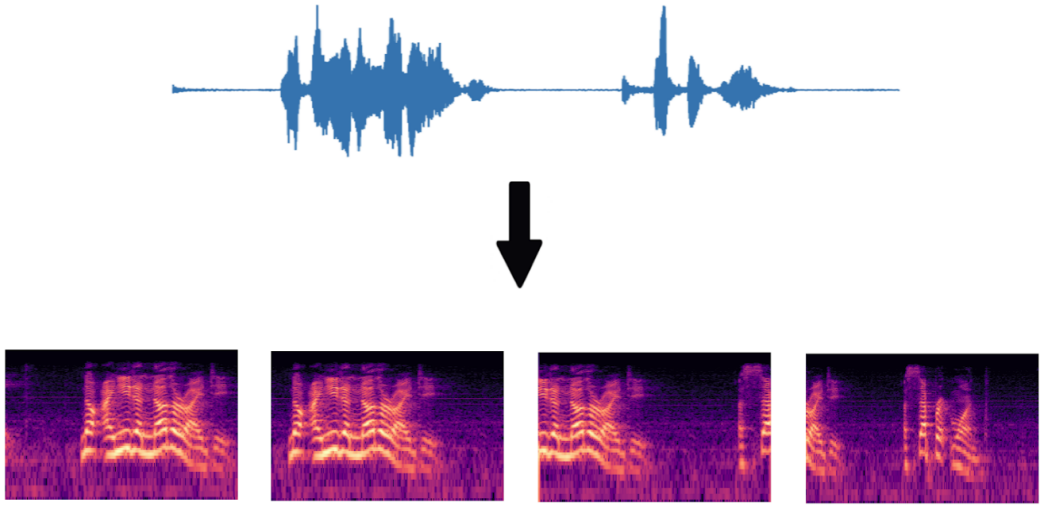
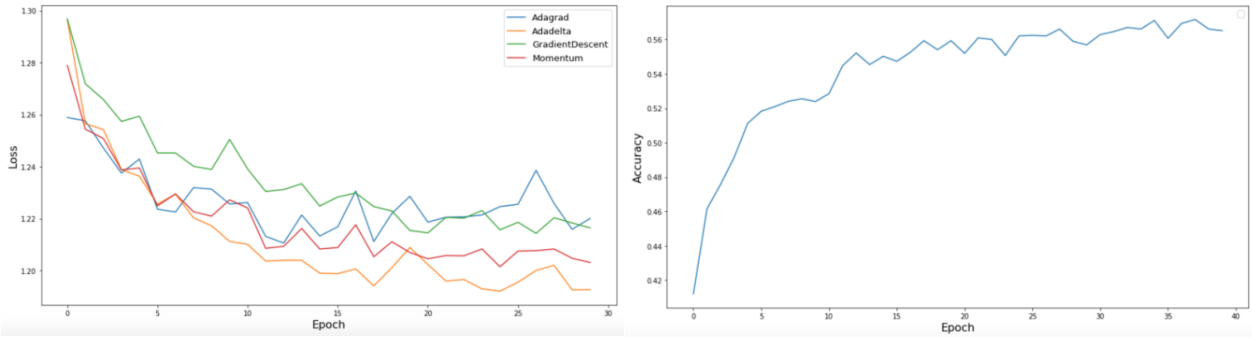


Figure 1: Process of segmenting audio files into spectrograms.

Individual samples in the IEMOCAP dataset display significant variability within each emotional class. As the consequence, convolutional neural network training was extremely unstable. VGG-19 network parameters were updated using Adadelta [7] optimizer with learning rate set to 0.001 and mini-batch size set to 64. The comparison of different optimizers performance is presented in Figure 2a The popular Adam optimizer [5] had failed to converge and has been omitted from the comparison.



(a) (b)

Figure 2: (a) Loss value depending on epoch number and selected optimizer (b) Weighted accuracy depending on epoch number during 10-fold cross validation

Final results were evaluated with 10-fold cross validation. Figure 2b presents the change of classification accuracy across training epochs. The highest accuracy score achieved by our model was 57%, which is 3% below state-of-the-art results [6]. However, our system is fully automatic and does not require any manual feature engineering, whereas best performing SER systems employ multiple custom feature extraction methods and audio signal parametrization.

4 Conclusion

In this work we show that raw speech spectrograms carry substantial information on affective and emotional contents of speech. Our approach of automatic *feature learning* applied to raw audio recordings allows us to abandon traditional methods based on manual feature engineering, leading to self-contained intelligent systems, without suffering a significant loss of classification accuracy. We avoid the dependence on the ASR system for spoken language transcription and we omit the process of sentiment analysis based on the semantics of utterances. Our method is completely language-agnostic as it does not depend on any upstream language model. We are also able to utilize affective signals contained in prosodic features and non-word utterances. A possible downside of our approach is the high computational cost and long training time. Also, spectrograms, represented as 224x224 float32 RGB matrices consume a lot of GPU memory, which negatively influences the learning process.

References

[1] Babak Basharirad and Mohammadreza Moradhaseli. Speech emotion recognition methods: A literature review. In *AIP Conference Proceedings*, volume 1891, page 020105. AIP Publishing, 2017.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[3] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[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. In *INTERSPEECH*, 2017.

[7] Matthew D. Zeiler. Adadelta: An adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.