A FIRST LOOK INTO A CONVOLUTIONAL NEURAL NETWORK FOR SPEECH EMOTION DETECTION

Dario Bertero, Pascale Fung

Human Language Technology Center
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
dbertero@connect.ust.hk, pascale@ece.ust.hk

ABSTRACT

We propose a real-time Convolutional Neural Network model for speech emotion detection. Our model is trained from raw audio on a small dataset of TED talks speech data, manually annotated into three emotion classes: "*Angry*", "*Happy*" and "*Sad*". It achieves an average accuracy of 66.1%, 5% higher than a feature-based SVM baseline, with an evaluation time of few hundred milliseconds. We also provide an in-depth model visualization and analysis. We show how our neural network effectively activates during the speech sections of the waveform regardless of the emotion, ignoring the silence parts which do not contain information. On the frequency domain the CNN filters distribute throughout all the spectrum range, with higher concentration around the average pitch range related to that emotion. Each filter also activates at multiple frequency intervals, presumably due to the additional contribution of amplitude-related feature learning. Our work will allow faster and more accurate emotion detection modules for human-machine empathetic dialog systems and other related applications.

Index Terms— deep learning, emotion detection, convolutional neural networks, neural network visualization

1. INTRODUCTION

Recognizing the emotions expressed in a speech signal, as well as in other modes, is an hard task. It is characterized by a level of subjectivity in defining and perceiving an emotion, as well as by a lack of a univocal definition of standard descriptors for each specific emotion [1, 2]. In recent years people have delegated the role of learning emotional models to deep neural networks, which have superseded the state-of-the-art methods. Several neural network variants were developed that take as input traditional prosodic features [3], spectrograms [4] or directly raw audio samples [5]. The latter are the most promising, as they entirely eliminate all the overhead required for the feature extraction step, while yielding equally good or superior performance. Deep learning from raw audio is now replacing traditional feature-based learning in all speech-related tasks, with Automatic Speech Recognition the most prominent field [6, 7, 8].

While deep learning is being applied to many different tasks, often superseding former state-of-the-art methods, researchers are somehow ignoring the issue of what the model is actually learning. Some timid attempts to visualize the neural network activation have been proposed for very well-known tasks in image recognition [9], Natural Language Processing [10] and ASR [8]. For other less widespread problems, to our knowledge, no studies of this kind exist. Emotion detection is one of those tasks, where although people have replaced shallow classifiers [11] with deep learning, there have not been enough attempts to understand what happens inside the DNNs. We believe it is not straightforward, as even for humans is not easy to define and cluster emotions. Being able to give proper interpretations is nevertheless an important challenge to tackle, in order afterwards to develop better and faster learning models.

In this paper we first propose a real-time, lightweight CNN model, able to process speech segments in a few hundred milliseconds on a low-end consumer notebook [12]. Fast processing of speech signal is extremely important for the development of machine dialog systems able to instantly react to the user inputs, either acoustic, textual or visual [13]. We then conduct an in-depth analysis of the model, showing where our emotion model activates in time and in frequency. Such analysis represents another important step towards our goal to build a empathetic robot able to feel and react to emotions like a human would do [13, 14, 15].

We concentrate on three basic emotions: *anger*, *happiness* and *sadness*. There is generally no agreement on which emotions constitute a fundamental set, or even if the concept of "fundamental emotion" can be defined [16]. Our empirical results on annotation and classification [17] suggested us that these three descriptor are sufficiently easy to annotate for humans and to distinguish for machines without having a large

This work was partially funded by the Hong Kong Phd Fellowship Scheme, and partially by grant #16214415 of the Hong Kong Research Grants Council.

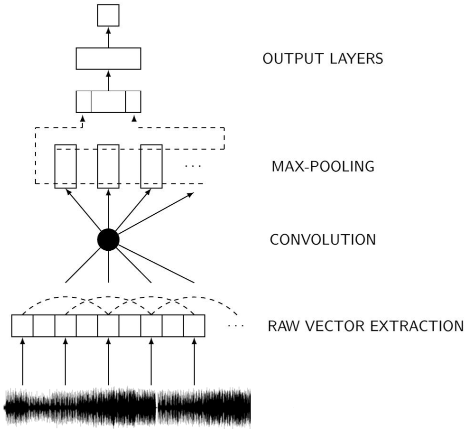


Fig. 1. Convolutional Neural Network for emotion detection.

dataset. We plan in the future to extend our analysis to other emotions and other data domains, after we collect and annotate more speech data.

2. CONVOLUTIONAL NEURAL NETWORK FOR EMOTION DETECTION

We train and analyze a Convolutional Neural Network (CNN) to detect emotions from raw audio. The network is designed with one convolutional layer to make it run very fast at evaluation time, a few milliseconds compared to several seconds of a two-layer structure in a standard desktop machine [12], thus making it very suitable for real-time applications [13, 14].

Our network layout is shown in Figure 1. It takes as input raw audio sampled at 8 kHz of arbitrary length. A convolution layer is run directly on the audio sample $\mathbf{x}$:

$$\mathbf{x}_i^{\mathrm{C}} = f(\mathbf{W}_{\mathrm{C}}\mathbf{x}_{[i,i+v]} + \mathbf{b}_{\mathrm{C}}) \tag{1}$$

where v is the convolution window size and f a non-linear function. We use a window size of 200, which at 8 kHz sampling rate corresponds to 25 ms, and move the convolution window with a step of 50, which corresponds to around 6 ms. The role of this layer is to extract the features for each frame, and evaluate the differences among overlapping frames. On top of the convolution outcome a max-pooling operation is applied:

$$\mathbf{x}_j^{\mathrm{MP}} = \max_i(x_{i,j}^{\mathrm{C}}) \tag{2}$$

where i is the window index, and j the vector index within each convolution window. The max-pooling allows to select the contributions from the most significant frames, and to combine them into a fixed size vector. It is then followed by a fully connected layer (of size 200) and a final softmax layer to perform the actual classification.

Emotion class	SVM	CNN
Angry	60.4	70.5
Happy	52.2	58.6
Sad	76.4	69.1
Average	63.0	66.1

Table 1. Average accuracy results over the three-folds obtained from the SVM and CNN experiments.

3. EXPERIMENTS

3.1. Corpus

We built a small corpus of data collected through an ongoing annotation project [17]. We split speeches obtained from the TEDLIUM v2 corpus [18] into segments of 13-15 s each. Among around 80K segments collected this way, we annotated 9879 segments. 8964 of them (around 90%) were annotated by students from our research group, while the other 915 (10%) through crowdsourcing from Amazon Mechanical Turk.

Each annotator was requested to select an emotion label for each sample among the following: "*Happy*", "*Sad*", "*Angry*", "*Neutral*", "*Garbage*", the latter to be used when the segment contained music or overlapped speech. For the data annotated through crowdsourcing, multiple annotations were retrieved for each sample, and we took the label chosen by the majority of the annotators. In case of a draw between multiple emotions we selected the neutral class for that sample.

We collected a total of 877 samples for the "*Sad*" class, 771 for the "*Angry*" class and 3498 for the "*Happy*" class[1], with the others classified into "*Neutral*" or "*Garbage*". For each emotion among "*Happy*", "*Sad*" and "*Angry*" we prepared a dataset for binary classification. We chose all the samples of that emotion for the positive set, and an equal proportion of samples for all the other classes except "*Garbage*" for the negative set. More samples for the "*Neutral*" class were chosen to balance the proportion when needed.

3.2. Experimental setup and classification results

We trained our network using standard backpropagation, with momentum set to 0.9 and initial learning rate of 10^{-4}. The learning rate was halved every 15 epochs, and we stopped the training when the error on the development set began to increase. As non-linear function we used the rectified linear function, since they gave better performance compared to $tanh$. The CNN was implemented with THEANO toolkit. Due to the limited size of the corpus we ran a 3-fold cross validation, each time randomly taking 80% of the data as training set, and 10% each for the development and test set. As a baseline system we trained a linear-kernel SVM with the

[1] In a preliminary phase of our project we concentrated on the annotation of happy samples obtained from an automatic API, thus the higher number of samples for this class compared to *angry* and *sad*.

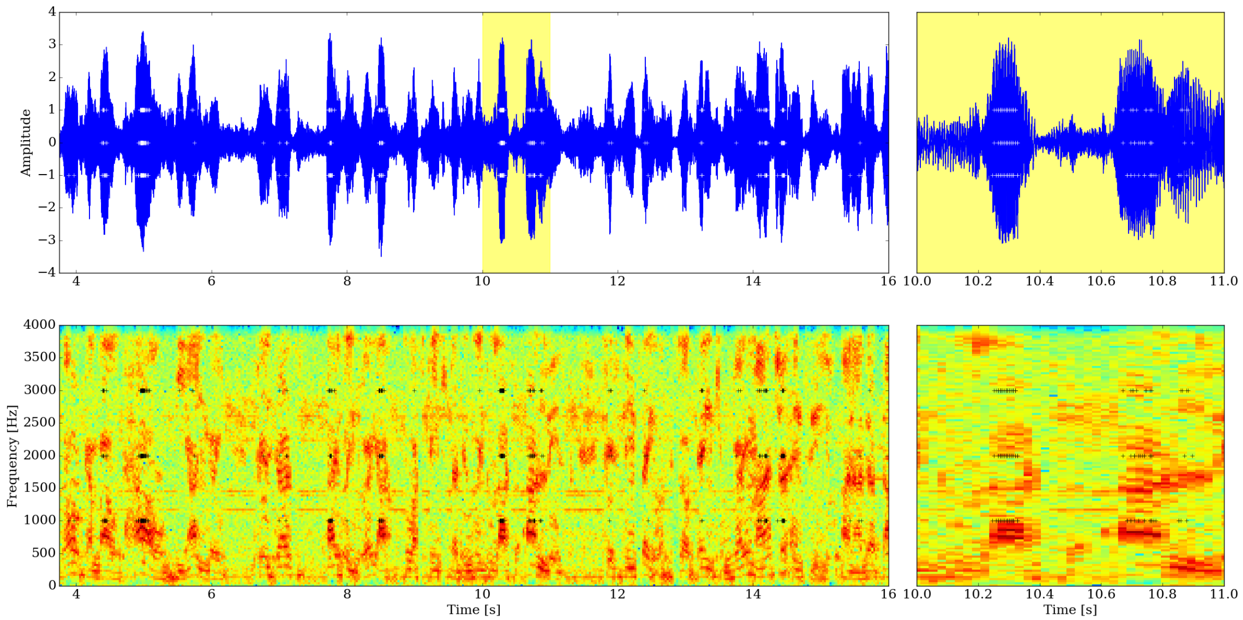


Fig. 2. Audio sample (above) and spectrogram (below) of a 16 s male speech, with detail of the yellow highlighted region on the right. White/black dots represent the max-pooling activation time instants. The top row of dots shows the activation of the "*Sad*" network, the middle row the activation of the "*Happy*" network, and the bottom row the activation of the "*Angry*" network.

INTERSPEECH 2009 emotion challenge feature set [11, 12], again applying three-fold cross-validation.

Overall average results for both methods are shown on Table 1. Our CNN obtains an average accuracy of 66.1%, over a SVM average result of 63.0%. The CNN yields higher results on average and for the "*Angry*" and "*Happy*" classes, while for "*Sad*" the SVM performs better. Results are generally lower for the "*Happy*" class, in spite of the more data available. One possible cause is that we noticed annotators from different backgrounds selected this class with different proportions, and it was generally harder to distinguish from "*Neutral*" [17].

4. NETWORK ANALYSIS

When using traditional feature-based classifiers, it is generally straightforward to analyze the contribution of each individual feature and identify the ones more representative for the task, as well as how their variation influences the classification outcome [19]. However deep learning based methods are often treated as "magical boxes" that simply yield good results. When deep learning is used to learn a feature representation in addition to just perform the classification, it is even more important to verify what happens under the hood.

4.1. Network activation

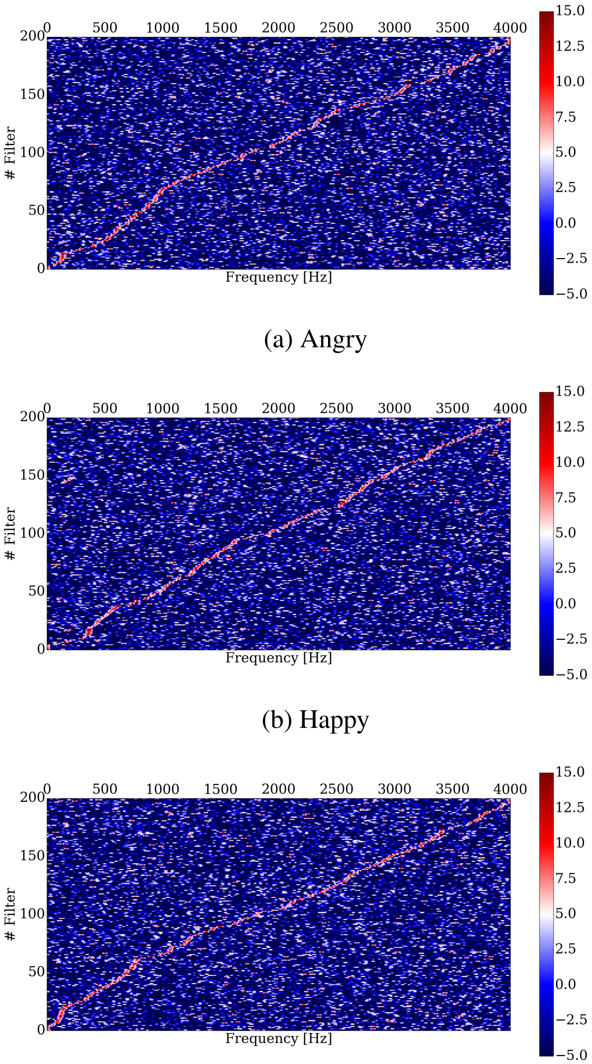
An important component of every CNN architecture is the max-pooling layer. It allows to select features coming from the dimensions where the convolution is applied, which is time in our case. We expect the network to activate during the speech intervals while ignoring the silences.

To show the role of the max-pooling layer in our application, for each convolution window i of our input signal we count the number of time it was selected by the max-pooling layer. We then retrieve the time instants of each window and highlight them on both a time-domain signal graph and the spectrogram of the same signal. Figure 2 shows the analysis of when the max-pooling was triggered on a long speech segment of around 16 s for all the three emotions considered. The CNN effectively picks the time instants where the content in frequency is higher, avoiding the silences. Although each emotion is modeled by a different binary classifier, there seems not to be much difference among them in the activation pattern.

4.2. Frequency analysis

Another important aspect to analyze is where the network activates in the frequency domain, and whether there are any differences among the three emotion models. The first layer of our CNN is dedicated to features extraction and learning. Each row of the parameter matrix $\mathbf{W}^c$ is a filtering function which is applied to each convolution window [8]. The contributions of all the filters (200 in our model) are then summed together. Each filter element $W^C_{i,j}$ is a time factor, spaced of the interval between one audio sample and the following of the discreet-time input signal. Thus a filter can be easily converted to a frequency spectrum, taking the absolute values of the FFT:

$$F(\mathbf{W}^C_i) = |\text{FFT}(\mathbf{W}^c_i)| \quad (3)$$



(a) Angry

(b) Happy

(c) Sad

Fig. 3. Frequency response of CNN filters. Activation is in logarithmic scale, red color means high activation.

where $i \in [0, 200]$ is the filter index.

Figure 3 shows the activation of each filter (rows in the diagrams) throughout the range analyzed, limited to $4\,\text{kHz}$ due to the sampling rate. The activation values have been converted to logarithmic scale with the following function:

$$a(i, f) = 20\ \log_{10}(F(W^{c}_{i,f})) \tag{4}$$

The filters are then sorted from bottom to top by ascending central frequency, which is the frequency with the highest activation value for each filter.

The first thing that can be noticed looking at Figure 2 is the activation path behavior in the range between 0 and $1\,\text{kHz}$. Human pitch lays typically in the range between $100\,\text{Hz}$ and $250\,\text{Hz}$, and around this range several CNN filters activate. The "*Sad*" emotion has the lowest activation point at around $125\,\text{Hz}$, followed by "*Angry*" at $140\,\text{Hz}$ and then by "*Happy*" at $360\,\text{Hz}$. This is consistent with the prior literature [2, 1], as "*Sad*" emotion is often characterized by a lower than average pitch, and "*Happy*" by an higher than average pitch. The "*Angry*" emotion is also sometimes characterized by a low pitch, but often exhibits higher energy at higher frequencies, and our neural network seems to reflect this, concentrating more filters in the range between 500 and $1000\,\text{Hz}$.

Another aspect evident from the figures is that the activation of each filter does not limit to only one frequency range, but includes multiple frequencies, sparse throughout all the spectrum. Multiple filters activates for each frequency value, even those ignored in the central frequency path. It is a very different behavior than what was shown for tasks like Automatic Speech Recognition [8]. In that case each filter limits to a very specific frequency range, and the filter distribution follows a logarithmic curve. While in ASR the information to retrieve is carried by the spectrogram, especially at low frequency values, in the emotion detection case higher frequencies seem to have an important role too and our network shows a linear activation pattern beyond $1\,\text{kHz}$. Moreover emotions are also described by features which do not depend on frequency, such as differences in amplitude of the audio sample. For example an angry speech has often great shifts in amplitude, while a sad one is monotone. This further explains the response of each filter to more than one frequency range.

5. CONCLUSION

We reported a real-time CNN model to detect emotion from speech. Our CNN system is able to achieve an average accuracy of 66.1% on three main emotions: "*Angry*", "*Happy*", "*Sad*". The result was achieved with a very small corpus of raw audio speech samples as training set. We also provided a deeper analysis of the model activation in time and frequency. We showed the max-pooling layer activates during the speech sections of the input, ignoring the silences. Compared to ASR tasks, the CNN filters concentrate around fundamental frequency values associated to the emotion they are trained on, and then distribute linearly for higher frequencies. Each filter also activates at multiple frequencies in order to learn features related to non-frequency related prosodic descriptors.

Our work included to our knowledge the first-ever analysis of a simple and fast CNN model trained to recognize emotions. It gives many insights on how the performance and the speed of an emotion classifier can be improved in the future, for example trying to separate the roles of amplitude and frequency. An accurate real-time emotion detection framework will be an important component of many speech-related application, in particular related to human-machine dialog systems [15].

6. REFERENCES

[1] Disa A Sauter, Frank Eisner, Andrew J Calder, and Sophie K Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *The Quarterly Journal of Experimental Psychology*, vol. 63, no. 11, pp. 2251–2272, 2010.

[2] Iain R Murray and John L Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[3] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech*, 2014, pp. 223–227.

[4] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[5] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Stefanos Zafeiriou, et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[6] Navdeep Jaitly and Geoffrey Hinton, "Learning a better representation of speech soundwaves using restricted boltzmann machines," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5884–5887.

[7] Dimitri Palaz, Ronan Collobert, et al., "Analysis of CNN-based speech recognition system using raw speech as input," in *Proceedings of INTERSPEECH*, 2015.

[8] Pavel Golik, Zoltán Tüske, Ralf Schlüter, and Hermann Ney, "Convolutional neural networks for acoustic modeling of raw time signal in lvcsr," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*. Springer, 2014, pp. 818–833.

[10] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky, "Visualizing and understanding neural models in NLP," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, June 2016, pp. 681–691, Association for Computational Linguistics.

[11] Björn Schuller, Stefan Steidl, Anton Batliner, et al., "The INTERSPEECH 2009 emotion challenge.," in *INTERSPEECH*, 2009, vol. 2009, pp. 312–315.

[12] Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," *Proceedings of the 2016 Conference of Empirical Methods for Natural Language Processing: Human Language Technologies*, 2016.

[13] Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan, "Zara the supergirl: An empathetic personality recognition system," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, San Diego, California, June 2016, pp. 87–91, Association for Computational Linguistics.

[14] Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin, "Towards empathetic human-robot interactions," *arXiv preprint arXiv:1605.04072*, 2016.

[15] Pascale Fung, "Robots with heart," *Scientific American*, vol. 313, no. 5, pp. 60–63, 2015.

[16] Andrew Ortony and Terence J Turner, "What's basic about basic emotions?," *Psychological review*, vol. 97, no. 3, pp. 315, 1990.

[17] Dario Bertero and Pascale Fung, "Towards a corpus of speech emotion for interactive dialog systems," *Oriental COCOSDA*, 2016.

[18] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," in *LREC*, 2014, pp. 3935–3939.

[19] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 582–596, 2009.