Language-independent Emotion Recognition from Speech: Performance of Activation Functions

Florian Strohm
University of Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart
Institute for Natural Language Processing
florian.strohm@ims.uni-stuttgart.de

Abstract

There are several areas of application for emotion detection systems for which it is important to reliably recognize expressed emotions. Since there are many different languages and for most of them little or no training data exist, it is favorable to have a model which can recognize emotions independently from the training languages. This work examines to what extend it is possible to recognize language-independent emotions by training a Convolutional Neural Network using different languages and especially analyzes the performance of different activation functions in this context. Our results show that language-independent emotion recognition is feasible and that the choice of the activation function is not crucial.

1 Introduction

Emotion recognition is the task to assign an emotion label to an utterance. This can be useful for many different domains and use-cases, for example:

- Customer service softwares like NICE Perform^{TM} recognize customer emotions and then perform prioritizing and routing of customers based on emotional content.
- In Human-Computer-Interaction, a computer can interact with a user differently, depending on the users emotional state.
- The information gained using emotion analysis can be valuable in different research domains such as psychology.

For all these domains a reliable classification is important, but the task to assign an emotion to a speech signal is very challenging due to the complex nature of emotions. For emotion classification we have to define a set of emotions, called basic emotions. It is not rigidly defined what basic emotions are. The most often used definitions are from Paul Ekman (Ekman, 92) and Robert Plutchik (Plutchik, 01).

In this work we use a Convolutional Neural Network since they already have been successfully used in speech recognition (NV, 17). We use three different approaches to emotion recognition:

(a) **mono-lingual:** train on a specific language L and test on the same language L

(b) **multi-lingual:** train on multiple languages $\mathbb{L}$ and test on a single language $L \in \mathbb{L}$

(c) **cross-lingual:** train on a specific language L_1 and test on another language L_2 with $L_1 \neq L_2$

The later seems rather difficult since the network will most likely learn language specific features, but (NV, 18) have shown that this is a feasible approach if we fine tune the trained model on a small subset of the target language. This is very useful if only a small amount of training data is available for a language.

2 Data

For our experiments we use two datasets in different language, namely the IEMOCAP (BBL, 07) and the RECOLA (RSS, 13) databases. The prior consists of 8029 files containing about 12 hours of audiovisual data comprising English actor performances specifically expressing emotions. The corpus is annotated by multiple annotators with multiple labels such as anger, happiness, sadness and neutrality as well as valence and activation. The later dataset consists of 1045 files containing about 3.8 hours of annotated audio data created by

different French speakers who were solving a task in collaboration. Among others, it was annotated with valence and activation labels by multiple annotators.

In total we have about 16 hours of annotated audio data in mixed English and French language allowing us to train the network on different languages and therefore learn language-independent features. We split both corpora into three subsets; training, validation and test set. Each training set contains 80% of the data, while the validation and test sets contain 10% each.

Since the RECOLA corpus is not annotated with categorical emotion labels like the IEMOCAP corpus, we are limited to valence and arousal for emotion classification. We map the continuous values of the valence and arousal labels to binary (low/high) labels and use these to define emotions, as can be seen in Table 1

Arousal/Valence	Positive	Negative
High	Joy	Angry
Low	Pleasure	Sad

Table 1: Mapping from arousal and valence to emotion labels

2.1 Feature Extraction

In order to feed the data into our neural network, we have to extract features from the source speech signal. We use the OpenSMILE software (EWS, 2010) to extract the first thirteen *Mel Frequency Cepstral Coefficients* (MFCC) for each frame (10ms) of the speech signal. Since each audio file in our dataset is of different length, we have to define a fixed number D of frames for our data. We set D to be the average number of frames of a dataset. For the IEMOCAP and RECOLA dataset this yields $D_1 = 444$ and $D_2 = 216$, respectively. When using a multi- or cross-language model we use $D_{\text{avg}} = 330$, being the average number of frames over both datasets. If a speech signal consist of more frames than D, the surplus is cut-off at the end and on the other side zero-padded, if it has less than D frames.

3 Baseline Network Architecture

As described in the previous section, a speech signal consists of several frames. We concatenate these frames which yields a matrix $M \in \mathbb{R}^{D\times 13}$, with the first axis representing the number of frames and the second axis the first thirteen MFCCs. This representation of a speech signal enables us to use a Convolutional Neural Network (CNN).

As can be seen in Figure 1, we feed our neural network with a speech signal represented by the matrix M. We use 50 kernels of size 10×13 on the convolution layer. Therefore, the filters span all 13 MFCC features over a period of 10 frames, which is equal to a time window of 100ms, enough to be able to recognize emotions. This step reduces the input to a new size $M' \in \mathbb{R}^{D\times 1}$. Afterwards max pooling is applied with a window size of 30 and stride of 3. Finally, we concatenate all samples to a dense layer, apply dropout (SHK, 14) and calculate the final network output. Table 2 shows the different hyperparameters of the network.

Hyperparameter	
Parameter	Baseline Value
Activation Function	ReLU
Loss Function	Cross Entropy
Optimizer	Adam (KB, 01)
Mini-Batch size	50
Dropout	50%
Epochs	50

Table 2: Results of our baseline network for all three given tasks and for both language.

4 Research Questions

In this work we compare different activation functions of a neural network (NN) and measure their impact on the emotion classification task. The different models are trained using the multi-lingual approach only since we want to compare their impact for language-independent emotion recognition. We use visualizations created by TensorFlow (ABJ, 16), to give a better overview of the results.

Activation Functions We investigate the performance of the following different activation functions in our NN:

- Sigmoid: $f(x) = \frac{1}{1+e^{-x}}$
- Hyperbolic Tangent: $f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
- Softplus: $f(x) = \log(1 + e^{x})$
- Rectified linear: $f(x) = \max(x, 0)$

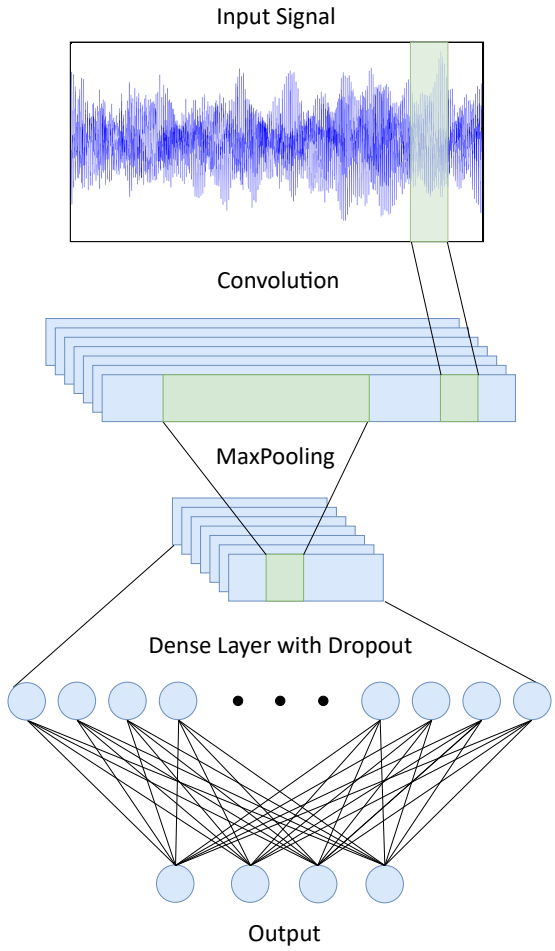

Figure 1: Architecture of the Convolutional Neural Network

- Exponential linear:
$$f(x) = \begin{cases} x & x \geq 0 \\ e^x - 1 & \text{otherwise} \end{cases}$$
- Leaky rectified linear:
$$f(x) = \begin{cases} x & x > 0 \\ \alpha x & \text{otherwise} \end{cases} \quad \alpha \leq 1$$

We compare the performances of the above listed activation functions and want to determine the best performing function in the context of language-independent emotion recognition from speech.

5 Experiments

Firstly, we create some baseline results with the network described in section 3. Table 3 shows the accuracy of this network for mono-, multi and cross-lingual testing. It reports the accuracy for each emotion and the micro average accuracy. The upper half of the table shows the results when trained on the English IEMOCAP dataset and the lower half for the French RECOLA dataset. For multi-lingual learning, each mini-batch contains 25 utterances from the RECOLA and 25 utterances from the IEMOCAP dataset.

		Accuracy		
	Emotion	Mono	Multi	Cross
EN	Sadness	0.00	**0.02**	**0.02**
	Anger	**0.02**	0.01	0.01
	Pleasure	0.04	0.01	**0.12**
	Joy	0.94	**0.99**	0.86
	Micro avg.	0.42	**0.43**	0.41
FR	Sadness	0.08	0.00	**0.23**
	Anger	0.20	0.20	0.20
	Pleasure	**0.36**	0.04	**0.36**
	Joy	0.75	**0.91**	0.40
	Micro avg.	**0.53**	0.52	0.36

Table 3: Results of our baseline network for all three given tasks and for both English (EN) and French (FR) language.

We can see, that the baseline model predicts the emotion *joy* the most accurate throughout all three evaluation approaches and for both languages. A reason for this is, that we have far more training data for *joy* than for other emotions. For multi-lingual learning, *joy* is almost the only emotion which can be recognized by the network.

The table also shows similar results for mono- and multi-language testing with about the same average accuracy, making multi-language training viable. We can see that other emotions than *joy* can be recognized by the network if we train and test on French utterances, but loose this ability when the network is multi-lingual trained (but retain comparable average accuracy).

The performance of cross-lingual testing is similar for the English, but drops for the French dataset. (NV, 18) showed that this drop can be limited if the model is fine tuned on the target language, only requiring a small amount of labeled data. Further, Cross-lingual testing shows a more spread performance across all emotions than mono- or multi-lingual testing.

We can observe a noticeable gap of about 9% in classification accuracy between the English and the French dataset, showing that the network can predict utterances from the RECOLA dataset more accurate in every case. This does not necessar-

ily mean, that emotions in French language are more easy to predict by a neural network than emotions in English language. The difference in performance may also be due to different creation process of the datasets. The French RECOLA dataset contains real dyadic interactions of participants who were solving a task in collaboration and therefore mostly real emotion expressions, while the English IEMOCAP dataset only contains acted dyadic sessions and therefore mostly imitations of emotion expressions.

Now we discuss the results of the next experiment. Table 4 shows the different accuracies for each activation function discussed in section 4. We only report accuracies for multi-lingual testing since we are interested in language-independent emotion recognition. As the table shows, all activation functions have a very similar average accuracy. Since the performance of the network varies in different training attempts due to random weights initializations, the small differences in the overall performance are not significant. However, we can see that the performances across specific emotions vary strongly. The ReLU activation function is best in recognizing the emotion *joy*. For the other emotions, ELU performs the best for English language and leaky ReLU for French language, which is also reflected by the notable higher macro average accuracy of ELU and leaky ReLu compared to the other activation functions.

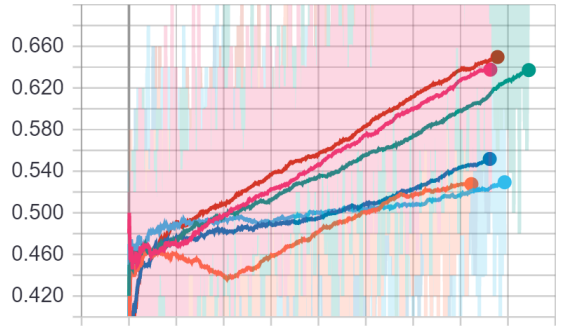

Figure 2: Smoothed training accuracies of all six activation functions on the y-axis and their relative training time on the x-axis. Legend:
Softplus ReLU Sigmoid
Leaky ReLU ELU Tanh

For more in-depth results, we plot the micro training accuracies and the cross-entropy loss of each activation function over time using TensorBoard. Figure 2 shows these training accuracies on the y-axis and the relative training duration on the x-axis. We can see that the training accuracies of the Tanh, ELU and Leaky ReLU activation functions are significantly higher compared to the remaining three activation functions. Since we know from our results in Table 4, that all activation functions perform similar, we can conclude that the model overfits the training data more with the three mentioned functions. Furthermore we can see, that with the ReLU function the network needed the least amount of time to finish the 50th epoch, which is not surprising since this function was specifically designed to be easily computable by computers.

Lastly, we have a look at the cross-entropy loss showed in Figure 3. We can see, that with the ELU and the Tanh function the network has a notable higher loss after training than with ReLU, Softplus or Sigmoid functions. This is consistent with the previous observations and strengthens our suspicion about the tendency of overfitting with ELU and Tanh, since they achieve a higher training accuracy even tough they have a higher loss.

Overall we can come to the conclusion, that none of the investigated activation functions achieves a noticeable higher performance than the others, yet ReLU is the best choice in this case since it is more resistant to overfitting and the fastest to compute.

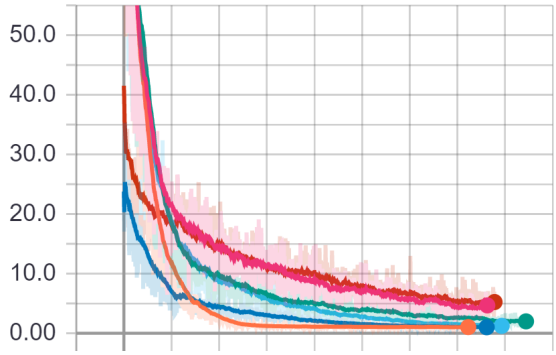

Figure 3: Smoothed cross-entropy loss of all six activation functions on the y-axis and their relative training time on the x-axis. Legend:
Softplus ReLU Sigmoid
Leaky ReLU ELU Tanh

		Accuracy					
	Emotion	Sigmoid	Tanh	Softplus	ReLU	ELU	Leaky ReLU
EN	Sadness	0.00	0.15	0.00	0.02	**0.17**	0.08
	Anger	0.29	0.34	0.12	0.01	**0.44**	0.28
	Pleasure	0.07	0.36	0.11	0.01	**0.43**	0.38
	Joy	0.85	0.61	0.87	**0.99**	0.48	0.56
	Micro avg.	**0.46**	0.45	0.43	0.43	0.43	0.41
FR	Sadness	0.00	0.23	0.15	0.00	0.00	**0.31**
	Anger	**0.20**	0.00	**0.20**	**0.20**	**0.20**	**0.20**
	Pleasure	0.29	0.29	0.14	0.04	**0.43**	**0.43**
	Joy	0.79	0.72	0.88	**0.91**	0.72	0.67
	Micro avg.	0.52	0.50	**0.55**	0.52	0.52	0.53

Table 4: Results of our network with different activation functions for the multi-lingual task and for both English (EN) and French (FR) language.

6 Summary & Outlook

In this work we have studied the performance of mono- multi- and cross-lingual testing and showed, that the multi-lingual approach achieves a very similar performance to the mono-lingual one. This allows to create a single model for emotion recognition in speech, independent of the spoken language. Furthermore, our results show that the choice of the activation function is not crucial for this domain.

For the future it would be interesting to see the performance of a model trained on multiple languages $\mathbb{L}$ and tested on a language $L \notin \mathbb{L}$. The model may learn language-independent features and thus be able to generalize to unseen languages. This was not able to test in this work since we are limited to only two different languages. Future work could also combine completely different languages like Khoisan languages, which uses click sounds, and Indo-European (e.g. English) languages. Doing so, we would expect that the multi-lingual accuracies drop below mono-lingual accuracies because of the very different articulation.

References

[NV17] Michael Neumann and Ngoc T. Vu. 2017. *Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech* Interspeech, Stockholm, August 2017.

[NV18] Michael Neumann and Ngoc T. Vu. 2018 . *Cross-lingual and Multilingual Speech Emotion Recognition on English and French* arXiv preprint arXiv:1803.00357.

[BBL07] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee and S. S. Narayanan. 2007. *IEMOCAP: Interactive emotional dyadic motion capture database* Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.

[RSS13] F. Ringeval, A. Sonderegger, J. Sauer and D. Lalanne. 2013. *Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions* 2nd International Workshop EmoSPACE, in Proc. of IEEE Face & Gestures 2013, Shanghai, April 2013.

[Kriesel07] David Kriesel 2007 . *Ein kleiner Überblick über Neuronale Netze*. http://www.dkriesel.com.

[ABJ16] M. Abadi, P. Barham, C. Jianmin, C. Zhifeng and others 2016 . *TensorFlow: A System for Large-Scale Machine Learning* in OSDI, volume 16, pages 265-283.

[EWS2010] F. Eyben, M. Wöllmerand, B. Schuller 2010 . *Opensmile: the munich versatile and fast open-source audio feature extractor* in Proceedings of the 18th ACM international conference on Multimedia, pages 1459-1462

[Plutchik01] Robert Plutchik 2001. *The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice* in American scientist, volume 89, pages 344-350

[Ekman92] Paul Ekman 1992. *An argument for basic emotions* in Cognition & emotion, volume 6, pages 169-200

[KB01] D. P. Kingma and J. Ba 2014. *Adam: A method for stochastic optimization* in International Conference for Learning Representations (ICLR), 2015.

[SHK14] n. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov 2014. *Dropout: A simple way to prevent neural networks from overfitting*. Journal of Machine Learning Research, vol. 15,no. 1, pp. 1929-1958, 2014.