Language-independent Emotion Recognition from Speech

Ayhan Kaplan
2886552

Cagri Tasci *
2884541

Florian Strohm
2868046

Strays Wonderland
University of Stuttgart

{FIRSTNAME.LASTNAME}@IMS.UNI-STUTTGART.DE

Abstract

This paper focuses on the task of language-independent emotion recognition from speech. In addition to language-independent, we also take a look into mono-lingual and cross-language training. In this work, we conduct several experiments using a convolutional neural network, in order to establish the best performance for the given tasks. We further compare different activation functions, optimizers and investigate the effect of alternating between mono-lingual batches of disparate languages. Our experimental results reveal that the mono- and multi-lingual tasks show similar accuracies across most classes as well as a similar micro-accuracy. Cross-language training on the other hand has generally shown notably lower accuracies.

1 Introduction

Recognition of the underlying emotion in speech is an important and, due to the complexity of emotional expressions and the lack of large annotated datasets, also a challenging task in the realm of human-computer-interaction (NV17). The most common approach to automatic classification of emotion in speech in deep learning, is to train and test a convolutional neural network on one annotated corpus of a single language (mono-lingual) (NT18). Instead, our work focuses on the task of language-independent emotion recognition from recorded speech. That is the classification of a speakers underlying emotion, independent from the language in which it is spoken. Thus, the network we have implemented is trained and tested on a collective dataset that is constructed by merging multiple corpora of different languages. Namely, the English IEMOCAP- and the French RECOLA-dataset. We further look to compare the accuracy of the multi-lingual task with two other common tasks: **(1)** training and testing on a single language (mono-lingual), **(3)** training on one language and testing on the other (cross-language). In this work, we conduct several experiments using a convolutional neural network (CNN), in order to establish the best performance for the given tasks. This serves as baseline for our work from which we will conduct further experiments in order to explore three diverse research topics. More precisely, we analyse the effect of using various activation functions, optimizers and the use varying mini-batches on the performance of our network. Hereby, the third research topic focuses on altering the ratio between the two languages in each batch, with the restriction that in each mini-batch, the ratio switches between either exclusively English or French data. This research topic is the main (individual)focus of this paper. The other two research topics are investigated by the other two group members.

2 Data

This section serves to outline the structure of our input data and how it was processed to use as input for the network. We operate on two languages, which are English and French, using two corpora which are freely available and frequently used in speech recognition. Therefore, our dataset consists of files containing recorded, spoken text in either of these languages. Our English dataset consists of 10.036 samples taken from the Interactive Emotional Dyadic Motion Capture (**IEMOCAP**) database (BBL$^+$08). Whereas our French data is taken from the Remote Collaborative and Affective Interactions (**RECOLA**) Database and consists of 1306 files (RSSL13). We split these datasets into train-, test-, and validation-sets with a ratio of 80-10-10 for both languages.

*author of this paper

2.1 Input features

We use the openSMILE toolkit (EWGS13) in order to extract 13 MFCC feature banks from each input file. These features are extracted frame-wise, where each frame corresponds to 10 ms. Furthermore, each file is either cut or padded to a specified number of frames. How this number is constructed is explained in detail in the following paragraph. Each file that is longer than this number is simply cut at the specified number of frames. Each frame that is shorter, is padded to the specified length with zero-frames containing a '0' for each feature.

Network Input: Thus, the input to our network consists of a total number of 11342 utterances for the multilingual task, where each utterance is represented by multidimensional tensor of dimension $F_{avg} \times 13$ with F_{avg} being the average number of frames-per-utterance and 13 the number of MFCC features. F_{avg} changes, depending on the given task (mono-lingual, multi-lingual, cross-language). In task **(1)** and **(3)**, F_{avg} is determined solely depending on the language that is currently trained, while in task **(2)**, this number is defined as the arithmetic middle of the two average frame values of both languages.

Network Output: Files of both language datasets are annotated with their respective binary valence and arousal values. We therefore construct 4 output-classes, representing each possible binary combination of these two values. These binary values together encode an underlying emotion, so each class represents a different emotion. The class encodings are listed in Table 1 and for clarity purposes. These will be used instead of the binary values throughout the paper. Both our datasets are split into these categories in order to calculate the accuracy for each class individually.

Valence \ Arousal	high	low
positive	Joy	Pleasure
negative	Anger	Sadness

Table 1: Encoding of output classes

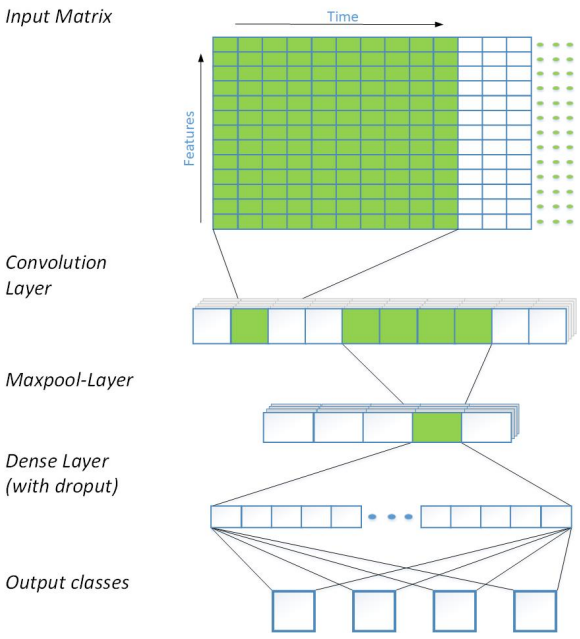

Figure 1: Network Architecture

3 Baseline Network Architecture

This section outlines the underlying structure of the implemented baseline network. This network will be slightly modified for each of the research topics. Our baseline network consists of one convolution-layer, followed by one maxpool-layer. After the maxpooling, regularisation in the form of dropout is applied to the dense-layer. The architecture of our network is depicted in Figure 1 and its hyperparameters are listed in Table 2, to offer a more compact overlook. The network was implemented using Tensorflow (ABh+16), which is a machine learning library for python.

Hyper-parameters: As mentioned in the previous section, the input for our network are matrices of dimension $(F_{avg} \times 13)$, with one such matrix per utterance. In our convolution layer, we apply 50 kernels with dimensionality of (10×13) each, to our input, spanning all 13 features of our input. As one frame corresponds to $10ms$, a filter of size (10×13) spans over $100ms$ of the input. This dimensionality ensures that enough of the valuable information in recorded speech is being considered in each kernel. As mini-batch size, we have chosen a size of 50 samples per mini-batch. Ergo, each mini-batch is a multi-dimensional matrix of size $(50 \times F_{avg} \times 13)$.

Hyper-parameters	
Paramter	Baseline Value
Activation Function	ReLU
Loss Function	Softmax - CE
Optimizer	ADAM
Init. Learn. Rate	0.001
Input matrices	$(F_{avg} \times 13)$
Mini-Batch size	50
Filter size	(10×13)
Number of Filters	50
Maxpool Filter	(30×1)
Stride	3
Dropout	0.5
Epoches	50

Table 2: Parameters used in the baseline-network

After the convolution layer, we apply maxpooling with a kernel-size of (30×1) and a stride of 3. Afterwards, dropout with *keep-probability* of 0.5 is applied for regularisation. To train our network, we run training iterations for 50 epochs in all our experiments and adjust the number of iterations accordingly.

As activation function, we chose rectified linear units (ReLU). The applied loss function is *Softmax-cross entropy* (KB05). During training, we optimize using stochastic gradient with an adaptive learning rate (*Adam*) (KB14).

4 Experimental Setup

Our baseline task is to train the network for binary classification of arousal and valence in speech. Furthermore, we decompose our baseline task into the following experiments:

(1) mono-lingual:
train and test the network on a single language.

(2) multi-lingual:
train and test the network by merging both language datasets into one set.

(3) cross-language:
train on one specified language/dataset L_1 and test on L_2.
(where $L_1, L_2 \in \{\text{RECOLA, IEMOCAP}\}$)

Our experiments are conducted by always testing on both languages, but changing how we construct the input data.

Meaning that in each task we compute the accuracy for each of our four output classes for both languages each, as well as the micro-accuracy for the entire dataset per language.

4.1 Individual Research Topics

After the network for the baseline task is established, each of us looks into a different research question in order to analyse and construct the best possible performance in each of the investigated aspects. The explored research topics are as follows:

(a) Comparison of optimizers
- A. Kaplan

(b) Comparison of activation functions
-F. Strohm

(c) Language-alternating-batches
-C. Tasci

Task (c) introduces a fourth experiment:

(4) language-alternating-batches:
training the network by alternating the language between each mini-batch.

To further clarify, in each epoch, we construct our input batches from scratch. But in assembling the mini-batches, we construct the first mini-batch from 50 (our specified batch-size) utterances, taken from the french RECOLA-dataset. Every mini-batch thenceforth, switches between containing utterances either only from the IEMOCAP- or the RECOLA-dataset. So each mini-batch itself is mono-lingual (either solely English or solely French utterances within the mini-batch), but during training we alternate between English and French mini-batches (IEMOCAP- and RECOLA-mini-batches). The network used for this research topic, or more accurately, its hyper-parameters, remain/s the same as in our baseline-tasks. The only constraint that is hereby needed, is the specification of the number of frames, that each utterance is cropped or padded to, since this value has to remain constant throughout the entire procedure of training the network. So instead of determining this number based on each current mini-batch; we define it as we did it in the multilingual task. Hence, this value is exactly the arithmetic middle of the average number of frames of each dataset. This is the task, that this paper focuses on and thus, subsection 5.1 will only provide results for this task.

		Mono-lingual	Multi-lingual	Cross-language
Test Language	Target Class		Accuracy	
En	Sadness	0.000	0.015	0.015
	Anger	0.019	0.014	0.014
	Pleasure	0.043	0.010	0.120
	Joy	0.942	0.985	0.864
	MICRO	0.421	0.432	0.405
Fr	Sadness	0.070	0.000	0.230
	Anger	0.200	0.200	0.200
	Pleasure	0.350	0.035	0.357
	Joy	0.754	0.912	0.403
	MICRO	0.533	0.524	0.359

Table 3: Results of our baseline network for all three given tasks. Note that for cross-language-training, the left-most column specifies the language which is tested, so the network is trained on the other, respectively.

5 Experimental Results

In this section, we want to have a look at and analyse the performance of our network. In subsection 5.1, we take a closer look at the weighted and unweighted accuracy for all given tasks in the baseline network. Afterwards, we gather results for the individual research topic in subsection 5.1.

5.1 Baseline Results

The results of our three baseline tasks will be summarised in Table 3. As performance measure, we list the accuracy of each individual class, as well as the micro-accuracy of the task itself for both datasets. We have chosen to not list the averaged unweighted accuracy, since these values are too low to draw significant conclusions, as these are lowered immensely by the classes *Sadness* and *Anger*. This is most likely due to the fact, that our input generally has the least amount of samples within these two classes. As can be seen in this table, across all tasks and across both languages, we generally observe the highest accuracy for the class *Joy* with a large margin in accuracy to the other classes. Sadness shows the lowest accuracy among all test cases. Again, this might be effect of having to few samples for these classes.

Another conclusion can be deducted, if we reconsider which classes are encoded by which binary combination of valence and arousal. By taking a look into Table 1, we reach following decompositions:

Sadness: negative valence ; low arousal,

Anger: negative valence ; high arousal,

Pleasure: positive valence ; low arousal,

Joy: positive valence ; high arousal,

Notice, that the first two classes, both encoding negative valence, generally have very low accuracy, while the two classes being encoded by positive valence show higher accuracies. So differences in valence show a significant change in accuracy while differences in arousal generally do not lead to a notable increase or decrease in accuracy. Though again, our datasets are highly imbalanced; the two negative-valence classes have only very few samples.

Furthemore, we can observe that overall, the accuracies within the task *mono-lingual* and the *multi-lingual* task, are similar to each other. Note, that the accuracy for the class *Joy* and the micro-accuracy even improve slightly. In other words, transforming a network designed for the mono-lingual task to a multi-lingual network, does not introduce a notable performance loss. Though we do observe a weighty decrease in accuracy for pleasure (high arousal) for both datasets. Accuracy of

the *cross-language* task on the other hand, is significantly lower than that of the other two tasks. Thus, cross-language emotion recognition is most-likely not as viable as the other two tasks. Though a notable difference in this task is the accuracy distribution among the four output classes for the French dataset. Training the network on English and testing it on the French dataset yields to a more evenly distributed accuracy across the four classes.

Language-Alternating Batches		
Test Language	Target Class	Accuracy
En	Sadness	0.000
	Anger	0.038
	Pleasure	0.245
	Joy	0.985
	MICRO	0.442
Fr	Sadness	0.025
	Anger	0.000
	Pleasure	0.440
	Joy	0.561
	MICRO	0.435

Table 4: Results of third research topic: language-alternating mini-batches.

5.2 Individual Results

This section will only list and discuss the findings and results of the third research topic and thus, the fourth task : language-alternating training. Hence, the results of the other two research topics will not be listed or discussed in this paper. Findings of these research topics will be listed and discussed in the papers of the respective group members (A. Kaplan & F. Strohm).

Language-alternating Batches Table 4 lists the achieved performance for both test sets, using language-alternating mini-batches. Note that the values in this table are the averages of achieved accuracies across multiple runs. We have conducted 5 independent training and test runs. Also; in each such run we have assigned the initial language, meaning the language which the first mini-batch will consist of, by random. Comparing the resulting values to those of our baseline tasks, we can make following observations: Firstly, we can observe, that the micro accuracy of the English-test-set remains roughly the same as in the multi-lingual task. Secondly, the micro-accuracy of the French-test-set shows a significant drop, compared to the multi-lingual. Accuracies of the individual output classes do not show a notable change, with the exception of the class *Joy* for the French-test-set Since the language-alternating-batch-training is a modification of the multi-lingual-task; we can compare these values to draw conclusions about the effect of our modification. As the accuracy does not change significantly, we can assume that our modification does not alter or does not have a meaningful effect on the performance of our system. Since our dataset is heavily unbalanced and the majority of utterances are from the English-dataset, further research here might be needed to justifiably exclude a possible effect of language-alternating-batch-training. This imbalance might be a large factor as to why the accuracy for the French-test-set decreases. However, both micro-accuracies show a small increase when compared to the cross-language task. This is important, since this research topic can also be regarded as a modification of the cross-language task and we have shown a slight improvement when alternating the language between each mini-batch. Though it can be argued, that the increase is too small to be significant.

6 Conclusion

In this paper, we have presented results for the classification of valence/arousal in recorded speech, not only using a mono-lingual set-up but also a multi-lingual and a cross-language one. The conducted experiments in the course of this work have shown that multi-lingual classification of emotion can be achieved without meaningful loss of accuracy compared to the mono-lingual classification, and using the same network. Yet the network we have implemented shows significantly lower accuracy when testing exclusively on one language and testing on another.

References

[ABh+16] Martín Abadi, P. Barham, J. hen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association.

[BBL+08] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. 42:335–359, 12 2008.

[EWGS13] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 835–838, New York, NY, USA, 2013. ACM.

[KB05] Douglas M. Kline and Victor L. Berardi. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing & Applications*, 14(4):310–318, Dec 2005.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[NT18] M. Neumann and N. Thang Vu. Cross-lingual and Multilingual Speech Emotion Recognition on English and French. *ArXiv e-prints*, March 2018.

[NV17] 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. *CoRR*, abs/1706.00612, 2017.

[RSSL13] Fabien Ringeval, Andreas Sonderegger, Jrgen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions, 04 2013.