CROSS-LINGUAL AND MULTILINGUAL SPEECH EMOTION RECOGNITION ON ENGLISH AND FRENCH

Michael Neumann, Ngoc Thang Vu

University of Stuttgart, Germany
{michael.neumann|thang.vu}@ims.uni-stuttgart.de

ABSTRACT

Research on multilingual speech emotion recognition faces the problem that most available speech corpora differ from each other in important ways, such as annotation methods or interaction scenarios. These inconsistencies complicate building a multilingual system. We present results for cross-lingual and multilingual emotion recognition on English and French speech data with similar characteristics in terms of interaction (human-human conversations). Further, we explore the possibility of fine-tuning a pre-trained cross-lingual model with only a small number of samples from the target language, which is of great interest for low-resource languages. To gain more insights in what is learned by the deployed convolutional neural network, we perform an analysis on the attention mechanism inside the network.

Index Terms— Speech Emotion Recognition, Multilingual, Cross-lingual, CNN, Attention

1. INTRODUCTION

The common approach to automatic emotion recognition is to train and test a classifier on one annotated (mostly mono-lingual) corpus, either by subdividing the data into train, validation and test sets or by means of cross-validation. This way, the system is highly specialized with respect to a number of factors, such as the speaker group, the recording situation, the language, and the type of speech (spontaneous or acted). Further, no conclusions can be drawn to what extend such a system can generalize across different interaction scenarios and languages. For this reason, we investigate cross-lingual and multilingual speech emotion recognition, as a step towards language-independent emotion recognition in natural speech.

In addition to the aforementioned reasons, cross-lingual classification can possibly facilitate emotion recognition for scenarios with no or only a small amount of annotated data in the target language, which we refer to as low-resource setting.

Various cross-corpus analyses have been conducted in recent years [1, 2, 3, 4, 5]. In an extensive study with six corpora, [1] examined many different combinations of corpora as training set, without focusing on one certain aspect of the data (e.g. different language or different interaction scenario). Although this study gives an overall impression on the performance of cross-corpus emotion recognition, it makes the interpretation of results difficult because it is not clear which factors have what kind of impact. Focusing on cross-language emotion recognition, [6] presented a comprehensive overview using 8 languages from 4 language families and showed that cross-language emotion recognition is feasible, but with notably lower accuracy than mono-lingual recognition.

An approach to multilingual emotion classification using language identification and model selection is presented in [7]. In contrast to this work where language-dependent models are trained and then selected accordingly, we examine the performance of one model trained on multiple languages. Another strategy to combine two languages for emotion recognition, described in [8], is to apply histogram equalization to remove cross-language variability. In [9], the authors compare automatic cross-lingual recognition with human perception of emotion.

Concerning classification performance it is difficult to compare to related research in this field, because there are no standards regarding several factors, including the number of classes, the division of corpora into train and test sets, the underlying emotion concepts (categorical emotions or continuous arousal/valence dimensions). Hence, we cannot discuss state-of-the-art performance in this study, because the aforementioned works differ in at least one of these respects, mostly in the number of classes, the utilized databases or the mapping between continuous and discrete annotations. The focus of the present research is on multilingual and cross-lingual speech emotion recognition compared to mono-lingual baselines trained on the respective corpora, as well as on an analysis of the attention mechanism used in the recognition system. We show that multilingual emotion recognition is feasible without adaptation to the language and present promising results for cross-lingual training followed by fine-tuning on the target language.

2. MODEL ARCHITECTURE

For this study we train an attentive convolutional neural network (ACNN) for binary classification of arousal and valence in speech. The model architecture is mainly adopted

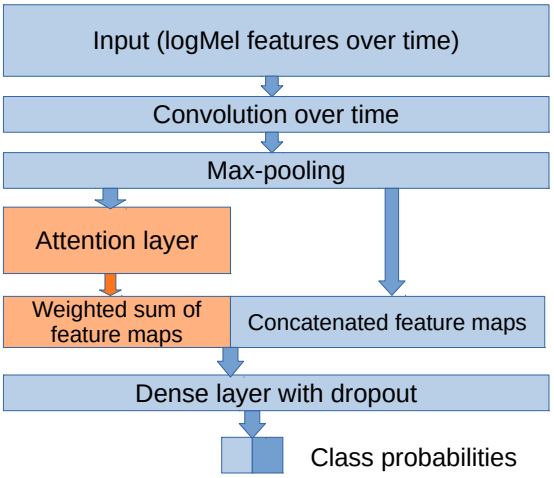

Fig. 1. Model topology.

from [10] and adjusted to this task and the cross-lingual setting. Figure 1 depicts the network topology schematically. As input features to the ACNN, 26 logMel filter-banks are extracted frame-wise from the segmented speech signal (frame size of 25ms and a window shift of 10ms). The input has fixed length of 7.5s, shorter utterances are padded with zeros at the end. The convolution kernels span all 26 features (1-D convolution over time). The output from the max-pooling layer is fed into an attention layer which computes a weighted sum of the information extracted from different parts of input. The input to the fully connected softmax layer at the end is the concatenation of attention vector and the feature maps from the pooling layer.

The attention layer computes attention weights α_i over all feature maps for each time step i. Equation 1 shows the computation of these attention weights α_i for an input sequence x consisting of vectors x_i, where $f(x) = W^T x$, with W being a trainable parameter.

$$\alpha_i = \frac{exp(f(x_i))}{\sum_j exp(f(x_j))} \tag{1}$$

The output of the attention layer, $attentive\_x$, is the weighted sum of the input sequence.

$$attentive\_x = \sum_i \alpha_i x_i \tag{2}$$

The intuition behind using an attention mechanism for emotion recognition is that emotional information is distributed differently over the signal. Therefore, we want to first weight the information extracted from different pieces of the input and then combine them in a weighted sum.

3. ENGLISH AND FRENCH EMOTIONAL SPEECH

We use two corpora of emotional speech which are frequently used and freely available. The main criterion for selecting the data was that the corpora contain the same type of speech in terms of conversation type (human-human) and naturalness.

The interactive emotional dyadic motion capture database (**IEMOCAP**) [11] is a multimodal database of English dyadic conversations containing both fixed speech (scripted dialogs) and free, spontaneous speech (improvised dialogs given a certain scenario and topic). The speakers are professional actors. IEMOCAP is annotated on turn level in two ways, with categorical emotion labels (such as 'anger', 'happiness', 'sadness') and with 5-point scales on the dimensions valence, arousal and dominance (1 - low/negative, 5 - high/positive). The corpus contains 10,039 utterances.

Recola [12] is a multimodal database of French speech consisting of dyadic conversations during a video conference where participants had to solve a collaborative task. From 46 speakers in total, we use the freely available portion of 23 speakers in this study, consisting of 1,308 utterances. Recola is annotated with continuous labels for arousal and valence in the range [-1, 1] on a 40ms rate. Annotation was done with ANNEMO [12], a tool similar to Feeltrace [13]. Since we are interested in recognition of emotions on utterance level, we calculated the mean of all values for one turn, and then took the average across all annotators as the final label.

To be able to train a model on several corpora, the problem of different annotation schemes has to be overcome. We decide to focus on a binary classification task of arousal (low/high) and valence (negative/positive).[1] The mapping of original annotations to a binary scheme is shown in Table 1.

	Low/Negative	**High/Positive**
IEMOCAP	range [1, 2.5]	range (2.5, 5]
Recola	range [-1, 0]	range (0, 1]

Table 1. Mappings to binary arousal/valence classes.

4. EXPERIMENTAL SETUP

We conduct the following four experiments: (a) mono-lingual (as baseline), (b) multilingual (merge Recola and IEMOCAP for training), (c) cross-lingual (train on one corpus, test on the other one), and (d) fine-tuning of a model trained in (c) in a simulated low-resource setting.

For (a) mono-lingual and (b) multilingual experiments we apply cross validation (CV) because there are no predefined train and test splits for these datasets. The IEMOCAP data consists of five sessions with one male and one female speaker each. We take data from four sessions to construct training and development sets and use the remaining session for testing, resulting in 5-fold CV. For Recola, we construct manually five splits so that they are balanced with respect to number of

[1] class distribution IEMOCAP: arousal - 3,121 low, 6,918 high; valence - 3,421 neg., 6,618 pos. — Recola: arousal - 520 low, 788 high; valence - 241 neg., 1,067 pos.

speakers and sex. This way, we ensure speaker-independent training (in contrast to random sampling).

The evaluation of (c) cross-lingual training is more straightforward, we take all data of one language as training set and all samples of the respective other as test set. For (d) fine-tuning (FT) in the simulated low-resource setup we take trained models from (c) as starting point. The model is then refined using 100 randomly selected samples from the target language for each CV split. Consequently, only 500 samples of the target language are used in total for FT.

In order to observe variations in the results due to non-deterministic operations on the GPU, we run all experiments five times and report the means.

Hyper-parameters

The ACNN model is implemented with the Tensorflow library [14]. We apply stochastic gradient descent with an adaptive learning rate (Adam [15]) for training. The systems hyper-parameters are the following: 200 kernels with a size of 26x10 in the convolutional layer (spanning all 26 logMel filter-banks); a mini-batch size of 32; and a pool size of 30 for max-pooling. For regularization we apply dropout ([16]) to the last hidden layer with a dropout rate of 0.5. We run training for 50 epochs in all experiments except for fine-tuning where the pre-trained models are refined with only 10 epochs.

The kernel width of 10 and the pool size of 30 were selected by tuning the mono-lingual models on the development set with a number of different parameter combinations. These relatively high values appear reasonable considering that emotions in speech are determined as long-term information. The kernel size of 10 corresponds to 100ms from the input signal. A large amount of overlap in the feature maps explains the large pooling window.

5. RESULTS

The performance measure used throughout all experiments is unweighted average recall (UAR, i.e. the average of the recall for each class). This measure best reflects the overall accuracy when the dataset is imbalanced with respect to the number of samples per class. The results are presented in Table 2.

The mono-lingual baselines for both IEMOCAP and Recola show that the prediction of valence is more difficult than arousal. This finding is in line with [3, 6, 17]. The performance for Recola is notably lower than for IEMOCAP. This is only partially due to the small size of Recola (1,308 samples). Using only 1,308 samples from IEMOCAP in comparison still leads to better results for English (68.20% arousal and 58.77% valence). Another possible reason is that the French data is highly imbalanced for valence (UAR of 52.30% is only slightly better than chance).

With multilingual training we want to investigate the effect of merging the two corpora and find out whether multi-lingual speech emotion recognition is possible without performance loss. The results show that we are able to use a system trained on both languages and achieve similar performance compared to the baselines. For arousal prediction, the additional training data even improves performance, whereas we observe a decrease in performance for valence. These findings are a first evidence that multilingual speech emotion recognition is viable without further adaptation.

	IEMOCAP (English)		Recola (French)	
	Arousal	Valence	Arousal	Valence
mono-lingual	68.09	**62.33**	60.77	**52.30**
multilingual	**70.06**	61.73	62.51	49.33
cross-lingual	59.32	49.08	61.27	47.52
CL + FT	67.03	50.42	**63.07**	49.81

Table 2. Results as unweighted average recall (UAR), cross-lingual: only trained on source language, CL + FT: pre-trained on source language and fine-tuned on 500 samples from target lanugage (CL - cross-lingual, FT - fine-tuning).

Cross-lingual training is useful in cases where no or only little training data in the target language is available. We therefore examine the performance of the system when trained on one language and tested on the other (and vice versa), given the same type of speech (human-human interaction). The results in Table 2 show that cross-lingual training works to some extend for arousal but not for valence prediction. For arousal, the performance drops notably for IEMOCAP (trained on Recola) compared to the mono-lingual baseline, achieving 59.32% UAR. For Recola (trained on IEMOCAP) it remains stable (60.77% mono-lingual, 61.27% cross-lingual). For valence, both results are below chance, suggesting that valence prediction might be more language-dependent than predicting arousal.

Fine-tuning the cross-lingual model with 10 training epochs on 500 samples from the target language produces promising results for arousal prediction. For IEMOCAP, the performance comes close to the baseline and for Recola, it is notably higher than the baseline. Again, the performance for valence remains approximately at chance level.

In summary, these results show that cross-lingual training can set a useful baseline. Especially for a target language with a small amount of annotated data, training a cross-lingual model and then fine-tuning it on the available target data appears to be a reasonable approach.

6. ANALYSIS OF ATTENTION WEIGHTS

To gain more insights about which parts of the input are important for classification, we analyze the attention weights α_i from the attention layer after the last training epoch. We focus on the mono-lingual baseline experiments in arousal predic-

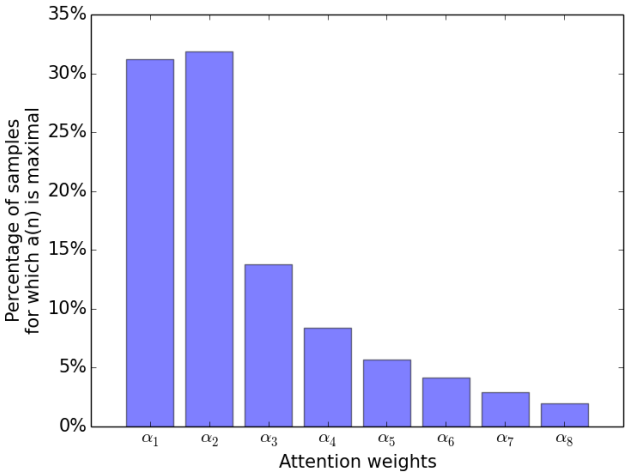

Fig. 2. Distribution of attention over time for arousal prediction on IEMOCAP.

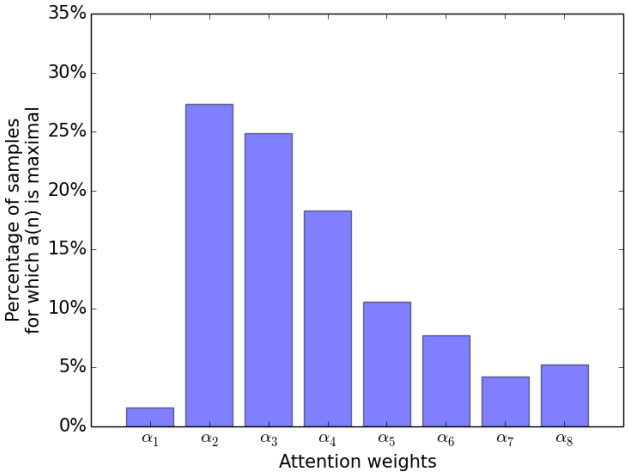

Fig. 3. Distribution of attention over time for arousal prediction on Recola.

tion. For each training sample, we output its attention weights and identify the maximum weight, i.e. the segment which appears to be most salient for this sample. Figures 2 and 3 show for every attention weight α_1 to α_8 the particular proportion of training samples for which this α_i yielded the maximum value. For example in Figure 2, for 31.2% of training samples, α_1 was highest, hence the first segment of the input to the attention layer is considered most salient. The number of attention weights corresponds to the output vector of the max-pooling layer and therefore depends on input signal length, kernel size and pool size. Figure 2 shows this distribution for the English data and Figure 3 for French.

From Figure 2 it can be observed that for a large majority of samples the attention lies at the beginning of the input. This finding is in line with the observation in [10] that a short snippet from the beginning of an utterance can be sufficient for a prediction in many cases. In addition to depicting the maximum attention weights, we took a closer look at the actual values of the maximum and the second highest weight to find out more about the weight distribution. Note, that the weights α_1 to α_8 sum up to 1.0 for every sample. For the English training data we found that for 73.1% of all samples the difference between highest and second highest attention weight is greater than 0.5. This means, for the majority of data one segment is weighted much higher than all others.

For the French data, the picture looks a bit different. Figure 3 reveals that α_2 to α_4 yield the maximum weight for a large proportion of data. Apart from α_1, the distribution exhibits similar characteristics as in Figure 2, that the beginning of the input is much more often considered important than the end. Our first hypothesis to explain the low rate for α_1 was that many samples contain silence at the beginning. However, using voice activity detection, we found that most signals contain speech straight from the beginning. Hence, further analysis is necessary to explain this difference. In the Recola dataset the difference between highest and second highest attention weights is only for 5% of training samples greater than 0.5. This overall flatter distribution suggests that it is more difficult to learn meaningful attention weights for the French data compared to English.

To conclude this analysis, we have found notable differences in the attention mechanism between the two datasets. But it is difficult to draw final conclusions about the languages themselves because the corpora are not recorded under same conditions (especially the underlying task for the participants). Hence, these findings point towards language-dependent characteristics in emotional speech, but are potentially skewed by language-independent factors such as recording situation or the lexical content of the conversations.

7. CONCLUSION

We presented results for binary arousal/valence classification using cross-lingual and multilingual training. We have shown that multilingual classification of emotions in speech is possible and can even enhance results for arousal prediction. This can be regarded as a valuable finding for research on code-switching speech. Further, we have shown that a model trained on a source language and fine-tuned with only a small number of samples from the target language can produce sound results for arousal prediction, whereas valence prediction appears to be more sensitive to cross-lingual training. These findings are potentially interesting for emotion research on low-resource languages.

8. ACKNOWLEDGEMENT

This work was funded by the German Science Foundation (DFG), Sonderforschungsbereich 732 Incremental Specification in Context, Project A8, at the University of Stuttgart.

9. REFERENCES

[1] Bjorn Schuller, Bogdan Vlasenko, et al., "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[2] Iulia Lefter, Leon JM Rothkrantz, et al., "Emotion recognition from speech by combining databases and fusion of classifiers," in *International Conference on Text, Speech and Dialogue*. Springer, 2010, pp. 353–360.

[3] Florian Eyben, Anton Batliner, et al., "Cross-corpus classification of realistic emotions–some pilot experiments," in *Proc. LREC workshop on Emotion Corpora, Valettea, Malta*, 2010, pp. 77–82.

[4] Björn W Schuller, Zixing Zhang, et al., "Using multiple databases for training in emotion recognition: To unite or to vote?," in *Interspeech*, 2011, pp. 1553–1556.

[5] Björn Schuller, Zixing Zhang, et al., "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proc. 2011 Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*, 2011.

[6] Silvia Monica Feraru, Dagmar Schuller, et al., "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 125–131.

[7] Hesam Sagha, Pavel Matejka, et al., "Enhancing multilingual recognition of emotion in speech by language identification," *Interspeech*, 2016.

[8] Bo-Chang Chiou and Chia-Ping Chen, "Speech emotion recognition with cross-lingual databases.," in *Interspeech*, 2014, pp. 558–561.

[9] Je Hun Jeon, Duc Le, et al., "A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception.," in *Interspeech*, 2013, pp. 2837–2840.

[10] 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," *Proc. Interspeech*, 2017.

[11] Carlos Busso, Murtaza Bulut, et al., "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[12] Fabien Ringeval, Andreas Sonderegger, et al., "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[13] Roddy Cowie, Ellen Douglas-Cowie, et al., "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[14] Martín Abadi, Ashish Agarwal, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[15] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations (ICLR)*, 2015.

[16] Nitish Srivastava, Geoffrey E Hinton, et al., "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] George Trigeorgis, Fabien Ringeval, et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.