University of Stuttgart
Germany

# Cross-lingual and Multilingual Speech Emotion Recognition on English and French

Michael Neumann, Thang Vu

{michael.neumann|thang.vu}@ims.uni-stuttgart.de

## Overview

**Are emotional expressions language-independent?**

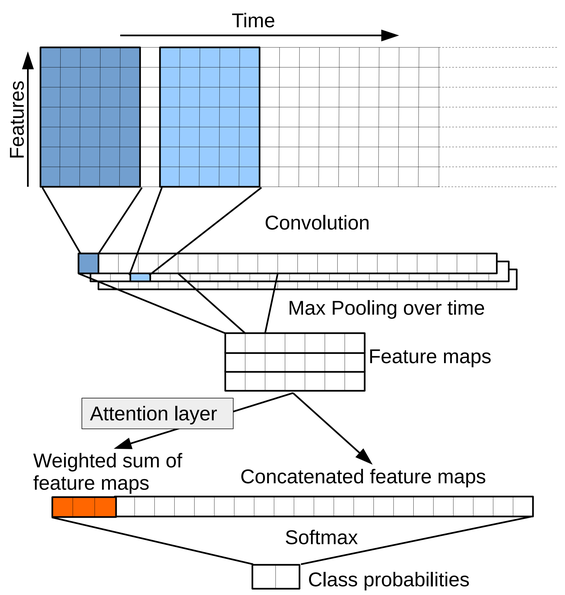
We present findings on multilingual and cross-lingual emotion recognition on English and French speech data with similar characteristics (human-human conversations).

**Main findings:**

- **Multilingual** emotion recognition **possible for arousal** prediction
- **Cross-lingual** training **plus fine-tuning** on target language → **sound results for arousal** prediction
- **Valence** prediction **more sensitive** to cross-lingual and multilingual training
- **Attention** mostly on **beginning of speech signal**

## Model Architecture

- Attentive convolutional neural network [1]
- Input: 26 logMel filter-banks (frame-wise, 25ms frames)
- Output: binary arousal / valence labels
- Trained with adaptive learning rate (Adam) over 50 epochs
- Hyper-parameters: 200 filters with size 26x10, mini-batch size 32, pool size 30, dropout 0.5



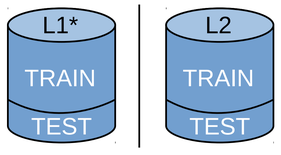

## Data

Main criterion for selecting data:
Natural human-human conversations

- English: IEMOCAP database [2]
  *10,039 utterances from 10 speakers*
- French: RECOLA database [3]
  *1,308 utterances from 23 speakers*

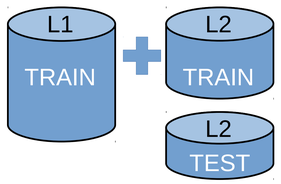
## Experimental Results

### (a) Monolingual Baselines




- Train and test on one language
- Arousal easier than Valence
- Results on Recola notably lower than on IEMOCAP

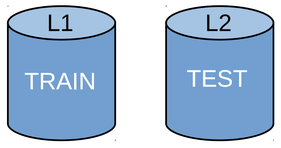
### (b) Multilingual Training




- Merge two languages for training
- Performance similar to baselines

→ Multilingual recognition is viable

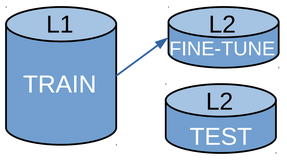
### (c) Cross-lingual Training




- Train on L1, test on L2
- Useful if no/little training data in target language available
- Works to some extend for arousal, not for valence

→ Valence more language-specific?

### (d) Cross-lingual (CL) + Fine-tuning (FT) on target language




- Fine-tune model from (c) on small portion of target language
- Promising results for arousal prediction
- For low-resource languages, cross-lingual pre-training is a reasonable approach

* L1 - language 1; L2 - language 2

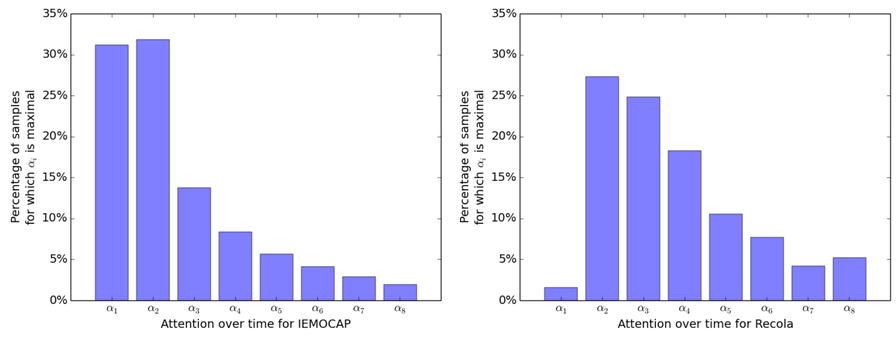
** For (a), (b) and (d) we applied 5-fold cross validation

**Results Overview (unweighted average recall)**

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

### Analysis of Attention Weights

- Identified maximum attention weight for each training sample (i.e. most salient segment for this sample)
- Plots show proportion of samples for which attention weight $\alpha_i$ yields the maximum value
- IEMOCAP: Attention lies at the beginning of the input ($\alpha_1 - \alpha_3$) for large majority of samples
- Recola: Apart from $\alpha_1$ the distribution looks similar, unclear whether difference is corpus- or language-specific




## Selected References

[1] 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 *Proc. of Interspeech*, 2017.

[2] Carlos Busso, Murtaza Bulut, et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, 2008.

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

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

[5] 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.

[6] Bo-Chang Chiou and Chia-Ping Chen, "Speech emotion recognition with cross-lingual databases.," in *Proc. of Interspeech*, 2014.

## Acknowledgements

This work was funded by the German Research Foundation (DFG) Sonderforschungsbereich 732 Incremental Specification in Context, Project A8.