A Comparative Cross language View On Acted Databases Portraying Basic Emotions Utilising Machine Learning

F. Burkhardt[1,2], A. Hacker[1], U. Reichel[2], H. Wierstorf[2], F. Eyben[2], B.W. Schuller[2,3,4]
[1]Technical University of Berlin, [2]audEERING GmbH, [3]University of Augsburg, [4]Imperial College London

Introduction & Summary

- Acted portrayals of basic emotions are still in the focus of research and technology
- We investigated *neutral*, *anger*, *happiness*, and *sadness*
- By comparing six natural and one synthetic database with two approaches:
  - machine learning cross corpus validation
  - expert acoustic feature comparison
- We found similarities as well as language constraints on acoustic speech features in encoding emotion
- Of course it must be noted that the databases are very different in many respects

Databases

Name	Language	Paper	#speak	#emo	#sent	#smpl
emodb	German	Burkhardt et al. 2005	10	7	10	484
emovo	Italian	Costantini et al. 2014	6	7	14	588
ravdess	English	Livingstone & Russo, 2018	24	8	2	1 440
Polish Emotional Speech	Polish	Powroźnik 2014	8	6	5	240
des	Danish	Engberg et al. 1997	4	5	13	260
buemodb	Turkish	Kaya et al. 2014	11	4	11	484
synthesised	German	Burkhardt 2022	6	4	720	720

Table: Overview of the emotional speech databases used

Analysis I: Machine learning

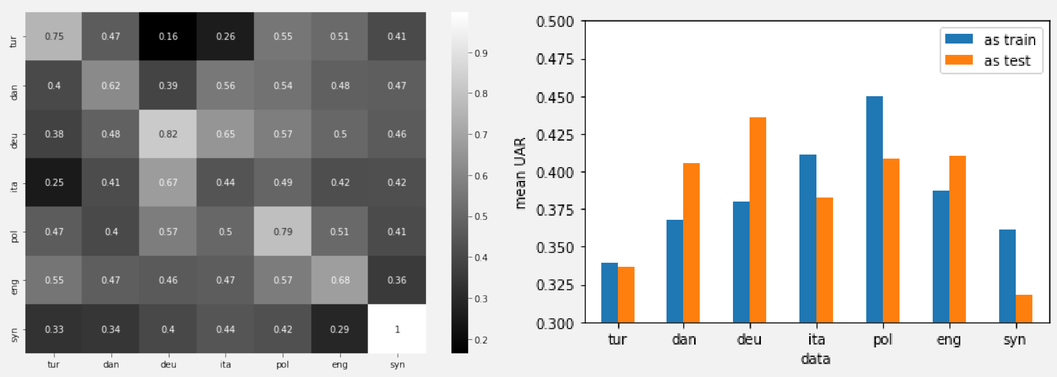
Classifier and Features

- Used the Nkululeko framework (`https://github.com/felixbur/nkululeko/`)
- Classifier: XGBoost (Chen and Guestrin, 2016) (eta = 0.3, max depth = 6, subsample = 1).
- Acoustic features (#88): extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al, 2015)

Discussion

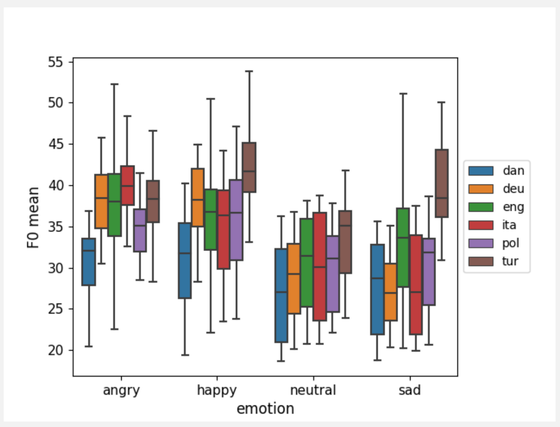
- See the figures for visualization of results
- All self performance values are clearly above chance level
- Most databases work better as either test or training set
- **Turkish** (buemodb) does not generalize very well, analyses indicate high arousal for all emotions
- **Danish** works reasonably well (but only 4 speakers)
- **German** (emodb) works quite well with all databases apart from the Turkish one, especially as a test set
- **Italian** (emovo) database does not perform very well in-domain (but a good model for others)
- **Polish** works quite well, especially when used as training.
- **English** (ravdess) works comparatively well, it's the largest database
- **synthesized** data works as a training for all natural databases, with the exception of Ravdess,

Figure: left) Heatmap (UAR) when used as test (rows) vs. train, diagonal is 50% speaker split. right) average (against all other databases) UAR of the databases



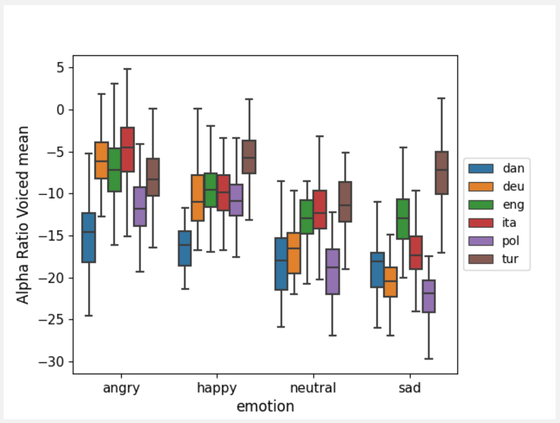
Analysis II: Feature Analysis

Prosody: F0 mean



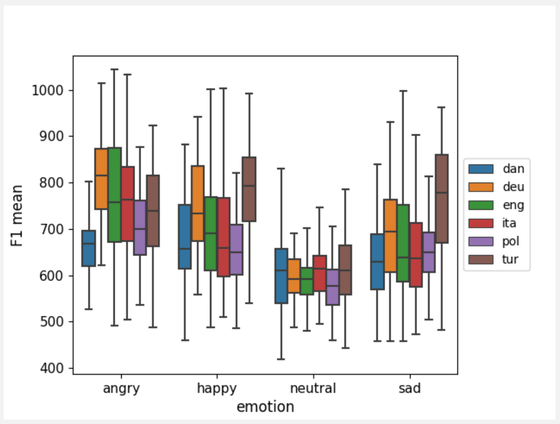
- **English, German, Italian, Polish:** distinguish between happy/angry and neutral/sad
- **Turkish:** distinguish between neutral and non-neutral
- **Danish:** less distinction between emotions

Voice quality: Alpha ratio



- ratio of high (1-5kHz) to low (50Hz-1kHz) spectral energy; higher values indicate increased vocal effort
- same cross-language pattern as for F0 mean

Articulation: F1 mean



- higher values indicate lowered jaw
- **all languages:** F1 lowest for neutral emotion
- lowered jaw reflecting acted non-neutral speech?

Discussion

- Turkish is different from the other databases
- Polish is very similar to most of the other languages, especially in terms of F0 mean and F1 mean.
- Danish shows low arousal in all emotions

Outlook

- Make the databases more alike, e.g. restrict samples to common set of speakers and text material
- Machine learning likeness: compare different classifiers or features
- Use pre-trained (transformer) embeddings as features to have better representation learning

Acknowledgements

This research has been partly funded by the European EASIER (Intelligent Automatic Sign Language Translation) project (Grant Agreement number: 101016982).