Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1917–1924
Marseille, 20-25 June 2022

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
Germany, Germany, Germany, UK
{f.burkhardt, anabell.hacker}@tu-berlin.de
{ureichel, hwierstorf, fe, bs}@audeering.com

Abstract

For several decades emotional databases have been recorded by various laboratories. Many of them contain acted portrays of Darwin's famous "big four" basic emotions. In this paper, we investigate in how far a selection of them are comparable by two approaches: on the one hand modeling similarity as performance in cross database machine learning experiments and on the other by analyzing a manually picked set of four acoustic features that represent different phonetic areas. It is interesting to see in how far specific databases (we added a synthetic one) perform well as a training set for others while some do not. Generally speaking, we found indications for both similarity as well as potential language-specific differences.
Keywords: emotional, database, language resource evaluation

1. Introduction

The way we speak is influenced by many factors. One of them is the emotional expression, a field that has been studied by numerous researchers over the past four decades (Schuller and Schuller, 2021; Gangamohan et al., 2016; Schröder, 2001). Many of these studies recorded their own emotional speech data, often acted versions of the "big four basic emotions": *anger*, *happiness*, *sadness* and *fear*.

In the theory of a "basic emotion" concept, these are the emotions that should be interpretable in a culturally universal manner and thus should be expressed in a comparable manner. In reality emotional expressions are influenced by culture and context (Scherer et al., 1999; Burkhardt et al., 2006; Barrett et al., 2019). In parts the outcomes of this investigation may answer the question, in how far the language (and underlying culture) of the speech resources influenced the acoustical correlates of the emotional portrayals, as has been done in, for example, (Tamulevičius et al., 2020; Neumann and Vu, 2018; Feraru et al., 2015; Schuller et al., 2010).

To investigate the matter, we analysed several databases from the literature that included at least the basic emotions *neutral*, *anger*, *happiness*, and *sadness* with respect to a set of hopefully meaningful acoustic parameters that we extract algorithmically.

In a first analysis step, we performed cross data supervised machine learning. Each database was used once as a training set and once as a test set. The underlying idea is that databases that portray culturally universal basic emotions should be able to act as a training or test set in cross database machine learning. We can then estimate the performance to learn emotional expression as portrayed in the other databases, thus resulting in an estimate of similarity. We added one synthesised database that uses rule-based prosody manipulations to express basic emotions as the result of a literature search for the best parameters (Burkhardt, 2005).

In a second investigation, we estimated the differences between the emotion portrayals by a statistical analysis of selected acoustic features.

In section 5, we compare and discuss the findings of the two experiments. The paper concludes with a summary and an outlook in section 6

The main contributions of this paper are:

- We compare a set of well known emotional "standard databases" systematically with respect to their phonetic qualities. As these databases are being used quite often in the literature, this study intends to help authors to estimate the generalisability of the respective database.
- We use machine learning performance as an indicator for likeness in emotional expression across cultures.

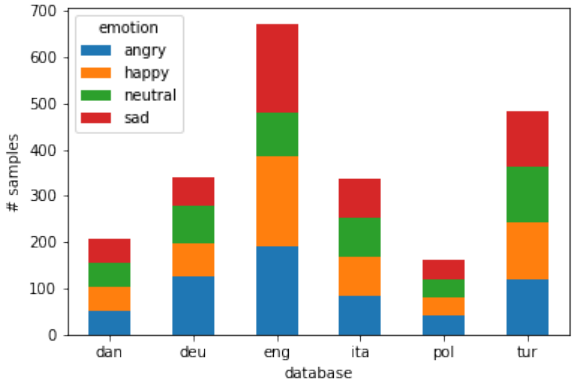
To our best knowledge this is the first time that this has been done.

2. Databases

We look at the following seven databases from different countries:

- 'emodb' (Germany, ISO 639-2 language code "deu")
- 'emovo' (Italy, "ita")
- 'ravdess' (USA, "eng")
- 'polish' (Poland, "pol")
- 'des' (Denmark, "dan")
- 'buemodb' (Turkey, "tur")
- 'synthesised' (Germany, "syn" – synthesised German speech)

Figure 1: Overview of databases with respect to basic emotion portrays



An overview of the databases is provided in Table 1.

The Berlin Emotional Speech Database (emodb)[1] (Burkhardt et al., 2005) is a well known studio recorded dataset. Ten (five female, five male) professional actors speak five longer and five shorter sentences German sentences, their semantics being neutral in emotion, in seven affective states (anger, boredom, disgust, fear, joy, neutral, and sadness). The final 494 phrases were selected using a perception test judging how well the emotions were acted.

Italian Emotional Speech EMOVO[2] (Costantini et al., 2014) is a database consisting of the voices of six actors (three female, three male) who utter 14 Italian sentences simulating seven emotional states: anger, disgust, fear, joy, neutral, sadness, and surprise. The Italian text material includes three short sentences, four long sentences, all emotionally neutral in semantics, and seven medium length non-sense sentences (including two questions).

The Ryerson Audio-Visual Database of Emotional Speech and Song (ravdess)[3] (Livingstone and Russo, 2018) contains recordings of 24 professional actors (12 female, 12 male), vocalising two short English statements in a neutral North American accent. Speech samples include angry, calm, disgust, fearful, happy, sad, and surprise expressions. Each expression is produced at two levels of emotional intensity (normal, strong) and in a neutral expression which makes 1 440 speech samples. The song recordings were not used in this experiment.

The Database of Polish Emotional Speech (Powroźnik, 2017) consists of speech from eight actors (four female, four male). Each speaker utters five short sentences with six types of emotional state: anger, boredom, fear, joy, neutral, and sadness. There is a total of 240 samples in this database.

The Danish Emotional Speech (des) (Engberg et al., 1997) database comprises acted emotions of four professional actors, two males and two females, for five emotional states: anger, happiness, neutral, sadness, and surprise. The material includes two short single words ("yes" and "no"), nine short sentences (including four questions) and two long passages of fluent speech. In total, there are 260 samples in the des database.

For the Turkish Emotional Database (buemodb) (Kaya et al., 2014) eleven amateur actors (eight female, three male) deliver eleven Turkish sentences with emotionally neutral content. Each sentence is recorded four times; each time with a different emotional state of the following four: angry, happy, neutral, and sad. Thus, the buemodb contains 484 utterances.

The synthesised database consists of samples that were generated with the rule based emotion simulation software "Emofilt" (Burkhardt, 2005). It utilises the diphone speech synthesiser MBROLA (Dutoit et al., 1996) to generate a wave form from a phonetic description (SAMPA symbols with duration values and fundamental frequency contours) and the MARY TTS system (Schröder and Trouvain, 2003) to generate the neutral phonetic description from text. Emofilt acts as a filter in between to 'emotionalise' the neutral phonetic description, the rules are based on a literature research described in (Burkhardt, 2000). All six available German MBROLA voices; *de2*, *de4*, and *de7* as female and *de1*, *de3*, and *de6* as male, were used. As text material, we used a German news corpus from the University of Leipzig[4]. No special preprocessing was applied. The selection used for this article is part of a larger set of 1000 samples per speaker and emotion (about 24000 samples). We restricted the data size to align the number of samples with the other databases. For the four target emotions (angry, happy, neutral, and sad), out of the 1000 we selected randomly 30 samples per speaker and emotion, getting 720 samples with distinct texts that might help to enhance the diversity of the data. The software framework that generated the samples has been made open source[5] (Burkhardt et al., 2022a).

The distribution of basic emotion portrays and biological sex of actors per emotion can be seen in Figures 1 and 2. The emotions are quite well balanced, but we have more women than men in our data because the Turkish database is gender imbalanced.

3. Analysis I: Machine Learning

In this section, we estimate the similarity of databases based on their mutual performance in cross-database machine learning experiments.

In supervised machine learning approaches, a classifier or regressor is trained on a training set of data samples

[1] https://www.tu.berlin/go22879/

[2] http://voice.fub.it/EMOVO

[3] https://doi.org/10.5281/zenodo.1188975

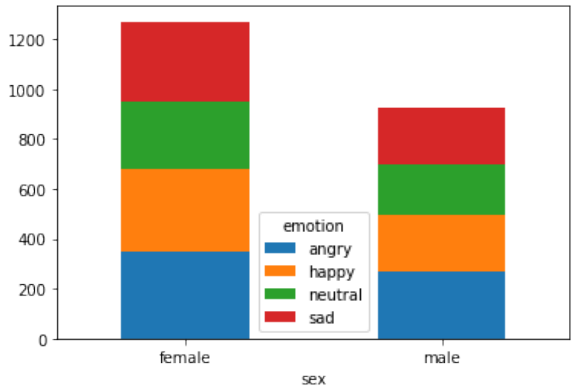
[4] https://wortschatz.uni-leipzig.de

[5] https://github.com/felixbur/syntAct

Name	Language	Year	#speakers	#emotions	#sentences	#samples
emodb	German	1999	10	7	10	484
emovo	Italian	2014	6	7	14	588
ravdess	N.-A. English	2018	24	8	2	1 440
Polish Emotional Speech	Polish	2014	8	6	5	240
des	Danish	1997	4	5	13	260
buemodb	Turkish	2014	11	4	11	484
synthesised	German	tba	6	4	720	720

Table 1: Overview of the emotional speech databases

Figure 2: Overview of acted emotions with respect to sex distribution



and then evaluated on a test set. The resulting evaluation metric can be seen as a measure of likeness, because the more similar the training is to the test set, the higher the accuracy of the machine learning predictions.

For the databases at hand, we thus set up a series of cross-database classification experiments by pairing them once as test and once as training sets.

If the database was evaluated against itself (the diagonal in Figure 4), we selected randomly half of the speakers as test and the other half as training set, irrespective of the number of samples. We think this approach is justified, because all databases come from controlled experimental data and the samples per speaker are nearly equally distributed. For reproducibility we provide a link to the speaker partitions here[6]

3.1. Classifier and Features

It must be noted that, prior to classification, we centered and scaled the features per speaker – an approach that is often not possible in real world applications because usually, the test speakers are unknown. We still decided for this approach, as such "speaker dependent" classification is not unheard of and the results become stronger.

For the experiments we employed the Nkululeko framework[7] (Burkhardt et al., 2022b) with an XGBoost classifier[8] (Chen and Guestrin, 2016) with the default meta parameters ($eta = 0.3$, $max\_depth = 6$, $subsample = 1$). This classifier is basically a very sophisticated algorithm based on classification trees and has been working quite well in many of our experiments (Burkhardt et al., 2021).

As acoustic features, we used the the eGeMAPS set (Eyben et al., 2015), an expert set of 88 acoustic features for the openSMILE feature extractor (Eyben et al., 2010) that were optimised to work well to explain speaker characteristics and in particular emotions. These features are being used in numerous articles in the literature as baseline features (e.g. (Ringeval et al., 2018; Schuller et al., 2016)) as they work reasonably well with many tasks and are easy to handle for most classifiers based on their small number.

3.2. Results

In Figure 4, the resulting heat map is displayed in form of the unweighted average recall (UAR) per experiment. The rows indicate the database that has been used as test set and the column the one as a training set. The diagonal represents the self performance, separating speakers evenly in test and train.

To better compare the performance as a "role model" for other databases, we add a plot that shows the mean UAR per database when used as test or train respectively in Figure 3. Interestingly, the databases differ quite strongly when used as test or as train. While the Polish data seems to deliver a good model for most of the other languages, the German data performs above average when being used as test set.

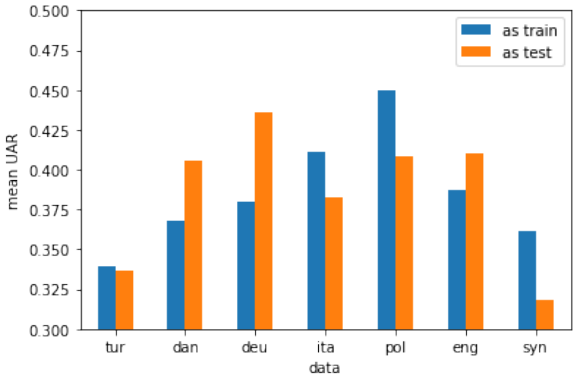
Because all these databases are laboratory data, the emotional expression is nearly evenly distributed and

[6] http://blog.syntheticspeech.de/2022/01/06/emotional-acted-database-comparison/

[7] https://github.com/felixbur/nkululeko/

[8] Actually using the Python XGBoost package: https://xgboost.readthedocs.io/

Figure 3: Mean (with respect to all other databases, excluding the self performance) UAR per database when used as the training or the test set



the chance UAR would be around .25 for four emotions, hence, all UARs that are substantially higher signify that the classifier learnt information related to emotions. We are happy to see that for all databases, even with a very low number of speakers like Danish (only two speakers in training and two in test), the self performance values (diagonal) are clearly above chance level.

In the order of descending self performance, we get the following ranking: synthesised, German, Polish, Turkish, English, Danish, Italian. The perfect in-domain performance of the synthesised samples is easy to explain: all emotional expression was simulated by the same rules and thus is easy to detect. But also for all other databases with the exception of English, the synthesised training works quite well to detect emotional expression, confirming our earlier findings (Schuller and Burkhardt, 2010). The low in-domain result for the Italian database seems strange but then these findings are merely peculiarities of the databases.

The German data seems to deliver a good model for prototypical emotion expression, especially in itself but also for other databases, apart from Turkish.

Strangely, the model trained on the Italian database performs with a higher UAR when it is tested on German, Danish and Polish data as when tested in-domain on Italian data. Similar: when Italian is used as a test set, the German model works better than Italian itself. We do not have an explanation yet.

The Turkish database on the other hand acts opposite to the Italian data set. The Turkish data works best when tested and trained on itself. It performs especially bad when tested with the German or Italian model. We informally listened with three listeners to the samples and the common impression is that the emotion display generally does seem to be more aroused than in the other databases on hand. But whether this is due to culture or the instructions for the actors, we cannot evaluate.

When listening to the Danish samples, it seems that the emotion portrays are, especially compared to the Turkish, rather restrained. Nonetheless, tests and trainings perform about twice the chance level and in-domain about 10 % better. It must be noted that there were only four speakers.

The Polish database performs quite well for all languages when used as a training set. Also, when used as a test set, it reaches clear above chance UARs with all other databases.

The English database is the largest one and works comparably well with all languages, but clearly not the synthetic samples.

4. Analysis II: Cross-Language Speech Feature Analysis

The results in Figure 4 show a performance decrease in cross-language emotion recognition for all databases except of Italian. In this section, we explore in how far this performance decrease may be explained by language constraints on acoustic speech features in encoding emotion. We focus on 4 feature functionals from the eGeMAPS set which cover voice quality, articulatory, and prosodic aspects of vocal emotion expression:

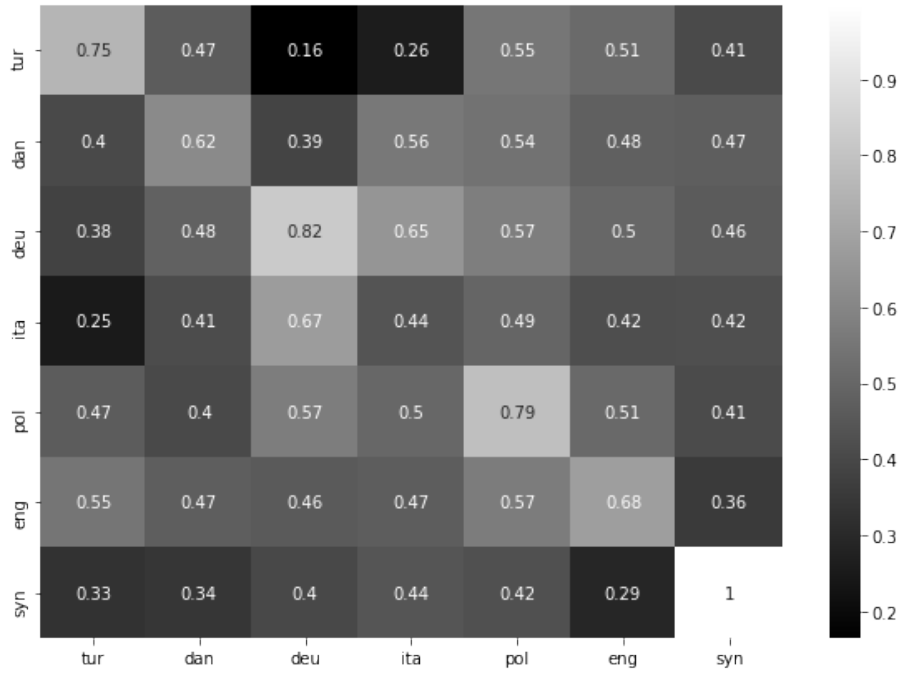
- F0 mean (prosody): the arithmetic mean of F0 converted to semitones relative to a base value of 27.5 Hz and smoothed by a moving average filter
- loudness peaks per second (prosody): measurement of energy peaks per second as a proxy for speaking rate
- alpha ratio (voice quality): ratio of high (1-5 kHz) to low (50-1000 Hz) spectral energy in voiced segments. Higher values indicate a flatter spectral slope which results from increased speed of glottal closure and which is an acoustic correlate of increased vocal excitation effort
- F1 mean (articulation): the arithmetic mean of the first formant, which is causally related to vertical jaw position. Higher values indicate a lowered jaw and a greater mouth opening

The covered acoustic cues, F0 mean, speaking rate, spectral slope, and F1 mean, have been shown to contribute to the vocal encoding of emotion e.g. by (Scherer, 2003) and (Goudbeek et al., 2009). The feature value distributions across emotions and languages are presented in Figures 5, 6, 7, and 8.

Next to visual inspection of distribution differences, we applied linear mixed effects (LME) models[9] with *emotion* and *language* as fixed effects, and *speaker* as a random effect added by random intercepts. We further included an interaction term of the fixed effects. The models were fitted by minimising the REML criterion. P-values were obtained by ANOVAs, and the significance level was set to .05.

[9]Python *statsmodels* package, version 0.12.2

Figure 4: Heat map for cross-database experiments. The values in the rows mean that the specific database has been used as the test, the values in the columns that it has been used as training database. The diagonal represents the outcomes of the self classification (randomly 50% of speakers split)

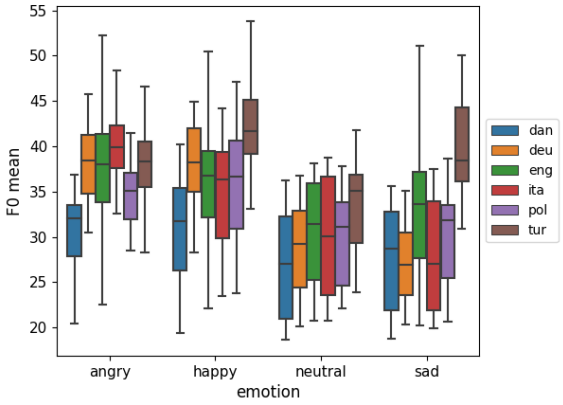


In the examined data sets, F0 mean (Figure 5) clearly shows a language impact on how this feature is utilised in emotion encoding. For English, German, Italian, and Polish speakers major F0 differences are observed between angry and happy with high F0, as opposed to neutral and sad with low F0. This is in line with the expectations, since the F0 level is correlated with arousal. Turkish speakers (brown boxes) in contrast show a major F0 difference between neutral (low F0) vs non-neutral emotions (high F0). Furthermore, due to the gender imbalance towards female speakers, F0 is overall higher in the Turkish dataset. Finally, for Danish speakers (blue boxes), F0 is overall low and shows much less pronounced differences across emotions. Language-dependency in vocal emotion expression is furthermore indicated by the LME model yielding significant results for 13 out of 15 emotion and language level interactions.

The distribution pattern of alpha ratios (Figure 6) is similar to the patterns observed for F0. Again, for our datasets, English, German, Italian, and Polish speakers distinguish between happy/angry and neutral/sad with more vocal effort for the former and less for the latter, while Turkish speakers rather distinguish between happy/angry/sad and neutral. Danish speakers again distinguish much less between the four emotions. Also here, the LME model gives significant results for the majority (12 out of 15) emotion and language level interactions indicating the influence of language in vocal emotion expression.

For the examined datasets F1 means show a more uniform pattern across languages (Figure 7). By and large,

Figure 5: Distribution of F0 means across emotions and languages. Languages: *dan* – Danish, *deu* – German, *eng* – English, *ita* – Italian, *pol* – Polish, *tur* – Turkish.



F1 values are lowest for utterances in neutral emotion. This might have been partly caused by the elicitation scenario: all non-neutral emotions the speakers needed to act out, for which they might have utilised 'stage speech' reflected in an increased mouth opening. Unfortunately, these findings cannot be further backed up by statistics, since the LME model did not converge (neither for alternative optimisation criteria).

Finally, loudness peaks per second (Figure 8) as a proxy of speaking rate shows a clear impact of language or database, but within each language no im-

Figure 6: Distribution of alpha ratios across emotions and languages. Languages: *dan* – Danish, *deu* – German, *eng* – English, *ita* – Italian, *pol* – Polish, *tur* – Turkish.

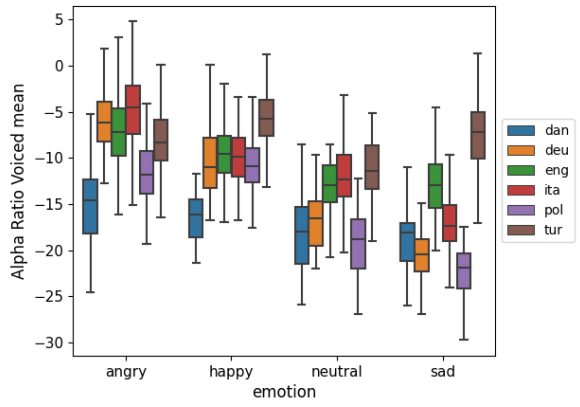
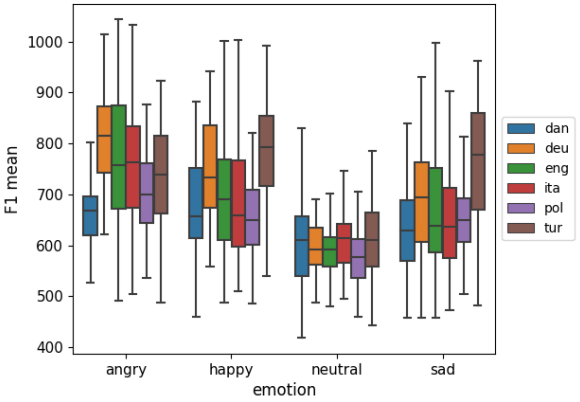


Figure 7: Distribution of F1 means across emotions and languages. Languages: *dan* – Danish, *deu* – German, *eng* – English, *ita* – Italian, *pol* – Polish, *tur* – Turkish.

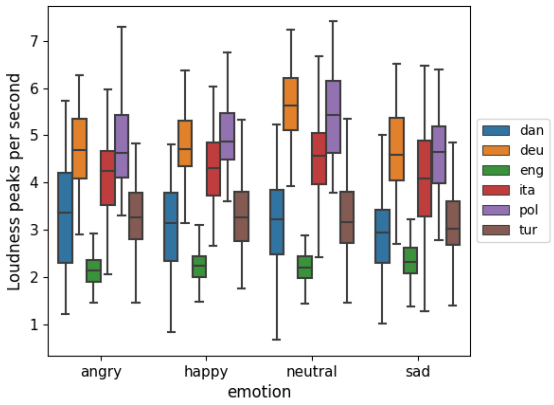


pact of emotion, since for each language these rates are stable across all emotion categories. Since it is well known that emotions can be distinguished by speaking rate (Murray and Arnott, 1993), it is likely that the used proxy feature is too sensitive to the channel characteristics of the different recording settings and thus not sufficiently robust for cross-database comparisons.

5. Discussion

The relatively low performance of the Turkish database in the machine learning experiment (see figures 3 and 4) might be backed in parts by the analysis of the acoustic features. When looking at the figures 5, 6, 7, and 8, the boxplots for the Turkish samples deviate from the other languages in many cases, especially for the emotions happy and sad. When listening to the audio files, the Turkish emotional expressions convey an impression of higher arousal than in other databases. It needs to be further investigated if this is a matter of cultural difference on how intense emotions are expressed or whether it is a specialty to the specific database used.

Figure 8: Distribution of loudness peaks per second across emotions and languages. Languages: *dan* – Danish, *deu* – German, *eng* – English, *ita* – Italian, *pol* – Polish, *tur* – Turkish.



The Polish database on the other hand performed especially well in the machine learning analysis and the acoustical analysis shows that Polish is very similar to most of the other languages, especially in terms of F0 mean and F1 mean.

In the Danish database, the emotional arousal is auditively low when listening to the samples. This is also visible in the acoustic data: In the high arousal emotions anger and happiness, the Danish data shows low values in all four feature functionals from the eGeMAPS set. This fact could be linked to overall less emotional arousal in Danish culture, but needs further research.

Interestingly, the Italian database performed relatively bad with itself in the machine learning experiment (see figure 4). The Italian database performed better with the German database as test and training database than with itself, for example.

6. Conclusion and Outlook

In this paper, we investigated a set of mostly European, well known data sets of acted emotional basic categories. Many of them were used in the literature and we were interested in how far they are comparable, given that they come from different languages / cultures. As expected, we find arguments/indications for both similarity as well as specificiality.

Future research would extend the field of databases and investigate other prominent acoustic features. Generic pre-trained speech representations that are more robust against acoustical changes (e.g. (Baevski et al., 2020)) might provide a better understanding of the contribu-

tion by language or by recording conditions to the findings.

We also might attempt a more strict comparison by limiting the number of speaker and samples for all databases to the number that is present in all databases.

7. Acknowledges

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

8. Bibliographical References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological science in the public interest*, 20(1):1–68.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of german emotional speech. In *9th European Conference on Speech Communication and Technology*, volume 5, pages 1517–1520, 09.

Burkhardt, F., Audibert, N., Malatesta, L., Türk, O., Arslan, L., and Auberge, V. (2006). Emotional prosody - does culture make a difference? In *Proceedings of the International Conference on Speech Prosody*.

Burkhardt, F., Brückl, M., and Schuller, B. (2021). Age classification: Comparison of human vs machine performance in prompted and spontaneous speech. In Stefan Hillmann, et al., editors, *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, pages 35–42. TUDpress, Dresden.

Burkhardt, F., Eyben, F., and Schuller, B. (2022a). Syntact: A synthesized database of basic emotions. In *Proc. of the Dataset Creation for Lower-Resourced Languages (DCLRL) workshop in conjunction with LREC 2022*.

Burkhardt, F., Wagner, J., Wierstorf, H., Eyben, F., and Schuller, B. (2022b). Nkululeko: A tool for rapid speaker characteristics detection. In *Proceedings of LREC 2022*.

Burkhardt, F. (2000). *Simulation emotionaler Sprechweise mit Sprachsynthesesystemen*. Shaker.

Burkhardt, F. (2005). Emofilt: The simulation of emotional speech by prosody-transformation. In *9th European Conference on Speech Communication and Technology*.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Costantini, G., Iaderola, I., Paoloni, A., and Todisco, M. (2014). Emovo corpus: an italian emotional speech database. In *LREC*.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and der Vreken, O. (1996). The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. ICSLP'96, Philadelphia*, 3:1393–1396.

Engberg, I. S., Hansen, A. V., Andersen, O., and Dalsgaard, P. (1997). Design, recording and verification of a danish emotional speech database. In *EUROSPEECH*.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). opensmile – the munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pages 1459–1462, 01.

Eyben, F., Scherer, K., Schuller, B., Sundberg, J., Andre, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., and Truong, K. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:1–1, 01.

Feraru, S. M., Schuller, D., and Schuller, B. (2015). Cross-language acoustic emotion recognition: An overview and some tendencies. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 125–131.

Gangamohan, P., Kadiri, S., and Yegnanarayana, B. (2016). Analysis of emotional speech – A review. *Toward Robotic Socially Believable Behaving Systems-Volume I*, pages 205–238.

Goudbeek, M., Goldman, J. P., and Scherer, K. R. (2009). Emotion dimensions and formant position. In *Tenth Annual Conference of the International Speech Communication Association*.

Kaya, H., Salah, A., Gurgen, F., and Ekenel, H. (2014). Protocol and baseline for experiments on bogazici university turkish emotional speech corpus. In *2014 22nd Signal Processing and Communications Applications Conference, SIU 2014 - Proceedings*, 04.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391.

Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108.

Neumann, M. and Vu, T. (2018). Cross-lingual and multilingual speech emotion recognition on english and french. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5769–5773, 02.

Powroźnik, P. (2017). Kohonen network as a classifier of polish emotional speech. *ITM Web of Conferences*, 15.

Ringeval, F., Cowie, R., Amiriparian, S., Michaud, A., Schuller, B., Kaya, H., Cummins, N., Çiftçi, E., Valstar, M., Schmitt, M., Lalanne, D., Güleç, H., Salah, A. A., and Pantic, M. (2018). AVEC 2018 Workshop and Challenge: Bipolar disorder and cross-cultural affect recognition. In *AVEC 2018 - Proceedings of the 2018 Audio/Visual Emotion Challenge and Workshop, co-located with MM 2018*.

Scherer, K. R., Banse, R., and Wallbott, H. G. (1999). Emotion Inferences from Vocal Expression Correlate across Languages and Cultures. *ICPhS 99*.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256.

Schröder, M. (2001). Emotional Speech Synthesis - A Review. In *Proc. Eurospeech 2001, Aalborg*, pages 561–564.

Schröder, M. and Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

Schuller, B. and Burkhardt, F. (2010). Learning with synthesized speech for automatic emotion recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5150–5153.

Schuller, D. M. and Schuller, B. W. (2021). A review on five recent and near-future developments in computational processing of emotion in the human voice. *Emotion Review*, 13(1):44–50.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131.

Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., and Evanini, K. (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity and native language. In *Proceedings Interspeech 2016*, pages 2001–2005, 09.

Tamulevičius, G., Korvel, G., Yayak, A. B., Treigys, P., Bernatavičienė, J., and Kostek, B. (2020). A study of cross-linguistic speech emotion recognition based on 2d feature spaces. *Electronics*, 9(10).