INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Nkululeko: Machine Learning Experiments on Speaker Characteristics Without Programming

Felix Burkhardt[1], *Florian Eyben*[1], *Björn Schuller*,[1,2,3]

[1]audEERING GmbH, Germany,
[2]Chair EIHW, University of Augsburg, Germany,
[3]GLAM, Imperial College London, UK

[fburkhardt, fe, bs]@audeering.com

Abstract

We would like to present Nkululeko, a template based system that lets users perform machine learning experiments in the speaker characteristics domain. It is mainly targeted on users not being familiar with machine learning, or computer programming at all, to being used as a teaching tool or a simple entry level tool to the field of artificial intelligence.

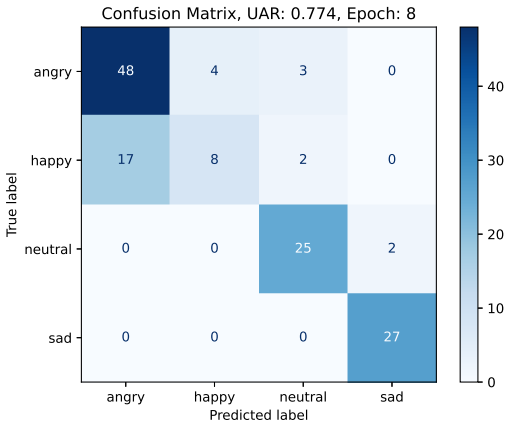

Figure 1: *Confusion matrix is usually the result of a typical Nkululeko experiment*

1. Description

In the past decades, the research community has been confronted with the tremendous success of approaches to estimate knowledge with artificial neural nets (ANN), predominantly under the label *deep learning*. Especially empirical sciences benefit from the opportunity to test hypotheses with machine learning experiments that are able to analyse statistically very large data quantities. Many empirical researchers, phoneticians, and linguists, did not study computer science and struggle with the necessary programming skills to set machine learning experiments up.

Nkululeko[1] is being developed preliminary as a tool for a series of machine learning seminars at the institute for speech communication at the Technical University of Berlin to enable students to conduct machine learning experiments with a very flat learning curve by simply filling configuration files. This makes it very easy to be used, compared to other high level frameworks for deep learning like Keras, Torch, Google AutoML or end2you [1, 2, 3, 4] while still keeping the flexibility as it is based on Torch.

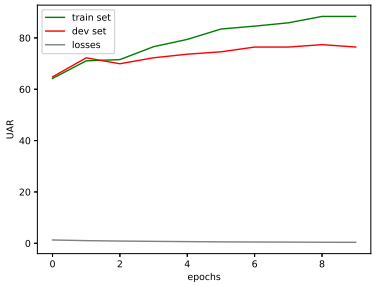

Figure 2: *With deep learning experiments, it makes sense to inspect the epoch progression.*

Nkululeko is open source software written in Python and hosted on github[2]. The data management is based on audformat[3], but a simpler CSV (comma separated values) format is also supported. All of the features can be seen in the description file for the templates[4], but there's also a blog featuring tutorials around Nkululeko[5]. The framework has first been described in [5]

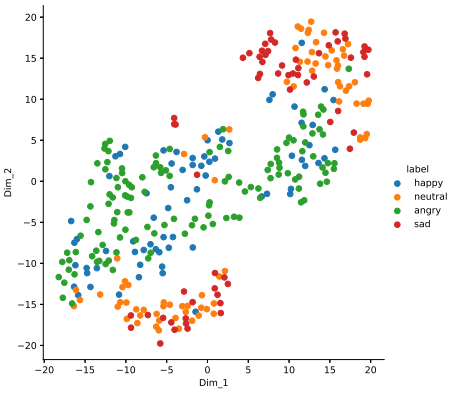

Figure 3: *A t-sne plot sometimes reveals problems in the data.*

[1]On the lookout for a distinctive name for this project we stumbled across an 1980ies punk album title. They tried new things out fastly, so this seemed fitting.

[2]https://github.com/felixbur/nkululeko/
[3]https://github.com/audeering/audformat
[4]https://github.com/felixbur/nkululeko/blob/main/ini_file.md
[5]http://blog.syntheticspeech.de/2022/12/01/nkululeko/

Currently, the following interfaces are implemented:

- **nkululeko**: doing experiments
- **demo**: demo the current best model on commandline
- **test**: predict a series of files with the current best model
- **explore**: perform data exploration
- **augment**: augment the current training data

Figure 4: *The distribution of features across specific categories can be very interesting.*

The main functionality is to combine a set of acoustic features with machine learning classifiers and regressors. For machine learners, Nkululeko mainly relies on python packages such as scikit-learn or pytorch. Figure 1 shows a confusion matrix and the evaluation per epoch for a typical Nkululeko experiment: investigating the performance of ANN embeddings in a cross database experiment for acted basic emotional vocal expressions. In Figure 2, a typical outcome of an experiment involving artificial neural nets is shown: the progression of performance per epoch for training and evaluation set. Points of over fitting can be detected.

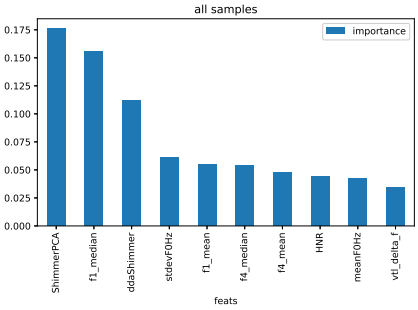

Figure 5: *The distribution of features across specific categories can be very interesting.*

There are several interfaces that can be used with Nkululeko, one of the is the explore module. Here one can, for example, specify to evaluate some given features by a t-SNE plot [6] (Figure 3 for example shows the t-SNE plot for the Berlin Emodb [7] and opensmile features [8]). The distribution of specific features per category can be visualized like the feature importance according to some model (see Figures 4 and 5). Lastly, category distribution in data can be displayed (Figure 6).

2. Conclusions

We would like to present Nkululeko – a free new open-source tool to set up machine learning experiments in the speech research domain that can be used without programming skills. Future works will include extension of its functionality.

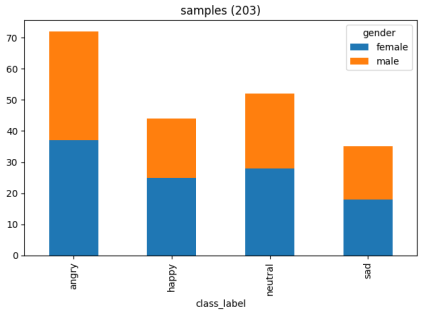

Figure 6: *The distribution of categories in the data can be shown.*

3. Acknowledgements

This research has been partly funded by the European EASIER (Intelligent Automatic Sign Language Translation) project (Grant Agreement number: 101016982), the European SHIFT (MetamorphoSis of cultural Heritage Into augmented hypermedia assets For enhanced accessibiliTy and inclusion) project (Grant Agreement number: 101060660) as well as the European MARVEL (Multimodal Extreme Scale Data Analytics for Smart Cities Environments) project (Grant Agreement ID: 975337).

4. References

[1] F. Chollet, *Deep Learning with Python & Keras*. Manning Publications, 2017, vol. 80.

[2] A. Chaudhary, K. S. Chouhan, J. Gajrani, and B. Sharma, *Deep learning with pytorch*, 2020. DOI: 10.4018/978-1-7998-3095-5.ch003.

[3] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2you–the imperial toolkit for multimodal profiling by end-to-end learning," *arXiv preprint arXiv:1802.01115*, 2018.

[4] E. Bisong, "Google automl: Cloud vision," in Sep. 2019, pp. 581–598, ISBN: 978-1-4842-4469-2. DOI: 10.1007/978-1-4842-4470-8_42.

[5] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller, "Nkululeko: A tool for rapid speaker characteristics detection," in *Proceedings of LREC 2022*, 2022.

[6] G. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in Neural Information Processing Systems*, vol. 15, Jun. 2003.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *9th European Conference on Speech Communication and Technology*, vol. 5, Sep. 2005, pp. 1517–1520. DOI: 10.21437/Interspeech.2005-446.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile – the munich versatile and fast open-source audio feature extractor," *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pp. 1459–1462, Jan. 2010. DOI: 10.1145/1873951.1874246.