Age Classification:
Comparison of Human and Machine Performance Using Different Utterance Types

Felix Burkhardt, Markus Brückl, Björn W. Schuller

Abstract

We report on the results of an investigation to

- classify speaker age in vocal utterances
- with state-of-the-art machine learning algorithms
- on a small data set.

We compare results

- of manual measurement, i. e., supervised automated extraction of phonetically interpretable measures and observation
- with the outcomes of experiments based on recent machine learning.

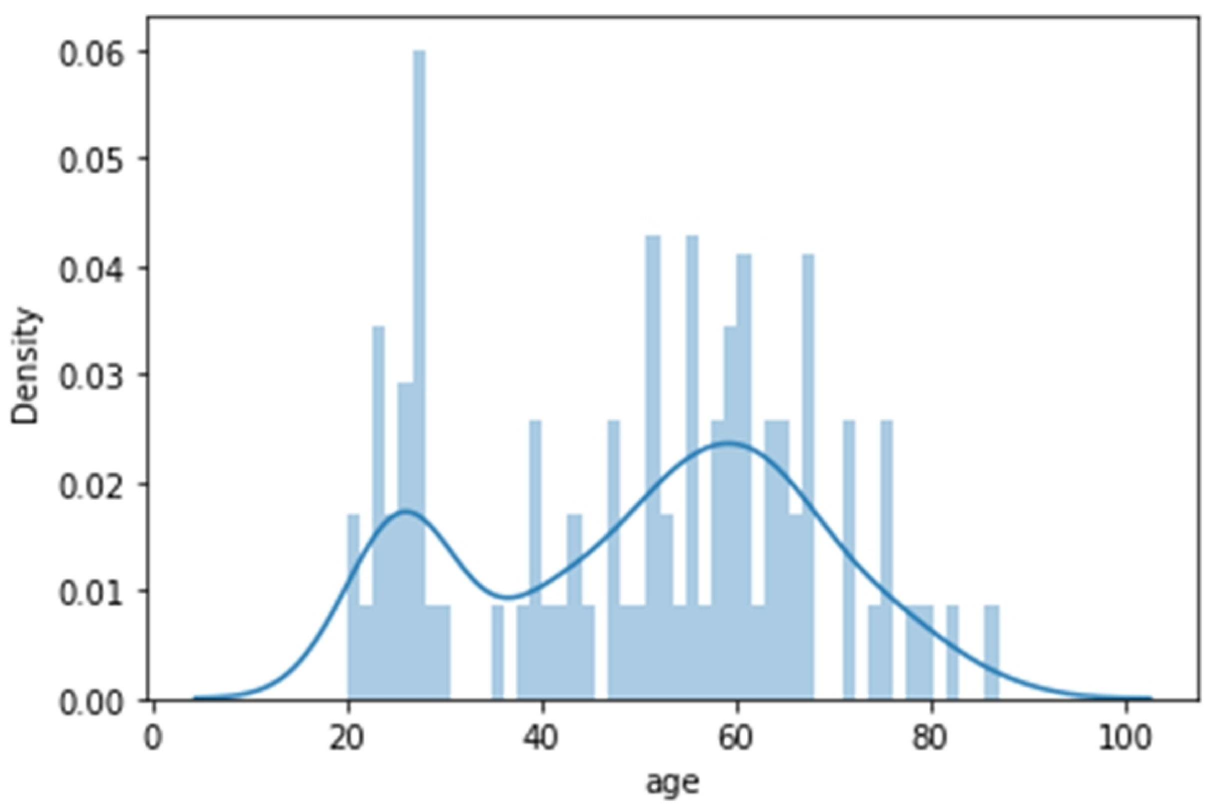
On isolated vowels the machine outperformed the human estimates.

Introduction

- Age can be seen as a paralinguistic speaker trait
- In contrast to emotion or personality it can be measured exactly
- There is not only the biological but also the perceptive age

The Database

DFG-project "Young and old voices", cf. [1]



Additional Databases

Deutsche Telekom Agender

- Telephone collected 8 kHz data
- Selected 1k female speakers per decade

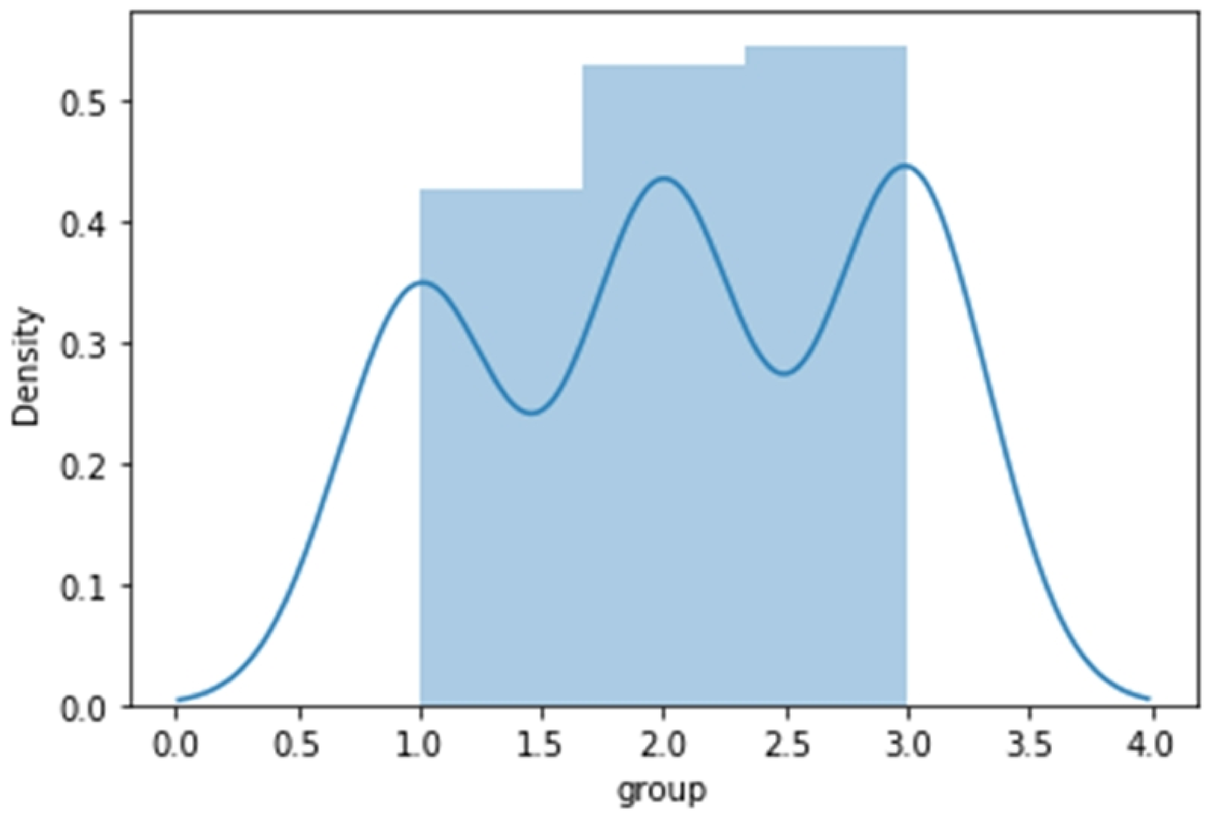
Mozilla common voice corpus

- Over the web donated speech samples
- Age stated in decades: 20-60 years old
- Selected randomly 2k samples per decade from female speaker

Age groups

Binned age into two groups

- a seven classes group representing the decades from twenties to eighties
  - performed oversampling done with the SMOTE (syntheticminority over-sampling technique) algorithm which adds samples by synthesizing them on a feature level based on distance to central class representatives.
- a three classes age group:
  - young(from zero to 40 years),
  - middle aged (from 40 to 60 years) and
  - elderly (above 60 years).



Classifiers

- Support Vector Machines
- XGBoost
- Multi Layer Perceptron, 2 hidden layers with 128 and 16 neutrons
- Convolutional Neural Network, pre-trained on speaker ID with Mel spectrograms as input

Features

- GeMAPS – 88 standard features set with OpenSmile, cf. [2]
- ComPARE 2016 feature set (6373)
- Compare top 512 features based on XGB classifier
- Trill features: embeddings from Google trained on several datasets for speaker, language, emotion and health classification
- Mel Spectrograms for the Conv Net

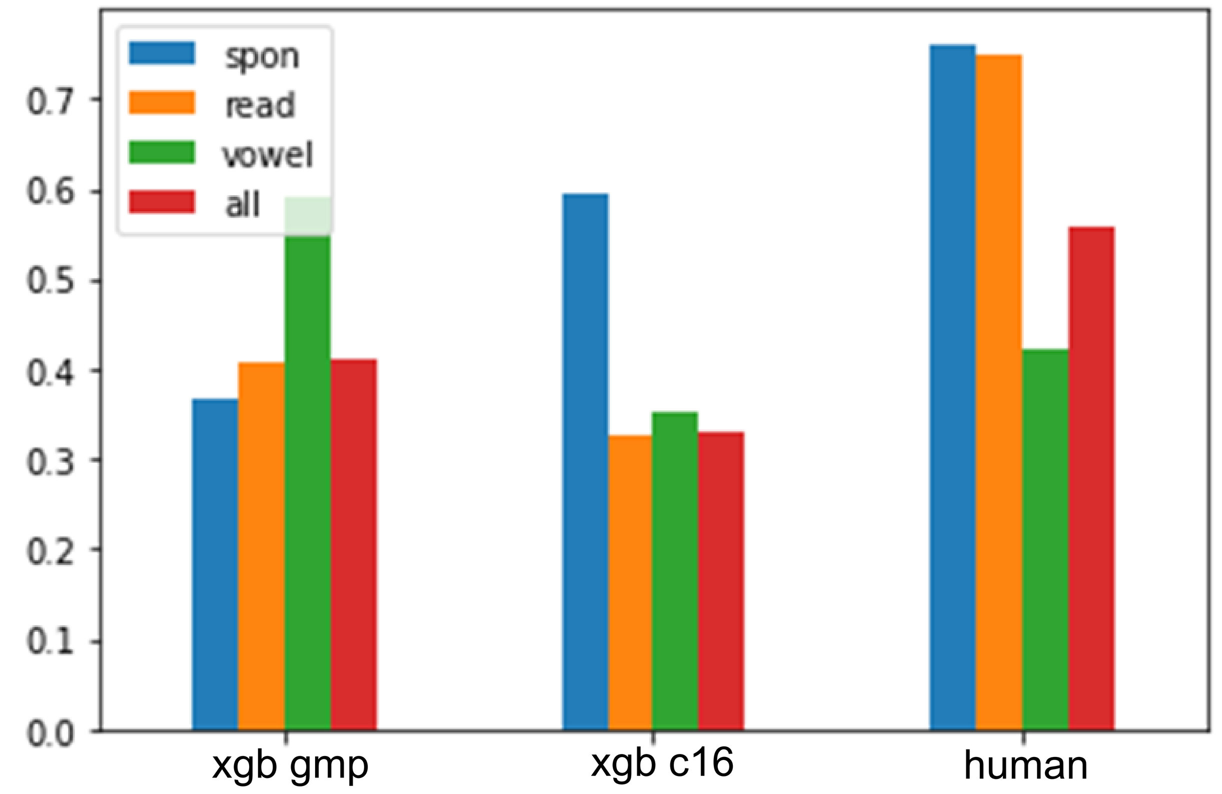
Results: Text Material

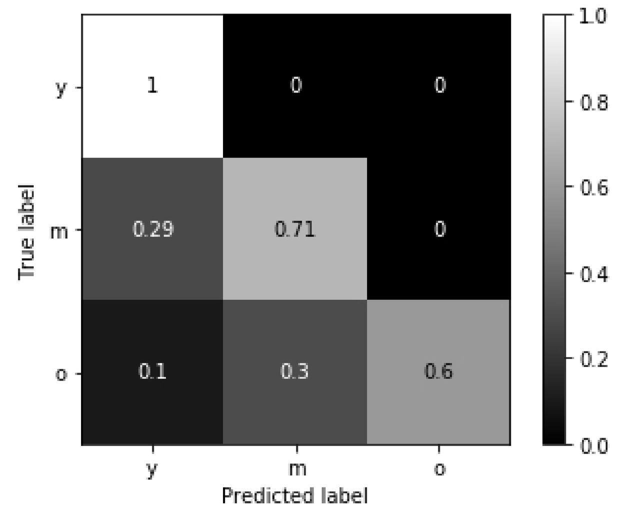
Comparing

- human performance
- on different text types
- with SVM and XGBoost classifiers
- for GeMAPS and Compare14 Features sets

SVM classifier did not converge (not enough data?)
Also for ANNs not enough data

Humans performed generally clearly better
XGBoost performs reasonably well on isolated vowels

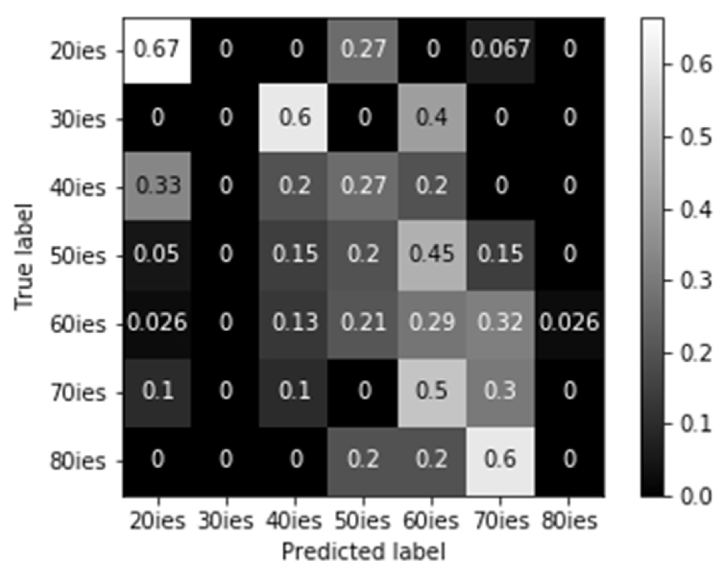




Confusion matrix for XGBoost with GeMAPS features for isolated vowels

Results for seven age classes

		feature set		
		top	all	trill
stat. classifier	SVM	.219	.210	.113
	XGB	.142	.222	.156
art. neural net	MLP mix	.148		.165
	MLP reg	.169		.173
	MLP class	.158		.172
	MLP+D1	.177		.255
	MLP+D2	.152		.171
	MLP+D1+D2	.161		**.237**
	MLP D1	.161		.194
	MLP D1	.200		.137
	MLP D1 and D2	.217		.217
	CNN		,233	
manual reg.	MLRP		.218	
Hum. group	HLP		**.299**	



Confusion Matrix based on Best CNN result (UAR .236)

Results: 10 best features

- 10 best performing features based on XGBoost classifier
- All of the most important features correspond to loudness in spectral bands
- The features don't correspond linearly to the age groups
- Does not match directly with best performing manual feature (vocal tremor)

Conclusion

We investigated the machine classification of speaker age on a small database.
With respect to our hypotheses, we could support only one of them:

- the machine performance is comparable to the human one
- but the most important features of the manual investigation do not correspond with the machine classifier.
- The lack of super performance is explainable by little data from similar domains and one should revisit this experiment with a more general age model as a background.
- On isolated vowels the machine outperformed the human estimates.

Acknowlegements

We thank the SEMULIN research project by the German BMWi (FK: 19A20012B) for partial funding of this work

References

[1] Brückl, M.: *Altersbedingte Veränderungen der Stimme und Sprechweise von Frauen*, W. Sendlmeier [Ed], Mündliche Kommunikation, Vol. 7, Logos Verlag, Berlin, 2011.

[2] Eyben, F., M. Wöllmer, and B. Schuller: openSMILE — the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462. 2010.