Towards Robust Voice Pathology Detection

Investigation of supervised deep learning, gradient boosting and anomaly detection approaches across four databases

Pavol Harar[1], Zoltan Galaz[1], Jesus B. Alonso-Hernandez[2], Jiri Mekyska[1], Radim Burget[1] and Zdenek Smekal[1]

[1]Dept. of Telecommunications, Brno University of Technology, [2]IDeTIC, University of Las Palmas de Gran Canaria

Objectives

In search towards a robust voice pathology detection system (VPD), we investigated three distinct classfiers within supervised learning and anomaly detection paradigms on data from four different databases with the aim to:

1. investigate whether it is possible to build a robust VPD system using currently available resources and mentioned classifiers
2. conduct a robust cross database experiment to eliminate possible overfitting of the classifiers on the recording conditions
3. measure the VPD performance using the recordings of all avalilable types of vocal pathologies unrestricted to just a subset
4. uncover the limitations of currently available resources with respect to this task

Introduction

Automatic objective non-invasive vocal pathology detection system can play an important role in early diagnosis, progression tracking and even effective treatment of pathological voices and even neurological and other disorders. In research it has been approached by subjective (perceptual examination) and objective evaluations (computer analysis of acoustic signals). The later is **faster, cheaper and free from the subjective bias**, but its performance depends on the available data and classifier capabilites and is often **harder to clinically interpret**. For this task, researchers frequently used sustained vowel phonation recordings from MEEI, SVD or AVPD databases. Due to the complexity of the task, authors restricted the experiments to just a subset of vocal pathologies (~3) from 1 to 3 databases. The results vary greatly between the published papers mainly due to differences in selected voice pathology samples, acoustic features, and classifiers that were used for the experiment.

Data

Table 1: Contents of the databases.

	AVPD	MEEI	PDA	SVD
H samples	188	53	239	687
P samples	178	657	200	1356
vowels	/a, e, o/	/a/	/a/	/a, i, u/
running speech	yes	yes	no	yes
EGG	no	no	no	yes
GRBAS	yes	no	no	no

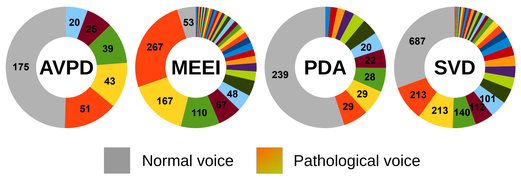

Figure 1: Inequality in no. of samples per pathology type

Data processing:

- recordings of **sustained vowel /a/**
- from speakers between 19 and 60
- shorter than 0.750 s excluded
- split into **0.750 s long chunks**
- $f_s = 16\,\text{kHz}$

Table 2: Number of chunks used in the experiments.

Database	H (M)	P (M)	H (F)	P (F)	Total
AVPD	625	509	872	804	2810
MEEI	126	114	185	168	593
PDA	1158	331	5	605	2099
SVD	400	645	624	871	2540
Total	2309	1599	1686	2448	**8042**

H – healthy, P – pathol., M – males, F – females.

We tested multiple input representations :

- raw waveforms
- spectral features
- cepstral features (MFCC)
- conventional speech features

Methodology

We utilized 3 different classifiers:

1. Isolation Forest (IF) - Anomaly detection
2. DenseNet (DNN) - Deep Neural Network
3. XGBoost (XGB) - Gradient Boosted Trees

To train and validate the models, we split the data into **training** (TR), **validation** (VA) and **testing** (TE) sets. Stratification of TR & VA was done by medical state, gender, age and gender-age group. Remaining unequalities were addresed by sample weights. In case of XGB & IF, 10-fold CV was used.

Results

Table 3: Testing confusion matrix for XGBoost

	true H	true P	total predicted
predicted H	**82**	26	108
predicted P	38	**94**	132
total true	120	120	accuracy: **0.733**

Table 4: Testing classification report for XGBoost

	precision	recall	f1-score	no. samples
class H	0.759	0.683	0.719	120
class P	0.712	0.783	0.746	120
avg. / total	0.736	0.733	**0.733**	240

Table 5: Testing classification report for DenseNet (MFCC)

	precision	recall	f1-score	no. samples
class H	0.624	0.608	0.616	120
class P	0.618	0.633	0.626	120
avg. / total	0.621	0.621	**0.621**	240

Table 6: Testing classification report for Isolation Forest

	precision	recall	f1-score	no. samples
class H	0.659	0.483	0.558	120
class P	0.592	0.750	0.662	120
avg. / total	0.626	0.617	**0.610**	240

Conclusions

Major limitations with respect to VPD:

- limited number of available databases
- different speech tasks in each database
- lack of clinically rated data
- missing well defined baseline for comparisons

Observations:

- XGBoost classifier achieved the best results and can perform its own feature selection
- Isolation Forest classifier showed to be sensitive on feature selection
- it was necessary to use inputs with reduced dimensionality instead of raw waveforms
- too few samples for Deep Neural Net
- MFCC alone are not reliable enough for VPD

Acknowledgements

This study was funded by the grant of the Czech Ministry of Health 16-30805A (Effects of non-invasive brain stimulation on hypokinetic dysarthria, micrographia, and brain plasticity in patients with Parkinson's disease) and project LO1401. For the research, infrastructure of the SIX Center was used. The authors (P. Harar, Z. Galaz) of this study also acknowledge the financial support of Erwin Schrödinger International Institute for Mathematics and Physics during their stay at the "Systematic approaches to deep learning methods for audio" workshop held from September 11, 2017 to September 15, 2017 in Vienna.

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION — department of telecommunications

Contact Information

Pavol Harar (pavol.harar@vut.cz)
Brno University of Technology
Technicka 3082/12, 61 600 Brno, Czech Republic