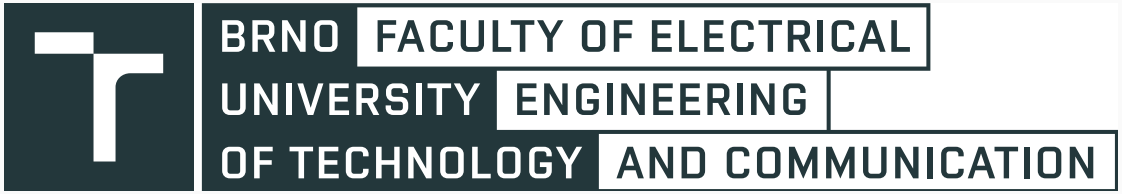


Voice Pathology Detection Using Deep Learning: a Preliminary Study

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

[1] Brno University of Technology [2] University of Las Palmas de Gran Canaria

Outline

1. Introduction
2. Methodology
3. Results
4. Conclusions

Introduction

Motivation

Lots of voice pathologies are not being diagnosed and treated due to a lack of training, time or appropriate equipment of general practicioners.

We are aiming for early diagnosis:

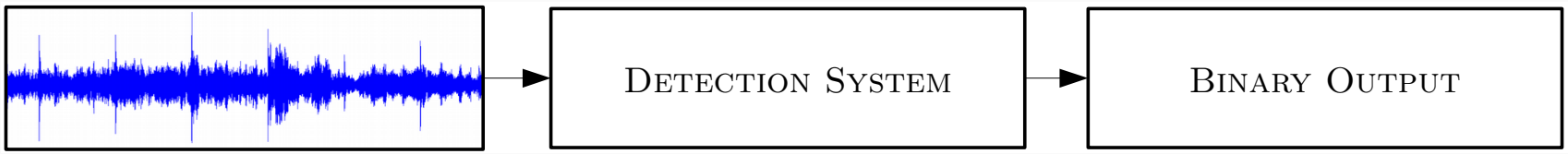
- using quick preventive tests
- without extensive training
- without the need for expensive equipment

Which would point the patient to a specialized voice therapist.

Motivation

Automatically distinguish between normal and pathological voice

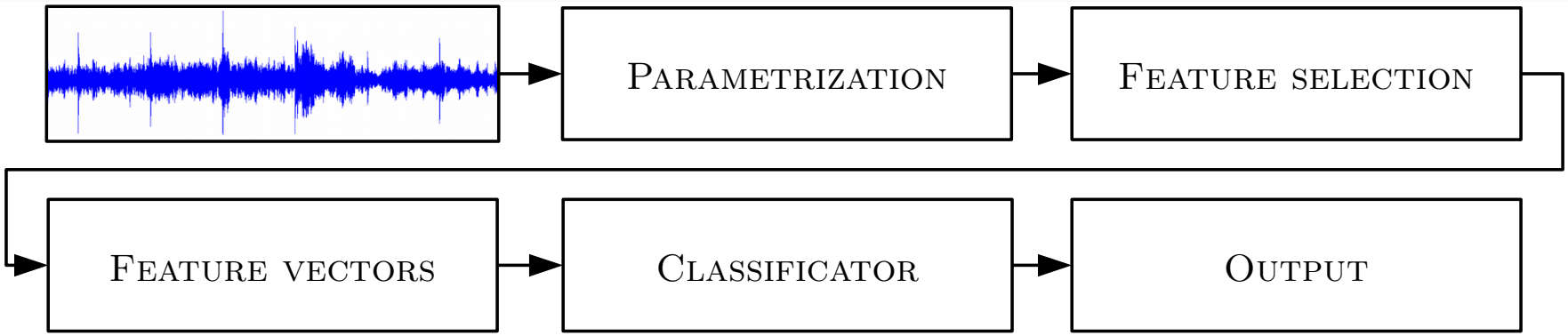
- end to end system
- raw audio signal input
- no parametrization



Related work

All authors relied upon feature extraction

- features from time, frequency and cepstral domains
- mostly utilized MFCC (mel-frequency cepstral coeficients)
- hard to compare experiments (different subsets of data)
- reported from $\approx 72\,\%$ to $\approx 99\,\%$ accuracy



Related work

Table 1: Overview of related works *

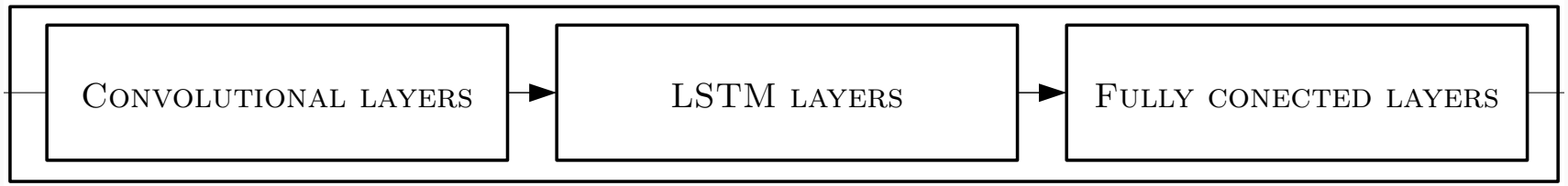
Article	Employed classifier	Accuracy	Notes
1.	KM, RF	100.00 %	Used combination of vowels /a/, /i/, /u/ F and M separately
2.	GMM	99.98 %	Used combination of voice and EGG signals
3.	SVM	99.68 %	Used subset containing 4 of 71 pathologies
4.	GMM	99.00 %	Used combination of vowels /a/, /i/, /u/
5.	SVM, ELM, GMM	95.00 %	Used mix of MEEI and SVD data
6.	SVM	93.20 %	Used subset containing 3 of 71 pathologies
7.	SVM	90.98 %	Used subset containing 4 of 71 pathologies
8.	ANN	87.82 %	Used subset containing 4 of 71 pathologies
9.	SVM	86.44 %	Used subset containing 4 of 71 pathologies
10.	GMM	79.40 %	Used combination of vowels /a/, /i/, /u/
11.	GMM, SVM	72.00 %	Used 200 samples of vowel /a/ at high pitch

* complete table with references to be found in the full article

Aim

Test the suitability and performance of Deep Neural Network with

- convolutional layers (for feature learning)
- recurrent LSTM layers (for understanding the sequence)
- fully connected layers (for final classification)



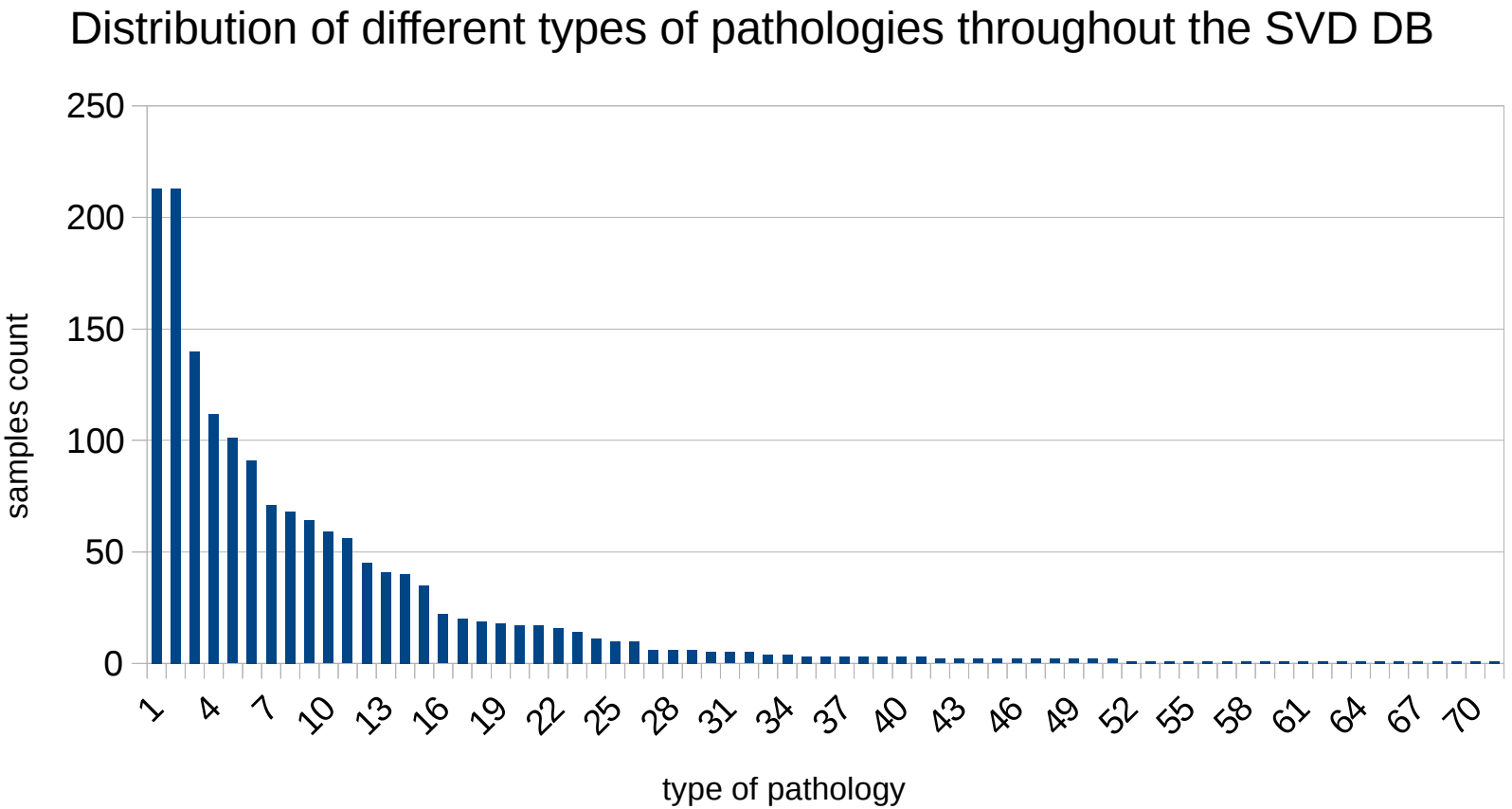
Methodology

Data

Saarbruecken Voice Database

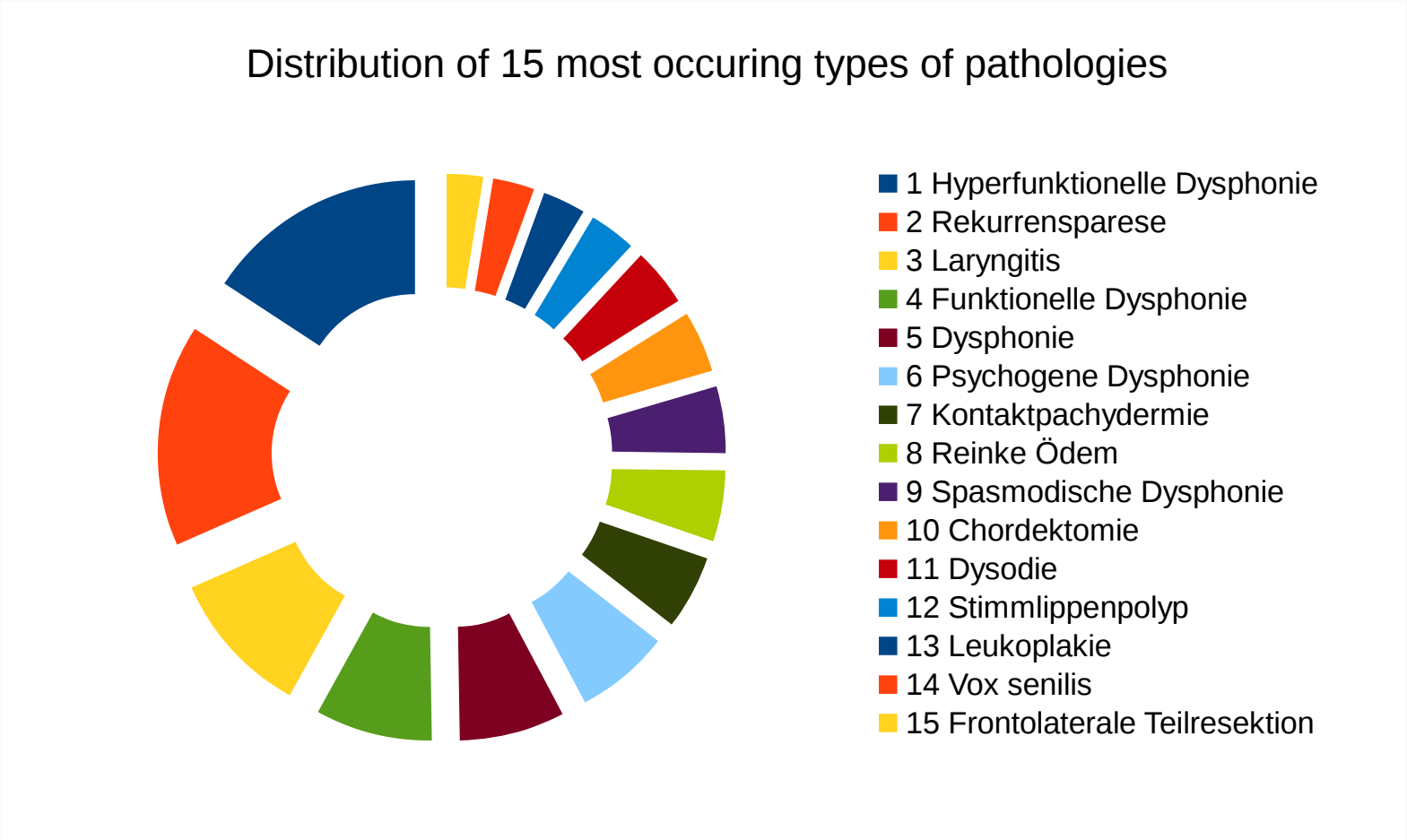
- sustained vowel /a/ at normal pitch
- more than 2000 speakers
- 71 different pathologies
- each sample approx. 1s long
- no onset or offset information

Data



- avg. no. of samples per pathology = 22.89
- 56 pathologies have less samples than avg.

Data



- these 15 pathologies contain 83 % of data

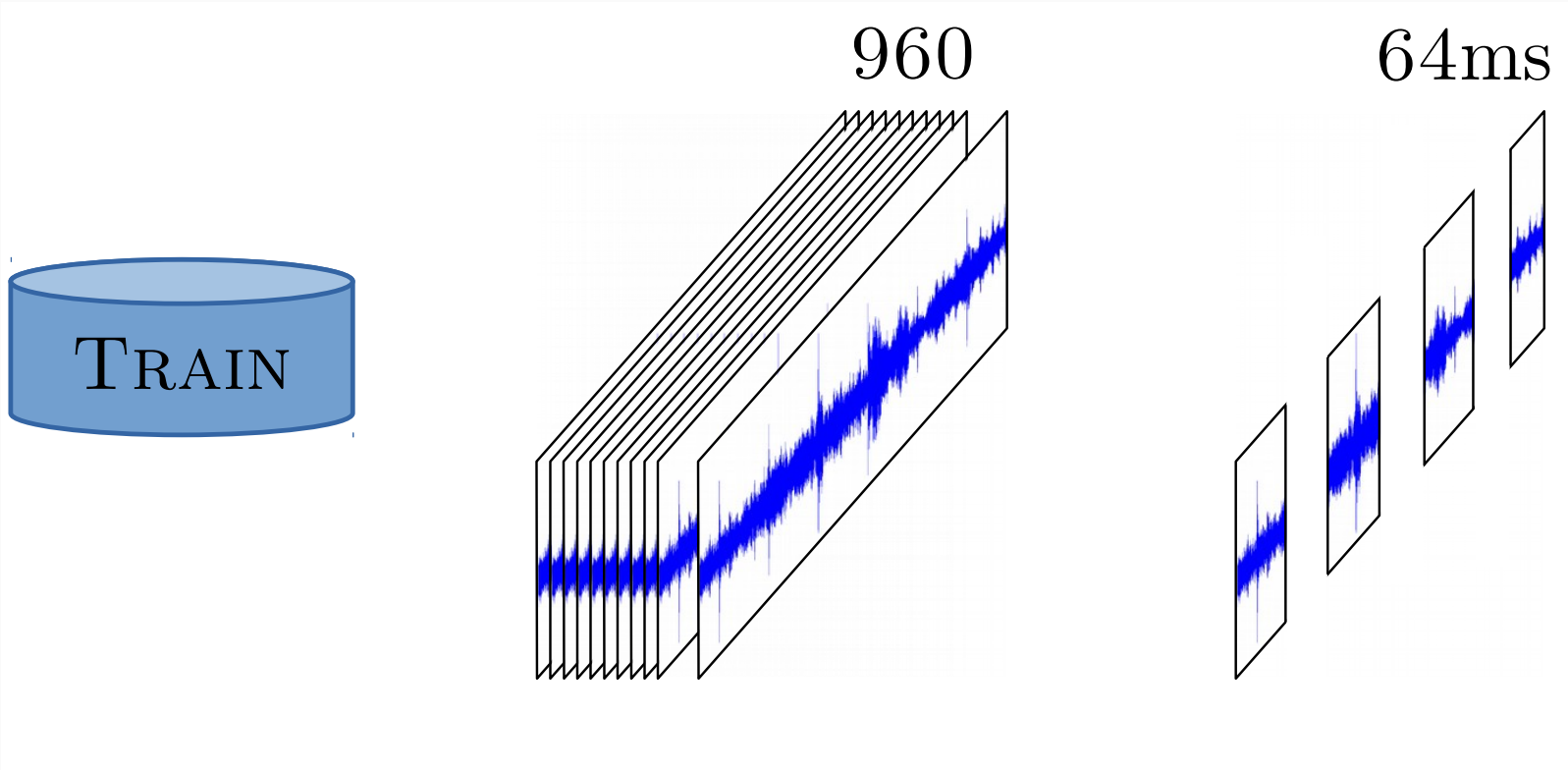
Data

Splitting into subsets:

- 70 % training (480 healthy, 480 pathological)
- 15 % validation (103 healthy, 103 pathological)
- 15 % testing (104 healthy, 770 pathological)*

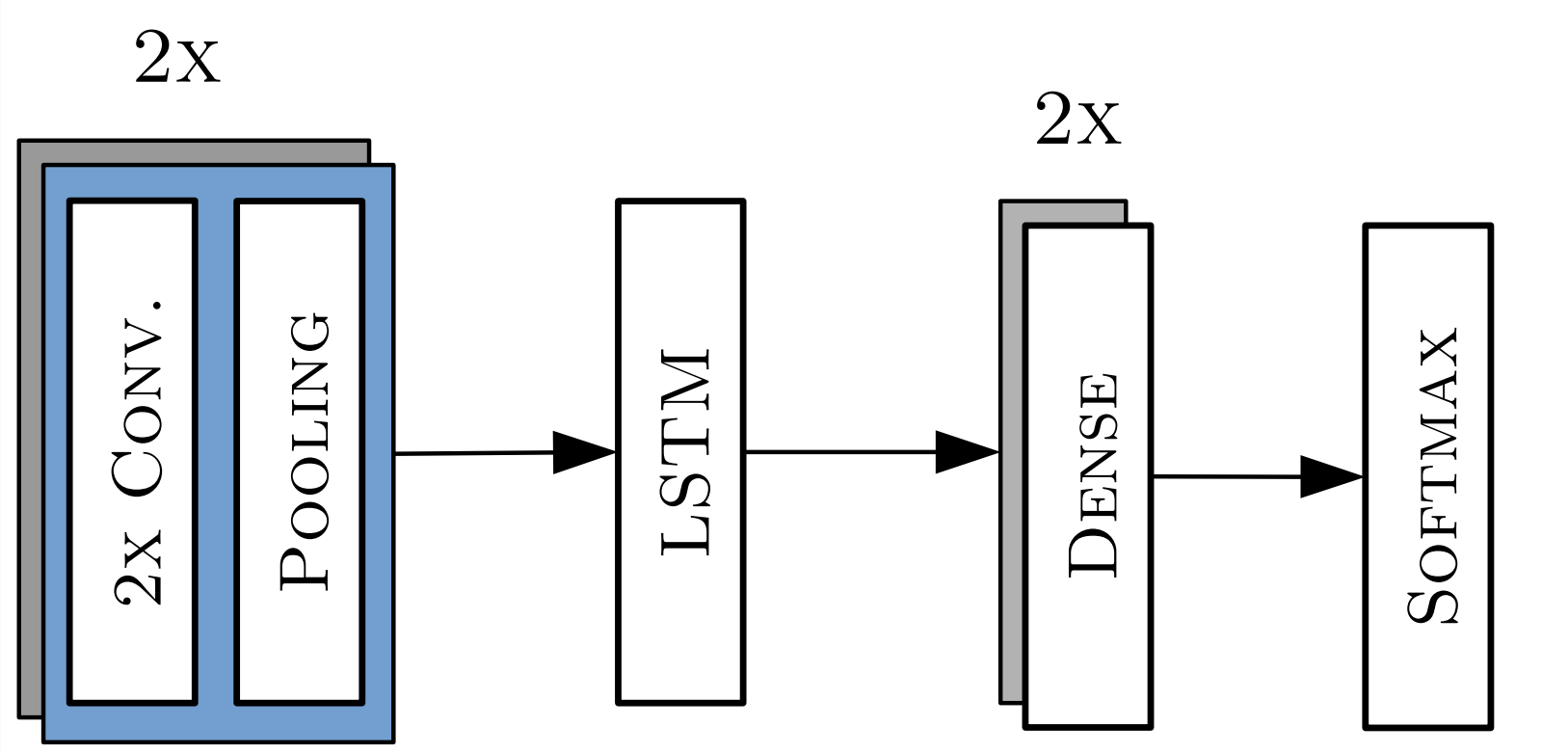
* percentages computed based on the less occurring class healthy

Data

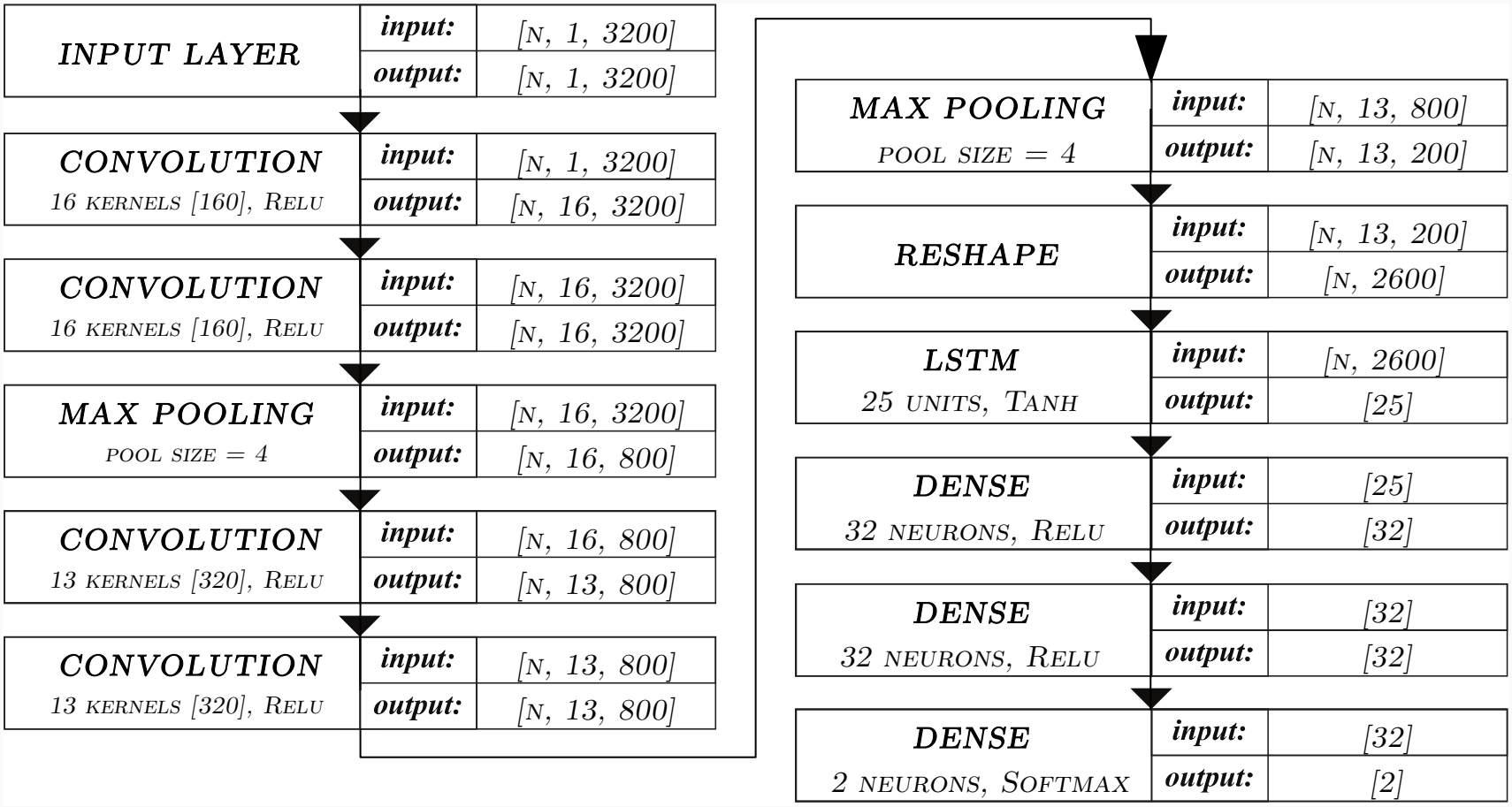


- 64 ms segment with 50 kHz sampling rate $=$ 3 200 features

DNN Architecture



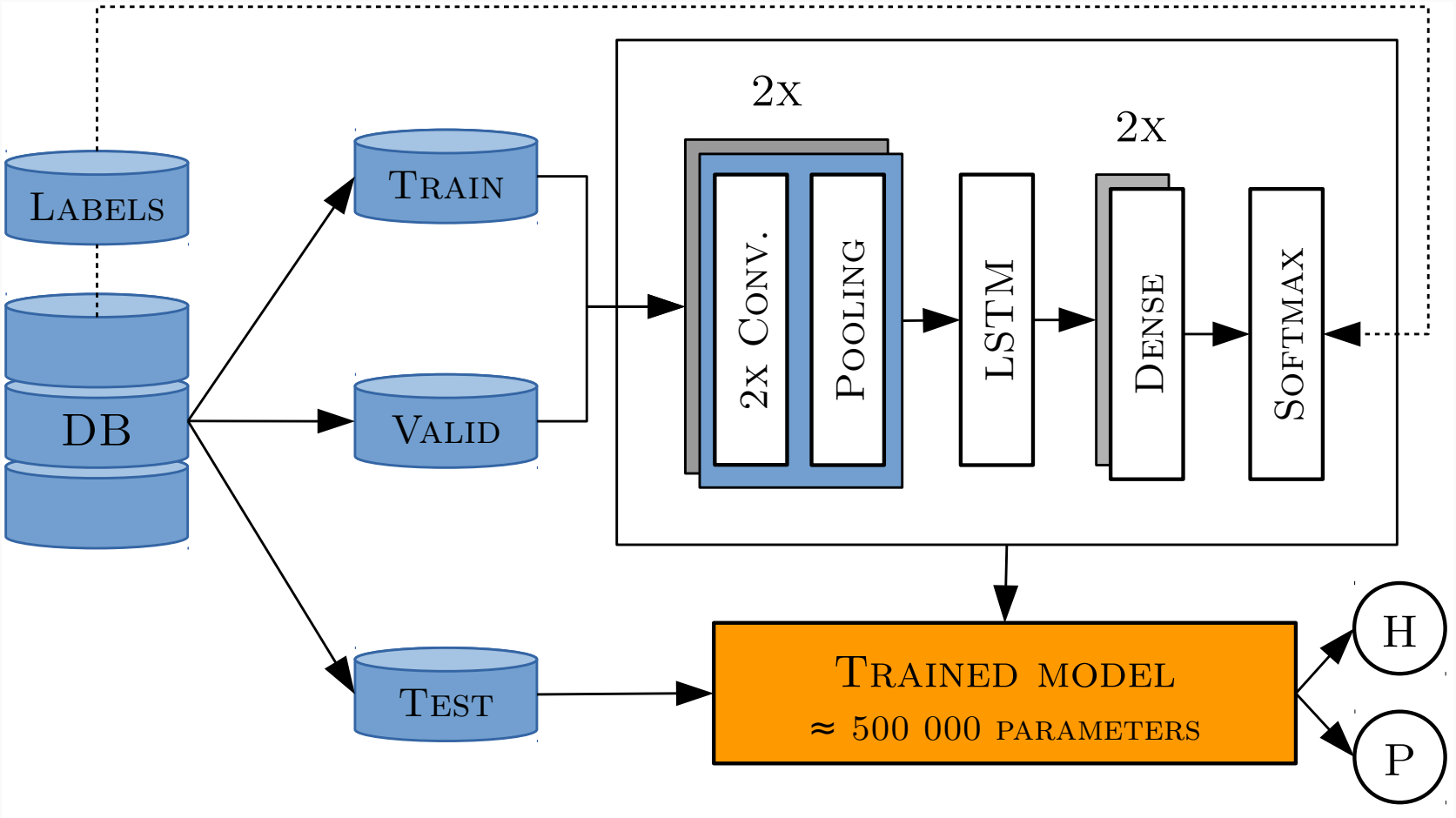
DNN Architecture - detailed



Hyper-parameters

- Relu activation for all CONV and DENSE layers
- Tanh activation for LSTM
- Softmax activation for output layer
- Glorot uniform initialization for all layers
- Categorical cross-entropy loss function
- Adam optimization algorithm with decreasing learning rate
- Batch size $= 1$ due to various sample lengths

Proposed system



Software

In order to built and train the DNN on GPU we used:

- Ubuntu 14.04 LTS x64 operation system
- Python 3.4 programming language
- Keras framework (for building the DNN)
- KEX library (for hyper-parameter search)

Get the source of the KEX library at

`https://gitlab.com/paloha/kex`

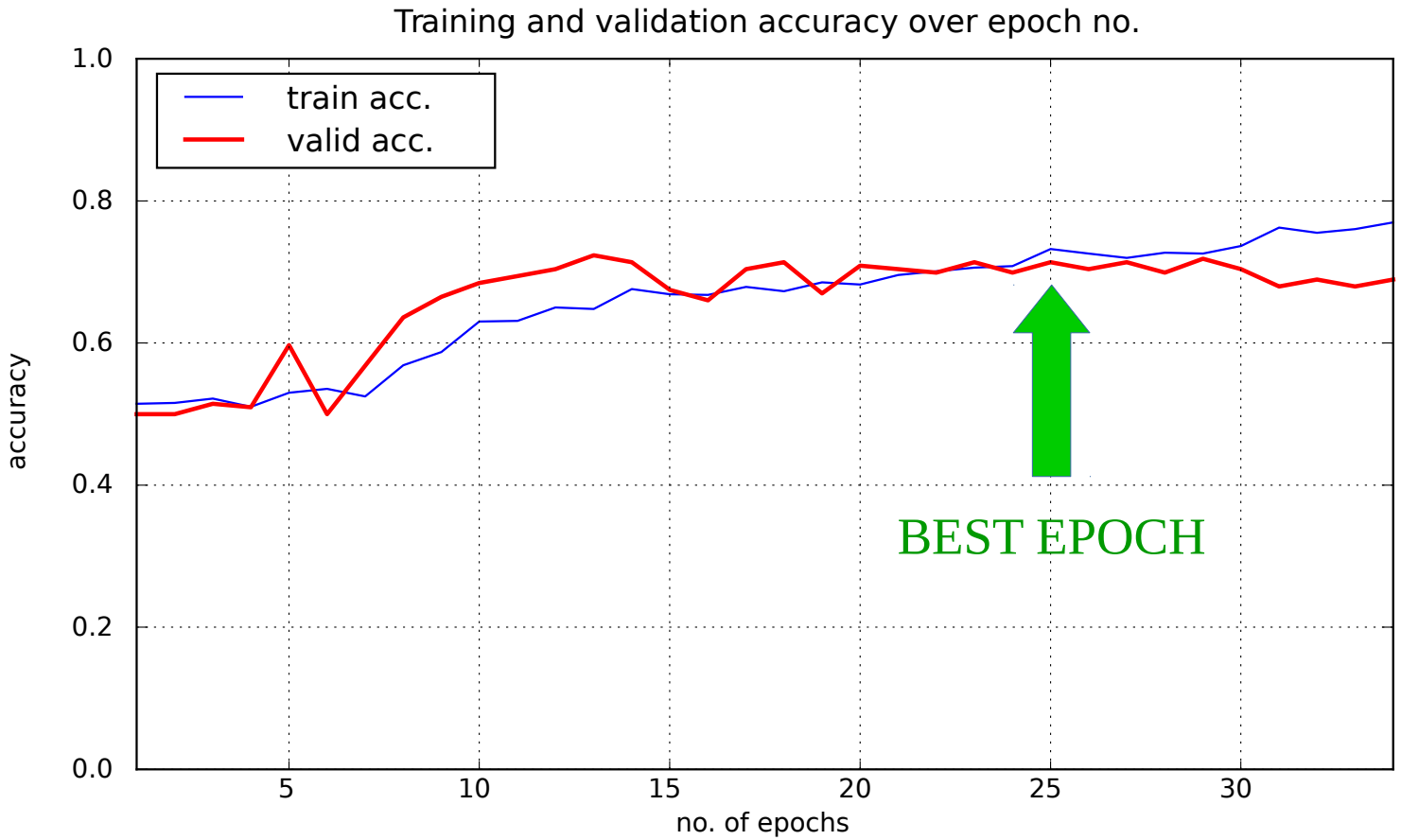
Hardware

In order to built and train the DNN on GPU we used:

- Intel Core i7-3770 CPU @ 3.40 GHz x 8 cores
- with 32 GB RAM
- nVidia GeForce GTX 690

IMG FROM: http://www.nvidia.in/content/product-detail-pages/geforce-gtx-690/geforce-gtx-690-front.png

Training history



- training of 25 epochs took 101 minutes
- hyper-parameters were tuned based on validation results

Results

Validation results

Table 2: VALIDATION confusion matrix

	true: pathological	true: healthy	no. of segments
pred: pathological	67	36	103
pred: healthy	23	80	103

Table 3: VALIDATION classification report

class	precision	f1-score	recall
pathological	0.74	0.69	0.65
healthy	0.69	0.73	0.78
overall accuracy:			71.36 %

Testing results

Table 4: TESTING confusion matrix

	true: pathological	true: healthy	no. of segments
pred: pathological	514	256	770
pred: healthy	23	81	104

Table 5: TESTING classification report

class	precision	f1-score	recall
pathological	0.96	0.79	0.67
healthy	0.24	0.37	0.78
overall accuracy:			68.08 %

Conclusions

Conclusions

Facts in favor of this approach

- using just recordings of sustained vowel /a/ on 71 pathologies
- $\approx 68\,\%$ accuracy is comparable result to similar experiment
- fully automatic without the need for an expert

Disadvantages

- lots of noise in used data
- needs a lot of data for training
- harder interpretation of the results

Conclusions

Possible improvements

- use only data of pathologies with enough samples
- choose pathologies which symptoms are present in used phonation
- combine multiple datasets to get more data
- introduce data warping
- create models for females and males separately
- create models based on age intervals for kids, adults, seniors

Conclusions

Future work

- extending this paper with mentioned improvements
- preparing a dataset for benchmark in voice pathology assessment
- system for classification of voice pathologies

Full paper

Citation:

HARAR, Pavol, et al. Voice Pathology Detection Using Deep Learning: a Preliminary Study. In: Bioinspired Intelligence (IWOBI), 2017 International Conference and Workshop on. IEEE, 2017. p. 1-4.

Available from: http://ieeexplore.ieee.org/abstract/document/7985525/

Thank you for your attention

-

pavol.harar@vut.cz

pavol.harar.eu

Questions?