Towards Robust Voice Pathology Detection

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

Pavol Harar · Zoltan Galaz · Jesus B. Alonso-Hernandez · Jiri Mekyska · Radim Burget · Zdenek Smekal

Abstract Automatic objective non-invasive detection of pathological voice based on computerized analysis of acoustic signals can play an important role in early diagnosis, progression tracking and even effective treatment of pathological voices. In search towards such a robust voice pathology detection system we investigated 3 distinct classifiers within supervised learning and anomaly detection paradigms. We conducted a set of experiments using a variety of input data such as raw waveforms, spectrograms, mel-frequency cepstral coefficients (MFCC) and conventional acoustic (dysphonic) features (AF). In comparison with previously published works, this article is the first to utilize combination of 4 different databases comprising normophonic and pathological recordings of sustained phonation of the vowel /a/ unrestricted to a subset of vocal pathologies. Furthermore, to our best knowledge, this article is the first to explore gradient boosted trees and deep learning for this application. The following best classification performances measured by F1 score on dedicated test set were achieved: XGBoost (0.733) using AF and MFCC, DenseNet (0.621) using MFCC, and Isolation Forest (0.610) using AF. Even though these results are of exploratory character, conducted experiments do show promising potential of gradient boosting and deep learning methods to robustly detect voice pathologies.

Keywords Voice pathology detection · deep learning · gradient boosting · anomaly detection

P. Harar
Brno University of Technology
Technicka 3082/12
61600, Brno, Czech Republic
E-mail: pavol.harar@vut.cz

1 Introduction

Voice pathology can be caused by the presence of tissue infection, systemic changes, mechanical stress, surface irritation, tissue changes, neurological and muscular changes, and other factors [59]. Due to vocal pathology, the mobility, functionality and shape of the vocal folds are affected resulting into irregular vibrations and increased acoustic noise. Such a voice sounds strained, harsh, weak, and breathy [58,27], which significantly contributes to the overall poor voice quality [10,39].

Up to this day, vocal pathology detection has been approached by subjective and objective evaluations [37]. The first category (subjective evaluation) consists of so called in-hospital auditory-perceptual and visual examination of the vocal folds [46,54]. For the visual examination laryngostroboscopy is commonly used [61]. For the auditory-perceptual examination several clinical rating scales to diagnose and rate severity of vocal pathologies have been developed [15,19,32,15,16]. Methods for subjective evaluation are, however, subject to inter-rater variability [9,21]. Furthermore, they require patients' presence at the clinic, which can be a serious problem especially in more severe stages of a disease. This type of evaluation is also time costly, and it requires careful evaluation and scoring by clinicians.

The second category (objective evaluation) is based on the automatic non-invasive computerized analysis of acoustic signals to quantify and identify the underlying vocal pathology that may not even be audible to a human being [39]. This type of evaluation is therefore inherently free from the subjective bias. Moreover, voice can be nowadays easily recorded using a variety of smart devices, and processed remotely using cloud technologies. From these reasons, works such as [17,

Table 1: Overview of related works focused on voice pathology detection.

First author	Year	Ref.	Database	Input vowels	Classifier	Accuracy [%]
Hemmerling	2016	26	SVD	/a, i, u/	KM, RF	100.00
Muhammad	2017	43	SVD	/a/	GMM	99.98
Al-nasheri	2017	2	MEEI, SVD, AVPD	/a/	SVM	99.81 (MEEI), 91.17 (AVPD), 90.98 (SVD)
Al-nasheri	2017	3	MEEI, SVD, AVPD	/a/	SVM	99.79 (AVPD), 99.69 (MEEI), 92.79 (SVD)
Al-nasheri	2017	4	MEEI, SVD, AVPD	/a/	SVM	99.68 (SVD), 88.21 (MEEI), 72.53 (AVPD)
Eskidere	2015	17	SVD	/a, i, u/	GMM	99.00
Amami	2017	7	MEEI	/a/	SVM	98.00
Sabir	2017	51	SVD	/a/	ANN	97.90
Hossain	2016	28	MEEI, SVD	/a, i, o/	SVM, ELM, GMM	95.00
Ali	2017	5	MEEI, SVD, AVPD	/a/	GMM	94.60 (MEEI), 83.65 (AVPD), 80.20 (SVD)
Muhammad	2017	44	MEEI, SVD, AVPD	/a/	SVM	99.40 (MEEI), 93.20 (SVD), 91.50 (AVPD)
Dahmani	2017	14	SVD	/a/	NB	90.00
Souissi	2016	56	SVD	/a/	ANN	87.82
Hemmerling	2017	25	SVD	/a/	ANN	87.50
Souissi	2015	55	SVD	/a/	SVM	86.44

Table notation: Ref. – reference, MEEI – Massachusetts Eye and Ear Infirmary Database [18,39], SVD – Saarbruecken Voice Database [62,44,2], AVPD – Arabic Voice Pathology Database [41,44], KM – K-means [23], RF – Random Forests [11], GMM – Gaussian Mixture Models [50], SVM – Support Vector Machines [24], NB – Naive Bayes [45], ELM – Extreme Learning Machine [30], and ANN – Artificial Neural Networks [53].

[26,44,2] focused on using signal processing techniques (to quantify vocal-manifestations of the pathology under focus) and machine learning algorithms (to automate the process of voice pathology detection) to build a system capable of accurate discrimination of healthy and pathological voices. In Table 1, we summarize recent (2015 – now) related works focused on the objective voice pathology detection.

For the purpose of the objective voice pathology evaluation, several databases have been frequently used by the researchers. Massachusetts Eye and Ear Infirmary Database (MEEI) [18,39], Saarbruecken Voice Database (SVD) [62,44,2], and Arabic Voice Pathology Database (AVPD) [41,44] are among the most commonly used ones. More specifically, most researchers have analyzed sustained phonation of the vowel /a/, e.g. [1,43,7,14] due to its presence in most databases (language-independent speech task [59]). Some researchers also analyzed a combination of the vowels, e.g. [36, 17,26], etc. From the voice pathologies point of view, most researchers restricted the dataset to a limited set of pathologies [7,43,14,25,51,44,5,3,4,2].

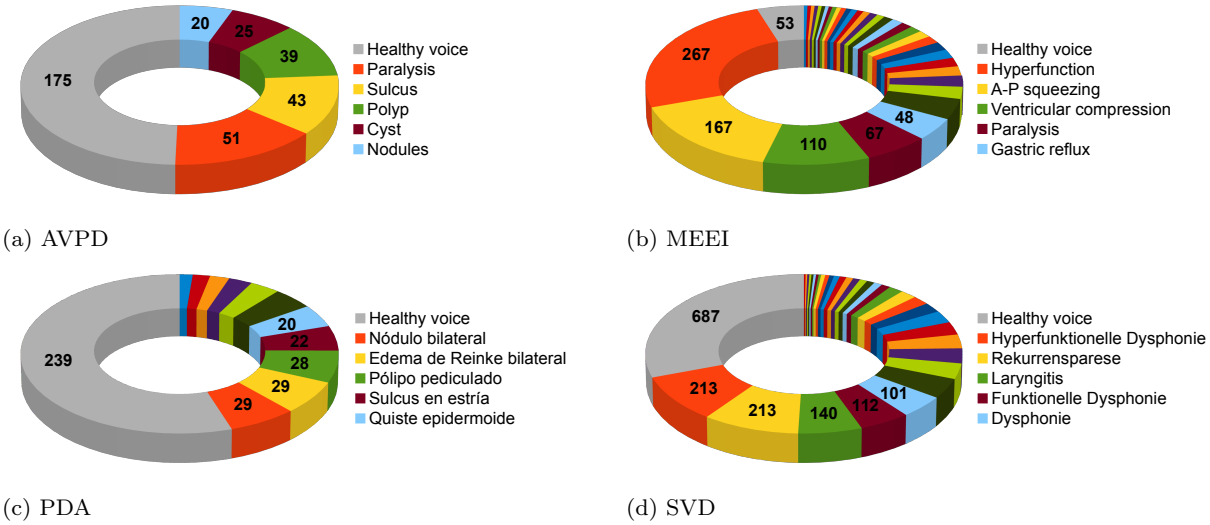
Next, conventional and clinically interpretable [10]) acoustic features were usually computed prior to pathology detection [43,14,51]. The acoustic features such as multidimensional voice program parameters (MDVP) [4], mel-frequency cepstral coefficients (MFCC) [52], glottal-to-noise excitation ratio (GNE) [42], etc. were usually extracted. For more information about methods for pathological speech parametrization, see [39]. After the feature extraction, multiple conventional classifiers have been used to detect the presence of voice pathology. Most authors relied on the following algorithms: Support Vector Machines (SVM), Gaussian Mixture Models (GMM), Random Forests (RF), and Artificial Neural Networks (ANN) [25,5,7,14], etc.

As can be seen, 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. However, we can conclude that:

1. most works analyzed a single speech task, mainly the sustained phonation of the vowel /a/ (language independent speech task)
2. most works analyzed datasets that were restricted to a subset of vocal pathologies from 1 to 3 databases (MEEI, SVD, AVPD)
3. most works extracted conventional dysphonic features to quantify major vocal-manifestations of specific vocal pathologies
4. most works employed conventional supervised learning algorithms such as the following: SVM, GMM, RF, ANN, and others

To propose results comparable with the previously published works, we analyze voice recordings of sustained phonation of the vowel /a/ as well. However, unlike the previous works, we analyze a larger dataset composed of 4 different databases, namely: MEEI [18, 39], SVD [62,44,44,2], AVPD [41,44] (these databases are commonly used by the community), and Príncipe de Asturias Database (PDA) [20,8,39]. Furthermore, to propose models capable of robust voice pathology detection, we do not restrict the dataset to only a subset of common vocal pathologies. With this approach, our dataset does contain a large number of pathologies with only few recordings. To see the sparsity of distribution and inequality of the number of pathologies in the databases, see Figures 1a (AVPD), 1b (MEEI), 1c (PDA), and 1d (SVD).

By using 4 different databases, we aim to increase the size of our dataset to enable exploring possibilities of using supervised deep learning techniques that de-



(a) AVPD (b) MEEI

(c) PDA (d) SVD

Fig. 1: Visualization of inequality of samples per vocal pathology in the datasets used in this work (only 5 most common pathologies in each database are present in the legend), and healthy samples. Databases: a) AVPD [41, 44], b) MEEI [18,39], PDA [20,8,39], and SVD [62,44,2].

livered state-of-the-art results in many domains including speech processing. To our best knowledge, despite our previous work [22], there are no other papers using deep learning algorithms for voice pathology detection. Next, we also employ the conventional voice pathology detection approach based on acoustic feature extraction procedure. However, unlike previous works, we use gradient boosting techniques for classification. To tackle the problem of sparse distribution of a variety of vocal pathologies with only few recordings across the databases, we also investigate usage of anomaly detection procedure.

The rest of this paper is organized as follows. Section 2 introduces databases utilized in this article. In Section 3, the methodology of the experiment is discussed. The results are presented in Section 4. Conclusions are drawn in Section 5.

2 Databases

As mentioned previously, we chose the following speech task: sustained phonation of the vowel /a/ as a basis for our experiments. During this particular speech task a speaker is asked to sustain phonation of a vowel, attempting to maintain steady frequency and amplitude at a comfortable level [59]. The advantage of this speech task in comparison with other common speech tasks such as reading tasks, or a running speech is that it is free of articulatory and other linguistic confounds [59]. This independence makes this task an ideal choice to construct a large dataset that is necessary for supervised deep learning algorithms. In fact, sustained /a/ vowel phonation is the only speech task that is present in all databases used in this work. The contents of the databases relevant to this work can be seen in Table 2.

2.1 AVPD database

Arabic Voice Pathology Database (AVPD) [41,44] was developed at the Communication and Swallowing Disorders Unit of King Abdul Aziz University Hospital, Riyadh, Saudi Arabia. The database contains recordings (366 samples: 188 healthy, 178 pathological) of sustained phonation of the vowels /a, e, o/, counting from 0-10, standardized Arabic passage, and reading three words. All recordings are sampled at $f_s = 48\,000$ Hz with a bit depth of 16 bits. The database comprises five organic voice disorders: vocal fold cysts, nodules, paralysis, polyps, and sulcus. Multiple recordings of the same vowel were taken to help model the intra-speaker variability.

2.2 MEEI database

Massachusetts Eye and Ear Infirmary Database (MEEI) [18,39] is one of the most popular and most widely-used database (used for many years as a benchmark in the field of pathological speech analysis). It contains more than 1 400 recordings of sustained phonation of the vowel /a/ (recorded from 657 pathological speakers with different types of pathologies and 53 healthy

speakers). This database has several disadvantages such as the fact that recordings of the normophonic voice were recorded in different conditions (e.g. different environment, recordings are sampled at: $f_s = 50\,000$ Hz, $f_s = 25\,000$ Hz, and $f_s = 10\,000$ Hz) when compared to pathological recordings. The database is also gender-unbalanced, etc.

2.3 PDA database

Príncipe de Asturias Database (PDA) [20,8,39] contains recordings of 200 pathological speakers with different types of organic pathologies (e.g. nodules, polyps, oedemas, and carcinomas, etc.) and 239 healthy speakers. For each speaker, sustained phonation of the vowel /a/ is recorded. All recordings are sampled at $f_s = 25\,000$ Hz. This database contains more speakers than a balanced version of MEEI database that according to [47] comprise only 173 recordings of pathological speakers.

2.4 SVD database

Saarbruecken Voice Database (SVD) [62,44,2] is a collection of voice recordings and electroglottography (EGG) signals from more than 2 000 speakers. It contains recordings of 687 healthy persons (428 females and 259 males) and 1356 patients (727 females and 629 males) with one or more of the 71 different pathologies. One session contains the recordings of the following components: a) vowels /i, a, u/ produced at normal, high and low pitch; vowels /i, a, u/ with rising-falling pitch; and c) sentence "Guten Morgen, wie geht es Ihnen?" ("Good morning, how are you?"). All samples of the sustained vowels are between 1 to 3 seconds long, sampled at $f_s = 50\,000$ Hz with 16-bit resolution [62]. In contrary to MEEI database, all audio samples (healthy and pathological) in SVD were recorded in the same environment.

3 Methodology

3.1 Data processing

We used 720 recordings from AVPD, 709 recordings from MEEI, 422 from PDA and 2 040 from SVD. We only excluded samples that were shorter than 0.750 s in length (removed 319 recordings). We also excluded all recordings of speakers bellow the age of 19 and also above the age of 60 (it is known that the most significant changes of voice happen during adulthood until the voice matures at around age of 20 and remains relatively stable until around age of 60) [57]. After these restrictions, the final number of samples equaled to 2 707.

Using SOX library (version 14.4.2), we re-sampled each recording to $f_s = 16\,000$ Hz. Then we trimmed each sample to exactly 0.750 s in duration. If a recording was below 0.950 s in duration, we extracted only one sample from the middle of it. For longer recordings we trimmed each end by 0.100 s and extracted as many 0.750 s long chunks as possible without overlap with stride of 0.375 s. Using this approach, the total number of 8 042 chunks was obtained. Further details regarding the number of chunks used can be found in Table 3.

Table 2: Contents of the databases used in this work.

	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

Table notation: PDA – Príncipe de Asturias Database (PDA) [20,8,39], MEEI – Massachusetts Eye and Ear Infirmary Database [18,39], SVD – Saarbruecken Voice Database [62,44,2], AVPD – Arabic Voice Pathology Database [41,44], H – healthy, P – pathological, and GRBAS – Grade, Roughness, Breathiness, Asthenia, Strain scale [15].

Table 3: 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**

Table notation: PDA – Príncipe de Asturias Database (PDA) [20,8,39], MEEI – Massachusetts Eye and Ear Infirmary Database [18,39], SVD – Saarbruecken Voice Database [62,44,2], AVPD – Arabic Voice Pathology Database [41,44], H – healthy, P – pathological, M – males, and F – females.

3.2 Feature extraction

At first, we considered raw audio samples as an input data for the voice pathology detection model. Each file (the 0.750 s chunk) was therefore inserted to the input of the neural network as a vector of 12 000 features (computed as: $0.750\,\text{s} \cdot 16\,000\,\text{Hz} = 12\,000$ features). Additionally, we normalized each sample using min-max scaling to a range $\langle 0, 1 \rangle$.

Next, we extracted a set of conventional commonly-used acoustic (dysphonic) features [10,39] using Neurological Disorder Analysis Tool (NDAT) [40,39] written in MATLAB and developed at the Brno University of Technology. Specifically, we computed the following acoustic features: pitch, jitter, shimmer, harmonic-to-noise ratio, detrended fluctuation analysis parameters, glottis quotients (open, closed), glottal-to-noise excitation ratio, Teager-Kaiser energy operator, modulation energy, and normalized noise energy. We further applied the following statistical properties: mean, standard deviation, coefficient of variation, quartiles (1^{st}, 2^{nd}, 3^{rd}), interquartile range, kurtosis, and skewness.

Moreover, we computed spectrograms using Matplotlib (version 2.1.2) library for Python. The computation setup: mode (power spectral density), no windowing, no overlap, and NFFT (512 samples). Following Ali et al. [6], we used data up to 1 500 Hz (1 150 features). Furthermore, we normalized the values of this matrix with min-max scaling to a range between 0 and 1 as well.

At last, we computed most commonly used perceptual [40] acoustic feature: MFCC using Python Speech Features library. The computation setup: length of a window function (0.025 s), step size (0.010 s), number of filters in the filter-bank (26), number of coefficients (13), and NFFT (512 samples). With this approach, we obtained a matrix consisting of 962 features (13 coefficients $\times$ 74 time steps). We also computed the mean and standard deviation of the 13 coefficients along the time axis, which resulted into additional 26 features per sample. Next, we scaled the MFCC feature matrix by min-max algorithm (means and standard deviations were computed before scaling). The statistical features were scaled separately to have 0 mean and unit variance before classification.

3.3 Experiments

As mentioned previously, there is a wide range of pathologies present in the databases used in this work. For more information, see Figures 1a (AVPD), 1b (MEEI), 1c (PDA), and 1d (SVD). Each database was labelled in different language and with different experts by different criteria. Therefore, it is almost impossible to combine these databases to obtain one consistent database of multiple pathologies with reasonable number of samples. Only feasible way of combining the samples seems to be the exhaustive manual pairing by an expert clinician, which is also rather difficult since lots of recordings are labelled with multiple pathologies. In order to conduct inter-database experiments, authors therefore usually pick a smaller sub-sample of 2 to 5 pathology types that are relatively easier to pair.

Next, most of these pathologies are very sparsely distributed across the databases. Searching for an ideal subset of acoustic features that would yield a good classification performance for each voice pathology is therefore almost impossible. Furthermore, it is not well-known if these pathologies present in the databases have similar vocal-manifestations.

In contrast to the previous works, we aim to investigate possibilities of robust voice pathology detection using a set of 4 almost unrestricted databases comprising a large number of pathologies. From these reasons, we decided to conduct several experiments: a) supervised learning (assuming the pathologies have similar manifestations and therefore the number of samples per pathology type is irrelevant), b) anomaly detection (assuming the pathologies do not have similar manifestations and therefore the number of samples per pathology type cannot be neglected).

Regarding the supervised learning approach, we used the state-of-the-art gradient boosting algorithm: XGBoost [12] (version 0.6) for its current successes in many Kaggle competitions, fast training and model interpretability. Additionally, we explored possibilities of deep learning approach for its ability to robustly model complex relationships when optimized using enough data. However, the equation for computing the sufficient size of training dataset has not been formally described yet. Generally established rule of thumb in machine learning community is to have more training samples than trainable parameters. For this reason, we used the DenseNet [29] architecture, which succeeded in overcoming the problem of having too many trainable parameters by densely connecting the convolutional layers. We adjusted Thibault de Boissiere's Keras implementation of the DenseNet (Keras framework [13], version 2.1.2), to process 1D signals treating raw audio as 1D vector. Spectrograms were processed as a matrix using the frequency bins not as y dimension, but rather as a stack of channels in the same way the 3 RGB channels are stacked in an image [63]. The MFCC were processed the same way as spectrograms. Since we are not able to say with 100% certainty that healthy examples are not polluted by deviant samples, we decided to use anomaly detection in favor of novelty detection in which case it is important to model the non-deviant samples. In this case, we chose Isolation Forest [34,35] classifier implemented in scikit-learn library [48] (version 0.19.1).

For the above mentioned experiments, we decided to analyze the performance of the voice pathology detection models using multiple types of input data: a) raw audio samples to follow our previous work [22] and fur-

ther explore possibilities of robust voice pathology detection without manually-selected features (DenseNet), b) conventional acoustic (dysphonic) features to follow the previously published works and quantify most common vocal pathologies (XGBoost, Isolation Forest), c) spectrograms to achieve a reasonable trade-off between dimensionality of the data and amount of information (DenseNet), and d) MFCC to follow the previous works focusing on voice and speech modelling, and voice pathology detection (all models).

3.4 Training and validation

To train and validate the models, we split the data to training, validation and testing sets. On top of that, we generated 10-fold validation indices using training and validation sets, so we can use exactly the same sets of data for each experiment. The test set was left for final evaluation of the models. Next, we stratified the testing and validation sets by medical state (healthy – H, pathological – P), gender, age, and gender-age group. Since the longer recordings were split into multiple chunks, we had to prevent the samples with chunks in the test or validation sets from leaking into the training set. Such chunks were carefully removed from the set. All other chunks were used in the training set.

At this point there is an unequal distribution of samples within the training set. We reacted to this fact by computing sample weights that can be used during training as a compensation measure for under-represented groups. The final sample weight is a product of 3 partial weights. Each of the partial weights quantifies the ratio between subgroups within the selected group (e.g. the ratio between the number of normophonic and pathological samples). For this purpose, we introduced a class weight α, gender weight β, and gender-age group weight γ resulting in final sample weight ω that is computed as $\omega = \alpha \cdot \beta \cdot \gamma$. Weight for a particular sample that belongs to subgroup α_i within group α, β_i within group β, and γ_i within group γ can be computed as follows:

$$\omega_{\alpha_i,\beta_i,\gamma_i} = \frac{n_{\alpha_i}}{\max(n_\alpha)} \cdot \frac{n_{\beta_i}}{\max(n_\beta)} \cdot \frac{n_{\gamma_i}}{\max(n_\gamma)} \tag{1}$$

where n represents the total number of samples: n_{α_i} is the number of samples in a particular class; n_{β_i} is the number of samples in a particular gender; and n_{γ_i} is the number of samples in a particular gender-age group.

We used 30 to 100 iterations of randomized cross-validation search for hyper-parameter optimization for both XGBoost and Isolation Forest classifiers. The number of iterations did depend on the fitting time. More specifically, in the case of XGBoost, we were searching over the following hyper-parameters: number of estimators $\langle 3, 300\rangle$, learning rate $\langle 0.006, 1\rangle$, gamma $\langle 10, 60\rangle$, maximum depth $\langle 0, 9\rangle$, minimum child weight $\langle 1, 3\rangle$, sub-sample ratio $\langle 0.3, 1\rangle$ and colsample bytree (sub-sample ratio of columns when constructing each tree) $\langle 0.1, 1\rangle$. Regarding Isolation Forest we were searching over the following hyper-parameters: number of estimators $\langle 6, 200\rangle$, maximum samples $\langle 8, 64\rangle$, contamination $\langle 0.40, 0.76\rangle$ and maximum features $\langle 0.05, 1\rangle$. As a performance measure, we used F1 micro score as a criteria of choosing the best hyper-parameters in the cross-validation setup. After the search for hyper-parameter, we re-fitted the models with the best hyper-parameters on the entire training set, and consequently evaluated on the testing set. The final results are presented in the form of confusion matrix (CM), and classification report (CR) tables. The formulas 2, 3 and 4 describe the way of computing the precision, recall and F1 score (weighted average of the precision and recall) metrics presented in CR tables.

$$precision = \frac{tp}{tp + fp}, \tag{2}$$

where tp denotes the number of correct predictions (observed class), and fp determines the number of incorrect predictions (observed class). The precision is a ratio between the number of correct predictions of the observed class and the total number of predictions of the observed class.

$$recall = \frac{tp}{tp + fn}, \tag{3}$$

where tp denotes the number of correct predictions (observed class), and fn determines the number of incorrect predictions (opposing class). The recall is a ratio between the number of correct predictions of the observed class and the total number of samples in the observed class.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

4 Results

XGBoost [12] trained (10-fold validation) with all features (consisting of 96 conventional dysphonic features and 26 MFCC coefficients) yielded an average F1 score of 0.922 ($\pm$ 0.004) on the training set, and 0.829 ($\pm$ 0.028) on the validation set. The final F1 score on the dedicated testing set was 0.733. Performance details (classification matrix and classification report) can be found in Table 4 and Table 5. Based upon the performance on the development set (training and validation sets) the 50 iterations of randomized cross-validated search

selected the following hyper-parameters: number of estimators (294), learning rate (0.3), gamma (10), max. depth (3), sub-sample (0.5), minimum child weight (1), colsample bytree (1). Details regarding the classification performance in relation to input data can be found in Table 6.

Table 4: Testing CM 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 5: Testing CR 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

Regarding deep learning approach, we used the adjusted DenseNet [29] architecture with the binary cross-entropy loss optimized using Adam optimizer [31]. The initial learning rate was set to 0.01 with decay of $1e-04$ on each epoch. Hyper-parameter optimization was done using training and validation sets, and the final parameters of the DenseNet network were set as follows: depth (4), number of dense blocks (2), growth rate (5), number of filters (10), drop-out rate (0.3), l2 weight decay ($1e-04$). The input shape of this network was (13×47) with one neuron in the last layer with sigmoid activation function, and the total of 1 629 trainable parameters. For this particular setup with MFCC as the input data, the system yielded F1 score of 0.595 on the training set, and 0.648 on the validation set. After the hyper-parameter optimization, we retrained the network on all data from the training and validation sets (the development set), and the system yielded the final F1 score on the dedicated testing set of 0.621. Performance details (classification matrix and classification report) can be found in Table 7 and Table 8.

DenseNet trained with spectrograms had input shape (46×25) and total of 301 trainable parameters. Even though this setup was considerably less complex, and regularized with drop-out (0.3) and l2 weight decay ($1e-04$), the network tended to over-fit after enough training epochs, which we prevented using early stopping that monitored changes in the validation accuracy.

Table 6: XGBoost performance related to input data

Input data	F1 CV train	F1 CV valid	F1 test
ALL	0.922 ($\pm$ 0.004)	0.829 ($\pm$ 0.028)	0.733
AF stats	0.886 ($\pm$ 0.004)	0.791 ($\pm$ 0.034)	0.686
AF	0.892 ($\pm$ 0.006)	0.798 ($\pm$ 0.025)	0.658
AF base	0.745 ($\pm$ 0.009)	0.689 ($\pm$ 0.036)	0.646
MFCC	0.680 ($\pm$ 0.010)	0.769 ($\pm$ 0.037)	0.623

Table notation and description of acoustic features used to build XGBoost model: MFCC – 26 Mel Frequency Spectral Coefficients (13 means & 13 standard deviations), AF base – 12 common acoustic (dysphonic) features, AF stats. – 84 acoustic features' statistics, AF – AF base & AF stats, ALL – AF & MFCC.

This system yielded F1 score of 0.635 and 0.531 on the training and validation sets, respectively. The performance on the testing set was 0.562 (F1 score 0.514 for class H and 0.609 for class P). After refitting on the whole development set, the final F1 score got worse on the dedicated testing set to 0.460 due to difficulties with classification of healthy voices (F1 score 0.239 for class H and 0.680 for class P). With raw input data, the network failed to learn any meaningful features (the size of out training dataset is still too small to provide deep learning algorithm to overcome more conventional approaches).

Hyper-parameter optimization for Isolation Forest trained (10-fold validation) with 96 speech parameters was done the same way as for XGBoost. The best parameters selected upon performance on the development set were as follows: number of estimators (200), contamination (0.4), maximum features (0.3), maximum samples was set to "auto". The system yielded F1 score of 0.576 ($\pm$ 0.005) on the training set and 0.578 ($\pm$ 0.023) on the validation set. The final F1 score on the dedicated testing set was 0.610. The performance details (classification matrix and classification report) can be found in Table 9 and Table 10. This system showed to be sensitive to the number of input features and the performance raised when we selected just a subset of them.

Table 7: Testing CM for DenseNet (MFCC)

	true H	true P	total predicted
predicted H	**73**	44	117
predicted P	47	**76**	123
total true	120	120	accuracy: **0.621**

Table 8: Testing CR 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 9: Testing CM for Isolation Forest

	true H	true P	total predicted
predicted H	**58**	30	88
predicted P	62	**90**	152
total true	120	120	accuracy: **0.617**

Table 10: Testing CR 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

5 Conclusions

In search towards robust voice pathology detection system using acoustic (voice) signals, researchers face a variety of problems. One of the major problems in this field of science is the limited number of available databases. Moreover, commonly used databases [18,44,41,20] are very hard to combine because of various distinctions such as: a) the databases are labeled in different languages, b) the databases do not comprise same set of speech tasks, c) there is a variety of voice pathologies unequally distributed across the databases, etc. For these reasons, researchers have used only a subset of the databases for their experiments providing results related to those carefully selected subset of data. However, this approach limits the possibilities of creating a robust voice pathology detector. Therefore, in this work, we have conducted experiments on recordings of sustained phonation of the vowel /a/ produced at normal pitch from 4 different databases trying to eliminate those limitation. To the best of our knowledge, this is the first work that uses such a large set of data to build mathematical models for computerized, objective voice pathology detection.

We researched 3 distinct classifiers within supervised learning and anomaly detection paradigms. We have explored raw waveforms, spectrograms, MFCC and conventional dysphonic features as input data. All experiments were evaluated by the same criteria on the same dedicated testing set. We observed that XGBoost classifier achieved the best results amongst DenseNet and Isolation Forest classifiers. We also observed that not only XGBoost provided the best performance, it could also handle the feature selection (input: all features) by itself in contrary to Isolation Forest classifier, which showed to be sensitive on the feature selection (input: manually selected subset of features). Overall advantage of using speech features and MFCC with XGBoost was the computation speed that allowed us to use exhaustive randomized cross-validated search to optimize the hyper-parameters, as well as the possibility to sort features by importance. This property is useful for clinical interpretability. Nevertheless, we consider these results exploratory due to the limitations of the databases. Reviewing the performances achieved in scenarios with MFCC as input data we conclude that MFCC alone are not reliable enough for robust voice pathology detection, which was also concluded by Ali et al. in [5]. Regarding the DenseNet, we conclude that in voice pathology detection scenarios with this little training data it is better to use inputs with reduced dimensionality in contrary to raw waveform inputs, or make use of transfer learning or data augmentation.

In this article there are several limitations. Firstly, there are limitations inherited from the databases along with new limitations caused by their combination. For instance, some databases have extremely unequal distribution of healthy and pathological classes, most of the databases have alarming inequalities between the number of samples per pathology type (e.g. many pathologies are present less than 3 times in the database), see Figures 1a (AVPD), 1b (MEEI), 1c (PDA), and 1d (SVD). Most databases have no information about severity of the pathology, nor they have information about manifestation of the pathology in phonation, which means that some of the samples might sound as healthy even though they are labelled as pathological and vice versa. Not to mention that recordings are labelled with more than 1 type of pathology, and in different languages, which makes it especially hard to combine or exclude the samples. Since we used 4 available databases, we utilized only the speech task available in all of them: sustained phonation of the vowel /a/ produced at normal pitch. Secondly, even though we have taken countermeasures to balance the classes with sample weights, we did not conduct our experiments separately on subsets of data for different genders.

Up to this point, most papers focused on voice pathology detection used conventional dysphonic features to quantify the underlying voice pathology. In general, these features are conceptually simple, which on one hand is an advantage as these features are clinically interpretable (i.e. clinicians are able to associate the val-

ues of the features with known physiological phenomena inside human body) [40], but on the other hand these features are often unable to describe the exact voice pathology under focus in a more complex way, especially in advanced stages of the disease (high level of acoustic noise, irregularity of voice, etc.). In future studies, researchers may consider exploring usage of a more sophisticated set of acoustic features to complexly and robustly describe the voice and speech production deterioration. For instance, such features have already been successfully applied in the field of non-invasive assessment of Parkinson's disease [33,60,38].

With the previously mentioned facts in mind, we think that recordings of the databases commonly used for automatic voice pathology detection should be consulted with clinicians to evaluate the severity of vocal manifestation of the present pathologies. There are standard metrics, which are used to evaluate the quality of voice that can be used for this purpose [15,19,32,15,16]. Addition of such information to the databases could provide researchers with a unique possibility to build models capable of classification and prediction emphasizing the severity of the exact vocal-manifestation (increased acoustic tremor, roughness, breathiness, etc.) of these pathologies.

We also anticipate that deep learning will play its role in robust voice pathology detection on the assumption that more data will be available, or at least reasonable combination of available databases will be made and limitations of these databases will be partially diminished by data augmentation and other countermeasures. In addition, we presume that use of deep learning methods for novelty detection such as deep autoencoder [49] for modelling the normophonic voice could be an interesting idea for future investigation with prospect to identify even disordered voices that are sparsely distributed across databases.

In summary, acoustic (voice) signals can nowadays be recorded using a variety of smart devices and processed remotely using modern cloud technologies. In comparison with the conventional perceptual voice quality examination, computerized acoustic analysis of voice signals can provide clinicians with fast, supportive methodology of objective voice pathology detection, assessment, and monitoring that can be used on everyday basis (see Health 4.0). However, to take advantage of such methodology, robust mathematical models capable of precise voice pathology detection must be introduced. Our work proposes the next step towards this goal using various state-of-the-art machine learning algorithms applied to the largest dataset that have been used for the purpose of automatic voice pathology detection.

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 the following projects: SIX (CZ.1.05/2.1.00/03.0072), and 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.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Al-nasheri, A., Ali, Z., Muhammad, G., Alsulaiman, M.: Voice pathology detection using auto-correlation of different filters bank. In: Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on, pp. 50–55. IEEE (2014)
2. Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z.: Investigation of voice pathology detection and classification on different frequency regions using correlation functions. Journal of Voice **31**(1), 3–15 (2017)
3. Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Malki, K., Mesallam, T., Farahat, M.: Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. IEEE Access **PP**(99), 1–1 (2017). DOI 10.1109/ACCESS.2017.2696056
4. Al-nasheri, A., Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T.A., Farahat, M., Malki, K.H., Bencherif, M.A.: An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. Journal of Voice **31**(1), 113–e9 (2017)
5. Ali, Z., Alsulaiman, M., Muhammad, G., Elamvazuthi, I., Al-nasheri, A., Mesallam, T.A., Farahat, M., Malki, K.H.: Intra-and inter-database study for arabic, english, and german databases: Do conventional speech features detect voice pathology? Journal of Voice **31**(3), 386–e1 (2017)
6. Ali, Z., Muhammad, G., Alhamid, M.F.: An automatic health monitoring system for patients suffering from voice complications in smart cities. IEEE Access **5**, 3900–3908 (2017)
7. Amami, R., Smiti, A.: An incremental method combining density clustering and support vector machines for voice pathology detection. Computers & Electrical Engineering **57**, 257–265 (2017)
8. Arias-Londoño, J.D., Godino-Llorente, J.I., Markaki, M., Stylianou, Y.: On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices. Logopedics Phoniatrics Vocology **36**(2), 60–69 (2011)
9. Armstrong, D., Gosling, A., Weinman, J., Marteau, T.: The place of inter-rater reliability in qualitative research: an empirical study. Sociology **31**(3), 597–606 (1997)

10. Brabenec, L., Mekyska, J., Galaz, Z., Rektorova, I.: Speech disorders in parkinsons disease: early diagnostics and effects of medication and brain stimulation. Journal of Neural Transmission **124**(3), 303–334 (2017)
11. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
12. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794. ACM (2016)
13. Chollet, F., et al.: Keras: Deep learning library for theano and tensorflow. URL: https://keras.io/ (2015)
14. Dahmani, M., Guerti, M.: Vocal folds pathologies classification using naïve bayes networks. In: Systems and Control (ICSC), 2017 6th International Conference on, pp. 426–432. IEEE (2017)
15. De Bodt, M.S., Wuyts, F.L., Van de Heyning, P.H., Croux, C.: Test-retest study of the grbas scale: influence of experience and professional background on perceptual rating of voice quality. Journal of Voice **11**(1), 74–80 (1997)
16. Dejonckere, P.H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., Van De Heyning, P., Remacle, M., Woisard, V.: A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Eur Arch Otorhinolaryngol. **258**(2), 77–82 (2001)
17. Eskidere, Ö., Gürhanlı, A.: Voice disorder classification based on multitaper mel frequency cepstral coefficients features. Computational and mathematical methods in medicine **2015** (2015)
18. Eye, M., Infirmary, E.: Voice disorders database, version. 1.03 (cd-rom). Lincoln Park, NJ: Kay Elemetrics Corporation (1994)
19. Gerratt, B.R., Kreiman, J., Antonanzas-Barroso, N., Berke, G.S.: Comparing internal and external standards in voice quality judgments. J Speech Hear. Res. **36**(1), 14–20 (1993)
20. Godino-Llorente, J.I., Gómez-Vilda, P., Cruz-Roldán, F., Blanco-Velasco, M., Fraile, R.: Pathological likelihood index as a measurement of the degree of voice normality and perceived hoarseness. Journal of Voice **24**(6), 667–677 (2010)
21. Gwet, K.L.: Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC (2014)
22. Harar, P., Alonso-Hernandezy, J.B., Mekyska, J., Galaz, Z., Burget, R., Smekal, Z.: Voice pathology detection using deep learning: a preliminary study. In: Bioinspired Intelligence (IWOBI), 2017 International Conference and Workshop on, pp. 1–4. IEEE (2017)
23. Hartigan, J.A., Wong, M.A.: Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1), 100–108 (1979)
24. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. IEEE Intelligent Systems and their applications **13**(4), 18–28 (1998)
25. Hemmerling, D.: Voice pathology distinction using autoassociative neural networks. In: Signal Processing Conference (EUSIPCO), 2017 25th European, pp. 1844–1847. IEEE (2017)
26. Hemmerling, D., Skalski, A., Gajda, J.: Voice data mining for laryngeal pathology assessment. Computers in biology and medicine **69**, 270–276 (2016)
27. Hillenbrand, J., Houde, R.A.: Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. J Speech Hear Res **39**(2), 311–21 (1996)
28. Hossain, M.S., Muhammad, G.: Healthcare big data voice pathology assessment framework. IEEE Access **4**, 7806–7815 (2016)
29. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)
30. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1), 489–501 (2006)
31. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
32. Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., Berke, G.S.: Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. J Speech Hear. Res. **36**(1), 21–40 (1993)
33. Little, M., McSharry, P., Hunter, E., Spielman, J., Ramig, L.: Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE T Bio-Med Eng **56**(4), 1015–1022 (2009)
34. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, pp. 413–422. IEEE (2008)
35. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) **6**(1), 3 (2012)
36. Martínez, D., Lleida, E., Ortega, A., Miguel, A., Villalba, J.: Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In: Advances in Speech and Language Technologies for Iberian Languages, pp. 99–109. Springer (2012)
37. Mehta, D.D., Hillman, R.E.: Voice assessment: updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. Current opinion in otolaryngology & head and neck surgery **16**(3), 211 (2008)
38. Mekyska, J., Galaz, Z., Mzourek, Z., Smekal, Z., Rektorova, I.: Assessing progress of Parkinson's using acoustic analysis of phonation. In: 2015 International Work Conference on Bioinspired Intelligence (IWOBI), pp. 115–122 (2015). DOI 10.1109/IWOBI.2015.7160153
39. Mekyska, J., Janousova, E., Gomez-Vilda, P., Smekal, Z., Rektorova, I., Eliasova, I., Kostalova, M., Mrackova, M., Alonso-Hernandez, J.B., Faundez-Zanuy, M., et al.: Robust and complex approach of pathological speech signal analysis. Neurocomputing **167**, 94–111 (2015)
40. Mekyska, J., Smekal, Z., Galaz, Z., Mzourek, Z., Rektorova, I., Faundez-Zanuy, M., López-de Ipiña, K.: Recent Advances in Nonlinear Speech Processing, chap. Perceptual Features as Markers of Parkinson's Disease: The Issue of Clinical Interpretability, pp. 83–91. Springer International Publishing, Cham (2016). DOI 10.1007/978-3-319-28109-4_9
41. Mesallam, T.A., Farahat, M., Malki, K.H., Alsulaiman, M., Ali, Z., Al-nasheri, A., Muhammad, G.: Development of the arabic voice pathology database and its evaluation by using speech features and machine learning algorithms. Journal of healthcare engineering **2017** (2017)
42. Michaelis, D., Gramss, T., Strube, H.W.: Glottal-to-noise excitation ratio–a new measure for describing pathological voices. Acta Acustica united with Acustica **83**(4), 700–706 (1997)
43. Muhammad, G., Alhamid, M.F., Hossain, M.S., Almogren, A.S., Vasilakos, A.V.: Enhanced living by assessing

voice pathology using a co-occurrence matrix. Sensors **17**(2), 267 (2017)
44. Muhammad, G., Alsulaiman, M., Ali, Z., Mesallam, T.A., Farahat, M., Malki, K.H., Al-nasheri, A., Bencherif, M.A.: Voice pathology detection using interlaced derivative pattern on glottal source excitation. Biomedical Signal Processing and Control **31**, 156–164 (2017)
45. Murphy, K.P.: Naive bayes classifiers. University of British Columbia (2006)
46. Oates, J.: Auditory-perceptual evaluation of disordered voice quality. Folia Phoniatrica et Logopaedica **61**(1), 49–56 (2009)
47. Parsa, V., Jamieson, D.G.: Identification of pathological voices using glottal noise measures. J Speech Lang. Hear. Res. **23**(2), 469–85 (2003)
48. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
49. Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Processing **99**, 215–249 (2014)
50. Reynolds, D.: Gaussian mixture models. Encyclopedia of biometrics pp. 827–832 (2015)
51. Sabir, B., Rouda, F., Khazri, Y., Touri, B., Moussetad, M.: Improved algorithm for pathological and normal voices identification. International Journal of Electrical and Computer Engineering (IJECE) **7**(1), 238–243 (2017)
52. Saldanha, J.C., Ananthakrishna, T., Pinto, R.: Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features. Journal of medical imaging and health informatics **4**(2), 168–173 (2014)
53. Schalkoff, R.J.: Artificial neural networks, vol. 1. McGraw-Hill New York (1997)
54. Song, P.: Assessment of vocal cord function and voice disorders. In: Principles and Practice of Interventional Pulmonology, pp. 137–149. Springer (2013)
55. Souissi, N., Cherif, A.: Dimensionality reduction for voice disorders identification system based on mel frequency cepstral coefficients and support vector machine. In: Modelling, Identification and Control (ICMIC), 2015 7th International Conference on, pp. 1–6. IEEE (2015)
56. Souissi, N., Cherif, A.: Speech recognition system based on short-term cepstral parameters, feature reduction method and artificial neural networks. In: Advanced Technologies for Signal and Image Processing (ATSIP), 2016 2nd International Conference on, pp. 667–671. IEEE (2016)
57. Stathopoulos, E.T., Huber, J.E., Sussman, J.E.: Changes in acoustic characteristics of the voice across the life span: measures from individuals 4–93 years of age. Journal of Speech, Language, and Hearing Research **54**(4), 1011–1021 (2011)
58. Teager, H.: Some observations on oral air flow during phonation. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(5), 599–601 (1980)
59. Titze, I.R.: Principles of voice production. Englewood Cliffs, N.J (1994)
60. Tsanas, A., Little, M.A., McSharry, P.E., Ramig, L.O.: Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. J. R. Soc. Interface **8**(59), 842–855 (2010)
61. Uloza, V., Vegiene, A., Saferis, V.: Correlation between the quantitative video laryngostroboscopic measurements and parameters of multidimensional voice assessment. Biomedical Signal Processing and Control **17**(Supplement C), 3–10 (2015)
62. Woldert-Jokisz, B.: Saarbruecken voice database (2007)
63. Wyse, L.: Audio spectrogram representations for processing with convolutional neural networks. arXiv preprint arXiv:1706.09559 (2017)