INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Automated Detection of Voice Disorder in the Saarbrücken Voice Database: Effects of Pathology Subset and Audio Materials

Mark Huckvale, Catinca Buciuleac

Speech, Hearing and Phonetic Sciences, University College London, UK
m.huckvale@ucl.ac.uk, catinca.buciuleac.16@alumni.ucl.ac.uk

Abstract

The Saarbrücken Voice Database contains speech and simultaneous electroglottography recordings of 1002 speakers exhibiting a wide range of voice disorders, together with recordings of 851 controls. Previous studies have used this database to build systems for automated detection of voice disorders and for differential diagnosis. These studies have varied considerably in the subset of pathologies tested, the audio materials analyzed, the cross-validation method used and the performance metric reported. This variation has made it hard to determine the most promising approaches to the problem of detecting voice disorders. In this study we re-implement three recently published systems that have been trained to detect pathology using the SVD and compare their performance on the same pathologies with the same audio materials using a common cross-validation protocol and performance metric. We show that under this approach, there is much less difference in performance across systems than in their original publication. We also show that voice disorder detection on the basis of a short phrase gives similar performance to that based on a sequence of vowels of different pitch. Our evaluation protocol may be useful for future studies on voice disorder detection with the SVD.

Index Terms: voice disorders, machine learning, health applications

1. Introduction

Automated systems for the detection and diagnosis of voice disorders from audio recordings could be useful for screening, for clinical assessment, and for monitoring the progress of patients after therapy. There have been many published studies that use machine learning approaches for the automated detection of voice disorders, but they vary considerably in materials and pathologies, such that they are difficult to compare and evaluate [3]. In this paper we compare the performance of previously published systems for the automated detection of voice disorders that have been trained and evaluated on the Saarbrücken Voice Database (SVD) [1]. Our aim is to explore how well these systems compare when evaluated using a common protocol. We first describe the contents of SVD, outline ways in which it has been used for automated detection of voice disorders and set out the objectives of our study.

1.1. Background to the corpus

The Saarbrücken Voice Database contains recordings of 1002 speakers exhibiting a wide range of voice disorders (454 male and 548 female), together with recordings of 851 controls (423 male and 428 female). The age of speakers varies from 6-94 years (pathological) and 9-84 years (control). There are an average of 1.2 recording sessions per speaker (max=24) leading to a total of 2225 sessions. Each recording session contains recordings of /i/, /a/ and /u/ vowels recorded on typical, higher, lower, rising and falling pitch, together with the short phrase "Guten Morgen, wie geht es Ihnen?". Audio and electroglottograph (EGG) recordings are available sampled at 16-bit precision at 50k samples/sec.

A wide range of pathologies are represented in the database [2]. There are 71 different pathology labels used, but 263 sessions are assigned more than one label. Some pathologies are much better represented than others. Of the 1093 pathological recordings with a single diagnostic label, the most frequent are: Vocal fold paralysis (197), Hyperfunctional dysphonia (143) and Laryngitis (82); while there are 19 pathologies which only occur once.

While the SVD is an extremely useful resource, it is not an easy database to partition for use in machine learning. The imbalance in the frequency of pathologies, the assignment of multiple pathologies to speakers, and the presence of multiple recordings per speaker could easily bias classification performance. For example, of the 62 examples of Spastic Dysphonia, 54 come from only three speakers. A system for detecting Spastic Dysphonia would do well if all it did was to recognize these speakers. Also, any cross-validation process that did not take speaker into account could allow the same speaker to be present in both training and testing partitions, artificially boosting performance.

1.2. Previous studies

Table 1 shows results of some selected studies of pathology detection on the SVD. The table shows that studies vary not only in terms of technique, but also in both the audio materials chosen for analysis, and how subsets of pathologies are used for classification. A more extensive review of previous studies can be found in [3].

Studies A, B and C in Table 1 demonstrate that extremely high disorder detection accuracy can be obtained from a single /a/ vowel when the data comes from a limited set of pathologies in the database. Studies B and C are limited to only the most significant structural disorders of the larynx. Studies D and E both use a convolutional neural network to analyze an /a/ vowel, and yet report very different success. This may also be due to a different choice of pathologies tested. Study F uses much more audio material per speaker than other studies, which should provide higher accuracy, and yet reports worse performance than studies A-E.

The database issues raised in section 1.1 may also be playing a role in this variation: the studies do not make clear how they are dealing with multiple recordings per speaker, or multiple diagnoses per recording. Overall, replication studies using a common protocol are required to adequately compare these studies.

http://dx.doi.org/10.21437/Interspeech.2021-1507

Table 1 *Selected previous studies on voice pathology detection with the SVD*

	Study	Pathology Subset	Material	Features	Classifier	Test Score
A	Hemmerling et al, 2016 [4]	Selected mix of pathologies	/a/ neutral	PCA on acoustic features	Random Forest	99%
B	Al-Nasheri et al, 2017 [5]	Cysts, polyps, paralysis	/a/ neutral	MDVP	SVM	99%
C	Muhammad et al ,2017 [6]	Cysts, polyps, paralysis	/a/ neutral	Spectro-temporal pattern	SVM (RBF)	93%
D	Wu & Lowit, 2018 [7]	Most common structural pathologies only	/a/ neutral	Spectrogram	CNN	77%
E	Mohammed et al, 2020 [8]	Mix of pathologies	/a/ neutral	Spectrogram	RESNET	94%
F	Barche et al, 2020 [9]	Structural, neurogenic, non-organic	/iau/ at low, mid, high pitch	SMILE	SVM (Poly)	83%

1.3. Goals of this study

The lack of consistency in the use of the recordings in the SVD by previous studies makes it impossible to determine the most promising approaches for the automated detection of voice disorders. In this study we aim to replicate in part three studies using a common testing protocol. We look at study F in Table 1 [9] since it uses a widely available feature set (OpenSMILE) and classifier (Support Vector Machine, SVM). We choose study D [7] since it uses a radically different approach with a spectrographic input into a convolutional neural network (CNN). We choose study E [8] since it uses a pretrained CNN with a proven architecture for image recognition (RESNET), and which also promises outstanding performance. This comparison will not only provide information about the relative performance of these methods on the same data but will also set out a standardized way of working with the multiple recordings and multiple diagnoses per speaker found in the SVD.

As secondary aims, we would like to explore the effect of pathology subset on performance, looking at all pathologies and a subset of organic pathologies. We will also evaluate the SVM method on single vowel, multiple vowels, and the short phrase to explore the impact of audio materials on performance.

The following sections will set out the methods used for evaluation, present comparable results for the three methods, discuss their implications, and suggest avenues for further work.

2. Methods

2.1. Database selection

We evaluate two different pathology subsets of the SVD. The first comprises all pathological recordings and all control recordings, including all repeated sessions. This is up to 869 control recordings and 1356 pathological recordings, although there are a few missing recordings depending on material type.

Secondly we construct a subset of pathologies that can be associated with organic damage or malfunction, in contrast to disorders that are psychological in origin. The motivation for this is that clinical therapies are distinct for organic vs non-organic disorders [9]. Organic pathologies include: Laryngitis, Leukoplakia, Polyps, Contact ulcer, Reinke's Oedema, Spastic Dysphonia, Paralysis, and Cancer. Non-organic pathologies excluded include Functional Dysphonia and Psychogenic Dysphonia. We also exclude diagnoses which are vague as to cause, for example: Dysphonia, Vox Senilis. Here we only include diagnoses with multiple diagnostic labels if both labels refer to organic disorders. This leads to a subset with up to 869 control recordings and 597 pathological recordings.

2.2. Audio selection and pre-processing

For these experiments, we have chosen three sets of audio materials: the /a/ vowel recording produced on a neutral pitch, the recording of all vowels on all pitch levels, and the recording of the spoken phrase. The /a/ vowel will be used in replicating the CNN & RESNET methods, while all three types of material will be used to replicate the SVM method.

Audio materials were downsampled to 20kHz for the SMILE feature extraction used for the SVM and 16kHz for the CNN spectrogram. Audio levels were also normalized to -20dB RMS re: full-scale.

2.3. Feature extraction and normalisation

For the SVM method, the OpenSMILE toolkit [10] was used to extract features using the ComParE feature set [11] as used in the 2013 Interspeech Computational Paralinguistics challenge. This delivers 6373 features computed as summative functionals over 126 low level signal features computed every 10ms. These features were normalized either by computing z-scores or by brute force Gaussianisation. Gaussianisation was performed by the bestNormalize package [12] in 'R', which maps the rank of each value into a sample from a cumulative gaussian pdf. Both types of normalization were performed as part of cross-validation, such that only the training data in each fold were used to define the normalizing transform.

For the spectrographic representation used with the CNN methods, the signal was pre-emphasised, divided into 50ms hamming windowed segments stepped by 4ms, and a 32768-point FFT was computed. The amplitude spectrum was then interpolated onto a log frequency scale between 62.5Hz and 8000Hz. Next the amplitude was converted to deciBels and the first two time derivatives computed. The amplitude spectrum was then floored to -50dB from the maximum amplitude and z-score normalisation was applied over all amplitude values and deltas. Finally the spectrogram was packed into a 224x224x3 image format. The red channel contained the amplitude, and the green and blue channels contained the first and second time derivatives. If the audio

Table 2 *Summary of SVM classifier performance. Baseline=All 6373 features, gender independent, z-score normalisation, allow multiple recordings per speaker, radial basis function kernel.*

UAR %	**All Pathologies**			**Organic Pathologies Only**			
Condition	**Vowel**	**IAU**	**Phrase**	**Vowel**	**IAU**	**Phrase**	**Average**
Baseline	67.93	81.93	79.84	73.49	84.58	85.66	78.91
+Best 1000 features	**69.74**	80.24	80.54	73.38	83.75	86.68	79.06
+Gender dependent	68.40	80.11	78.93	**76.44**	82.93	85.20	78.67
+Gaussianisation	69.30	81.86	**80.71**	74.08	**85.69**	**86.73**	**79.73**
+Polynomial kernel	68.33	**82.30**	80.38	72.67	84.15	86.18	79.00
+Single recording	69.05	82.15	80.36	72.85	83.59	84.86	78.81

recording contained fewer than 224 frames, then the image was padded with silence. If the audio recording was greater than 224 frames, then the central 224 frames were selected. Analysis parameters were chosen to ensure that a 1s recording fitted into 224 pixels. Examples of spectrographic images are shown in Figure 1.

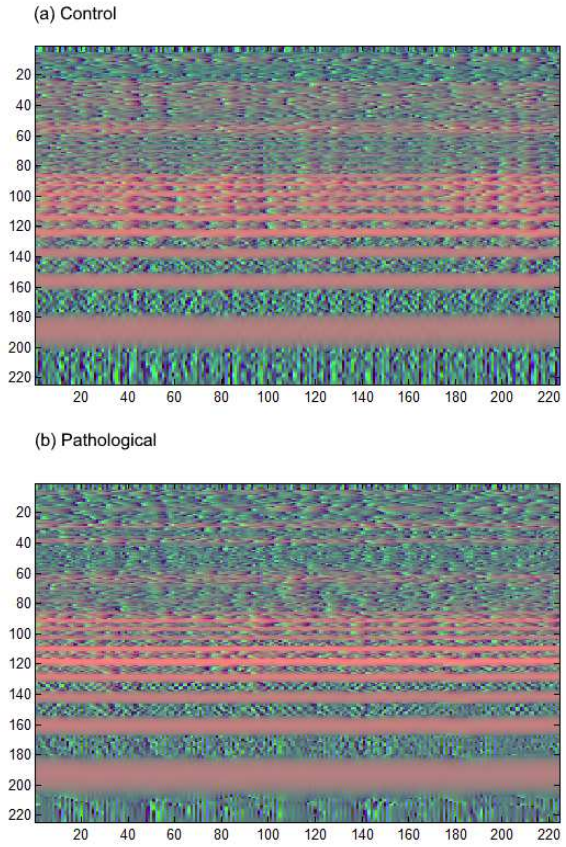

Figure 1 *Spectrographic representation used for CNN methods*

2.4. Feature selection

For the SMILE features, we also tested the value of a simple feature selection method. This is based on calculating the F-ratio for each feature, that is the ratio of variance of the feature values across all data to variance within the separate control and pathology subsets. Features with the largest F-ratio vary most between control and pathology classes, and the best N features can be selected for classification. We performed feature selection as part of cross-validation, such that for each fold features are selected from the training set alone.

2.5. Classifier construction

For the Support Vector Machine (SVM) classifier, this was implemented in R using the e1071 package [13]. Both radial basis function and polynomial kernels were tested, and a simple grid-search established to find optimum values for the cost parameter and (for the polynomial) the degree and constant parameters. The gamma parameter was set to the reciprocal of the number of features.

For the CNN classifier replicated from [7], a six layer CNN network was constructed in the Keras toolkit [14]. Input to the lower convolutional layers was the spectrographic image format described in 2.3. Convolutional kernels were set to 8x8, with a stride of 1 in layers 1 and 2, and set to 2 and 4 in layers 3 and 4. Max pooling with a kernel of 4x4 and a stride of 1 was placed between convolutional layers. The flattened output of the last convolutional layer was sent to a dense layer of 16 units and a single output unit. This is a slightly simplified implementation of the CNN used in the original study.

For the CNN classifier replicated from [8], the pre-trained implementation of RESNET50 available in the Keras toolkit was used. This CNN is designed for image classification and contains a large number of convolutional, pass-through and pooling layers. The highest layers of the network are removed exposing a globally pooled layer of 2048 units. To this are added a dense layer of 16 units and a single output unit with a sigmoid activation delivering the probability of a pathological sample.

Both CNN networks were trained using binary cross-entropy and the adam optimizer with a learning weight of 0.0001. Training is performed for 25 epochs, with 10% of the training data held out for validation and a batch size of 32. Sample weights were used to compensate for class imbalance. The training epoch that delivered the highest area-under-curve (AUC) on the validation data was chosen for testing.

2.6. Cross-validation and performance statistic

Both SVM and the CNNs were evaluated using five-fold cross-validation. Assignment of recordings to the cross-validation fold was done on the basis of speaker number, to ensure that the same speaker did not appear in more than one fold.

Performance is reported as unweighted average recall (UAR), which is just the average of the accuracies in correctly labelling the control recordings and pathological recordings. This would be the accuracy of the system if the test data had equal numbers of normal and pathological cases.

3. Results

3.1. Replication of SVM experiment

For evaluation of the SVM method with the SMILE features, a number of test variants were compared against a baseline configuration. For the baseline, all of the 6373 SMILE features were used, with a single model for both genders, using z-score normalisation, and a radial basis function kernel, using all recordings, including multiple recordings per speaker. Variations included the use of the best 1000 features, separate models for male and female speakers, Gaussianisation normalisation, use of the degree 2 polynomial kernel, and single recordings per speaker were evaluated. A summary of test scores after cross validation is shown in Table 2.

3.2. Replication of CNN experiments

For evaluation of the two CNN methods, only the vowel recordings were used, with the input being the spectrographic image as described in 2.3. We explored many variations of training data augmentation to reduce over-training. The results shown in Table 3 were derived from training with random gaussian noise with sd=0.05 added to the input patterns, together with random multiplicative scaling between 0.8 and 1.25.

Table 3 *Summary of CNN classifier performance based on single vowel compared to published accuracy.*

UAR%	All Pathologies	Organic Pathologies	Published
CNN	67.81	69.71	77
RESNET pre-trained weights	69.27	70.87	94

4. Discussion

This paper has looked at re-implementing three previous systems for detecting disordered voice using materials from the SVD. These have been evaluated on common subsets of the database, and using the same protocol for cross-validation, to try and generate results that can be compared against one another.

For the SMILE+SVM method, results show that performance is better on the organic pathology subset compared to all pathologies. Also performance using all vowels from a speaker are better than when only one vowel per speaker is used. Performance from the short phrase alone is as good as the performance from all vowels and pitches. Overall the results match the results published in [9], with about 85% accuracy for an organic pathology subset from all vowels.

We found that feature selection, changing to a gender dependent classifier, Gaussian normalisation, or changes to the SVM kernel had very little effect on performance. No one combination gave the best performance in all test conditions. The variant providing the best value was the use of a gender-dependent system for classification of a single vowel. When the method was limited to operating from one recording per speaker, performance decreased slightly. It is not possible to tell whether this is due to the removal of repeated speakers in testing, or because of a reduction in the overall amount of data available for training.

The two CNN methods gave very similar performance with each other and with the SVM methods from the single vowel on all pathologies. The CNN methods improved in performance on the organic pathology subset, but not as much as the SVM method. This may be because these methods are more severely affected by the reduction in number of training samples. The performance of our CNN re-implementation is slightly worse than that published in [7]. This may be due to differences in the spectrographic format and number of network layers and training protocol, as well as differences in pathology subset and cross-validation. The performance of our RESNET re-implementation is significantly worse than that published in [8]. In training the RESNET architecture, we found over-training to be a big problem since there are over 23M parameters in the network and only ~2000 training samples. It may be that the authors of [8] had particular ways to train their network to avoid this. Visual inspection of the spectrographic images used as input to the CNN methods (shown in Fig.1) do not show clear differences between control and pathological samples. Thus it is not surprising that the CNN methods proved difficult to train.

Taking our results overall, there is agreement that voice disorder can be recognized with about 70% accuracy from a single vowel if all pathologies are included – *independently of the classifier algorithm*. Using all the recorded vowels or using the short phrase, this accuracy rises to about 80% with the SVM classifier. If the pathologies are limited to only organic pathologies, then accuracy rises by about 5% for the SVM classifier, but less than 2% for the CNN classifiers.

In this work we have set out a particular way of using the SVD for training and evaluating automated methods for the detection of voice disorders. The fact that three very different approaches achieve similar performance once evaluated using the same pathologies and cross-validation protocol shows that these aspects are essential when comparing different studies. We hope that future studies will copy the approach set out here. Please contact the first author for details of the organic pathology subset.

There are clearly many opportunities for further work on the SVD to establish in more detail which acoustic properties of the recordings are most useful in both detecting disorder and discriminating between different types of disorder. In particular the EGG recordings seem very underexploited, although some preliminary analysis is available in [6]. Fundamentally, it is yet to be established whether the EGG waveforms in the SVD contain information that aids detection or diagnosis of disorder that is not present in the audio recording. If EGG waveforms do contain additional information, then a subsequent question would be whether a system which imputed the EGG waveform from the Speech signal would also extract that information. There is recent work in [15] which goes in this direction.

5. Acknowledgements

The authors would like to thank those involved in the creation of the Saarbrücken Voice Database, without which this study would not have been possible.

6. References

[1] M. Putzer, W. Barry, "Saarbrucken Voice Database", Institute of Phonetics, Univ. of Saarland, Accessed March 2021 from http://www.stimmdatenbank.coli.uni-saarland.de/ .

[2] M. Putzer, J. Koreman, "A German database of pathological vocal fold vibration," *Phonus 3* Institute of Phonetics, University of the Saarland, 1997, 143-153.

[3] S. Hedge, S. Shetty, S. Rai, T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders", *J. Voice*, 33 2019.

[4] D. Hemmerling, A. Skalski, J. Gajda, "Voice data mining for laryngeal pathology assessment", *Computers in Biology and Medicine* 69 2016.

[5] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, M. Farahat, K. Malki, "An investigation of Multidimensional Voice Program parameters in three different databases for voice pathology detection and classification*". J. Voice* 31 2017.

[6] G. Muhammad, M. Alhamid, M. Shamim Hossain, A. Almogren, A. Vasilakos, "Enhanced living by assessing voice pathology using a co-occurrence matrix", *Sensors* 17 2017.

[7] H. Wu, A. Lowit, "A deep learning method for pathological voice detection using convolutional deep belief networks", *Interspeech* 2018.

[8] M. Mohammed, K. Abdulkareem, S. Mostafa, M. Khanapi Abd Ghani, M. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, F. Taha Al-Dhief, "Voice Pathology Detection and Classification Using Convolutional Neural Network Model", *Applied Sciences* 10 (2020)

[9] P. Barche, K. Gurugubelli, A. Kumar Vuppala, "Towards automatic assessment of voice disorders: A clinical approach", *Interspeech 2020*, Shanghai, China.

[10] F. Eyben, M. Wollmer, B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[11] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi et al., "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, 2013.

[12] R. Peterson, "bestNormalize: Normalizing transformation functions", Version 1.7.0, Retrieved March 2021 from https://cran.r-project.org/web/packages/bestNormalize/

[13] D. Meyer et al, "E1071: Miscellaneous functions of department of statistics TU Wien", Version 1.7-6, Retrieved March 2021 from https://cran.r-project.org/web/packages/e1071/

[14] F. Chollet, et al, "Keras" Retrieved March 2021 from https://github.com/fchollet/keras

[15] I. Howard, J. McGlashan, A. Fourcin, "Machine learning analysis of speech and EGG for the diagnosis of speech pathology", *Proc. Electronic Speech Signal Processing Conference ESSV*, 2021.