VPS VIBRATIONS IN PHYSICAL SYSTEMS

DOI: 10.21008/j.0860-6897.2022.3.07
Vibrations in Physical Systems, 2022, 33(3), 2022307

# GFCC-based x-vectors for Reinke's Edema Detection

**Katarzyna KOTARBA**

AGH University of Science and Technology, Department of Mechanics and Vibroacoustics,
al. Mickiewicza 30, 30-059 Krakow, Poland

**Corresponding author:** Katarzyna KOTARBA, email: urbaniec@agh.edu.pl

**Abstract** Automatic assessment of voice disorders is one of the most important applications of speech signal analysis. Various algorithms utilizing both sustained vowels and continuous speech have been successfully used to perform detection of many voice pathologies, e.g. dysphonia, laryngitis, and vocal folds paralysis. However, algorithms described in literature used for classification of Reinke's edema – one of the most severe smoking-induced voice conditions – are scarce and rely mostly on speech signals containing sustained vowels. In this paper, a method incorporating gammatone frequency cepstral coefficients (GFCC) based x-vectors extracted from continuous speech is presented. The extracted x-vectors are used to train a SGD classifier performing Reinke's edema detection. For validation folds, the proposed method yielded AUC ROC, accuracy, recall, and specificity of 0.96 (±0.03), 0.94 (±0.02), 0.92 (±0.03), and 0.94 (±0.02), respectively. For testing set, the method yielded AUC ROC, accuracy, recall, and specificity of 0.98, 0.89, 0.88, and 0.89, respectively.

**Keywords:** x-vectors, Reinke's edema, voice pathology classification.

## 1. Introduction

Voice pathology detection is one of the most crucial applications of machine learning methods in medical acoustics. Recent epidemiological situation caused increasing need for telemedicine technologies and solutions helping healthcare workers with diagnosing patients remotely. In case of laryngeal pathologies such a solution may be based on automatic speech signal analysis and classification.

Many attempts have been made to classify speech signals into two categories: being uttered by healthy person or person suffering from laryngeal pathology. Most of the studies did not focus on any specific disease and presented methods capable of differentiating healthy and sick people. Hemmerling et al. presented method based on random forest classifier and sustained vowel /a/ obtaining classification accuracy rate as high as 100% [1]. Vasilakis and Stylianou [2] performed classification based on short-term jitter extracted from continuous speech and yielded area under ROC curve of 0.88. Cordeiro et al. [3] proposed a classifier used for detection of vocal fold edema and unilateral vocal folds paralysis based on 12-dimensional mel-frequency cepstral coefficients (MFCC) and Gaussian Mixture Models (GMM) obtaining the accuracy rate of 74%. All these studies used German or English speech corpora. It is worth noting, however, that some studies incorporated speech corpora containing utterances in other languages are also available, e.g. Wszołek et al. [4] proposed a method based on Self-Organizing Maps (SOM) incorporating Polish speech samples uttered by healthy children and children suffering from palatoschisis (a cleft palate). Another example of Polish speech corpus usage is work presented by Engel et al. [5], in which method of pathological speech assessment based on zero-crossing rate (ZCR), mel-frequency cepstral coefficients (MFCC), and short-time Fourier transform features extracted from Polish speech is proposed.

Some of the abovementioned studies used speech corpora, in which speech of patients diagnosed with Reinke's edema – a voice condition occurring in smokers – were present, e.g. [1, 4]. However, only one study focusing specifically on Reinke's edema detection can be found in literature. Madruga et al. [6] presented method based on acoustic features, e.g. ZCR, shimmer, jitter, and MFCC extracted from recordings of sustained vowel /a/ included in Massachusetts Eye and Ear Infirmary (MEEI) Voice Disorders Database [7] and in unshared database created by the authors. The authors reported classification accuracy exceeding 90% for MEEI dataset and almost 90% for self-created dataset.

The main objective of this study is to assess applicability of x-vectors – fixed-length speaker embeddings used usually for speaker recognition – to Reinke's edema detection based on continuous speech. The main advantage of x-vectors in comparison to other frequently used acoustic features, e.g formant frequencies, jitter, and shimmer is their repeatability – according to Stegmann et al. [8], repeatability of these acoustic features is poor, and their usage may therefore lead to obtaining misleading classification results. Moreover, x-vectors have been proven to be suitable for application in diagnosing both voice pathologies [9], and other diseases, e.g. Parkinson's disease [10, 11].

The rest of this paper is organized as follows. In section 2 the speech corpus used in this study, data pre-processing, feature extraction and models' training are presented. In section 3 obtained results are reported and discussed. In section 4 the work is concluded, and future perspectives are discussed.

## 2. Proposed method

### 2.1. Saarbruecken Voice Database

The dataset used in this study is the Saarbruecken Voice Database [12] – a German speech corpus containing speech samples from healthy people and patients suffering from various otholaryngeal pathologies. The corpus consists of recordings of the phrase '*Guten Morgen, wie geht es Ihnen?*' and sustained vowels /a/, /i/, and /u/. Since the objective of this study is to investigate methods of Reinke's edema detection based on continuous speech, only the recordings of the phrase uttered by patients diagnosed with this disease and healthy people were used.

Recordings selected for further analysis were divided into two stratified sets: training set consisting of 80% of available audio files, and testing set consisting of remaining 20% of the files. The details on the datasets are listed in Tab. 1.

**Table 1.** Details of the two subsets of the speech corpus used for models' training and evaluation: training set and testing set.

| | Number of utterances | |
|---|---|---|
| Set | Healthy | Sick |
| Training | 507 | 31 |
| Testing | 127 | 8 |

### 2.2. Data pre-processing and feature extraction

The audio files containing continuous speech of healthy people and patients diagnosed with Reinke's edema were pre-processed and used for 30-dimensional gammatone frequency cepstral coefficients (GFCC) extraction (Fig. 1) following the recipe described in author's previous work – for details, see [9]. The GFCC features were then fed to x-vectors extractor based on time-delay neural network (TDNN) architecture provided by Kumar et al. (Fig. 2) [13]. However, the loss function used during neural network's training was changed from *softmax* function used in [9, 13] to additive angular margin loss (Eq. 1) introduced by Deng et al. [14], leading to enhancement of intra-class variance and inter-class similarity. The x-vectors extractor was trained on VoxCeleb 2 dataset [15], following the recipe described in [9].

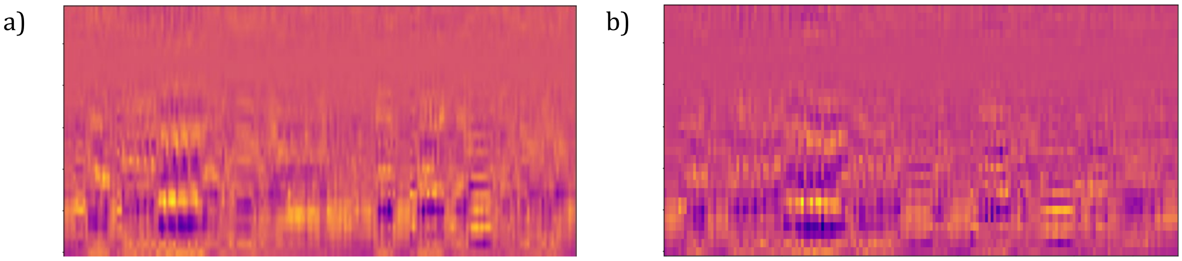

**Figure 1.** Sample GFCC features extracted from phrase '*Guten Morgen, wie geht es Ihnen?*' uttered by a) healthy person, b) person diagnosed with Reinke's edema.

$$L\big(y_i,\ f(x_i)\big) \ = \ -\frac{1}{N}\sum_{i=1}^{N}\log\ \frac{\exp\Big(s\big(\cos(\theta_{y_i}+m)\big)\Big)}{\exp\Big(s\big(\cos(\theta_{y_i}+m)\big)\Big)+\sum_{j=1,\,j\neq y_i}^{n}\exp\Big(s\big(cos\theta_j\big)\Big)} \qquad (1)$$

where $y_i$ denotes true label of $i$-th sample, $f(x_i)$ stands for model's prediction based on feature $x_i$, $\theta_j$ denotes the angle between the $j$-th column of the weights' matrix and feature $x_i$, $m$ stands for margin penalty between

$x_i$ and the $j$-th column of the weights' matrix, and $s$ denotes a radius of the hypersphere, on which the learned speaker embedding features are distributed.

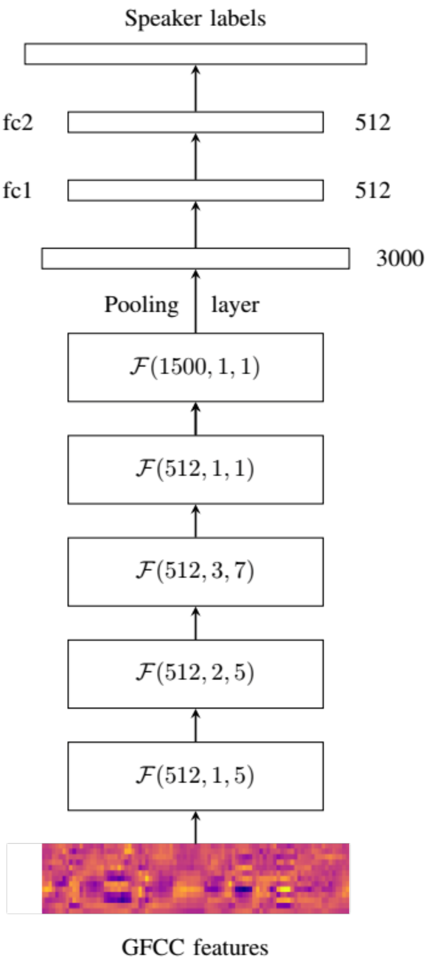

**Figure 2.** Overview of architecture of neural network used for x-vectors extraction (based on [13]). F(N, D, K) denotes time-delay layer, where N stands for output embedding dimension, D stands for dilatation, and K stands for context width used during calculation of layer activations. fc1 and fc2 are fully connected layers, from which the x-vectors can be extracted.

### 2.3. Models training

Four linear models fitted by minimizing a regularized empirical loss with stochastic gradient descent (SGD) algorithm implemented in Python's scikit-learn library [16] were trained on x-vectors described above using a 5-fold cross-validation. Each model used different loss function, namely:

- epsilon-insensitive loss:

$$L\big(y_i, f(x_i)\big) = \max(0, |y_i - f(x_i)| - \varepsilon), \tag{2}$$

- Huber loss:

$$L\big(y_i, f(x_i)\big) = \begin{cases} \varepsilon|y_i - f(x_i)| - \frac{1}{2}\varepsilon^2 when|y_i - f(x_i)| > \varepsilon \\ \frac{1}{2}\big(y_i - f(x_i)\big)^2 when|y_i - f(x_i)| \leq \varepsilon \end{cases}, \tag{3}$$

- squared loss:

$$L\big(y_i, f(x_i)\big) = \frac{1}{2}\big(y_i - f(x_i)\big)^2, \tag{4}$$

- squared epsilon-insensitive loss:

$$L\big(y_i, f(x_i)\big) = \max\left(0, \frac{1}{2}\big(y_i - f(x_i)\big)^2 - \varepsilon\right), \tag{5}$$

where $y_i$ denotes true label of $i$-th sample, $f(x_i)$ stands for model's prediction based on feature $x_i$, and $\varepsilon$ stands for the threshold used during model's validation – if model's prediction differs from true label by less than ε, the difference between them is ignored [16].

Moreover, feature dimensionality reduction technique called truncated singular value decomposition (tSVD) [17] was applied to the feature matrix to reduce model's overfitting and training time. Number of components used during model's training and model's hyperparameters were optimized using Optuna framework [18].

## 3. Experimental results

Classification models are usually evaluated using accuracy, recall, precision, and F1 score, as they are considered to reliably assess model's performance. However, some of these metrics are not easily understandable to people not specializing in machine learning algorithms. Since presented study regards application of SGD classifier to classification of medical data, metrics recommended for presenting classification results to clinicians should be provided instead [19]. Based on recommendation of Sidey-Gibbons [20], recall (also known as sensitivity, Eq. 6), specificity (Eq. 7), accuracy (Eq. 8), and area under a ROC curve (Eq. 11) were chosen to evaluate proposed models' performance [21, 22]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}, \tag{6}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}}, \tag{7}$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}, \tag{8}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}, \tag{9}$$

$$\text{ROC}(\cdot) = \{\text{FPR}(c),\ \text{Recall}(c),\ c \in (-\infty, +\infty)\} = \big\{\big(t, \text{ROC}(t)\big), t \in (0,1)\big\}, \tag{10}$$

$$\text{AUC ROC} = \int_0^1 \text{ROC}(t)dt, \tag{11}$$

where TN denotes true negative, TP denotes true positive, FN denotes false negative, FP denotes false positive, ROC stands for receiver operating characteristic curve constructed as a plot of recall versus false positive rate (FPR). The ROC function maps $t$ to recall(c), and $c$ stands for the cut-off value corresponding to FPR(c) = t. The area under a ROC curve (AUC ROC) can be interpreted as the probability that the binary classifier will yield a higher value for a randomly chosen positive instance than for a randomly chosen negative instance [21].

The AUC ROC scores obtained by the trained models are listed in Tab. 2. The model with the highest score was chosen for further evaluation by abovementioned binary metrics, i.e. accuracy, recall, and specificity. However, it is worth noting, that the differences between models' performance were minor: the best model, i.e. model using Huber loss function, yielded AUC ROC of 0.98, while the worst model, i.e. model trained with epsilon-insensitive loss function, yielded AUC ROC of 0.95. It can be therefore assumed, that loss function influence on the model's performance is not significant in this particular case.

The best model yielded promising results – as can be seen in Tab. 3, all the performance metrics exceed 0.9 for validation folds. What is more, the standard deviation of performance metrics values obtained for validation folds are small and do not exceed 0.02, which implies model's stability. The model performed slightly worse on the testing set – in this case the metrics reached values ranging from 0.88 to 0.89. Nevertheless, the small difference between AUC ROC obtained for testing set and validation folds proves model's good generalization ability.

**Table 2.** AUC ROC scores yielded for testing set by evaluated SGD models. The best results obtained by evaluated models are boldface.

| Loss function | AUC ROC |
|---|---|
| Epsilon-insensitive (Eq. 1) | 0.95 |
| Huber (Eq. 2) | **0.98** |
| Squared (Eq. 3) | 0.97 |
| Squared epsilon-insensitive (Eq. 4) | 0.96 |

**Table 3.** Classification performance metrics obtained for validation folds and testing set by proposed SGD model. The standard deviation of each metric yielded for validation folds is reported in the brackets.

| Metric | Validation folds | Testing set |
|---|---|---|
| Accuracy | 0.94 (± 0.02) | 0.89 |
| Recall | 0.92 (± 0.03) | 0.88 |
| Specificity | 0.94 (± 0.02) | 0.89 |
| AUC ROC | 0.96 (± 0.03) | 0.98 |

Unfortunately, it is not possible to directly compare the results presented in this study with results provided in the literature. There is only one paper regarding Reinke's edema detection available and the method presented in it is based on sustained vowel /a/ and incorporates different speech corpus [6]. Moreover, Madruga et al. provided only the results obtained during the cross-validation process, while in this study the results obtained for testing set are also presented. According to Rao et al. [23] and Tabe-Bordbar et al. [24], using both cross-validation approach and isolated testing set allows better understanding of model's generalization abilities and is the recommended approach to classification results analysis. Using only n-folds cross-validation may provide misleading results and lead to overestimation of model's performance.

## 4. Conclusions

In this study a method of automatic Reinke's edema detection based on x-vectors extracted from continuous speech is presented. The method incorporates a linear classifier optimized by stochastic gradient descent (SGD) algorithm. The x-vectors – speaker embeddings suitable for classification of speech signals differing in length and content – were based on 30-dimensional GFCC features extracted from German speech samples contained in Saarbruecken Voice Database. Four loss functions were used during model's training and optimization, namely Huber loss, squared loss, epsilon-insensitive loss, and squared epsilon-insensitive loss, leading to comparable models' performance – all models yielded AUC ROC scores in range of 0.95-0.98, the highest being obtained by classifier with Huber loss. The results obtained by the best model are promising – mean accuracy of 0.94, mean recall of 0.92, and mean specificity of 0.94 on validation folds and accuracy of 0.89, recall of 0.88, and specificity of 0.89 on the testing set were yielded. These values are comparable with values presented in literature - Madruga et al. Reported reaching accuracy of approximately 0.98 for Massachusetts Eye and Ear Infirmary Database (MEEI) [7] and approximately 0.88 for self-created database. It is worth noting, however, that in both studies assessment of model's performance was carried out using different approaches – while Madruga et al. presented only mean accuracies yielded for validation folds, model proposed in this study was assessed using both values obtained for validation folds and for isolated testing set, the latter being considered to be of higher importance.

Even though the results presented in this study are satisfactory, it is worth noting that the dataset used is small and further analysis using more data, e.g. MEEI dataset, should be performed to assess their significance. Additional analyses, e.g. carried out in gender-dependant scenario, could also provide a valuable insight into Reinke's edema detection problem.

## Acknowledgements

Research financed by a grant of the European Union Founding Project POWR.03.05.00 00 Z307/17 00, hold on the Faculty of Mechanical Engineering and Robotics in the AGH University of Science and Technology, Krakow, Poland.

## Additional information

The author(s) declare: no competing financial interests and that all material taken from other sources (including their own published works) is clearly cited and that appropriate permits are obtained.

## References

1. D. Hemmerling, A. Skalski, J. Gajda; Voice data mining for laryngeal pathology assessment; Computers in Biology and Medicine 2015, 69, 270-276. DOI: doi.org/10.1016/j.compbiomed.2015.07.026
2. M. Vasilakis, Y. Stylianou; Voice pathology detection based on short-term jitter estimations in running speech; Folia phoniatrica et logopaedica: official organ of the International Association of Logopedics and Phoniatrics (IALP) 2009, 61(3), 153–70. DOI: 10.1159/000219951
3. H. Cordeiro, C. Meneses, J. Fonseca; Continuous speech classification systems for voice pathologies identification; IFIP Advances in Information and Communication Technology 2015, 450, 217–224.
4. W. Wszołek, A. Izworski, G. Izworski; Signal processing and analysis of pathological speech using artificial intelligence and learning systems methods; Acta Physica Polonica A 2013, 123(6), 995-1000.
5. Z. W. Engel, M. Kłaczyński, W. Wszołek; A vibroacoustic model of selected human larynx diseases; International Journal of Occupational Safety and Ergonomics 2007, 13(4), 367–379. DOI: 10.1080/10803548.2007.11105094
6. M. Madruga, Y. Campos-Roca, C. K. Pérez; Robustness Assessment of Automatic Reinke's Edema Diagnosis Systems; Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 2020.
7. Massachusetts Eye and Ear Infirmary; Voice disorders database, ver. 1.03. Kay Elemetrics Corp., Lincoln Park, New Jersey, USA, 1994.
8. G. Stegmann et al.; Repeatability of Commonly Used Speech and Language Features for Clinical Applications; Digital Biomarkers 2020, 4(3), 109–122. DOI: 10.1159/000511671
9. K. Kotarba, M. Kotarba; Voice pathology assessment using x-vectors approach; Vibrations in Physical Systems 2021, 32(1), 2021108. DOI: 10.21008/j.0860-6897.2021.1.08
10. L. Jeancolas et al.; X-vectors: New quantitative biomarkers for early Parkinson's disease detection from speech; Front.Neuroinform. 2021, 15, Article 578369. DOI: 10.3389/fninf.2021.578369
11. C. Botelho et al.; Pathological speech detection using x-vector embeddings; arXiv: 2003.00864 [eess.AS] 2020. DOI: 10.48550/arXiv.2003.00864
12. W. Barry, M. Pützer; Saarbrücken voice database; Institute of Phonetics, Univ. of Saarland. Online. Available: http://www.stimmdatenbank.coli.uni-saarland.de
13. M. Kumar et al.; Designing neural speaker embeddings with meta learning; arXiv: 2007.16196 [eess.AS] 2020. DOI: 10.48550/arXiv.2007.16196
14. J. Deng et al.; ArcFace: Additive Angular Margin Loss for Deep Face Recognition; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, 2019.
15. J. S. Chung, A. Nagrani, A. Zisserman; Voxceleb2: Deep speaker recognition; Proceedings of the Interspeech, Hyderabad, India, 2018.
16. F. Pedregosa et al.; Scikit-learn: Machine learning in Python; Journal of Machine Learning Research 2011, 12, 2825–2830.
17. N. Halko, P.-G. Martinsson, J. A. Tropp; Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions; SIAM Review 2011, 53(2), 217-288. DOI: 10.1137/090771806
18. T. Akiba et al.; Optuna: A next-generation hyperparameter optimization framework; Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining, Anchorage, Alaska, USA, 2019.
19. J. Olczak et al.; Pesenting artificial intelligence, deep learning, and machine learning studies to clinicians and healthcare stakeholders: an introductory reference with a guideline and a Clinical AI Research (CAIR) checklist proposal; Acta Orthopaedica 2021, 92(5), 513-525. DOI: 10.1080/17453674.2021.1918389
20. J. Sidey-Gibbons, Ch. Sidey-Gibbons; Machine learning in medicine: a practical introduction; BMC Med. Res. Methodol., 2019, 19, Article 64. DOI: 10.1186/s12874-019-0681-4
21. T. Fawcett; An introduction to ROC analysis; Pattern Recognition Letters, 2006, 27(8), 861–874. DOI: 10.1016/j.patrec.2005.10.010
22. C. Calì, M. Longobardi; Some mathematical properties of the ROC curve and their applications; Ricerche di Matematica 2015, 64(2), 391-402.

23. R. Rao, G. Fung; On the dangers of cross-validation. An experimental evaluation; Proceedings of the SIAM International Conference on Data Mining, Atlanta, Georgia, USA, 2008.
24. S. Tabe-Bordbar et al.; A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models; Scientific Reports 2018, 8, Article 6620. DOI: 10.1038/s41598-018-24937-4