INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-class Detection of Pathological Speech with Latent Features: How does it perform on unseen data?

Dominik Wagner[1], *Ilja Baumann*[1], *Franziska Braun*[1], *Sebastian P. Bayerl*[1], *Elmar Nöth*[2], *Korbinian Riedhammer*[1], *Tobias Bocklet*[1,3]

[1]Technische Hochschule Nürnberg Georg Simon Ohm, Germany
[2]Friedrich-Alexander-University Erlangen-Nürnberg, Pattern Recognition Lab, Erlangen, Germany
[3]Intel Labs

firstname.lastname@th-nuernberg.de

Abstract

The detection of pathologies from speech features is usually defined as a binary classification task with one class representing a specific pathology and the other class representing healthy speech. In this work, we train neural networks, large margin classifiers, and tree boosting machines to distinguish between four pathologies: Parkinson's disease, laryngeal cancer, cleft lip and palate, and oral squamous cell carcinoma. We show that latent representations extracted at different layers of a pre-trained wav2vec 2.0 system can be effectively used to classify these types of pathological voices. We evaluate the robustness of our classifiers by adding room impulse responses to the test data and by applying them to unseen speech corpora. Our approach achieves unweighted average F1-Scores between 74.1% and 97.0%, depending on the model and the noise conditions used. The systems generalize and perform well on unseen data of healthy speakers sampled from a variety of different sources.
Index Terms: pathological speech, latent representations, multi-class, classification

1. Introduction

Latent features such as GMM-supervectors [1], i-vectors [2] and x-vectors [3] have been found useful for the analysis of pathologies from speech samples. Bocklet et al. use GMM-supervectors for the assessment of various speech pathologies [4], for intelligibility evaluation of laryngeal cancer patients [5], and for automatic ratings of Parkinson's patients [6]. Laaridh et al. use i-vectors for automatic intelligibility ratings of dysarthric speech [7] and head and neck cancer patients [8]. In [9], x-vectors are used to classify between patients with Parkinson's disease and a healthy control group. In [10], x-vectors obtained from speech samples of laryngeal cancer patients are used as regression model inputs to predict perceptual rating scores.

Latent representations obtained from wav2vec 2.0 [11] (W2V2) models have been successfully applied in dementia, dysfluency, and vocal fatigue detection [12, 13, 14]. These studies indicate that W2V2 embeddings are not only well suited to encode speaker and language characteristics, but also general characteristics of atypical speech. However, recent work has called into question, whether it is possible to mix different corpora under varying recording conditions for the detection of multiple pathologies in a single system [15]. In [15], models are trained to discriminate between six different healthy speech corpora. The strong classification results led to the conclusion that a large portion of the models' predictive power can be attributed to feature sets encoding information about recording conditions and speaker demographics.

In this paper, we continue to address these issues by carefully matching the control group to the speech pathologies in terms of speaker demographics such as age and gender distribution, as well as recording conditions. Furthermore, we account for class imbalances by applying an over-sampling algorithm to the minority classes. To further evaluate the robustness of our approach, we experiment on reverberated speech data and apply the trained models to unseen speech corpora. We utilize latent speech representations extracted from a pre-trained wav2vec 2.0 encoder as inputs to large margin classifiers, tree-based methods, and neural networks.

Our contributions are: We are the first to use W2V2 embeddings for the automatic detection of voice and speech pathologies in a multi-class setting. We address the issue of involuntarily encoding information about recording conditions and speaker demographics in cross-corpora studies. We conduct systematic robustness tests on different W2V2 representations (at various layers), noise conditions, and unseen speech corpora. Finally, we demonstrate that multi-class detection performs similar to binary classification on each individual corpus and can even lead to more robust results on unseen data.

2. Data

We use five different speech corpora to train our classifiers. Four corpora contain recordings of patients with different voice and speech pathologies, three of which also include matching control groups. These corpora are briefly introduced in Section 2.1. The full control group for this study is comprised of the control groups from the three corpora containing recordings of healthy subjects and an additional corpus of 110 healthy speakers in an age cohort similar to three of the pathological speech corpora. All utterances in our training data are spoken by native German speakers. The speech of some subjects exhibits strong forms of local dialects. Four additional datasets, unrelated to the training data, are used to evaluate the robustness of the proposed approach w.r.t. variations in recording conditions and speaker demographics. These datasets are briefly described in Section 2.3.

2.1. Pathological speech corpora

Laryng41: The tracheoesophageal (TE) substitute voice is a common treatment to restore the patient's ability to speak after laryngectomy, i.e., the removal of the entire larynx. The Laryng41 (LAR) [16] corpus is a collection of tracheoesophageal speakers, reading the German version of the "The North Wind and the Sun" (NWS) text passage [17], a phonetically rich text that is widely used in speech therapy. The corpus contains 41 laryngectomees ($\mu = 62.0 \pm 7.7$ years old, 2 female and 39 male) with TE substitute voice.

Oral squamous cell carcinoma: Oral squamous cell carcinoma (OSCC) and its treatment impair speech intelligibility by alteration of the vocal tract. The OSCC dataset is comprised

10.21437/Interspeech.2023-464

of 71 patients ($\mu = 59.9 \pm 10.1$ years old, 16 female and 55 male) with OSCC in various stages. Tumors were located on the lower alveolar crest ($n = 27$), tongue ($n = 23$), and floor of the mouth ($n = 21$) [18]. The patients were recorded reading the German version of the NWS text passage 14-20 days after the tumor resection.

Parkinson's disease: Parkinson's disease (PD) is a degenerative disorder of the central nervous system. It arises from the death of dopamine-containing cells in a region of the midbrain. The full PD corpus contains 88 native German speakers diagnosed with PD ($\mu = 66.6 \pm 9.0$ years old, 44 female and 44 male) and a healthy control group comprised of another 88 speakers ($\mu = 58.1 \pm 14.2$ years old 43 female, 45 male) [6]. The dataset contains speech of different tasks (e.g. read text, read words, repetition of syllables, sustained vowels etc.). We selected the task of reading a phonetically rich text for the experiments in this work.

Erlangen-CLP: The Erlangen-CLP corpus [19] is a speech database of 818 children ($\mu = 8.7 \pm 13.3$ years old, 355 female and 463 male) with cleft lip and palate (CLP) and 380 age-matched control speakers ($\mu = 7.8 \pm 10.4$, 185 female and 195 male) who spoke the PLAKSS (Psycholinguistische Analyse Kindlicher Sprechstörungen) test. The PLAKSS test is a semi-standardized test that consists of words with all German phonemes in different positions and is used by speech therapists in German speaking countries. We use a subset consisting of 598 CLP speakers to reduce the dominance of the corpus relative to the rest of the data.

2.2. Healthy control corpora

The healthy control group (CTL) for our experiments is comprised of the control groups from the CLP and PD corpora, as well as a collection of 110 elderly ($\mu = 75.7 \pm 9.6$ years old, 79 female and 31 male) native German speakers reading the NWS text passage. We call this corpus AgedVoices110. The conditions in AgedVoices110 are similar to those in the LAR and OSCC corpora in several ways: the participants read the same phonetically rich text, belong to a similar age cohort (> 60 years old), the same recording equipment was used, and the recordings were made in the same hospital. Furthermore, the participants live in the same region of Germany and speak the same local dialect. By including the AgedVoices110 corpus, we add a substantial number of speakers similar to the LAR, OSCC, and PD corpora, thereby counterbalancing the 380 control speakers from the CLP corpus.

2.3. Additional corpora

Partial resection: The PR85 dataset [20] also consists of speech samples recorded after laryngeal cancer treatment. In this case, 85 patients ($\mu = 60.7 \pm 9.7$ years old, 10 female, 75 male), who underwent a partial resection (PR) that allowed the preservation of at least one vocal fold, were recorded reading the NWS text passage. Partial resection is a less invasive procedure than total laryngectomy. Consequently, the intelligibility of the speakers in the PR85 corpus is much higher. The lack of pathologic features in this corpus has been noted and analyzed in [21]. Therefore, we assign the healthy control group label to all utterances in the PR85 corpus in our experiments.

Tuda-De100: The Tuda-De100 data is comprised of 100 utterances from the Tuda distant speech corpus [22]. The corpus was recorded with 5 different microphones in parallel focusing on distant speech recognition. The speech data was collected in the same room with a 1 meter distance between speaker and microphone. Most speakers in the corpus are between 21 and 30 years old. We randomly selected a small subset of utterances from 10 speakers (5 female, 5 male) recorded with the Yamaha PSG-01S microphone. Each speaker reads 10 sentences from the German Wikipedia.

MLS100: The Multilingual LibriSpeech (MLS) dataset [23] is a large multilingual speech corpus of read audiobooks from LibriVox. The German portion of the dataset comprises approximately 3.3k hours of speech from 244 speakers. Similar to the Tuda-De100 data, we randomly selected a subset of 100 utterances from 10 speakers (5 female, 5 male), each reading 10 sentences.

NWS reading: NWS Reading (NWSR) is a small corpus of 8 native German speakers (1 female and 7 male) reading the NWS text passage multiple times throughout a period of approximately one year. Six speakers were more than 50 years old at the time of recording and two speakers were 12 and 23 years old.

3. Method

Pre-trained wav2vec 2.0 (W2V2) models are widely used as a general-purpose feature extractor for downstream tasks such as acoustic model training for automatic speech recognition (ASR). In our experiments, we use a model pre-trained on 960 hours of unlabeled speech from the LibriSpeech corpus [24], finetuned for ASR on the transcripts of the same data. We use the 768-dimensional intermediate representations after each of its 12 transformer blocks. The system provides a vector for every 20 ms of raw audio input. We extract those vectors for each utterance in our training and test corpora and compute the mean along the time axis, resulting in a single vector that represents each utterance.

The W2V2 embeddings serve as input to three different classes of models; large margin classifiers, tree-based methods, and neural networks. We choose Support Vector Machine (SVM), XGBoost (XGB), and a feedforward neural network with fully connected layers (FFN) for the task. The optimal hyperparameters for each estimator are determined in a 5-fold cross-validation on the training set using the grid search method.

SVMs are trained using radial basis function (RBF) kernels. The kernel parameter γ is selected from $\gamma \in \{10^{-k} \mid k = 5, \ldots, 1\} \subset \mathbb{R}_{>0}$, and the penalty parameter of the error term C is selected from $C \in \{5, 10, 20, 50\} \subset \mathbb{N}_{>0}$.

XGB models are trained using a tree booster as the underlying learner. The maximum tree depth parameter d is chosen from $d \in \{2^k \mid k = 1, \ldots, 4\} \subset \mathbb{N}_{>0}$. The learning rate η is chosen from the set $\eta \in \{\frac{k}{10} \mid k = 1, \ldots, 5\} \subset \mathbb{R}_{>0}$ and and the minimum sum of an instance weight w is chosen from $w \in \{2^k \mid k = 0, \ldots, 3\} \subset \mathbb{N}_{>0}$.

The FFN employs the Adam [25] optimizer with exponential decay rates of $\beta_1 = 0.90$, $\beta_2 = 0.99$ and a $\mathcal{L}_2$ regularization term of 10^{-4}. The FFN learning rate α is chosen from $\alpha \in \{10^{-k} \mid k = 1, \ldots, 4\} \subset \mathbb{R}_{>0}$. The activation function is either Tanh or ReLU, the number of hidden layers is either 2 or 3, and the number of hidden units h is selected from $h \in \{2^k \mid k = 5, \ldots, 8\} \subset \mathbb{N}_{>0}$.

3.1. Reverberation

We evaluate the robustness of our classifiers by convolving the speech signals with simulated room impulse responses (RIRs). We use the RIRs simulated with the technique described in [26] in the "medium room" setting (room width and length between

10m and 30m, room height between 2m and 5m). Each RIR is linearly scaled with a factor of 2 to make it more dominant.

3.2. Class imbalances

The number of available utterances for each pathology is imbalanced. LAR, PD, and OSCC account for 3.0%, 6.4%, and 5.1% of the overall data. The majority classes are the control group (42.0%) and CLP (43.5%). We employ the SMOTE [27] method in all our experiments to over-sample the minority classes, thereby generating a more balanced data distribution.

4. Experiments and results

We conduct our experiments using SVM, FFN, and XGB models for classification. First, we extract and aggregate the W2V2 features for each utterance in the speech corpora. The extracted embeddings are then divided into a training and test set with an 80/20 split ratio and passed as inputs to each of the three classification models. For the experiments on reverberated data, the RIRs are applied to the raw audio before they are passed to the wav2vec 2.0 model for embedding retrieval.

4.1. Visualization of latent representations

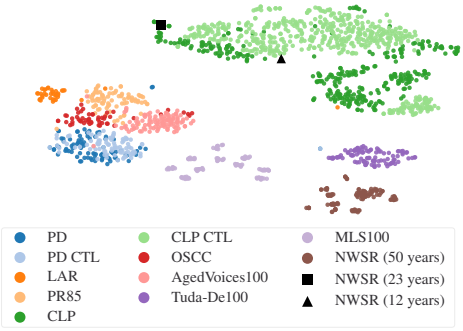

Figure 1: *2-dimensional t-SNE projection of 768-dimensional latent features extracted at W2V2 layer 4 for all datasets used in this study (perplexity = 30).*

Figure 1 depicts a t-distributed stochastic neighbor embedding (t-SNE) [28] projection of W2V2 features extracted at layer 4 for all datasets used in this study. The data is roughly separated into older and younger speakers. The CLP corpus, which contains children at an average age of $\mu = 8.7$ years, forms a group on the right. The pathological speakers (OSCC, LAR, PR85, and PD), as well as their corresponding control groups (PD CTL, AgedVoices110) are grouped on the left side of the figure. The closeness between these corpora can be explained by the similar age structure and the same read sentences, i.e., the NWS text passage. The older speakers in the NWSR corpus form a group of their own, whereas the two younger speakers from the same corpus can be found among the CLP data (black triangle and square in Figure 1). In this case, the age factor seemingly outweighed other decisive components, such as recording environment and reading task, for mapping the data. This indicates that a multitude of different speech- and speaker-related characteristics are encoded in each embedding.

4.2. Classification results

All three classifiers yield similar performance with minor differences depending on the W2V2 layer used for feature extraction. The lower right diagram in Figure 2 shows that the performance across all 12 W2V2 layers remains relatively stable in earlier layers and drops in layer 10-12. We assume that the use of multiple pathologies in a single classifier can have a strong regularizing effect, making the overall system less dependent on variations arising from features extracted at different W2V2 encoder layers. The confusion matrices in Figure 2 show that misclassifications occur mainly between CTL and PD.

Table 1: *Unweighted accuracy and F1-Scores for combinations of reverberated (REV) and clean (CLN) training- and test-sets. The column "Layer" indicates the W2V2 encoder layer at which the classifier yielded the best performance in terms of unweighted average F1-Score. The F1-Scores were balanced w.r.t. Precision and Recall. The rightmost column contains average and standard deviation of the unweighted average F1-Scores across all 12 W2V2 layers.*

#	Data		Model	Layer	Accuracy	F1-Score	
	Train	Test		(best)	(best layer)	Unw. Avg (best layer)	Avg±Std (all layers)
1	CLN	CLN	SVM	4	97.7	96.9	93.4 ± 4.6
			FFN	3	98.5	97.0	93.0 ± 4.8
			XGB	4	97.7	95.7	92.2 ± 4.0
2	CLN	REV	SVM	3	84.3	84.3	70.6 ± 10.8
			FFN	6	83.5	78.5	65.8 ± 9.1
			XGB	3	69.7	75.3	61.6 ± 8.8
3	REV	REV	SVM	1	98.5	95.7	89.1 ± 6.8
			FFN	1	97.3	93.5	86.4 ± 6.4
			XGB	1	96.9	93.5	85.1 ± 6.6
4	REV	CLN	SVM	8	86.6	76.0	68.0 ± 9.3
			FFN	8	89.7	80.8	66.6 ± 8.1
			XGB	7	78.5	74.1	60.1 ± 9.2

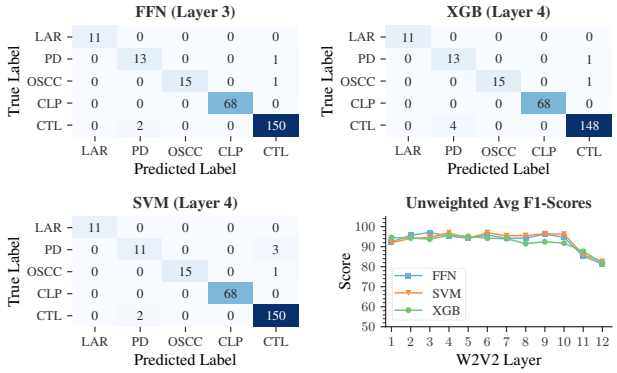

Figure 2: *Confusion matrices for the best-performing variants of all three classifiers and unweighted average F1-Scores w.r.t the 12 W2V2 layers.*

Table 1 shows a drop in performance, when models trained on clean (CLN) data are used to predict pathologies based on features obtained from reverberated (REV) utterances. For example, the unweighted average F1-Score for the FFN drops from 97.0% to 78.5% and the best performing layer shifts from 3 to 6. Testing CLN models on REV data also exhibits more variation in the results achievable by features extracted at different layers, as indicated by mean and standard deviation (see the rightmost column). However, the drop in performance can be mitigated by training and testing on REV data (see experiment #3).

The classification results of our models on the additional datasets described in Section 2.3 are summarized in Table 2. The classifiers perform best on MLS100 and NWSR. Correct

Table 2: *Performance of classifiers trained with CLN data on additional datasets. The percentage of correct predictions is the number of instances classified correctly to the total number of instances in the dataset. The column "Layer" indicates the W2V2 encoder layer at which the classifier yielded the best performance. The column next to it shows the classification results that are achieved, when features from the W2V2 layer indicated in column "Layer" are used as inputs. The rightmost column shows the percentage of correct predictions, when the best layers for training and testing on CLN data (cf. experiment 1 in Table 1) are used, i.e, layer 4 (SVM), 3 (FFN), and 4 (XGB).*

#	Dataset	Model	Layer	Percent Correct	
			(best)	(best layer)	(best layer #1 in Tab. 1)
1	Tuda-De100	SVM	6	86.0	84.0
		FFN	6	85.0	77.0
		XGB	5	87.0	86.0
2	MLS100	SVM	5	98.0	91.0
		FFN	6	96.0	95.0
		XGB	7	99.0	95.0
3	NWSR	SVM	7	90.9	90.7
		FFN	6	91.9	87.7
		XGB	8	95.5	94.0

classification rates up to 95% (MLS100) are achieved, when the best layer from experiment #1 in Table 1 is used to predict utterances from the additional corpora (cf. rightmost column of Table 2). The best performing layer for each dataset yields accuracy rates between 85% and 99%.

The sample-weighted average F1-Scores for binary classifiers trained on each of the four pathologies individually are 96.0% (SVM), 94.9% (FFN), and 92.7% (XGB). The best binary classifier (SVM) yields the following F1-Scores for the individual pathologies: $77.9\% \pm 5.3\%$ (PD), $100\% \pm 0\%$ (LAR), $96.6\% \pm 5.0\%$ (CLP), and $98.7\% \pm 2.5\%$ (OSCC). The binary classification baseline is much less robust, when confronted with unseen healthy speaker data: Three of the four binary SVMs have correct prediction rates of at most 49.2% across all 12 W2V2 layers on the Tuda-De100 data. Only the SVM trained on LAR data achieves above chance-level results. On the MLS100 dataset, only two binary SVM models achieve above chance-level results (those trained on OSCC and LAR data), the other two yield at most 49.5% accuracy. Similarly, the speakers in the NWSR corpus are correctly predicted above chance-level by two SVMs (those trained on PD and LAR data), the other two yield at most 49.8% accuracy.

4.3. Partial resection

The PR85 dataset sets itself apart from the other unseen datasets as it contains pathological speech, albeit the pathological features are much less pronounced than in the other pathological speech corpora used in this work. The classification results for the PR85 data are summarized in Table 3. Under the assumption that all 85 utterances from PR85 belong to the healthy control group, the FFN performs best with a share of correctly classified instances of 84.7%, when features extracted at W2V2 layer 7 are used. Figure 3 illustrates the results for the PR85 dataset in more detail. It shows the class predictions for each of the 85 utterances using the FFN model. Most instances not classified as the control group, were classified as "LAR". These 7 instances exhibit a worse intelligibility than the rest of the corpus. The intelligibility for each recording in the PR85 corpus was rated by speech experts with at least 5 years experience. The average intelligibility rating of utterances classified as "LAR" is $\mu = 3.9$ on an inverted five-point Likert-scale, i.e., 1 being the most intelligible and 5 being the least intelligible, whereas the average intelligibility rating of the other 78 utterances is $\mu = 2.9$. These results indicate that the model is capable of detecting the few patients in the PR85 corpus, who underwent a stronger medical intervention, which led to less intelligible speech after surgery, and therefore making them more similar to patients after total laryngectomy in the LAR corpus.

Table 3: *Performance of CLN models on PR85.*

Dataset	Model	Layer	Percent Correct	
		(best)	(best layer)	(best layer #1 in Tab. 1)
PR85	SVM	3	72.9	68.2
Assumption:	FFN	7	84.7	68.2
CTL is the "correct" class	XGB	2	69.4	64.7

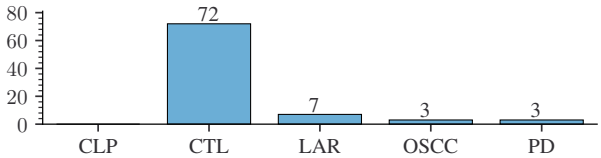

Figure 3: *Predictions for 85 patients who underwent a partial resection of the larynx (PR85) using the FFN model with features extracted at W2V2 layer 7.*

5. Conclusion

We show that W2V2 features are well-suited to encode characteristics of various speech pathologies. We achieve unweighted average F1-Scores between 96% and 97% on clean test data. The performance drops between ≈ 13 (SVM) and ≈ 20 (XGB) percentage points, when reverberated test data is applied on models trained on clean data. Nevertheless, the results remain far above chance-level and can be largely mitigated by using reverberated data during training as well. We also find that all three classifiers perform similarly well on the task.

Control groups that match the pathologies in question are important for the robustness of pathological speech classifiers. We agree with the concerns raised in [15] that improperly selected recordings from healthy speakers might cause classifiers to learn to distinguish between differences in the demographics and recording conditions of the underlying data, rather than the distinct characteristics of healthy and pathological voices. However, we show that these issues can be mitigated by taking prior knowledge about the data distribution into account when choosing a control group, as well as by training systems on multiple pathologies at once.

We demonstrate the robustness of our approach by testing the trained classifiers on unseen datasets and by reverberating the training data with room impulse responses. Our models generalize well to unseen data, which becomes especially apparent, when tested on an unseen corpus of patients who received a partial resection as treatment for laryngeal cancer. The few patients who underwent a more invasive surgical treatment, which led to less intelligible speech, were correctly classified into "LAR", whereas the other patients with almost no discernible speech disorder were mainly classified into "CTL".

Exogenous conditions such as the recording environment might seemingly be the dominant factors, when multiple datasets are used. Nevertheless, the features of interest (i.e., the ones that encode a pathology), are still present in the data and can be leveraged to build well-performing and robust classification systems.

6. References

[1] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, 2006, pp. I–I.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[4] T. Bocklet, T. Haderlein, F. Hönig, F. Rosanowski, and E. Nöth, "Evaluation and Assessment of Speech Intelligibility on Pathologic Voices Based Upon Acoustic Speaker Models," in *Proceedings of the 3rd Advanced Voice Function Assessment International Workshop*, 3rd Advanced Voice Function Assessment International Workshop, Ed., 2009, pp. 89–92.

[5] T. Bocklet, K. Riedhammer, E. Nöth, and U. Eysholdt, "Automatic Intelligibility Assessment of Speakers After Laryngeal Cancer by Means of Acoustic Modeling," *Journal of Voice*, vol. 26, pp. 390–397, 2012.

[6] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of parkinson's speech – acoustic, prosodic and voice related cues," in *Proc. Interspeech 2013*, 2013, pp. 1149–1153.

[7] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech," in *Proc. Interspeech 2017*, 2017, pp. 1834–1838.

[8] I. Laaridh, C. Fredouille, A. Ghio, M. Lalain, and V. Woisard, "Automatic Evaluation of Speech Intelligibility Based on I-vectors in the Context of Head and Neck Cancers," in *Proc. Interspeech 2018*, 2018, pp. 2943–2947.

[9] L. Moro-Velazquez, J. Villalba, and N. Dehak, "Using x-vectors to automatically detect parkinson's disease from speech," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1155–1159.

[10] R. Scheuerer, T. Haderlein, E. Nöth, and T. Bocklet, "Applying x-vectors on pathological speech after larynx removal," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1079–1086.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[12] F. Braun, A. Erzigkeit, H. Lehfeld, T. Hillemacher, K. Riedhammer, and S. P. Bayerl, "Going beyond the cookie theft picture test: Detecting cognitive impairments using acoustic features," in *Text, Speech, and Dialogue*. Cham: Springer International Publishing, 2022, pp. 437–448.

[13] S. P. Bayerl, D. Wagner, E. Nöth, and K. Riedhammer, "Detecting Dysfluencies in Stuttering Therapy Using wav2vec 2.0," in *Proc. Interspeech 2022*, 2022, pp. 2868–2872.

[14] S. P. Bayerl, D. Wagner, I. Baumann, T. Bocklet, and K. Riedhammer, "Detecting vocal fatigue with neural embeddings," *Journal of Voice*, 2023.

[15] C. Botelho, T. Schultz, A. Abad, and I. Trancoso, "Challenges of using longitudinal and cross-domain corpora on studies of pathological speech," in *Proc. Interspeech 2022*. ISCA, 2022, pp. 1921–1925.

[16] T. Haderlein, *Studien zur Mustererkennung*. Berlin: Logos Verlag, 2007, vol. 25, ch. Automatic Evaluation of Tracheoesophageal Substitute Voices, p. 238.

[17] K. Kohler, "German," *Journal of the International Phonetic Association*, vol. 20, no. 1, p. 48–50, 1990.

[18] F. Stelzle, C. Knipfer, M. Schuster, T. Bocklet, E. Nöth, W. Adler, L. Schempf, P. Vieler, M. Riemann, F. Neukam, and E. Nkenke, "Factors influencing relative speech intelligibility in patients with oral squamous cell carcinoma: a prospective study using automatic, computer-based speech analysis," *International Journal of Oral and Maxillofacial Surgery*, vol. 42, no. 11, pp. 1377–1384, Nov. 2013.

[19] T. Bocklet, A. Maier, K. Riedhammer, U. Eysholdt, and E. Nöth, "Erlangen-CLP: A large annotated corpus of speech from children with cleft lip and palate," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 2671–2674.

[20] T. Bocklet, E. Nöth, and G. Stemmer, "Voice assessment of speakers with laryngeal cancer by glottal excitation modeling based on a 2-mass model," in *Text, Speech and Dialogue*, I. Habernal and V. Matoušek, Eds., 2011, pp. 348–355.

[21] T. Haderlein, A. Maier, E. Nöth, F. Rosanowski, and U. Eysholdt, "Automatische Verständlichkeitsbewertung von Telefonaufnahmen Larynxteilresezierter mittels prosodischer Analyse," in *Aktuelle phoniatrisch-pädaudiologische Aspekte 2010*, a. Z.-D. A. Gross Manfred, Ed., 2010, pp. 165–167.

[22] S. Radeck-Arneth, B. Milde, A. Lange, E. Gouvêa, S. Radomski, M. Mühlhäuser, and C. Biemann, "Open source German distant speech recognition: Corpus and acoustic model," in *Text, Speech, and Dialogue*, P. Král and V. Matoušek, Eds. Springer International Publishing, 2015, pp. 480–488.

[23] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "MLS: A Large-Scale Multilingual Dataset for Speech Research," in *Proc. Interspeech 2020*. ISCA, 2020, pp. 2757–2761.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. South Brisbane, Queensland, Australia: IEEE, Apr. 2015, pp. 5206–5210.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[26] R. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

[28] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.