Tirronen, Saska; Javanmardi, Farhad; Kodali, Manila; Kadiri, Sudarsana; Alku, Paavo

Utilizing WAV2VEC in database-independent voice disorder detection

Published in:
Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'23)

DOI:
10.1109/ICASSP49357.2023.10094798

Published: 01/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

UTILIZING WAV2VEC IN DATABASE-INDEPENDENT VOICE DISORDER DETECTION

Saska Tirronen, Farhad Javanmardi, Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku

Department of Information and Communications Engineering, Aalto University, Finland.

ABSTRACT

Automatic detection of voice disorders from acoustic speech signals can help to improve reliability of medical diagnosis. However, the real-life environment in which speech signals are recorded for diagnosis can be different from the environment in which the detection system's training data was originally collected. This mismatch between the recording conditions can decrease detection performance in practical scenarios. In this work, we propose to use a pre-trained wav2vec 2.0 model as a feature extractor to build automatic detection systems for voice disorders. The embeddings from the first layers of the context network contain information about phones, and these features are useful in voice disorder detection. We evaluate the performance of the wav2vec features in single-database and cross-database scenarios to study their generalizability to unseen speakers and recording conditions. The results indicate that the wav2vec features generalize better than popular spectral and cepstral baseline features.

***Index Terms*—** Pathological Voices, Voice disorders, Cross-database evaluation, wav2vec

1. INTRODUCTION

Voice disorders are caused by physiological and psychological disorders, infections, misuse, or due to the vocal abuse or surgery [1]. Automatic detection of voice disorders from acoustic speech signals is a widely studied research topic [2]. Automatic detection systems can be utilized in medical practice to improve the objectivity and reliability of the diagnosis. The traditional detection approach corresponds to using so-called pipeline systems that consist of two steps: feature extraction and classification [2, 3]. As an alternative, modern deep learning architectures have inspired the usage of so-called end-to-end systems that do not require a separate feature extraction step [4–6]. In both system architectures, the disorder detection is made by a supervised machine learning algorithm, which has to be trained on a set of training data before it can be used in detection.

Publicly available voice disorder databases contain much less data compared to repositories that are used in, for example, automatic speech recognition (ASR) or speech synthesis. The limited amount of training data effectively limits the ability of the detection system to generalize to data from unseen speakers. Moreover, the available databases are usually recorded by using professional recording equipment in noise-controlled laboratory environments. When an automatic detection system is used for real medical diagnosis, the recording equipment and environment can be different from those used in the system's training phase. This mismatch in recording conditions may further decrease the detection performance. A popular solution is to use data augmentation (DA) to increase the amount and variability of training data [7]. Ideally, the synthetic new data generated in DA should enrich the training data by adding natural variability between different speakers and recording environments [7–9].

Another approach is the usage of deep learning models that have been pre-trained using a large database on some other speech-related task (e.g., ASR) than voice disorder detection. During pre-training, the earlier layers of the model may learn to extract speech features that are useful for a wide variety of speech-related and voice-related tasks, including voice disorder detection. Therefore, pre-trained models can be used as feature extractors in pipeline systems [10–15]. It is also common to fine-tune the model on the target database. This increases the utility of the later layers and also allows the usage of the pre-trained models in building end-to-end systems [16–18].

Despite the fact that the problem of data sparcity has been addressed in earlier studies on detection of voice disorders [2, 19, 20], only a few previous investigations have studied training and evaluating the system using different databases. Such cross-database evaluation would ensure that the performance measures also capture the effect of different recording environments and equipment. To the best of our knowledge, only few studies have reported cross-database performances in voice disorder detection [2,19,20]. In [19], authors explored the usage of modulation spectral features for cross-database experiments. Mel frequency cepstral coefficients (MFCCs) along with perturbation features were explored in [2], and the experiments of [20] explored a few specific diseases (such as paralysis, cyst and polyp) in cross-database experiments using auto-correlation and entropy features in different frequency regions. Importantly, the usage of pre-trained feature extractors has not been studied in cross-database scenarios.

In this work, we compare a wav2vec [10] feature extractor, pre-trained for the purposes of ASR using a very large database, with popular baseline features. The baseline features used in the comparison are the MFCCs, spectrogram and mel-spectrogram. We use all these features in a pipeline system together with the support vector machine (SVM) classifier, and evaluate the performances of the different feature extraction approaches in single-database and cross-database scenarios. Hypothetically, pre-training on a larger database may improve the generalizability, and therefore the performance, in both scenarios.

The main contributions of this study are:

- Layer-by-layer analysis of the utility of pre-trained features in the context of voice disorder detection.
- Evaluation of pre-trained features in single-database and cross-database scenarios, to observe their effect on generalizability.

2. CLASSIFICATION SETUP

In this section, the technical details of the classification setup are described, including the databases, features, classifiers, and the evaluation methodology.

2.1. Databases

In the current study, two popular dysphonia databases containing healthy and disordered voice samples are used. The first one is the Hospital Universitario Príncipe de Asturias (HUPA) [21, 22] database and the second one is the Saarbrücken Voice Disorders (SVD) [23, 24] database. These two repositories are briefly described below.

2.1.1. The HUPA database

HUPA [21, 22] was recorded at Universidad Politécnica de Madrid and it contains samples from 239 healthy speakers and 200 speakers with a voice disorder. The database contains a wide range of organic disorders such as nodules, polyps, and Reinke's edema. In collecting the data, every subject was asked to utter a sustained phonation of the vowel /a/ in a constant pitch. The voices were recorded with a sampling frequency of 50 kHz. The current study included samples of all available disorders in HUPA. A balanced subset was selected such that the number of healthy and disordered voice samples were both 200.

2.1.2. The SVD database

SVD [23,24] contains sustained vowels /a/, /i/ and /u/ in low, normal, high, and rising-falling pitches. Apart from vowels, the data also contains the sentence "Guten Morgen, wie geht es Ihnen" ("Good morning, how are you"). This database was recorded at the Institut für Phonetik at Saarland University. The sampling frequency of the data is 50 kHz. The repository contains samples from 2225 German speakers consisting of 869 healthy controls and 1365 speakers with a voice disorder. The pathological data covers altogether 71 different laryngeal voice disorders. The current study uses only the /a/ vowels of normal pitch from SVD. This choice was made, because voice samples representing the vowel /a/ are available both in SVD and HUPA and because the sustained production of /a/ is a popular speaking task that is available in many voice disorder databases. The laryngeal disorders used in this study include vocal fold paresis, laryngitis, and all 8 different types of dysphonia that are available in SVD. We selected a balanced subset that included 631 samples from healthy speakers and 631 samples from speakers with a disorder. Figure 1 shows the number of the included samples from healthy speakers and speakers with a disorder from the HUPA and SVD databases. The data was resampled to 16 kHz in this study.

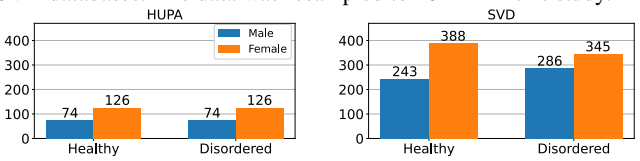

Fig. 1. Number of samples from healthy speakers and speakers with a disorder from the HUPA and SVD databases.

2.2. Features

In this study, we investigate the use of the pre-trained wav2vec model [10, 25] as a feature extractor in a pipeline system to automatically detect voice disorders from acoustic voice signals. The pre-training was conducted on a combination of three ASR databases, containing 56,000 hours of audio from 53 different languages. In order to perform an ASR task, the model extracts contextualized embeddings. The computation is done by first mapping speech segments of 20 milliseconds into feature vectors by a convolutional neural network (CNN) -based latent feature encoder architecture. Then, the features are projected to the correct embedding size (1024), the relative positional embeddings are computed by grouped 1D convolution and added to the features vector, and layer normalization is performed. The resulting vectors are then fed as input to a context network, which consists of 24 Transformer layers. During training, the outputs of the context network are used, together with the outputs of a separate quantization module, to optimize the model by using contrastive loss. Figure 2 shows the architecture of wav2vec. The wav2vec model has been proven to yield embeddings that effectively capture the information regarding the phonemes, therefore performing well in ASR [10,25]. However, especially the outputs of the first Transformer layers contain information regarding phones as well, and they can therefore be utilized as features in a wide range of speech related tasks in addition to ASR [11, 15, 26]. In addition, as the pre-training is conducted by using large training data sets and a lot of computational resources, the features can effectively generalize to unseen speakers [27].

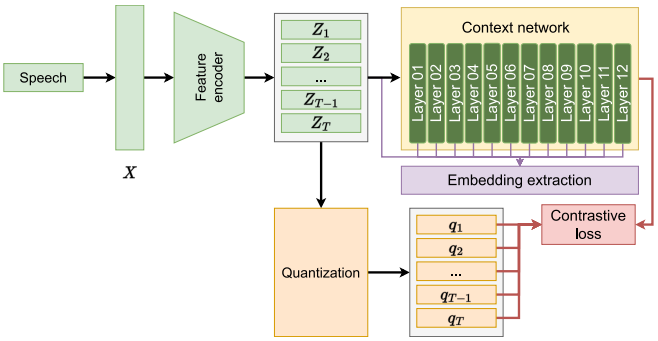

Fig. 2. Block diagram of a wav2vec 2.0 architecture with 12 transformer layers.

Due to the benefits of wav2vec described above, we use the Transformer-layer outputs of the model as features in voice disorder detection in the current study. More specifically, we compute the temporal average of the inputs to the first Transformer layer, as well as of the outputs of all 24 Transformer layers. This results in extracting the total of 25 feature vectors for each voice signal, each vector being 1024-dimensional. In the remaining part of this work, these features are referred to as the $wav2vec\ features$, or alternatively as $wav2vec-N$, when referring to the features from the layer number N.

To compare the performance of the wav2vec features, we also extracted three conventional features, namely, spectrograms (Spec), Mel-spectrogram (Mel-spec), and MFCCs. Voice signals were processed in 25 ms frames with a 5 ms shift using the Hamming window. The number of FFT bins was 1024, and the number of mel-filters banks used in computing the mel-spectrogram was 80. Along with the 13-dimensional MFCCs, their derivatives and double derivatives were considered. The features were averaged across all frames resulting in 513-dimensional, 80-dimensional, and 39-dimensional feature vectors for Spec, Mel-spec, and MFCCs, respectively.

2.3. Classifiers and Evaluation

In the current study, SVM was used as a classifier for the automatic classification of voice disorders. SVM is a well-known and popularly used ML classifier for the detection and regression tasks. In this study, we used linear kernel function, fixed the regularization parameter to 1, and computed gamma as $\gamma = 1/(D \cdot \widehat{Var}(X))$, where D is the dimensionality of the features and $\widehat{Var}(X)$ is the estimated variance of the features in training data.

In order to train and evaluate the systems, a 5-fold cross-validation (CV) strategy was used. The samples from each speaker were contained within a single fold to ensure that the classifier could not learn speaker identity. For every iteration, four folds were used for training and one fold was used for testing. Balanced classification accuracy, precision, recall and F1-score were used as the evaluation metrics.

2.4. Experiments

The evaluation of the wav2vec features was conducted in two parts consisting of a single-database evaluation and a cross-database evaluation. In the single-database evaluation, SVM classifiers were trained and evaluated using only the data from an individual database. This was repeated for both SVD and HUPA. The single-database evaluation has two main purposes. First, it allows to observe the effectiveness of the pre-trained wav2vec features in a scenario where the recording environment and equipment are constant between the training and testing phases. In this scenario, the main difference between the training and testing data is due to the speakers that are included. Therefore, the single-database evaluation measures the generalizability of the wav2vec features to unseen speakers. It also provides a baseline for comparison with the cross-database scenario.

The second part of the experiments was a cross-database evaluation, in which the training and testing were conducted by using data from different databases. We first trained the detection system using the SVD data and then evaluated it using the HUPA data, and then trained the system using the HUPA data and evaluated it using the SVD data. The cross-database evaluation allows to observe the generalizability of the wav2vec features in conditions where also the recording environment and equipment are different between training and testing. A comparison with the single-database scenarios enables assessing the sensitivity of the features to the mismatch between training and testing caused by environmental factors.

In both the single-database and cross-database scenarios, the wav2vec features are compared with the baseline features described in Section 2.2. In the cross-database experiments, the cross-evaluation is performed once per every fold of the 5-fold CV by using all data in the other database as evaluation data.

3. RESULTS

This section reports the results of the experiments. First, the results of the single-database experiments are presented in Section 3.1 and followed by the results of the cross-database experiments in Section 3.2.

3.1. Detection Results in the Single-Database Scenarios

As described in Section 3, the evaluation of all the 25 wav2vec features was first conducted in a single-database setting. The resulting classification accuracies are shown in Figure 3 for all the wav2vec features and baselines. Other metrics are displayed in Table 3.1, for the baselines and the best wav2vec features.

The results show that the wav2vec features clearly outperform all the baseline features. For HUPA, wav2vec-1 outperforms the best baseline (Mel-spec) by an absolute improvement of 12.70%. For SVD, wav2vec-1 outperforms the best baseline (Mel-spec) by an absolute improvement of 3.19%. The best accuracies are 83.11% for HUPA and 68.55% for SVD.

Table 1. Performance metrics of the single-database scenarios. All the baseline features and the best wav2vec features are included. The wav2vec feature names are written as $wav2vec - N$, where N is the number of the corresponding layer. Standard deviations over the folds are reported for classification accuracies (ACC).

SVD				
Feature	**ACC [%]**	**Precision**	**Recall**	**F1**
Reference features				
Spec	61.52 ± 1.04	0.63	0.63	0.63
Mel-spec	65.36 ± 2.26	0.67	0.63	0.65
MFCC	63.24 ± 1.74	0.65	0.61	0.63
wav2vec features				
wav2vec-1	68.55 ± 2.45	0.70	0.67	0.68
HUPA				
Feature	**ACC [%]**	**Precision**	**Recall**	**F1**
Reference features				
Spec	66.01 ± 8.33	0.64	0.70	0.67
Mel-spec	70.41 ± 2.42	0.70	0.72	0.71
MFCC	63.63 ± 5.06	0.64	0.63	0.63
wav2vec features				
wav2vec-1	83.11 ± 4.56	0.82	0.85	0.83

3.2. Detection Results in the Cross-database Scenario

The classification accuracies of the cross-database experiments are shown in Figure 4 for all the wav2vec features and baselines. Table 3.2 shows the other metrics for the baselines and the best wav2vec features.

Again, the wav2vec features clearly outperform the baseline features. For HUPA, the best accuracy was 59.35%, and it was achieved by using the wav2vec-2 features. This corresponds to an absolute improvement of 4.7% compared to the best baseline (Mel-spec). For SVD, the best accuracy is 57.61%, which was achieved by using the wav2vec-3 features. Compared to the best baseline (Spec), the absolute improvement is 5.31%. Compared to the accuracies in the single-database experiments, the best performances decreased by 23.76% (absolute) for HUPA and by 10.94% (absolute) for SVD.

4. SUMMARY AND CONCLUSIONS

In this paper, a comparison of the pre-trained wav2vec features and the popularly used spectral and cepstral baseline features (spectrogram, mel-spectrogram, and MFCCs) was conducted in voice pathology detection by studying single-database and cross-database scenarios. The goal of the study was to examine how well the pre-trained wav2vec speech embeddings from ASR tasks are generalizable to the detection of voice disorders, and in particular to examine the effectiveness of the wav2vec features to generalize to unseen speakers and recording environments.

The results show that the wav2vec features outperformed the baselines in the single-database scenario for both included databases (SVD and HUPA). This shows that the wav2vec model has learned to extract useful features that can be used also for the detection of voice disorders. In addition, the wav2vec features help the detection system to generalize to unseen speakers due to the pre-training on large datasets. The utility of the wav2vec features was particularly high for the HUPA database. In the single-database scenarios, the wav2vec features gave an absolute improvement of 12.7% for HUPA, and 3.19% for SVD in comparison to best baseline (mel-spectrogram) feature.

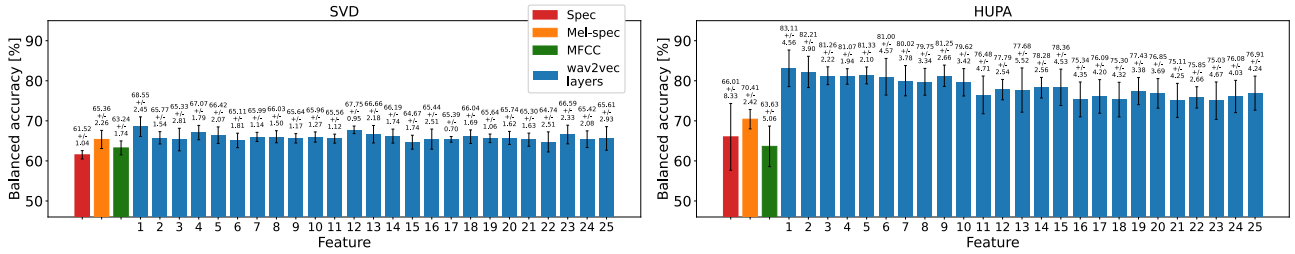

Fig. 3. Classification accuracies of the baseline and wav2vec features in the single-database scenarios. The blue bars represent the features derived from wav2vec, with the tick labels indicating the index of the corresponding layer. Heights of the bars represent the mean accuracies over the folds. The tails represent the standard deviations.

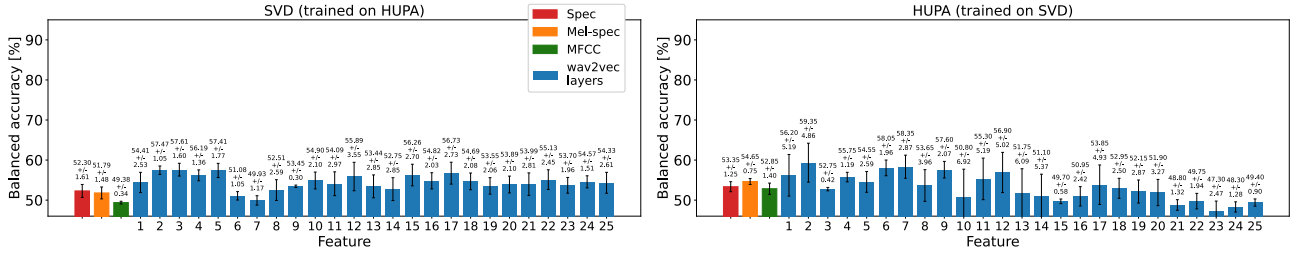

Fig. 4. Classification accuracies of the baseline and wav2vec features in the cross-database scenario. The blue bars represent the features derived from wav2vec, with the tick labels indicating the index of the corresponding layer. Heights of the bars represent the mean accuracies over the folds. The tails represent the standard deviations

Table 2. Performance metrics of the cross-database scenarios. All the baseline features and the best wav2vec features are included. The wav2vec feature names are written as $wav2vec - N$, where N is the number of the corresponding layer. Standard deviations over the folds are reported for classification accuracies (ACC).

SVD (trained on HUPA)				
Feature	**ACC**	**Precision**	**Recall**	**F1**
Reference features				
Spec	52.30 +/- 1.61	0.53	0.79	0.63
Mel-spec	51.79 +/- 1.48	0.53	0.76	0.62
MFCC	49.38 +/- 0.34	0.51	0.51	0.51
wav2vec features				
wav2vec-3	57.61 +/- 1.60	0.59	0.60	0.59
HUPA (trained on SVD)				
Feature	**ACC**	**Precision**	**Recall**	**F1**
Reference features				
Spec	53.35 +/- 1.25	0.52	0.76	0.62
Mel-spec	54.65 +/- 0.75	0.53	0.72	0.61
MFCC	52.85 +/- 1.40	0.52	0.67	0.59
wav2vec features				
wav2vec-2	59.35 +/- 4.86	0.65	0.60	0.62

In the cross-database scenario, the performance decreased remarkably when compared to the single-database evaluations. In many cases, the baseline performances decreased nearly to the chance level (50%). The wav2vec features outperformed the baselines in the cross-database experiments for both databases, which implies that they are more generalizable to different recording environments and equipment than the baseline features. However, even for them, the decrease in performance was large. For HUPA, the accuracy of the best performing wav2vec features (wav2vec-2) in the cross-database scenario was 23.76% (absolute) lower than for the best performing wav2vec features (wav2vec-1) in the individual-database scenario. Similarly in the case of SVD, the accuracy decreased by 10.94%, and it occured between wav2vec-1 in the single-database scenario and wav2vec-3 in the cross-database scenario. This suggests that even the current popular pre-trained models may not alone be sufficient to develop fully database-independent systems for detection of voice disorders.

In all scenarios, the wav2vec features from the first layers of the network performed better than the features from the last layers. This is because the layers that are located closer to the output layer have learned embeddings that carry information of the phoneme identity, which obviously is important in ASR but is not necessarily useful for voice disorder detection. Therefore, to fine-tune the model to voice disorder data remains a topic for future work, by potentially also utilizing the model to build end-to-end classifiers. Another future topic would be to perform the fine-tuning by using glottal source waveforms that are extracted by using inverse filtering methods [3]. This may further help generalizability, by removing vocal tract information that is redundant particularly in the detection of laryngeal voice disorders.

5. REFERENCES

[1] J. J. Ballenger and J. B. Snow, *Ballenger's otorhinolaryngology: head and neck surgery*. Pmph-usa, 2003.

[2] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and

study on the effects of different variability factors," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019.

[3] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2019.

[4] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: a preliminary study," in *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*. IEEE, 2017, pp. 1–4.

[5] H. Wu, J. Soraghan, A. Lowit, and G. Di-Caterina, "A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief Networks," in *Proc. Interspeech 2018*, 2018, pp. 446–450.

[6] T. Kourkounakis, A. Hajavi, and A. Etemad, "Fluentnet: End-to-end detection of stuttered speech disfluencies with deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2986–2999, 2021.

[7] Y. Jiao, M. Tu, V. Berisha, and J. M. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6009–6013.

[8] M. Hireš, M. Gazda, L. Vavrek, and P. Drotár, "Voice-specific augmentations for Parkinson's disease detection using deep convolutional neural network," in *20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2022, pp. 000 213–000 218.

[9] I. Miliaresi, K. Poutos, and A. Pikrakis, "Combining acoustic features and medical data in deep learning networks for voice pathology classification," in *28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1190–1194.

[10] 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.

[11] A. Hernandez, P. A. Pérez-Toro, E. Nöth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual self-supervised speech representations for improved dysarthric speech recognition," *arXiv preprint arXiv:2204.01670*, 2022.

[12] O. Mohamed and S. A. Aly, "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset," *arXiv preprint arXiv:2110.04425*, 2021.

[13] T. Grósz, D. Porjazovski, Y. Getman, S. Kadiri, and M. Kurimo, "Wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7026–7029.

[14] J. Zuluaga-Gomez, A. Prasad, I. Nigmatulina, S. Sarfjoo, P. Motlicek, M. Kleinert, H. Helmke, O. Ohneiser, and Q. Zhan, "How does pre-trained wav2vec2. 0 perform on domain shifted asr? an extensive benchmark on air traffic control communications," *arXiv preprint arXiv:2203.16822*, 2022.

[15] S. A. Sheikh, M. Sahidullah, F. Hirsch, and S. Ouni, "Introducing ecapa-tdnn and wav2vec2. 0 embeddings to stuttering detection," *arXiv preprint arXiv:2204.01564*, 2022.

[16] N. Vaessen and D. A. Van Leeuwen, "Fine-tuning wav2vec2 for speaker recognition," in *ICASSP*. IEEE, 2022, pp. 7967–7971.

[17] M. K. Baskar, T. Herzig, D. Nguyen, M. Diez, T. Polzehl, L. Burget, J. Černockỳ *et al.*, "Speaker adaptation for wav2vec2 based dysarthric asr," *arXiv preprint arXiv:2204.00770*, 2022.

[18] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[19] M. Markaki and Y. Stylianou, "Normalized modulation spectral features for cross-database voice pathology detection," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[20] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. F. Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *Ieee Access*, vol. 6, pp. 6961–6974, 2017.

[21] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 60–69, 2011.

[22] L. Moro-Velázquez, J. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale," *BioMed Research International*, vol. 2015, p. 259239, 2015.

[23] M. Pützer and W. J. Barry, "Saarbrücken voice database, institute of phonetics, univ. of saarland," 2010, http://www.stimmdatenbank.coli.uni-saarland.de/ (Last viewed November 26, 2021).

[24] M. Pützer and W. J. Barry, "Instrumental dimensioning of normal and pathological phonation using acoustic measurements," *Clinical Linguistics & Phonetics*, vol. 22, no. 6, pp. 407–420, 2008.

[25] Q. Xu, A. Baevski, and M. Auli, "Simple and effective zero-shot cross-lingual phoneme recognition," 2021. [Online]. Available: https://arxiv.org/abs/2109.11680

[26] L. Gauder, L. Pepino, L. Ferrer, and P. Riera, "Alzheimer disease recognition using speech-based embeddings from pre-trained models." in *Interspeech*, 2021, pp. 3795–3799.

[27] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP*. IEEE, 2021, pp. 5939–5943.