Pathological Voice Detection From Sustained Vowels: Handcrafted vs. Self-supervised Learning

Bagus Tris Atmaja
AIST
Tsukuba, Japan
b-atmaja@aist.go.jp

Akira Sasou
AIST
Tsukuba, Japan
a-sasou@aist.go.jp

***Abstract*—Pathological voice detection aims to detect voice disorders from speech samples. With the recent development of self-supervised learning (SSL), most studies in the past years of voice disorder employ that technique. Handcrafted features may suggest better prediction since they contain physical information and are more interpretable than SSL. We evaluated different handcrafted acoustic features and SSL approaches for pathological voice detection tasks using sustained vowels. We extracted 88 and 39 dimensional handcrafted features using openSMILE and Praat feature extractors. For SSL, we evaluated wav2vec 2.0, HuBERT, and WavLM, both for feature extractors and finetuning. Results showed that handcrafted features are consistently competitive with SSL features. An ensemble model combining handcrafted and SSL features achieved the best performance with an F1-score of 0.8739 on the test set, outperforming previous studies on the same dataset under the test set. This finding suggests that handcrafted features are still competitive for voice disorder detection tasks, and combining them with SSL features can further improve performance.**

***Index Terms*—Pathological voice detection, voice disorders, self-supervised learning, handcrafted features, ensemble learning**

I. Introduction

Automatic voice condition analysis (AVCA) provides an objective and quantitative assessment of voice disorders. It aims to detect and classify voice pathologies from speech samples to assist clinicians. The classical approach required instrumental (objective) and perceptual (subjective) evaluation with other types of examination to determine the existence and level of voice impairment [1]. Accurate automatic analysis can potentially reduce workload and subjectivity in clinical voice assessment.

Research on AVCA can be divided into two categories: detection and identification. Voice disorder detection determines whether a voice sample contains pathology or not (e.g., [2], [3]). The task is binary classification with F1-score and area under curve (AUC) were the most cited metrics [4]. Identification further classifies the type of voice disorder if detected as pathological, e.g., if a voice falls into one type of structural, neurogenic, functional, and psychogenic [5]. This task is multiclass classification with accuracies as the common metric. This study focuses on the detection of voice disorders from sustained vowels. Research on pathological voice disorder detection can be traced back to the availability of its datasets. Saarbrüecken Voice Database (SVD) [6], [7] perhaps is the first publicly available voice disorder dataset. HUPA [8] is a dataset of sustained phonation of the vowel /a/ of 366 adult Spanish speakers (169 pathological and 197 normophonic) recorded in Spain. AVFAD [9] is a dataset developed in Portugal containing 346 clinically diagnosed samples with vocal pathology and 363 samples with no vocal alterations. VOICED [10] is a database that includes 208 voice samples from 150 pathological and 58 healthy voices recorded in Italy. PVQD [11] is 296 high-quality audio files consisting of sustained /a/ and /i/ vowels and sentences from Consensus Auditory-Perceptual Evaluation of Voice that were recorded in the US. Among these datasets, SVD remains as the most widely used dataset for pathological voice detection research due to its large number of samples and availability of demographic information.

In pathological voice disorder detection, two important factors should be considered: input speech and decision blocks [12]. Input speech can refer to the type of audio materials (e.g., sustained /a/, /i, /u/ vowels, sentences) and/or the type of acoustic feature extractors to extract features from that audio material (e.g., MFCC, prosodic features, deep learning features). Decision blocks refer to the steps taken to make the final diagnosis, i.e., the classifiers. This study focuses on the first part to evaluate different input speech and acoustic features for voice disorder detection.

This study contributes to the pathological voice disorder detection in two aspects. First, we focused on normal sustained vowels as input speech and explored different vowels (/a/, /i/, /u/, /aiu/) for voice disorder detection tasks with handcrafted and self-supervised learning (SSL) features. Second, we evaluated different combinations of handcrafted and SSL features through ensemble learning to further improve the performance in different metrics.

II. Dataset and Materials

The Saarbrüecken Voice Database was examined in this study. The SVD includes recordings of vowels /a/, /i/, and /u/ phonated by healthy and pathological voices. Healthy refers to normophonic condition where speech sounds are produced with normal, typical phonetic characteristics. Pathological voices contain organic (e.g., benign, malignant) and non-organic voice disorders (hyper- and hypofunctional dysphonia,

Supported by NEDO Project JPNP20006 and JSPS KAKENHI 24K0296

phononeurosis, and dysodia). In addition to normal speech, vowel materials are also available in lower, high, and raising-falling pitches, as well as phrases of speech data; however, normal speech is considered following the previous studies [13]–[15]. We also evaluated the concatenation of /a/, /i/, and /u/ vowels as separate training data instead of combining them into a single recording.

The database is divided into training, development, and test sets following the previous evaluation [16]; each set contains recordings from different subjects. A total number of 2032 samples are included, with 1650 samples for training (550 normal and 1100 pathological), 192 for development (68 normal and 124 pathological) and 190 for test (69 normal and 123 pathological). The recordings are single channels with an original sampling frequency of 50 kHz but were downsampled to 16 kHz for feature extraction. More details about the database can be found in [7].

III. Methods

A. Acoustic Features

We evaluated two handcrafted acoustic features and three SSLs. In addition to using SSL as feature extractors, we also finetune the SSL models for classification tasks. The choice of these features are based on the previous studies [15], [17], [18]. The following are description of these features used in the experiments:

os is a set of features extracted using **o**pen**S**MILE toolkit [19] with eGeMAPSv02 feature set (88 dimensionals) [20]. It includes low-level descriptors like MFCCs, spectral features, voicing probability, F0, and others; however, only their statistic functionals are extracted.

praat [21] is a set of acoustic feature from Praat toolkit [22], which are extracted via Parselmouth [23]. It contains 39 dimensional features, including measures of fundamental frequency (F0), formants (F1-F3), harmonics-to-noise ratio, jitter, and shimmer.

w2v is the robust version of wav2vec 2.0 model [24]. The original model ID (hugging face) is 'facebook/wav2vec2-large-robust-ft-swbd-300h'.

hubert is the large version of HuBERT model [25]. The original model ID is 'facebook/hubert-large-ll60k'.

wavlm is the large version of WavLM model [26]. The original model ID is 'facebook/wavlm-large'.

ft-w2v, **ft-hubert**, and **ft-wavlm** are the finetuning versions of w2v, hubert and wavlm respectively.

We evaluated the effect of scaling and balancing on feature extraction and classification performance. We experimented with standard and robust scalers for scaling and SMOTE [27] for data balancing. Furthermore, we reported the best results either with or without scaling; balancing is used in all reported results except the finetuning models. If a vowel is not written after the feature, it implies that the /a/ vowel is examined.

B. Classifier

We evaluated SVM, KNN, Bayes, MLP, and Tree models. However, only XGB is reported in this study since it consistently achieved the best performance across all features. We used the default hyperparameters for the XGB classifier [28].

We further combine the results of different XGB models from different vowels and different features via ensemble learning. Feature concatenation (early fusion) is also evaluated. We experimented with averaging the prediction probabilities based on the previous study [18] for ensemble learning.

The configuration files (in INI format) to define the acoustic feature, classifier models, hyperparameters, and other settings are accessible from the GitHub repository.[1] We used the Nkululeko toolkit [29] for experiments, which provides an easy and fast setup for studying machine learning speaker characteristics. The repository also contains a partition of training, development, and test sets to benchmark this study.

C. Metrics

We evaluated the results from different vowels and features using the following metrics:

- Unweighted Accuracy (UA): average accuracy for each class, equal to unweighted average recall (UAR), or recall macro.
- Weighted Accuracy (WA): overall accuracy, equal to recall weighted.
- F1: F1 score of the positive class (pathological as a positive label) in binary classification. This is the metric for comparison with the previous study [16].
- F1-ma(cro): F1 score by calculating the score for each label and finding their unweighted mean.
- F1-we(ighted): F1 score by calculating the score for each label and finding their weighted average based on the number of samples for each label.
- AUC: Area under the curve of true positive rate vs. false positive rate curve based on true values and predicted probabilities of positive labels.

IV. Results and Discussion

A. Evaluation of Different Vowels

Evaluation of the type of vowel materials (/a/, /i/, /u/) is important to find which type of vowel contains the most discriminative information for pathological voice detection. We evaluated several features (handcrafted and SSL) and report samples of results in Table I with os and hubert features. In contrast to previous study [1], [13], [15], we found that using only /u/ vowel with os achieved the best F1 score across all metrics compared to using other vowels or combining all three vowels (/aiu/) for handcrafted features; however, the average score (from UA to F1-we) of os and praat are same for /a/ and /u/. For SSL-based feature extractors (evaluated on w2v, hubert, and wavlm), /i/ vowel consistently achieved higher scores than /a/ and /u/ vowels, except for wavlm where /u/ obtained the best. This suggests that close-front vowel /i/ may contain more information that could be captured better by SSL models compared to other types of vowels.

[1] https://github.com/bagustris/svd-exploration

TABLE I
PERFORMANCE OF DIFFERENT VOWELS (CLASSIFIER: XGB)

Vowels	UA	WA	F1	F1-ma	F1-we	AUC
			os			
a	**0.771**	0.773	0.817	0.760	0.777	**0.854**
i	0.609	0.700	0.799	0.606	0.662	0.609
u	0.754	**0.779**	**0.831**	**0.756**	**0.778**	0.836
aiu	0.706	0.742	0.806	0.710	0.739	0.819
			hubert			
a	0.717	0.753	0.814	0.722	0.749	0.717
i	**0.729**	**0.763**	**0.822**	**0.734**	**0.760**	**0.729**
u	0.655	0.695	0.770	0.658	0.691	0.656
aiu	0.575	0.647	0.751	0.574	0.626	0.575

Figure 1 shows the distribution of selected handcrafted features for normal and pathological voices from the praat feature set. It can be seen that features such as HNR, mean F0, jitter, and shimmer have better distinctions between the two classes compared to other features like formants. This validates the importance of these prosodic and perturbation features for pathological voice detection tasks. These important features were calculated by the tree model. We found that reducing the number of features from 39 (praat) to such numbers (e.g., 10, 20, and 30) did not improve the performance of the model. Instead, balancing the dataset by SMOTE oversampling helped to improve the performance.

B. Evaluation of Different Features

Several classifiers have been evaluated, i.e., SVM, KNN, Bayes, Tree, MLP, and XGB; XGB consistently achieved the best performance across all features; hence, it was used to evaluate different features. Table II shows the performance of different features using XGB classifier. The scores are obtained either from /a/, /i/, or /u/ vowel, which vowel obtained the highest average score across all metrics. Again, we found a similar pattern that for handcrafted features, the /a/ vowel achieved the best results (F1) compared to other vowels or combinations of vowels. Meanwhile, for SSL-based features, /i/ vowels performed better (average scores across metrics) than /a/ vowel (including the result for wavlm, which is not shown). This suggests that different types of features (handcrafted vs. SSL-based) may capture different information from different vowels.

TABLE II
PERFORMANCE OF DIFFERENT FEATURES FROM WITH XGB

Features	Vowel	UA	WA	F1	F1-ma	F1-we
os	/a/	0.771	0.773	0.817	0.760	0.777
praat	/a/	**0.806**	**0.784**	0.815	**0.778**	**0.789**
w2v	/i/	0.732	0.758	0.815	0.733	0.757
hubert	/i/	0.729	0.763	**0.822**	0.734	0.760
wavlm	/u/	0.714	0.753	0.816	0.720	0.748
ft-w2v	/i/	0.721	0.758	0.819	0.727	0.754
ft-hubert	/i/	0.682	0.663	0.704	0.657	0.671
ft-wavlm	/i/	0.680	0.726	0.798	0.686	0.719

C. Ensemble Learning Results

Fusing different types of features and vowels can potentially improve the performance of pathological voice detection by capturing complementary information. We evaluated ensemble learning by combining predictions from different features and vowels. Table III shows the performance of ensemble learning of (1) early fusion (feature concatenation) handcrafted features, (2) early fusion (ef) of SSL-based features (both are from /a/ vowels), (3) late fusion (lf) of the same /a/ vowels extracted using os and praat, (4) wav2vec 2.0 feature extractors on /a/ and /u/ vowels, (5) fusion of different features from /a/ vowels with os, praat, w2v and (6) from os, praat, hubert. The last three fusions are performed at the decision stage via averaging (mean ensemble in [18]). Although we have tried to fuse more models, these four results are our best for each type of fusion (early fusion, handcrafted, SSL, and mix handcrafted + SSL). This demonstrates that integrating information from different representations can boost the detection accuracy in pathological voice detection.

TABLE III
PERFORMANCE OF ENSEMBLE LEARNING

Method	UA	WA	F1	F1-ma	F1-we	AUC
early fusion						
os+praat	0.780	0.789	0.833	0.774	0.791	0.872
w2v+h+w	0.731	0.753	0.808	0.730	0.753	0.809
late fusion						
os+praat	0.797	0.795	0.833	0.784	0.798	**0.876**
w2v /a/+/u/	0.752	0.789	0.844	0.761	0.785	0.831
os+praat+w2v	0.814	0.816	0.852	0.804	0.818	0.869
os+praat+h	**0.841**	**0.842**	**0.874**	**0.831**	**0.842**	0.867

Figure 2 shows the receiving operator characteristics (ROC) curves of predictors from the top-4 performing ensemble methods with `os+praat ef` achieved the best. Although ROC AUC is a common performance metric for binary classification, it is difficult to judge the best method based on AUC and that plot. To better understand the predictive performance, we also include UA, WA, F1, F1-macro, and F1-weighted scores, in which `os+praat+h` achieved the best performance across all these metrics.

D. Benchmarks

A benchmark with previous studies on the same test set and F1-score is shown in Table IV. The proposed method using an ensemble of os, praat, and w2v features with XGB classifier, including its balancing methods, achieved an F1 score of 0.8739, significantly outperforming the previous state-of-the-art methods of CAR-HMM (F1=0.7527) and modified CPC (F1=0.7421) in [16]. Event the single features without fusion outperform the previous best results (e.g., os and hubert). These results demonstrate the effectiveness of the proposed feature engineering and machine learning approach for pathological voice detection over the previous scheme, paving the way for developing more accurate voice screening tools.

Additionally, a benchmark of AUC scores also can be done, as shown in V. This benchmark compares the highest AUC score from ensemble method (os+praat late fusion on /a/ vowel) to previous studies on the same dataset. Previous works evaluated raw speech perturbation, MFCC features on

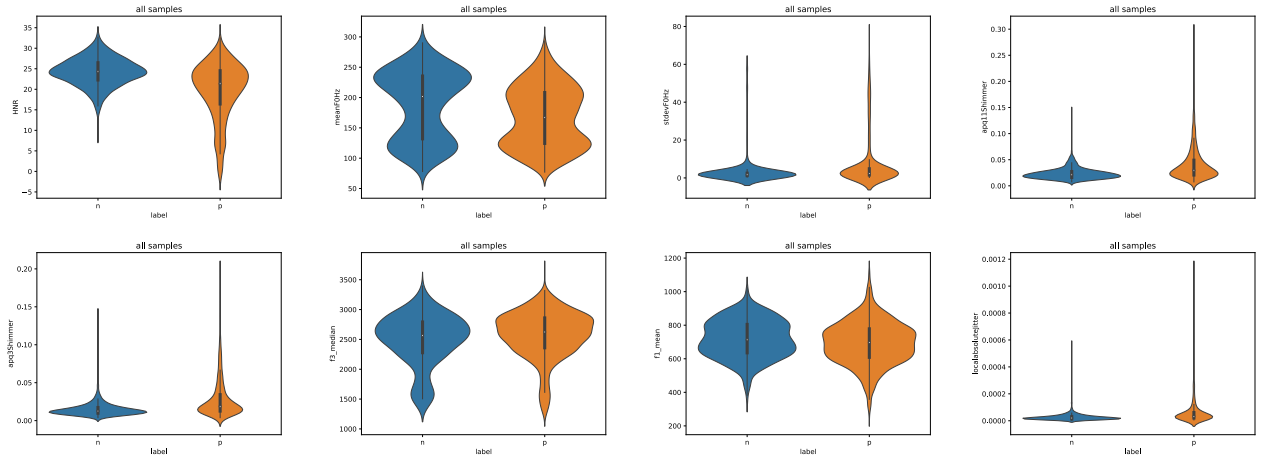

Fig. 1. Distribution of top-8 important features from praat /a/ vowel based on the tree model.

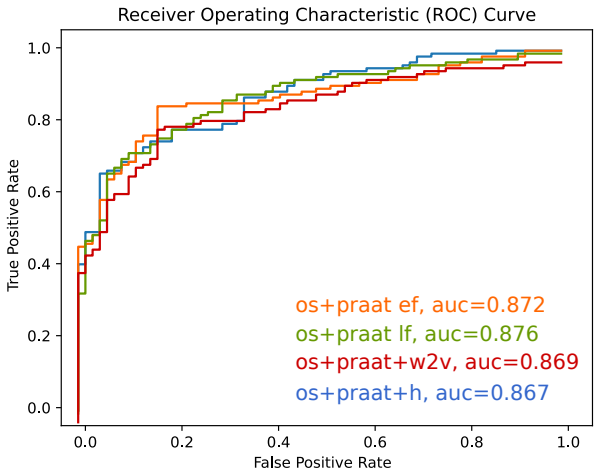

Fig. 2. ROC curve of predictors from top-4 ensemble methods

TABLE IV
F1-SCORE BENCHMARKS WITH PREVIOUS STUDY ON THE SAME TEST SET

Method	F1 score
CAR-HMM [16]	0.7527
modidifed CPC [16]	0.7421
os /u/ (ours)	0.8306
hubert /i/ (ours)	0.8221
os+praat+hubert (ours)	**0.8739**

raw speech/extracted vowels, ComParE system with mean and inter-quartile, wav2vec2, and Transformer-based models. Our evaluated method only obtained a lower score than the Transformers-based approach, which used extensive data augmentation from both phrases and vowels. Note that although all reported results are based on the same SVD dataset, the exact test/validation splits may differ across studies, which could account for some performance variations.

TABLE V
AUC BENCHMARKS WITH PREVIOUS STUDIES ON THE SAME DATASET

Method	Vowel/Phrases	Augmentation	AUC
Perturbation [1]	/a/	No	0.78
MFCC raw speech [1]	phrases	No	0.86
MFCC extracted vowel [1]	phrases	No	0.84
ComParE [3]	phrases	No	0.72
wav2vec2 + CNN [14]	phrases	No	0.87
Transformers [30]	phrases + /aiu/	Yes	0.91
os+praat + XGB (ours)	/a/	No	0.88

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we evaluated handcrafted and self-supervised features for pathological voice detection tasks via different approaches: evaluation of single features, early fusion, and late fusion. Results of single feature evaluation reveal that handcrafted features are still competitive to SSL-based acoustic features; *SSL features generally obtained higher scores with /i/ vowel while handcrafted features performed best with /a/ vowels*. We showed that ensembles of different features and vowels significantly improved the multi-metric performance compared to using a single feature or vowel. The proposed method achieved state-of-the-art performance (F1-score) based on benchmarks with previous studies on the same test set on the SVD dataset.

It is difficult to find which material and feature are the most important for pathological voice detection based on different metrics. While we evaluated six metrics for binary classification, each metric has its own limitation. For instance, F1 score only take account positive label while AUC includes probabilities. Future work could tackle the limitation of the F1 score AUC by using more recent metrics like the Matthew correlation coefficient (MCC), which considers true and false positives and negatives.

Since the nature of the problem of detecting pathological voices can be classified as anomaly detection, future work can also be accomplished to observe the effectiveness of anomaly detection methods for pathological voice detection.

REFERENCES

[1] 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," *Biomed. Signal Process. Control*, vol. 48, pp. 128–143, 2019.

[2] A. Sasou, "Automatic identification of pathological voice quality based on the GRBAS categorization," in *Proc. - 9th Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2017*, vol. 2018-Febru, no. December, 2018, pp. 1243–1247.

[3] M. Huckvale, Z. Liu, and C. Buciuleac, "Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech," *Biomed. Signal Process. Control*, vol. 86, no. PB, p. 105201, 2023.

[4] E. C. Nunes, "Anomalous Sound Detection with Machine Learning: A Systematic Review," no. January, 2021.

[5] P. Barche, K. Gurugubelli, and A. K. Vuppala, "Comparative Study of Filter Banks to Improve the Performance of Voice Disorder Assessment Systems using LTAS Features," *Apsipa*, no. December, pp. 1–6, 2021.

[6] M. Pützer and J. Koreman, "A German database of patterns of pathological vocal fold vibration," *Phonus*, vol. 3, pp. 143–153, 1997.

[7] M. Pützer and W. J. Barry, "Saarbrucken Voice Database," 2007.

[8] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo, "Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program," *Eur. Arch. Oto-Rhino-Laryngology*, vol. 265, no. 4, pp. 465–476, apr 2008.

[9] L. M. Jesus, I. Belo, J. Machado, and A. Hall, "The Advanced Voice Function Assessment Databases (AVFAD): Tools for Voice Clinicians and Speech Research," in *Adv. Speech-language Pathol.* InTech, sep 2017, vol. 32, no. tourism, pp. 137–144.

[10] U. Cesari, G. De Pietro, E. Marciano, C. Niri, G. Sannino, and L. Verde, "A new database of healthy and pathological voices," *Comput. Electr. Eng.*, vol. 68, no. December 2017, pp. 310–321, 2018.

[11] P. R. Walden, "Perceptual Voice Qualities Database (PVQD): Database Characteristics," *J. Voice*, vol. 36, no. 6, pp. 875.e15–875.e23, 2022.

[12] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art," *Biomed. Signal Process. Control*, vol. 51, pp. 181–199, 2019.

[13] M. Huckvale and C. Buciuleac, "Automated detection of voice disorder in the Saarbrücken voice database: Effects of pathology subset and audio materials," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 6, pp. 4850–4854, 2021.

[14] D. Ribas, M. A. P. Yoldi, A. Miguel, D. Martínez, A. Ortega, and E. Lleida, "S3prl-Disorder: Open-Source Voice Disorder Detection System based in the Framework of S3PRL-toolkit," in *IberSPEECH 2022*, no. November. ISCA: ISCA, nov 2022, pp. 136–140.

[15] D. Ribas, M. A. Pastor, A. Miguel, D. Martinez, A. Ortega, and E. Lleida, "Automatic Voice Disorder Detection Using Self-Supervised Representations," *IEEE Access*, vol. 11, no. February, pp. 14 915–14 927, 2023.

[16] A. Sasou and Y. Chen, "Comparison of GIF- and SSL-based Features in Pathological-voice Detection," in *INTERSPEECH 2023*, no. August. ISCA: ISCA, aug 2023, pp. 2893–2897.

[17] B. T. Atmaja and A. Sasou, "Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition," *IEEE Access*, vol. 10, pp. 124 396–124 407, 2022.

[18] B. T. Atmaja, A. Sasou, and F. Burkhardt, "Uncertainty-Based Ensemble Learning For Speech Classification," in *Oriental COCOSDA*, 2024, pp. 1–6.

[19] F. Eyben, F. Weninger, M. Wollmer, B. B. Schuller, M. Wöllmer, and B. B. Schuller, *OpenSMILE - The Munich versatile and fast open-source audio feature extractor*, 2015, no. December.

[20] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[21] D. R. Feinberg, "Parselmouth Praat Scripts in Python," 2018.

[22] P. Boersma and V. van Heuven, "Speak and unSpeak with Praat," *Glot Int.*, vol. 5, no. 9-10, pp. 341–347, 2001.

[23] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *J. Phon.*, vol. 71, no. 2018, pp. 1–15, 2018.

[24] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," in *Interspeech 2021*, vol. 3. ISCA: ISCA, aug 2021, pp. 721–725.

[25] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[26] S. Chen, C. Wang, Z. Z. Chen, Y. Wu, S. Liu, Z. Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Y. Qian, Y. Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2021.

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 2, pp. 321–357, jun 2002.

[28] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," mar 2016.

[29] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. W. Schuller, "Nkululeko: A Tool For Rapid Speaker Characteristics Detection," in *Proc. Lr.*, 2022, pp. 395–410.

[30] A. Koudounas, G. Ciravegna, M. Fantini, G. Succo, E. Crosetti, T. Cerquitelli, and E. Baralis, "Voice Disorder Analysis: a Transformer-based Approach," in *interspeech 2024*, 2024.