Unsupervised Anomalous Sound Detection Using Timbral and Human Voice Disorder-Related Acoustic Features

Malik Akbar Hashemi Rafsanjani[1], Candy Olivia Mawalim[2], Dessi Puji Lestari[1], Sakriani Sakti[3], Masashi Unoki[2]
[1] Bandung Institute of Technology, Bandung, Indonesia
E-mail: {13520105@std, dessipuji@staff}.stei.itb.ac.id
[2] Japan Advanced Institute of Science and Technology, Nomi, Japan
E-mail: {candylim, unoki}@jaist.ac.jp
[3] Nara Institute of Science and Technology, Nara, Japan
E-mail: ssakti@is.naist.jp

***Abstract*—Anomalous sound detection (ASD) crucially prevents industrial accidents by distinguishing normal and abnormal machine sounds. Previous research utilizing timbral and short-term features attained a notable F1 score of 0.920. However, relying solely on supervised learning models is impractical due to the difficulty of acquiring anomaly data. This study focuses on developing an unsupervised learning model for ASD, emphasizing prominent timbral features. We also investigate human voice disorder (HVD)-related features, which are potentially linked to human perception of anomalous sounds in machines. We conducted a comparative analysis using 5-fold cross-validation to evaluate our proposed method, with the area under receiver operating characteristic (ROC AUC) as the metric. The proposed ASD method using timbral and HVD-related features significantly improved AUC by 10.87% compared to the baseline system in the DCASE Challenge 2020.**

I. Introduction

Anomalous sound detection (ASD) is the task of recognizing whether the sound emitted from a machine is normal or not [1]. ASD is crucial for early detection, preventing industrial accidents and mechanical failures [1], [2]. These objectives are crucial since industrial machines are complex and prone to malfunctions.

Supervised learning requires labeled data, but accurate anomaly labels are often difficult to obtain [3]. As a result, many ASD methods use unsupervised learning, which trains only on normal data. This approach can handle large amounts of unlabeled data from continuous monitoring systems [4].

ASD studies often use spectrograms and Mel frequency cepstral coefficients (MFCCs) as features [2]. Yet, recent research has explored timbral features, which are associated with human perception of sound differences [5]. Using Support Vector Machines (SVMs), this research achieved an average accuracy of 0.984 using only 7 features. However, this study uses supervised learning which may not be ideal for ASD.

This paper aims to address two issues in earlier studies. First, the performance of utilizing timbral features in unsupervised learning is unknown. Second, research on utilizing human perception-related features, particularly those associated with human voice disorder (HVD), including pathological features, for ASD remains underexplored.

Human perception is highly attuned to subtle variations in sound quality associated with HVD, such as asthenia, roughness, and strain. These variations are often accompanied by changes in specific acoustic features, such as shimmer, jitter, and fundamental frequency (F0) [6]. Since anomalous machine sounds also produce changes in acoustic features, similar features, which capture analogous variations in industrial machine sounds, could be valuable for ASD.

We propose an unsupervised ASD approach using a Gaussian Mixture Model (GMM) with timbral and HVD-related features. GMM is a widely used model for unsupervised anomaly detection and is chosen for its ability to model complex distributions. We hypothesize that both timbral and HVD features could significantly enhance ASD performance.

This paper is part of the ASEAN IVO 2023 project, 'Spoof Detection for Automatic Speaker Verification', which focuses on distinguishing between normal and spoofed sounds. In alignment with this goal, the paper examines methods for improving the detection of normal versus anomalous machine sounds. Furthermore, the project's objective to explore pathological features for spoof detection is reflected here, as pathological features are also investigated for anomalous sound detection.

II. Related Work

A. Anomalous Sound Detection

ASD is the task of identifying whether a sound is normal or anomalous [2]. ASD can be applied in security, network monitoring, and machine maintenance [7]. Unsupervised techniques are very suitable for use in the context of anomaly detection as they only require normal data for training. These approaches are advantageous given the challenges in obtaining accurate labels for anomalous sounds [3]. Additionally, unsupervised methods can handle large amounts of unlabeled data from continuous monitoring systems [4].

ASD studies often use MIMII and ToyADMOS datasets [2], which contain anomalous and normal sounds of machine operations [8], [9]. Inspectors, who manage machine maintenance, rely on their senses, especially hearing, to detect anomalies [5]. This approach drives the usage of timbral

features, which refer to the attribute of sound that allows humans to distinguish among different sound sources [10]. This human-centric approach aligns the technology with practical inspection experience, making it more intuitive and effective.

B. Human Voice Disorder-Related Features

HVD-related features are crucial in various fields, including healthcare and speech therapy. These features encompass a wide range of characteristics related to voice quality and may offer unique insights into sound abnormalities in ASD. The GRBAS scale is a well-known HVD measurement tool [11]. It rates five parameters—grade, roughness, breathiness, asthenia, and strain—on a scale from 0 (normal) to 3 (most severe), correlating well with features like shimmer and jitter.

Studies have shown that shimmer and jitter are closely linked to speech impairments from Parkinson's disease [12], and the standard deviation of F0 and asthenia are negatively correlated [6]. Additionally, jitter local and breathiness are significantly correlated [13].

Acoustic features each represent specific audio properties: jitter indicates fundamental frequency perturbation [12], shimmer shows amplitude perturbation [12], HNR measures the harmonic-to-noise ratio [14], and fundamental frequency (F0) is the smallest true period in the audio [15]. Details of these features and their relation to HVD are summarized in Table I. The following equations express the derivation of HVD-related acoustic features [14], [16], [17]:

$$\text{Shimmer(local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i} \times 100, \quad (1)$$

$$\text{Shimmer(local\_dB)} = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(\frac{A_{i+1}}{A_i})|, \quad (2)$$

$$\text{Shimmer(apqL)} = \frac{\frac{100}{N-L+1}\sum_{i=1+\frac{L-1}{2}}^{N-\frac{L-1}{2}}\left|A_i - \frac{1}{L}\sum_{k=i-\frac{L-1}{2}}^{i+\frac{L-1}{2}} A_k\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i}, \quad (3)$$

$$\text{Shimmer(dda)} = 3 \times \text{Shimmer(apq3)}, \quad (4)$$

$$\text{Jitter(local)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i} \times 100, \quad (5)$$

$$\text{Jitter(local\_absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|, \quad (6)$$

$$\text{Jitter(ppqL)} = \frac{\frac{100}{N-L+1}\sum_{i=1+\frac{L-1}{2}}^{N-\frac{L-1}{2}}\left|T_i - \frac{1}{L}\sum_{k=i-\frac{L-1}{2}}^{i+\frac{L-1}{2}} T_k\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i}, \quad (7)$$

$$\text{Jitter(rap)} = \text{Jitter(ppq3)}, \quad (8)$$

$$\text{Jitter(ddp)} = 3 \times \text{Jitter(rap)}, \quad (9)$$

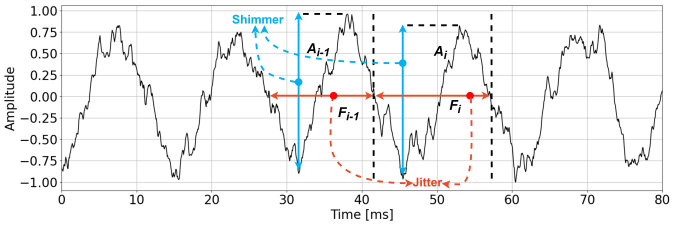


Fig. 1: Jitter and shimmer illustration

TABLE I: Several acoustic features and their relation to HVD

Features	Relation to the HVD
Shimmers	Grade[6], [18], [13], Roughness [6], [13], Breathiness [6], [13], Asthenia [18], [13], laryngeal diseases [19], hoarseness and lesions of larynx [20]
Jitters	Grade[18], Breathiness [6], [18], [13], Asthenia [6], [18], [13], laryngeal diseases [19], hoarseness and lesions of larynx [20]
Fundamental Frequency (F0)	Asthenia [6], Strain [6], background noise-related hearing impairment [21]
Harmonic-to-Noise Ratio (HNR)	Grade[6], [18], Breathiness [6], Asthenia [6], [18], Strain [18]

where A_i, T_i, and N are the i-th amplitude, fundamental period of the input signal of the sound, and the number of samples, respectively. Those features are visualized in Fig. 1.

Additional features include the directional perturbation factor (DPF) and the phonatory frequency range (PFR). DPF indicates vocal fold vibratory perturbation by measuring waveform period changes [22]. PFR is the maximum and minimum F0 that the speaker can produce [23], and is related to the superior laryngeal nerve paresis or paralysis[24]. We also use pitch period entropy (PPE), which is suitable for predicting Parkinson's disease by detecting dysphonia [25].

Another potential HVD-related feature is the harmonic structure of sustained vowels, which has been recognized as an essential feature for voice pathology identification [14]. In that research, the mean and standard deviation of the harmonic structure of the sustained vowels H_p and $RelH_p$, which represent the inverse of the sum of the absolute value of the mean and standard deviation of H_p, respectively, are used as features to classify amyotrophic lateral sclerosis (ALS) patients. Lastly, the pathological vibrato index (PVI) is useful for detecting bulbar dysfunction in ALS patients [26].

III. Proposed Method

The proposed method involves three main steps: feature extraction, model training and optimization, and model testing, as shown in Fig. 2. The data is split into train, validation, and test data. During the training step, the model only uses normal data, leading to higher anomaly scores for anomalous data.

A. Features Extraction

We use the Troparion library [26], a pathological voice analysis tool, to extract HVD-related acoustical features. It uses the instantaneous robust algorithm for pitch tracking (IRAPT) algorithm as a pitch estimation technique. This technique is immune to any pitch modulations and provides more accurate instantaneous pitch values due to time-warping [27].

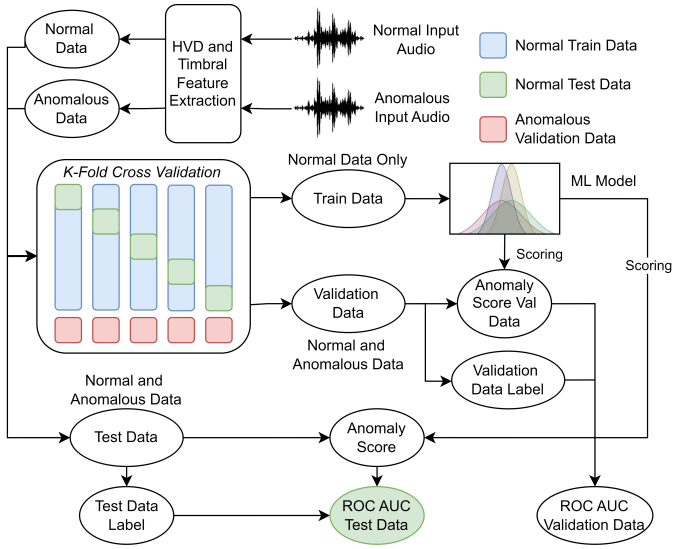


Fig. 2: Proposed method using HVD-related and timbral features for unsupervised anomalous sound detection (ASD)

Additionally, we also implement several features, such as Jitter(ddp), Shimmer(dda), Jitter(local absolute), Shimmer(local dB), and fundamental frequency, given formulas from these studies [14], [16], [17]. In addition, we use the Parselmouth library[1] and Signal Analysis library[2] to calculate two variants of HNR features.

Timbral features are extracted using a model from the Audio Commons project[3] [28]. Using this library, we can extract eight timbral features: hardness, depth, brightness, roughness, warmth, sharpness, booming, and reverberation.

B. Unsupervised Detection Model

We chose a GMM for its effectiveness in probability-based anomaly detection [29]. The GMM, defined by a weighted sum of Gaussian components [30], models complex distributions and identifies data points that deviate from expected patterns. It assumes normal data aligns with the distributions, while anomalies do not fit any clusters. We also trained using a one-class SVM, local outlier factor, and isolation forest, but GMM proved the most suitable, yielding the highest ROC AUC.

C. Model Training

The dataset, which consists of normal and anomaly audios, was extracted to obtain the feature values. These features were divided into training, validation, and test sets. The test set included half of the anomalous data, with an equal amount of normal data. Models were trained solely on normal data and tested on both normal and anomalous data.

The remaining data was split into training and validation sets using K-fold cross-validation ($K = 5$), as suggested by Rodriguez et al. [31], to minimize bias. This method comprehensively evaluates the model's generalizability. The validation set was used for parameter optimization through feature scaling, selection, and hyperparameter tuning.

Three feature groups were evaluated: timbral features, HVD-related acoustic features (HVD features), and their combination. Combining timbral and HVD features involved appending them together. By comparing HVD features with timbral features, we can analyze the capabilities of HVD features for ASD, compared to timbral features, which can be used for ASD. By comparing the combinations of these two groups of features, we can also analyze whether those two groups of features can complement each other.

IV. Experiments

A. Dataset

The Malfunctioning Industrial Machine Investigation and Inspection (MIMII) dataset [8] is a publicly available collection of audio recordings from industrial machines under normal and faulty conditions. This research used the 6 dB SNR subset, chosen for its minimal noise as we are not focused on noise handling. The dataset includes sounds from various machines like valves, pumps, fans, and sliders, with anomalous sounds such as contamination, leakage, rotation imbalance, and rail damage recorded to emulate real scenarios.

B. Evaluation Metrics

The model's performance was evaluated using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The ROC curve depicts the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at various thresholds. ROC AUC is a global measure of a test's ability to discriminate whether a specific condition is present or not [32]. In this context, the AUC provides a single metric summarizing the model's ability to discriminate between normal and anomalous sounds, with higher AUC values indicating better performance. The ROC AUC metric formula can be expressed as

$$\mathrm{AUC}_{m,n,d} = \frac{1}{N_d^- N_n^+} \sum_{i=1}^{N_d^-} \sum_{j=1}^{N_n^+} H(A_\theta(x_\mathrm{j}^+) - A_\theta(x_\mathrm{i}^-)) \quad (10)$$

where m represents the index of a machine type, n represents the index of a section, and $d = \{source, target\}$ represents a domain. The N_d^- is the number of normal test clips in domain d, N_n^- is the number of normal test clips in section n, and N_n^+ is the number of anomalous test clips in section n. Here, $\{x_i^-\}_{i=1}^{N_d^-}$ are normal test clips in domain d in section n and $\{x_j^+\}_{j=1}^{N_n^+}$ are anomalous test clips in section n in machine type m. $H(x)$ is the hard-threshold function that returns 1 when $x > 0$ and 0 otherwise [33].

C. Model Optimization

We used feature scaling, feature selection, and hyperparameter tuning for the optimization because the features extracted are numerical data and contain no NaN values. We evaluated the models' performance using validation data to determine the best technique. For feature scaling, we tried four techniques, with standard scaling producing the best ROC AUC.

[1]https://parselmouth.readthedocs.io/en/stable/
[2]https://pypi.org/project/Signal_Analysis/
[3]https://github.com/AudioCommons/timbral_models

TABLE II: Selected features after feature selection

Machine Type	Timbral Features	HVD Features
Fan	boominess, roughness, warmth	Jitter PPQ5, Jitter absolute, Jitter local, PPE, PPF, Praat.HNR, RelHp, Shimmer local dB, Signal_Analysis.HNR, mean Hp, stdev F0
Pump	brightness, reverb, sharpness, warmth	DPF, Jitter RAP, Jitter local, PPE, PPF, Praat.HNR, RelHp, Shimmer APQ55, Shimmer local dB, mean F0, stdev F0, stdev Hp
Slider	brightness, depth, reverb, roughness, warmth	Jitter PPQ55, Praat.HNR, RelHp, Shimmer APQ11, Shimmer APQ5, Shimmer APQ55, Shimmer local dB, Signal_Analysis.HNR
Valve	brightness, hardness, roughness, sharpness, warmth	Jitter PPQ55, PPE, PPF, PVI, Praat.HNR, Shimmer APQ11, Shimmer APQ3, Shimmer APQ5, Shimmer APQ55, Shimmer local dB, Signal_Analysis.HNR, mean F0, stdev F0

To reduce redundant features and identify beneficial ones, we used feature selection. We utilized the sequential feature search (SFS) technique due to its low computational requirements. We determined the granularity of the feature selection on the machine type level, to incorporate shared characteristics of each machine type and prevent overfitting on the machine ID or model level. Selected features after feature selection are shown in Table II.

For hyperparameter tuning, we used grid search on various parameters, including the number of components, covariance type, convergence threshold, max iteration, and parameter initializer. We found the best combination for the timbral features are 3 mixture components, covariance type full, convergence threshold 0.001, max iteration 100, and K-means as a parameter initializer. We used 5 mixture components, covariance type full, convergence threshold 0.01, max iteration 300, and randomly selected data points as the parameter initializer for HVD features. We used the same parameter as timbral features for the combination of both features.

D. Baseline Model

The ASD model by Morita et al. [34] was used as a baseline model because it uses GMM and the MIMII dataset, the same model and dataset that we use. This ASD model extracts log-mel spectrogram features from the audio input and reduces the dimension using PCA. The output of PCA is aggregated using the mean or variance of the output and then is used as the model input. We chose the GMM and variance methods to be replicated as the model produces better ROC AUC than other combinations.

V. Results

A. Proposed Features Produce Excellent ROC AUC

The overall results are shown in Table III. The results show the performance evaluation of different machine types (fan, pump, slider, and valve) based on three different feature sets: timbral features, HVD features, and combined features. The replicated baseline results are also shown. Both feature sets performed well individually, but their effectiveness varied across different machine types and IDs. On average, these two features were comparable for valves, but HVD features excelled for sliders, while timbral features excelled for fans and pumps. The aggregate average performances across all machine types—0.921 for timbral and 0.912 for HVD—highlight a marginal and slight advantage for the timbral features in detecting anomalous sound.

TABLE III: ROC AUCs for each machine type with GMM and using timbral and HVD features, compared to baseline

Machine Type	Machine ID	Replicated Baseline [34]	Timbral Features	HVD Features	Combined Features
Fan	00	0.768	**0.903**	0.769	0.879
Fan	02	0.854	0.978	0.973	**0.981**
Fan	04	0.754	0.947	0.967	**0.981**
Fan	06	0.847	**0.999**	0.994	0.994
Fan	Avg	0.806	0.957	0.926	**0.959**
Pump	00	0.951	**0.967**	0.960	0.963
Pump	02	**0.917**	0.839	0.891	0.852
Pump	04	0.758	0.948	0.938	**0.978**
Pump	06	0.825	**0.990**	0.847	0.950
Pump	Avg	0.863	**0.936**	0.909	0.936
Slider	00	0.999	0.977	0.994	**1.000**
Slider	02	0.930	0.960	0.983	**0.998**
Slider	04	0.759	0.908	**0.969**	0.962
Slider	06	0.840	0.660	0.680	**0.707**
Slider	Avg	0.882	0.876	0.907	**0.917**
Valve	00	0.933	**1.000**	0.942	**1.000**
Valve	02	0.786	**0.980**	0.912	0.976
Valve	04	0.840	**0.968**	0.957	0.964
Valve	06	0.775	0.717	0.814	**0.820**
Valve	Avg	0.833	0.916	0.906	**0.940**
All	Avg	0.846	0.921	0.912	**0.938**

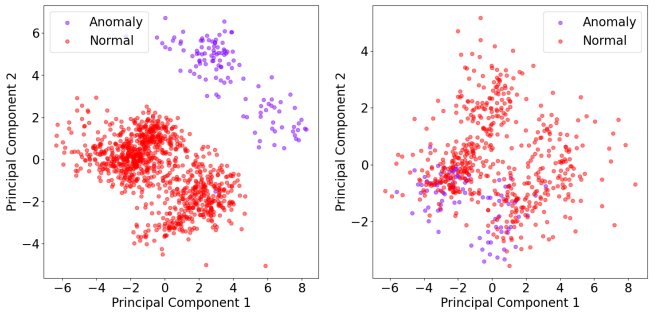
Moreover, the data indicate a notable variance in the performance of individual machine IDs within each machine type, implying that the effectiveness of the feature sets might be influenced by specific machine characteristics. The valve type, for example, performed optimally with timbral features for ID 00 but performed significantly worse for valve ID 06, possibly due to inherent differences in the machine IDs, as different IDs may come from different product models.

Overall, the fact that combined features were better than timbral features and HVD features for every average score of machine types and the aggregated averages across all machine types (0.921 for timbral, 0.912 for HVD, and 0.938 for combined features) demonstrate that while each feature set individually offers substantial discriminative power, the combined features consistently yielded better performance. This result suggests that the integration of multiple feature sets significantly enhanced the anomaly detection capability.

We also compare the result with a recent unsupervised ASD study. A newly proposed unsupervised method, Denoising Sparse Wavelet Network (DeSpaWN), achieves a very good performance for ASD [35]. DeSpaWN is a deep-learning architecture designed to perform denoising and feature extraction from high-frequency time-series signals. It produces an AUC of 0.946 with the 6 dB MIMII dataset. This study's proposed method gives almost similar results using the combined features, which is 0.938. Although only GMM, a shallow learning model, is used, the proposed method produces a very comparable result to that of deep learning. This finding suggests the proposed method gives excellent results, comparable with state-of-the-art methods, while only using shallow learning.

TABLE IV: Average F1 score for each machine type

Machine Type	Timbral Features	HVD Features	Combined Features
Fan	0.901	0.875	**0.905**
Pump	**0.871**	0.825	0.868
Slider	0.835	0.856	**0.889**
Valve	**0.875**	0.812	0.865
All	0.870	0.842	**0.882**



(a) Valve ID 00 (b) Slider ID 06

Fig. 3: Plot of PCA results on the best and the worst performing machine ID using combined features.

B. Operational Prediction Performance

To evaluate the effectiveness of the proposed feature in real-world situations, particularly in predicting whether data is normal or anomalous, the F1 score of the models was tested. This involved finding the optimal threshold for classifying anomaly scores using validation data. The granularity of the threshold was specific to each machine type and model due to their shared characteristics.

The results of the average F1 score of each machine type using the three groups of features are shown in Table IV. Although the F1 score is lower than the ROC AUC score due to the exact threshold from the validation data, the results are still very good. The models produced an F1 score of 0.870 using timbral features, 0.842 using HVD features, and 0.882 using combined features. These results have the same patterns as the ROC AUC results: the best result was achieved by the combined features and the timbral features result was better than the HVD features result. From these findings, we can conclude that the proposed approach can perform very well in real-world situations.

C. PCA Visualization

To understand the varying performance observed across different machine IDs, we compared the visualization of the PCA of scaled combined features data with the standard scaler. We visualize the best (valve ID 00, AUC: 1.000) and the worst (slider ID 06, AUC: 0.707) machine ID results in Fig 3. For valve ID 00, there is a clear separation between the normal (red) and anomalous (purple) data, whereas for slider ID 06, a significant overlap is observed between the two. These results suggest that the varying performance stems from the model's difficulty in distinguishing anomalies in certain machine IDs, where normal and anomalous data exhibit greater characteristic overlap compared to others.

TABLE V: Ablation study results

Feature Group	Relative ROC AUC Score	Features
Shimmer	-0.050	Shimmer.APQ11, Shimmer.APQ3, Shimmer.APQ5, Shimmer.APQ55, Shimmer.local, Shimmer.DDA, Shimmer.local_db
Harmonicity	-0.033	Praat.HNR, Signal_Analysis,HNR
Pathology	-0.024	DPF, mean Hp, stdev Hp, RelHp, PPE, PPF, PVI
F0	-0.003	mean F0, stdev F0
Jitter	-0.003	Jitter PPQ55, Jitter PPQ5, Jitter RAP, Jitter local, Jitter DDP, Jitter absolute

D. Ablation Study

We conducted an ablation study to assess the importance of each newly proposed HVD feature. By grouping related features, removing each group, and calculating the ROC AUC, we determined how critical each group was. The results when no features were excluded were then subtracted from the results. The more negative the result, the more important the group of features is. We conducted the ablation study using the HVD features, and the results are presented in Table V. The findings suggest that the most crucial feature groups are Shimmer, Harmonicity, and Pathology.

Removing the shimmer group resulted in a -0.050 decrease in ROC AUC, indicating that amplitude perturbation can play a crucial role in ASD. Despite the harmonicity group having only two features, its removal caused a -0.033 decrease in ROC AUC, showing the importance of the HNR feature in classifying anomalous sound. In contrast, removing F0 (fundamental frequency) and Jitter, which represents perturbation in fundamental frequency, groups only resulted in a minimal decrease of ROC AUC, indicating modest contributions to ASD performance. These findings suggest that the amplitude and harmonicity aspects are more useful than the frequency aspect for classifying machine sound anomalies. In addition, the pathology group that contains a lot of features that directly correlate to HVD also produced a high decrease in ROC AUC when removed, strengthening our hypothesis that features related to HVD enhance and are beneficial to ASD performance.

VI. Conclusion and Future Work

This study explored human voice disorder (HVD)-related features for unsupervised anomalous sound detection (ASD). Our experimental results showed the feasibility of using HVD-related features for ASD, with an average area under the curve (AUC) of 0.912, as these features are comparable to the timbral features that can be used to detect anomalous sound, with an average AUC of 0.921. The results also highlight the enhanced detection capability achieved by combining HVD-related features with timbral features with an average AUC of 0.938. This work contributes to the development of more robust and efficient unsupervised ASD methods. Future work could investigate the application of different algorithms for estimating the fundamental frequency (F0), such as the YIN and SWIPE algorithms. Utilizing these alternative algorithms may yield more robust HVD-related acoustic features and uncover new aspects of sound anomalies that were not captured by the IRAPT algorithm.

ACKNOWLEDGMENTS

This work was supported by the JST Sakura Science Exchange Program and JSPS KAKENHI grant (23H04344, 23K18491, 22K21304). The ASEAN IVO (http://www.nict.go.jp/en/asean ivo/index.html) project, 'Spoof Detection for Automatic Speaker Verification', was involved in the production of the contents of this publication and financially supported by NICT (http://www.nict.go.jp/en/index.html).

REFERENCES

[1] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma," in *25th European Signal Processing Conference, EUSIPCO 2017, Kos, Greece, August 28 - September 2, 2017*. IEEE, 2017, pp. 698–702.

[2] E. C. Nunes, "Anomalous sound detection with machine learning: A systematic review," *CoRR*, vol. abs/2102.07820, 2021.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, 2009.

[4] G. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2168–2187, 2022.

[5] Y. Ota and M. Unoki, "Anomalous sound detection for industrial machines using acoustical features related to timbral metrics," *IEEE Access*, vol. 11, pp. 70 884–70 897, 2023.

[6] S. V. Freitas, P. M. Pestana, V. Almeida, and A. Ferreira, "Integrating voice evaluation: correlation between acoustic and audio-perceptual measures," *Journal of Voice*, vol. 29, no. 3, pp. 390–e1, 2015.

[7] E. D. Fiore, A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "An anomalous sound detection methodology for predictive maintenance," *Expert Syst. Appl.*, vol. 209, p. 118324, 2022.

[8] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," pp. 209–213, 2019.

[9] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2019, New Paltz, NY, USA, October 20-23, 2019*. IEEE, 2019, pp. 313–317.

[10] K. Patil, D. Pressnitzer, S. A. Shamma, and M. Elhilali, "Music in our ears: The biological bases of musical timbre perception," *PLoS Comput. Biol.*, vol. 8, no. 11, 2012.

[11] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS scale," in *28th International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2006, New York City, NY, USA, August 30 - September 3, 2006, Main Volume*. IEEE, 2006, pp. 2478–2481.

[12] S. S. Upadhya, A. Cheeran, and J. Nirmal, "Statistical comparison of Jitter and Shimmer voice features for healthy and Parkinson affected persons," in *2017 second international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2017, pp. 1–6.

[13] B. Sabir, B. Touri, and M. Moussetad, "Correlation between acoustic measures, voice handicap index and GRBAS scales scores among Moroccan students." *Current Pediatric Research*, vol. 21, pp. 343–353, 04 2017.

[14] M. Vashkevich and Y. Rushkevich, "Classification of ALS patients based on acoustic analysis of sustained vowel phonations," *Biomed. Signal Process. Control.*, vol. 65, p. 102350, 2021.

[15] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[16] J. P. Teixeira and A. Gonçalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Computer Science*, vol. 100, pp. 271–279, 2016.

[17] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.

[18] R. Fujiki and S. Thibeault, "Examining relationships between GRBAS ratings and acoustic, aerodynamic and patient-reported voice measures in adults with voice disorders," *Journal of Voice*, vol. 37, 03 2021.

[19] A. Gelzinis, A. Verikas, and M. Bacauskiene, "Automated speech analysis applied to laryngeal disease categorization," *Comput. Methods Programs Biomed.*, vol. 91, no. 1, pp. 36–47, 2008.

[20] H. Lathadevi and S. P. Guggarigoudar, "Objective acoustic analysis and comparison of normal and abnormal voices." *Journal of Clinical & Diagnostic Research*, vol. 12, no. 12, 2018.

[21] C. A. Brown and S. P. Bacon, "Fundamental frequency and speech intelligibility in background noise," *Hearing research*, vol. 266, no. 1-2, pp. 52–59, 2010.

[22] M. B. Higgins and J. H. Saxman, "A comparison of intrasubject variation across sessions of three vocal frequency perturbation indices," *The Journal of the Acoustical Society of America*, vol. 86, no. 3, pp. 911–916, 1989.

[23] G. J. Cler, V. S. McKenna, K. L. Dahl, and C. E. Stepp, "Longitudinal case study of transgender voice changes under testosterone hormone therapy," *Journal of Voice*, vol. 34, no. 5, pp. 748–762, 2020.

[24] C. A. Eckley, R. T. Sataloff, M. Hawkshaw, J. R. Spiegel, and S. Mandel, "Voice range in superior laryngealnerve paresis and paralysis," *Journal of Voice*, vol. 12, no. 3, pp. 340–348, 1998.

[25] M. A. Little, P. E. McSharry, E. J. Hunter, J. L. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, 2009.

[26] M. Vashkevich, A. Petrovsky, and Y. Rushkevich, "Bulbar ALS detection based on analysis of voice perturbation and vibrato," in *Signal Processing: Algorithms, Architectures, Arrangements, and Applications, SPA 2019, Poznan, Poland, September 18-20, 2019*. IEEE, 2019, pp. 267–272.

[27] E. Azarov, M. Vashkevich, and A. A. Petrovsky, "Instantaneous pitch estimation based on RAPT framework," in *Proceedings of the 20th European Signal Processing Conference, EUSIPCO 2012, Bucharest, Romania, August 27-31, 2012*. IEEE, 2012, pp. 2787–2791.

[28] A. Pearce, B. Tim, and M. Russell, "Release of timbral characterisation tools for semantically annotating non-musical content," 2020.

[29] R. Blanco, P. Malagón, S. Briongos, and J. M. Moya, "Anomaly detection using gaussian mixture probability model to implement intrusion detection system," in *Hybrid Artificial Intelligent Systems: 14th International Conference, HAIS 2019, León, Spain, September 4–6, 2019, Proceedings 14*. Springer, 2019, pp. 648–659.

[30] D. A. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics, Second Edition*, S. Z. Li and A. K. Jain, Eds. Springer US, 2015, pp. 827–832.

[31] J. D. Rodríguez, A. P. Martínez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, 2010.

[32] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?" pp. 357–359, 2017.

[33] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *CoRR*, vol. abs/2305.07828, 2023.

[34] K. Morita, T. Yano, and K. Q. Tran, "Anomalous sound detection by using local outlier factor and gaussian mixture model," in *Proceedings of the 5th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE), Tokyo, Japan*, 2020, pp. 2–4.

[35] G. Michau, G. Frusque, and O. Fink, "Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series," *Proceedings of the National Academy of Sciences*, vol. 119, no. 8, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1073/pnas.2106598119