IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 June 2023, accepted 7 July 2023, date of publication 11 July 2023, date of current version 18 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294334

RESEARCH ARTICLE

Anomalous Sound Detection for Industrial Machines Using Acoustical Features Related to Timbral Metrics

YASUJI OTA, (Member, IEEE), AND MASASHI UNOKI, (Member, IEEE)
School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 932-1292, Japan

Corresponding author: Yasuji Ota (y_ota@jaist.ac.jp)

This work was supported by the Fund for the Promotion of Joint International Research (Fostering Joint International Research (B)) (20KK0233) and Grant-in-Aid for Transformative Research Areas (A) (23H04344).

ABSTRACT This paper proposes an anomalous sound detection (ASD) method that uses a combination of timbral metrics and short-term features tailored to industrial machine faults to identify whether the sound emitted from a target machine is anomalous. The timbral-feature-based ASD (TF-ASD) method involves using five timbral metrics and two developed features as auditory features and a support vector machine (SVM) for classification. We develop two types of short-term features to estimate the change in the fluctuation of sound waves and pitch in terms of harmonics to improve the time resolution of the timbral analysis. This combination of timbral metrics and our two short-term features is based on an investigation of timbral association with industrial machine malfunction from the viewpoint of ‘‘noticeable difference in hearing’’ that is the human ability to discriminate differences in sounds. We evaluated the ASD performance of our method in terms of SVM classification using the MIMII (Malfunctioning Industrial Machine Investigation and Inspection) dataset. The results indicate that the proposed method has excellent classification performance with an accuracy of 0.984 on average for emitted sounds of 16 machine types and models. This demonstrates that the combination of timbral metrics and our short-term features in accordance with the ‘‘noticeable difference in hearing’’ is effective for ASD.

INDEX TERMS Anomalous sound detection, timbral metrics, industrial machine faults, support vector machine.

I. INTRODUCTION

Daily maintenance of industrial machines is essential to ensure safe operation for efficient production and business management. Inspectors, who are in charge of the maintenance of industrial machines, use their knowledge to detect anomalous situations by using their senses, i.e., sight, sound, smell, and touch. Monitoring based on acoustics, i.e., hearing is excellent due to its instantaneousness, wide angle of acceptance, and large dynamic range of sensitivity. Inspectors have excellent skills in discriminating differences in sounds using their ‘‘noticeable difference in hearing’’.

However, there are practical issues with fault diagnosis using human hearing due to ambiguous criteria and performance variations that heavily depend on individual skills and knowledge. Therefore, it is difficult to manage human resources dynamically to balance workload and cost while maintaining stable manufacturing. Therefore, the Japanese government states their requirements for technologies for automation solutions of operation and inspection in their future vision, and anomalous sound detection (ASD) is expected to help inspectors identify whether the sound emitted from a target machine is anomalous [1].

One of the most well-known worldwide ASD competitions for industrial machines is Task 2 of Detection and Classification of Acoustic Scenes and Events (DCASE) [2], [3]. In this task, a common dataset of operation sounds emitted

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Zia Ur Rahman.

from real industrial machines is provided for anomalous detection. The basic ASD framework tends to be a cascade of acoustical feature extraction from operation sounds and machine-learning-based classification of anomalies using the extracted features. As a result of this competition, several promising classification methods have been developed based on machine learning [4], [5], [6], [7]. Fundamental but primitive analysis metrics, such as Mel-frequency cepstrum coefficients or log-Mel energies, are generally used as acoustic features. Therefore, the development of acoustical features could be seen as making relatively little progress compared with that of machine-learning techniques.

Anomalous sound detection aims to develop computational methods to detect anomalies that deviate from the spatial or temporal regularity that the "normal" acoustical sounds follow. Therefore, the approaches can be categorized into two areas; one is to deal with outliers in acoustical-feature space for detection. The other is to detect the anomaly change in sound as its time series. In both categories, it is important how to deal with features for efficient detection. In recent works, temporal modulation features on the gamma-tone auditory filterbank are used to detect outliers [8]. From the change in time series perspective, meta-features are derived as 1-D sequence to describe the local dynamics with arbitrary length [9] or multivariate time series are projected to a lower dimensional space with a space-embedding strategy [10]. These two approaches are designed to capture abrupt change with relatively low computational cost. We think that it is important to develop practical common features that can represent the local dynamics with different temporal resolutions and then captures the temporal change that can detect incipient failures using the common features. In this study, we focus on the issues to pursue the common features of ASD on industrial machines.

From the machine-fault-analysis point of view, diagnosis methods have been proposed for detecting faults of specified machine types or their parts, e.g., rotary machines or bearings, by using perceptive features. One of these methods derived a wide set of statistical features, such as mean, standard deviation, skewness, and kurtosis of vibration signals, was to feed an anomalous classifier of a support vector machine (SVM) using vibration signature [11], [12]. A more intuitive approach is mimicking human diagnosis. Acoustical features related to timbre are important classifiers to detect anomalies because timbre is a promising indicator of anomalous sound perception. There have been several studies on acoustical features regarding timbre. One used the difference of rotational period of acoustic signals in the adjacent frame, which was derived based on the discrete Fourier transform spectrum, as a timbral feature [13]. Another method uses a combination of timbre, roughness, and sharpness, and other audio metrics, such as loudness and fluctuation strength. These metrics are highly correlated with audible perception [14]. By specifying the machine type or its parts, these methods statistically discover the relation between observed perceptive signals and various machine faults using machine-learning techniques. However, to address the factory inspection issue, an ASD method that has an explainable decision criterion and is widely applicable to primary industrial machinery is required.

When we consider the implementation of ASD to address issues of industrial-maintenance workload, a human-centric ASD method, with which the detection point matches what anomalous sounds the inspector can hear from the target machine, is substantial and practical. The following question arises, "What is the key to discriminating differences in anomalous sounds from the perspective of timbre?" To answer this question, we investigate using timbre as the key to deriving related metrics and features for classification in a signal-processing manner.

We, therefore, propose a timbral-feature-based ASD (TF-ASD) method that involves the use of a combination of timbral metrics (TMs) and two short-term features to identify whether the sound emitted from a target machine is anomalous. TF-ASD is comprised of timbral feature extraction from sounds emitted from industrial machines, forming a proper combination of the features to fit the machine's property, training with a combination of features, and classification of anomaly based on the training model by incorporating machine learning technology. To be a cogent approach for users like inspectors, we focused on their "noticeable difference in hearing" from the sound emitted from industrial machines. We investigated typical causes of malfunctioning of four types of industrial machinery, then selected several timbres that relate perception from the emitting sound of malfunctioning from the "noticeable difference in hearing" perspective. We used "onomatopoeia" [15] as a mediator to bridge human "noticeable difference in hearing" and their auditory perception. We assume that inspectors intuitively manipulate onomatopoeia to accumulate and exchange maintenance knowledge. Based on this assumption, an adequate number of timbres plays an important role in classifying machine anomalous sounds [16].

Based on the original investigation, we used five timbral metrics (TMs), boominess, brightness, depth, roughness, and sharpness, developed by the University of Surry. Since the TMs express the overall characteristics with long-term analysis, we developed two types of short-term features as complementary characteristics for the TMs, the first feature estimates the change in the fluctuation of sound waveforms and the second feature estimates changes in pitch, in other words, tone-height perception in terms of harmonics. To maximize the significance of applicable timbre-related features for ASD to be associated with the human "noticeable difference in hearing", four or five TMs are first selected to fit the target machine type, and the short-term features are added to the selected TMs to improve the time resolution of the timbral analysis. The combination of TMs and short-term features is incorporated to detect anomalous sounds emitted from a target machine by using machine-learning techniques.

On the premise that both "normal" and "abnormal" operating sounds are already known, we verify the effectiveness of the anomalous detection of timbral-related features as binary classification performance. We incorporated a support vector machine as the classification technique because SVM classifiers perform well in high-dimensional space and have excellent accuracy.

We then evaluated the TF-ASD performance of our proposed method as an SVM classification assessment using a publicly available industrial sound dataset. The evaluation results showed that our proposed ASD method can provide a stable classification performance with accuracy over 0.97 for all four typical industrial types of machinery and demonstrated our TF-ASD is superior to the recent conventional method dedicated to bearing faults using sound quality metrics.

The rest of this paper is organized as follows. In Section II, we briefly introduce TMs. In Section III, we present our proposed method and the evaluation of its performance in Section IV. After discussing the evaluation results in Section V, we conclude the paper in Section VI.

II. TIMBRAL METRICS AND THEIR IMPLEMENTATION

Humans obtain information about their surroundings from sound, i.e., auditory perception, and many psycho-acoustic studies have been conducted to determine the relation between acoustical analysis and timbre, which is a perceptual variable [17]. Since timbre is multidimensional, it is separated into several attributes corresponding to adjectival phrases, such as sharpness or roughness [18].

Many studies have been conducted on timbral modeling and implementation of each timbral attribute as objective metrics [19]. In one seminal study, the University of Surry developed timbral models in the Audio Commons project, and these models are widely used in psycho-acoustic research [20]. The algorithm is based on kinds of literature describing exemplary computational models and subjective experiments. Furthermore, it is useful for statistical analysis in that the calculated metric can be dealt with as an indicator.

A. SHARPNESS

Sharpness, which is a metric related to sharp or shrill sensation, increases in magnitude with shifting the center frequency to a higher region. From this perspective, Zwicker defined 1 acum as a unit of a narrow-band noise centered at 1, 000 Hz with a loudness level of 60 phon [18]. A sharpness model was then constructed based on the acum and expressed as

$$S = 0.11 \, \frac{\int_0^{24\mathrm{Bark}} N'(z)\, g(z)\, z\, dz}{\int_0^{24\mathrm{Bark}} N'\, dz} \tag{1}$$

where S is sharpness, $N'(z)$ is the loudness density in the critical-band rate z, and $g(z)$ is the weighting factor of S at z. Loudness is the intensive attribute of an auditory sensation. The loudness level of a sound, in phons, is the sound pressure level in dB of a pure tone of frequency 1 kHz that is judged to be equivalent in loudness [21]. From psycho-acoustic experiments, the weighting factor was defined as unity (1.0) in the frequency range up to 3, 000 Hz, and increased rapidly up to four (4.0) for higher frequencies.

B. ROUGHNESS

Roughness, describing buzzing, harsh, raspy sound quality, is strongly related to the change in the modulation frequency of loudness [18], [22]. From studies on a sinusoidal model approach, a roughness-calculation model was proposed [23]. This model is constructed from three elements based on the assumption that a signal has two sinusoidal components. The components are specified with frequencies f_1, f_2 and amplitudes A_1, A_2, where $f_{\mathrm{min}} = \min(f_1, f_2)$, $f_{\mathrm{max}} = \max(f_1, f_2)$, $A_{\mathrm{min}} = \min(A_1, A_2)$, $A_{\mathrm{max}} = \max(A_1, A_2)$.

Roughness in a 50 ms frame, R_{frame} is calculated as

$$R_{\mathrm{frame}} = X^{0.1} \times 0.5\,(Y^{3.11}) \times Z \tag{2}$$

where

$$X = A_{\mathrm{min}} \times A_{\mathrm{max}} \tag{3}$$

The term $X^{0.1}$ is the dependence of roughness on intensity related to the amplitude of two sinusoidal components.

$$Y = 2 \times A_{\mathrm{min}} \,/\, (A_{\mathrm{min}} + A_{\mathrm{max}}) \tag{4}$$

The second term $Y^{3.11}$ is the dependence of roughness on the amplitude-fluctuation degree related to the amplitude of two sinusoidal components.

$$Z = e^{-3.5c\,(f_{\mathrm{max}} - f_{\mathrm{min}})} - e^{-5.75c\,(f_{\mathrm{max}} - f_{\mathrm{min}})} \tag{5}$$

where

$$c = \frac{0.24}{0.0207 f_{\mathrm{min}} + 18.96} \tag{6}$$

The third term Z is the dependence of roughness on the amplitude-fluctuation rate, which is the frequency difference between two sinusoidal components.

Finally, the overall roughness is calculated using a regression formula with the mean of all frame roughness values [24].

C. BOOMINESS

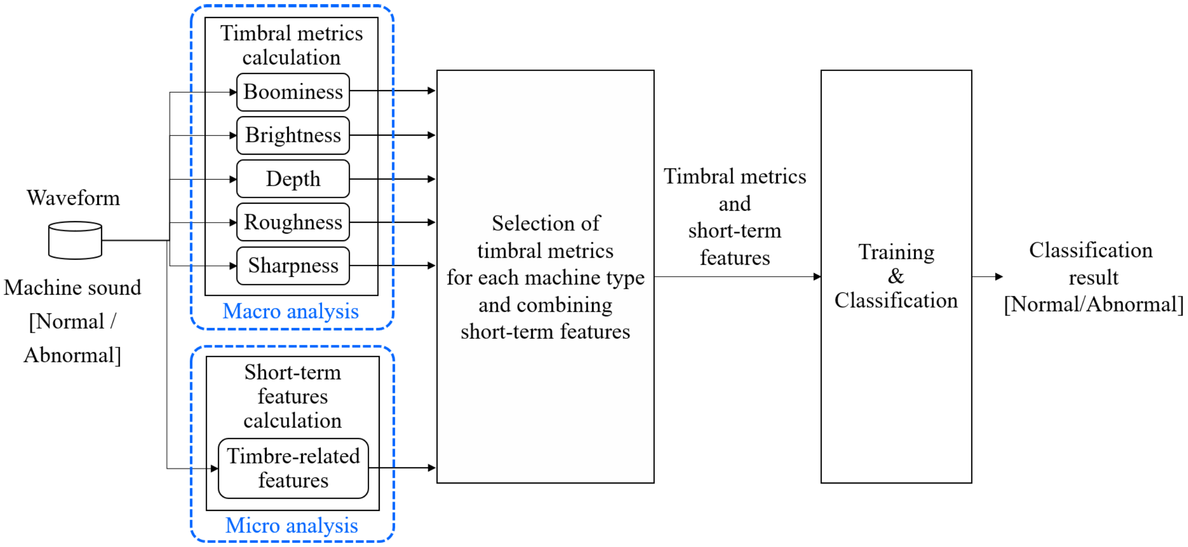
Boominess is measured to evaluate booming sensation by using a method that is based on the power summation of the 1/3 octave band signals [25]. Since booming might be often perceived as a low-pitch vibration, it uses the ratio of loudness below 280 Hz to the total bandwidth.

The booming degree is calculated as the booming index by using the following equation.

$$\text{Booming index} = \mathit{Bandsum} \times (S_{\mathrm{l}}/S_{\mathrm{t}}) \tag{7}$$

where "*Bandsum*" is the power summation of the 1/3 octave band signals, S_{t} is the loudness of the total band, and S_{l} is the loudness of the band below 280 Hz.

In the implementation of the timbral models, linear regression is used to obtain the final boominess value [26].



Waveform
Machine sound [Normal / Abnormal]
Timbral metrics calculation
Boominess
Brightness
Depth
Roughness
Sharpness
Macro analysis
Short-term features calculation
Timbre-related features
Micro analysis
Selection of timbral metrics for each machine type and combining short-term features
Timbral metrics and short-term features
Training & Classification
Classification result [Normal/Abnormal]

FIGURE 1. Block diagram of the proposed method.

D. BRIGHTNESS

It has been reported that the spectral centroid and ratio of high frequencies to the overall energy correlates with perceived brightness [27]. Therefore, a brightness model was previously developed to incorporate both a spectral centroid variant and spectral energy ratio.

The frequency-limited spectral centroid, SC_{br}, was calculated in bandwidth over 2, 000 Hz and the energy ratio, ER_{br}, was calculated as the proportion of energy over 2, 000 Hz to energy over 20 Hz.

$$SC_{\mathrm{br}} = \frac{\sum_{k\in\{2{,}000\text{ to }8{,}000\text{ Hz}\}} f(k)\,m(k)}{\sum_{k\in\{2{,}000\text{ to }8{,}000\text{ Hz}\}} m(k)} \tag{8}$$

$$ER_{\mathrm{br}} = \frac{\sum_{k\in\{2{,}000\text{ to }8{,}000\text{ Hz}\}} m(k)}{\sum_{k\in\{20\text{ to }8{,}000\text{ Hz}\}} m(k)} \tag{9}$$

where $f(k)$ is the frequency of the k^{th} bin, and $m(k)$ is the magnitude of the k^{th} bin.

Finally linear regression with the SC_{br} and ER_{br} is used to obtain the final brightness value [24].

E. DEPTH

Depth, defined as a timbral, not a spatial attribute, has been reported to be related to an emphasized low-frequency component [24]. Therefore, a model was developed for calculating the depth metric by conducting linear regression with mainly two elements, a spectral centroid, SC_{dp}, in the range of 20 to 2, 000 Hz, and the ratio of energy, ER_{dp}, between 20 and 500 Hz.

$$SC_{\mathrm{dp}} = \frac{\sum_{k\in\{20\text{ to }8{,}000\text{ Hz}\}} f(k)\,m(k)}{\sum_{k\in\{20\text{ to }8{,}000\text{ Hz}\}} m(k)} \tag{10}$$

$$ER_{\mathrm{dp}} = \frac{\sum_{k\in\{20\text{ to }500\text{ Hz}\}} m(k)}{\sum_{k\in\{20\text{ to }8{,}000\text{ Hz}\}} m(k)} \tag{11}$$

where $f(k)$ is the frequency of the k^{th} bin, and $m(k)$ is the magnitude of the k^{th} bin [24].

III. PROPOSED METHOD

A. BASIC CONFIGURATION

The proposed method is configured as a sequence of TM calculation, short-term features calculation, selection of the TMs to fit each machine type, combining the selected TMs and short-term features, and training / classification with a set of selected TMs and short-term features, as shown in Fig. 1.

We introduce multiple temporal analysis in which TMs extract the overall characteristics, that is "macro analysis" and short-term features extract the local dynamics by timbral-related characteristics with a relatively high temporal resolution, that is "micro analysis".

B. MACHINE FAULT AND TIMBRE FROM THE NOTICEABLE DIFFERENCE IN HEARING

In a factory, various machines are continuously working, while the operation sound expresses their status. Inspectors use their ability to discriminate differences in machine sounds, named "noticeable difference in hearing", as a non-intrusive diagnosis. From interviews we conducted with inspectors of a Japanese factory, we found that they tend to describe their impression of manufacturing-machine sound with onomatopoeia [28]. The on-site observation indicated that onomatopoeia can be a useful interpreter to associate the inspector's noticeable difference in hearing in industrial machine sounds from the perspective of timbre. Therefore, we consider the association between industrial machine fault and timbre perceived as audible differences by describing impressions using onomatopoeia.

We exemplify the association among machine fault, noticeable difference in hearing, and perception as timbre with four types of industrial machinery as representatives, i.e., rotating machinery, such as fans or motors; sliding machinery, such as sliders or grinders; striking machinery, such as valves or pressers; and liquid manipulators such as pumps or compressors.

1) ROTATING MACHINERY

The most typical fault in rotating machinery is rotor unbalance [29]. It has been reported that the rotational energy is transferred into vibration due to its inertia centrifugal force. Other faults, such as rotor-to-stator rubbing or rotor cracking, increase the friction between the inner parts of rotating machinery. The maintenance procedure for rotating and moving parts is using lubricant, which helps make the parts move smoothly. However, a lack of lubricant causes several faults due to the increase in friction.

Industrial fans emit booming and whir or whizz sounds due to rotating the blades. Faults with these types of machines generate scratching and rattling sounds due to unbalance or misalignment of the rotor. This perception as timbre correlates generated sound and machine faults from the viewpoint of the noticeable difference in hearing. Roughness, sharpness, and brightness can be used for measuring scratching and rattling sounds, and boominess and depth can be used for measuring masked sounds from the blades rotating.

2) SLIDING MACHINERY

A typical sliding-type machine is a linear slide rail, in which a metal base moves periodically back and forth on a metal rail. Such a machine uses many bearings for smooth reciprocation. Common bearing faults are the lack of lubricant or misalignment, which result in rolling fatigue and may lead to flaking, which is one of the most severe failures of bearings [30]. Flaking is small pieces of the bearings split off from the surface, making the surface rougher and coarser and increasing friction.

A linear sliding rail emits hissing and clicking sounds in normal operation and will emit squealing and grinding sounds when there is a malfunction due to an increase in friction. To correlate the impression of sound with timbre, sharpness and brightness can be used for measuring normal operating sounds and roughness can be used for measuring squealing and grinding sounds. Depth can help measure anti-brightness impressions in timbre.

3) STRIKING MACHINERY

Striking machinery, such as casting presses or solenoid valves, strikes metal elements to other objects to mold original materials into a shape or control the flow of liquid or gas [31]. Solenoid valves open and close the main valve orifice, which is the only flow path in the valve to control flow. Therefore, it generates clicking sounds regularly, which correspond to opening and closing operations.

Due to wear, tear, or damage to the valve orifice, misalignment could lead to degradation of hammering. In an anomalous situation, clicking sounds become dull and obscured, and beating sounds may occur. This sound change can be detected through timbre in terms of a decrease in brightness and sharpness, and an increase in roughness from the beating. Boominess and depth can help measure the dull impression of clicking sounds.

TABLE 1. Relation among machine fault, noticeable difference in hearing, and associated timbre.

Machine type	Machine faults (Malfunction)	Noticeable difference in hearing	Associated timbre
Rotating machinery	Unbalance, Misalignment, Friction increase	Scratching, Rattling, Noisy hurtling	Boominess, Brightness, Depth, Roughness, Sharpness
Sliding machinery	Misalignment, Flaking, Friction increase	Squealing, Grinding, Noisy rasping	Brightness, Depth, Roughness, Sharpness
Striking machinery	Misalignment, Degradation of hammering	Beat noise, Obscure click	Boominess, Brightness, Depth, Roughness, Sharpness
Liquid manipulator	Clogged, Unstable flow	Gurgle, Disappearance of splashing	Boominess, Brightness, Roughness, Sharpness

4) LIQUID MANIPULATOR

The most typical liquid manipulator is a pump, which is a device that moves liquids. Common faults with this type of machine are caused by wear or cavitation, leading to leakage or clogging [32]. During a malfunction, liquid cannot be supplied constantly; thus, irregular suction and discharge repeatedly occur.

A pump emits similar sounds as those from rotating machinery, but since it is the most commonly used in water discharge, it mainly generates splashing sounds in normal operation and generates gurgling sounds when malfunctioning due to clogging. The sound under normal operation can be discriminated through timbre as booming regarding rotation, brightness, sharpness regarding splashing, and roughness regarding gurgling.

Table 1 summarizes the relation among machine faults, noticeable difference in hearing with the faults, and associated timbre.

C. FEATURE EXTRACTION

1) TIMBRAL METRICS

Five TMs (boominess, brightness, depth, roughness, and sharpness) are calculated as an acoustical feature using the timbral models developed by the University of Surry, as described in Section II.

Since TMs express the degree to index the overall impression of hearing, each calculation algorithm is designed to

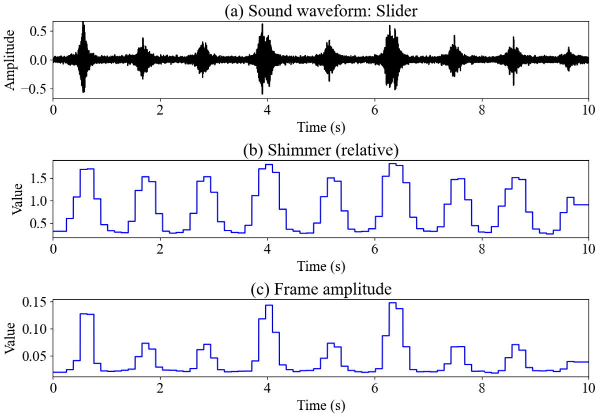


FIGURE 2. (a) Waveform of a sound emitting from slider machine under normal operation, (b) Shimmer value calculated from the sound, and (c) Frame amplitude calculated from the sound.

derive long-term-frame-wise characteristics. In this study, all TMs were calculated in a 1,024 ms frame, the duration of which is determined in the range from 700 ms and 1.5 s. It has been reported that a minimum of 700 ms is required to determine the time order of sounds [33] and unitary perception of duration occurs up to a maximum of approximately 1.5 s [34].

To tailor TMs to meet the perception of anomalous sound related to each machine's faults, we assigned four or five TMs to each machine type based on our investigation result shown in Table 1.

2) SHORT-TERM FEATURE: AMPLIFIED SHIMMER

As described above, TMs can be used to detect the "macro" psycho-acoustical characteristics of machine sounds. It is useful to introduce techniques to derive "micro" psycho-acoustical characteristics associated with timbre. Thus, we developed two short-term features related to TMs.

The first feature is amplified shimmer (AS), which measures the fluctuation in sound waveforms. This is a modification of the acoustical feature "shimmer", which is used to derive the differences in the amplitude of adjacent samples by using the following equation [35].

$$\text{Shimmer (relative)} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}|A_i|} \tag{12}$$

where A_i is the i th amplitude of the input signal of a machine sound, and N is the number of samples.

Shimmer has been reported to strongly relate to roughness on the GRBAS scale. The GRBAS scale is used to rate the overall severity of hoarseness psycho-acoustically and is accepted as standard by the Japanese Society of Logopedics and Phoniatrics [36]. GRBAS comprises five rating elements, Grade of dysphonia, roughness, breathiness, asthenia, and strain, and has been reported as being the most reliable and relevant perceptual-voice-quality rating. Prior research investigating the relation between the GRBAS scale and other acoustical measures reported a high correlation between roughness and shimmer [37].

Fig. 2 shows a waveform of sound emitting from a sliding machine (linear slider), shimmer calculated using (12) at a 256-ms-long frame, and mean amplitude of the waveform at the same frame length. As shown in the figure, the calculated shimmer increases and decreases, conforming with the waveform. This behavior is synchronized with the movement of the sliding platform and fits the perception of roughness. In consideration that the strength of perception must depend on the amplitude of the waveform, we newly define AS as a complementary feature for timbre, in which the differences in the amplitude of adjacent samples are amplified with the amplitude.

We use AS to estimate the time-variant nature associated with the roughness perception of a sound emitted from a machine. AS is calculated as

$$\mathrm{AS} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i-1}|}{\frac{1}{N}\sum_{i=1}^{N}|A_i|} \times \frac{1}{N}\sum_{i=1}^{N}|A_i| \tag{13}$$

$$= \frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i-1}| \tag{14}$$

where A_i is the ith amplitude of the input signal of a machine sound, and N is the number of samples in a 256 ms frame.

3) SHORT-TERM FEATURE: AMPLIFIED PREDOMINANT FREQUENCY

Rotating machinery, such as a fan or motor, tends to emit sounds characterized by timbral pitch, in other words, tone height, which corresponds to its rotating speed. Since an emitting sound caused by machine fault also depends on the rotating speed i.e., pitch, a change in pitch can be an important key for inspectors to diagnose the malfunction using their noticeable difference in hearing. It was reported that the ambiguity of the pitch of a complex signal can be considered as virtual pitch [38]. The perception of pitch is strongly related to a series of sub-harmonics [39].

Under normal fan operation, where components are moving smoothly, emitting sounds tend to have arranged harmonics in their spectra and stable vibration of pitch can be heard, which corresponds to the rotating speed. In an anomalous situation, the sound becomes unstable in pitch or becomes noisy due to eccentricity or friction. The change in perception might be represented as wakening the degree of harmonicity, and the basic frequency may be shifted lower due to the reduction in rotating speed. Therefore, to discover the harmonic change from a machine's sound, we also introduce a feature based on sub-harmonic summation [40].

The sub-harmonic summation is conducted as follows:

1) Spectrum calculation
A 2,048-point FFT is applied to the input machine's sound, then the amplitude spectrum is calculated. Consecutive spectra are obtained for the same frame length of 2,048 samples (128 ms at 16 kHz sampling) with shifted 1,024 samples (64 ms at 16 kHz sampling).

TABLE 2. Number of samples of MIMII dataset.

Number of samples	Fan	Pump	Slider	Valve	Total
Normal	4,075	3,749	3,204	3,691	14,719
Abnormal	1,475	456	890	479	3,300
Total	5,550	4,205	4,094	4,170	18,019

2) Derive a spectral peak from the spectrum
A spectral peak is derived as the highest amplitude in the $1,024$-point spectrum at lower frequencies below $2,000$ Hz. The frequency having a peak is treated as predominant frequency, f_{pdf}. Thus, an amplitude in f_{pdf} is expressed as $A_{mp}(f_{\text{pdf}})$.

3) Sub-harmonics summation
Sub-harmonic summation is conducted by adding the amplitude corresponding to n times the frequency of f_{pdf}. Before summing up, each amplitude is multiplied by the factor $h(n)$ to decrease the contribution along with a distance from f_{pdf}. The summation value "HS(f_{pdf})" is calculated as

$$\text{HS}(f_{\text{pdf}}) = \sum_{n=1}^{K} h(n)\, A_{mp}(nf_{\text{pdf}}), \tag{15}$$

where $A_{mp}(f_{\text{pdf}})$ is the spectral amplitude at f_{pdf} Hz, K is the order of summation, which is 5, and $h(n)$ was set to 0.84^{n-1} in this study.

The precise virtual-pitch magnitude is specified by sub-harmonic pitch magnitude within the crucial interval [39]. We consider that intensity of the virtual-pitch perception deeply depends on its magnitude i.e., amplitude. Therefore, to measure the predominant pitch frequency with the intensity of the virtual perception, we newly define the harmonic-related feature "amplified predominant frequency (APF)", which is calculated as

$$\text{APF} = f_{\text{pdf}} \times \text{HS}(f_{\text{pdf}}) \tag{16}$$

D. COMBINATION OF TIMBRAL METRICS AND SHORT-TERM FEATURES

Five or four types of TMs are selected to fit the association of timbre with assumed industrial machine fault in each machine type. Assignment between related timbre and machine type is determined from the summary in Table 1. If a target machine type does not fall into the four types of machinery, all five TMs are selected.

Since our two short-term features are designed to derive signal fluctuation and spectral harmonics as common features, we add AS and APF to a set of TMs for all machinery. Finally, a combination of TMs and short-term features is configured and used for classification.

E. CLASSIFICATION OF ANOMALOUS SOUND

We used a support vector machine (SVM) as the binary classification technique because SVM classifiers perform well in high-dimensional space and have excellent accuracy. To apply the SVM for binary classification, the extracted combination of TMs and short-term features (combi-features) with normal / abnormal flags are used to train a model. In the training process, the combi-features from abnormal sounds are assigned "true", and those from normal sounds are "negative". The SVM then obtains the separation plane to classify two categories, true and negative, to maximize the gap between the two categories.

The training process generates a classification model, then the classification of unknown data of each machine is conducted based on the model dedicated to each machine type.

IV. EVALUATION

A. SOUND DATA SETUP

We used the MIMII dataset for our experiment [41]. This dataset contains four machines, i.e., fan, pump, slider (slide rail), and valve, which correspond to the machine types listed in Table 1. The fan represents an industrial fan, which provides a continuous flow in normal operation, and in abnormal situations, unbalanced or clogging occurs. The pump is a water pump that discharges water to a pool continuously, and in abnormal situations, leakage or clogging occurs. The slide rail, or slider, is a linear slider that consists of a moving platform and a stage base, and in abnormal situations, rail damage or no grease occurs. The valve is a solenoid valve that is repeatedly opened and closed, and in abnormal situations, various contaminations occur. Normal and abnormal sounds were recorded as 10 s-long sound files at 16 kHz sampling in a reverberant environment. Background noise which was recorded in multiple real factories was mixed with the recorded machine sound files. The total number of sound files is $14,719$ for normal conditions and $3,300$ for abnormal conditions (see Table 2). Four models with ID 00, 02, 04, and 06 were used for all machine types where a signal-to-noise ratio of the background noise was 6 dB.

B. CLASSIFICATION EVALUATION

We used accuracy, F-measure, and Matthews Correlation Coefficient (MCC) to evaluate ASD performance in terms of SVM classification performance.

Accuracy measures the proportion of correctly classified data instances over the total number of data instances. The calculation formula is

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{17}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

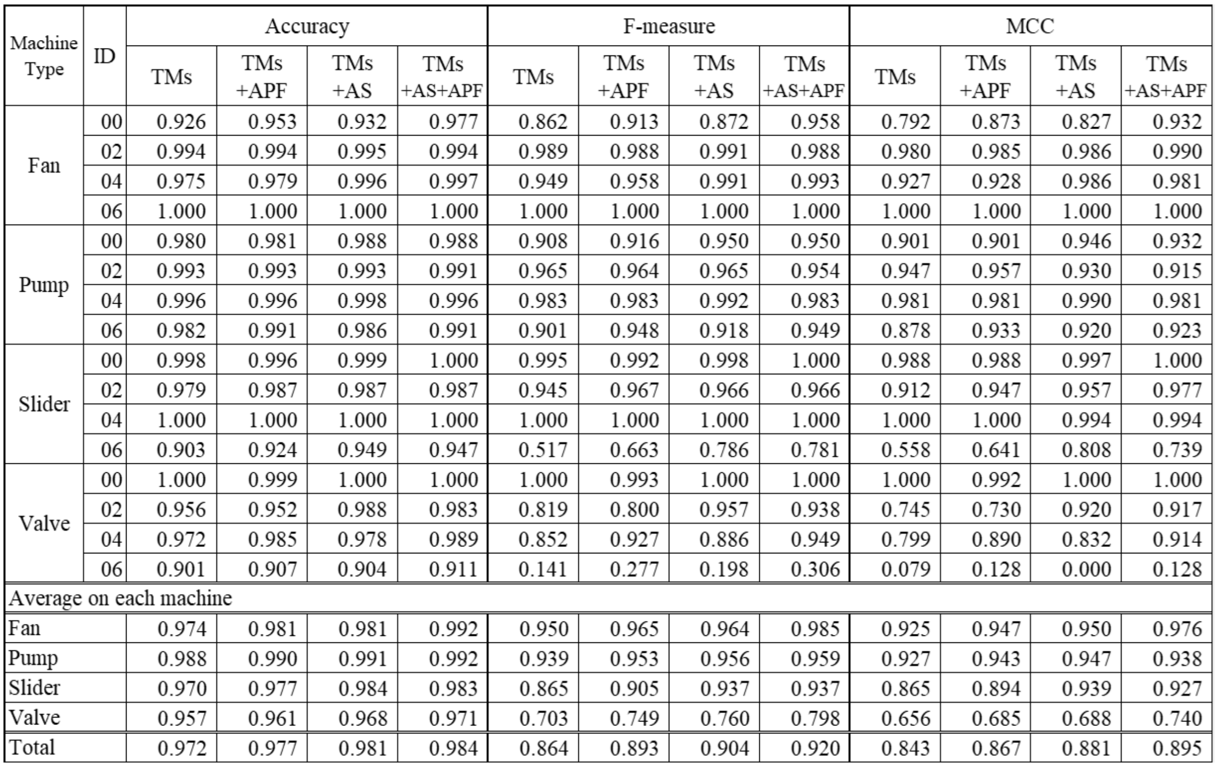
F-measure is an error metric for measuring model performance by calculating the harmonic mean of precision and recall of the model.

The formula for F-measure is as follows:

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

Precision is used to measure the proportion of positive class predictions that belong to the positive class. Thus, precision

TABLE 3. Classification performance evaluation on each machine type.



Machine Type	ID	Accuracy				F-measure				MCC			
		TMs	TMs +APF	TMs +AS	TMs +AS+APF	TMs	TMs +APF	TMs +AS	TMs +AS+APF	TMs	TMs +APF	TMs +AS	TMs +AS+APF
Fan	00	0.926	0.953	0.932	0.977	0.862	0.913	0.872	0.958	0.792	0.873	0.827	0.932
	02	0.994	0.994	0.995	0.994	0.989	0.988	0.991	0.988	0.980	0.985	0.986	0.990
	04	0.975	0.979	0.996	0.997	0.949	0.958	0.991	0.993	0.927	0.928	0.986	0.981
	06	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Pump	00	0.980	0.981	0.988	0.988	0.908	0.916	0.950	0.950	0.901	0.901	0.946	0.932
	02	0.993	0.993	0.993	0.991	0.965	0.964	0.965	0.954	0.947	0.957	0.930	0.915
	04	0.996	0.996	0.998	0.996	0.983	0.983	0.992	0.983	0.981	0.981	0.990	0.981
	06	0.982	0.991	0.986	0.991	0.901	0.948	0.918	0.949	0.878	0.933	0.920	0.923
Slider	00	0.998	0.996	0.999	1.000	0.995	0.992	0.998	1.000	0.988	0.988	0.997	1.000
	02	0.979	0.987	0.987	0.987	0.945	0.967	0.966	0.966	0.912	0.947	0.957	0.977
	04	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.994
	06	0.903	0.924	0.949	0.947	0.517	0.663	0.786	0.781	0.558	0.641	0.808	0.739
Valve	00	1.000	0.999	1.000	1.000	1.000	0.993	1.000	1.000	1.000	0.992	1.000	1.000
	02	0.956	0.952	0.988	0.983	0.819	0.800	0.957	0.938	0.745	0.730	0.920	0.917
	04	0.972	0.985	0.978	0.989	0.852	0.927	0.886	0.949	0.799	0.890	0.832	0.914
	06	0.901	0.907	0.904	0.911	0.141	0.277	0.198	0.306	0.079	0.128	0.000	0.128
Average on each machine													
Fan		0.974	0.981	0.981	0.992	0.950	0.965	0.964	0.985	0.925	0.947	0.950	0.976
Pump		0.988	0.990	0.991	0.992	0.939	0.953	0.956	0.959	0.927	0.943	0.947	0.938
Slider		0.970	0.977	0.984	0.983	0.865	0.905	0.937	0.937	0.865	0.894	0.939	0.927
Valve		0.957	0.961	0.968	0.971	0.703	0.749	0.760	0.798	0.656	0.685	0.688	0.740
Total		0.972	0.977	0.981	0.984	0.864	0.893	0.904	0.920	0.843	0.867	0.881	0.895

is calculated as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{19}$$

Recall is used to measure the proportion of positive class predictions out of all positive instances and calculated as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

MCC measures a correlation between predicted classes and ground truth, calculated by following the formula.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{21}$$

C. EVALUATION RESULT

Table 3 lists the evaluation results of the SVM classification performance of the proposed method using the MIMII dataset. Four combinations of features were compared to verify how much each addition of the feature contributes to improving classification performance. The first combination is the selected TMs for each machine type ("TMs"). The second is APF added to the TMs ("TMs+APF"). The third is AS added to the TMs ("TMs+AS"). The fourth combination is APF added to the TMs and AS ("TMs+AS+APF"). Since the evaluation was conducted for each machine type and model which is identified as ID 00, 02, 04, and 06, the number of conditions was 16 in total, as listed in Table 3.

The results indicate that:

1) The classification performance by applying only TMs showed rather a high-performance rate, over 0.926 in accuracy, over 0.819 in F-measure, and over 0.792 in MCC except under two conditions. This demonstrates that selected TMs are effective for ASD in classifying anomalous sounds since these metrics were selected to fit "noticeable difference in hearing" from anomalous machine sounds.
2) Adding the short-term features, AS and APF, improves the classification performance. The results indicate the proposed method's excellent classification performance with an accuracy of 0.984, an F-measure of 0.920, and an MCC of 0.895 on average for the 16 conditions. The evaluation results of the F-measure and MCC indicated nearly identical trends. The results of the accuracy ranging from 0.971 to 0.992 in each machine type are comparable to those of the recent conventional method [14] with which the classification accuracies range from 97.0 to 99.7 %. The conventional method was designed for diagnosing bearing faults using sound-quality metrics. Our proposed method has superiority in capability of a wide range of machinery including rotating and sliding machinery that use bearings.
3) The contribution of adding short-term features seems to be effective for all machine types. AS seems to improve the classification for slider and valve rather than fan and pump and APF improves the classification for fan and pump rather than slider and valve.

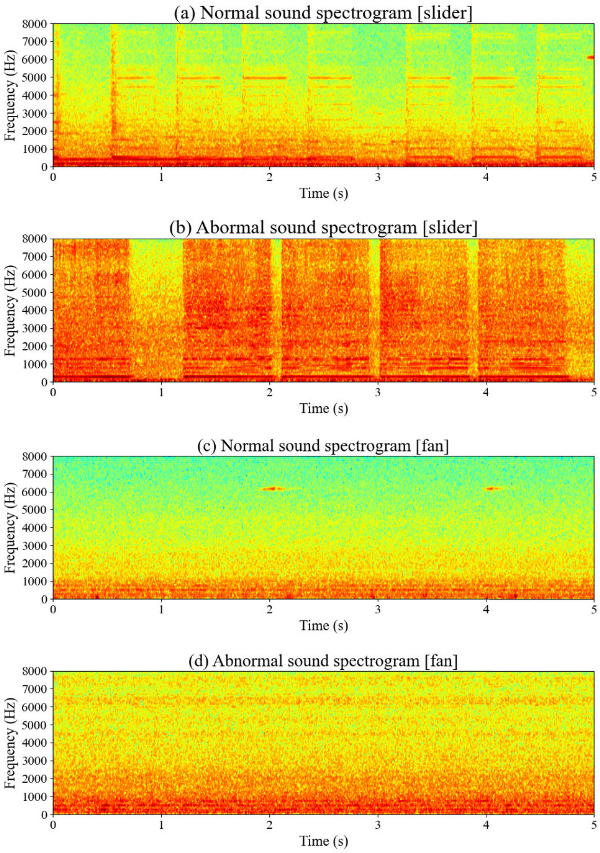


FIGURE 3. Spectrograms: (a) Normal sound of a slider, (b) Abnormal sound of a slider, (c) Normal sound of a fan, and (d) Abnormal sound of a fan.

4) Only two models, valve ID 06 and slider ID 06, did not improve enough regarding F-measure. Since the performances of the other 14 models had scores higher than 0.938 in F-measure, it suggests that different TMs should be added, especially for valve ID 06 and slider ID 06. Since only the F-measure was low, e.g., 0.306, for valve ID 06 under TM+AS+APF and accuracy remained rather high, e.g., 0.911, under the same condition, a decrease in the number of FNs for SVM training seems to be necessary. From these results, we can analyze the false reason and consider a solution by investigating the correlation of each timbre (see Section V-D). This can demonstrate the effectiveness of the proposed method in detecting various faults in terms of timbre.

V. DISCUSSION

A. MACHINE SOUND CHANGE AND THE NOTICEABLE DIFFERENCE IN HEARING

We verify the relation between machine-sound change and the noticeable difference in hearing on the MIMII dataset. We exemplify a slider and fan as representative machines and investigate to illustrate the change point between their ‘‘normal’’ and ‘‘abnormal’’ operating sounds.

For the slider, a slight scratching sound is emitted under normal operation due to the metal base moving back and forth on the metal rail periodically. The sound is emphasized or increased in its magnitude when the base stops at one edge of the rail and moves away to the opposite edge. This sound is expressed as a bar that separates the periodical sound area in its spectrogram, as shown in Fig. 3 (a). During a malfunction, a strong creaking or squealing sound is emitted due to the increase in friction between the base and rail. This sound is expressed as a uniform component or flat frequency spectrum in red in the spectrogram, as shown in Fig. 3 (b), where red means higher magnitude or amplitude than green or yellow. This change can be heard as increasing roughness, sharpness, and brightness and decreasing depth from the timbral perspective.

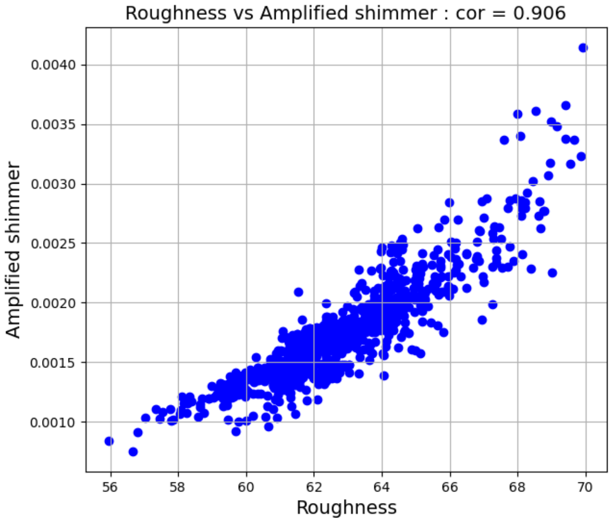


FIGURE 4. Correlation between roughness and amplified shimmer.

For the fan under normal operation, a slight booming sound is caused by the blades rotating in the air and pushing them forward as wind is emitted. The sound forms multiple horizontal bars or spectral harmonics corresponding to the period of rotating in the lower frequency area below 1,000 Hz in its spectrogram, as shown in Fig. 3 (c). There is also a slight scratching sound caused by the rotation of the blade shaft. During a malfunction, a strong squealing sound is emitted due to eccentricity or an increase in friction. Therefore, high-amplitude components appear in the upper-frequency area in the spectrogram, as shown in Fig. 3 (d). This change can be heard as an increase in roughness and brightness and a relative decrease in booming and depth.

B. CORRELATION BETWEEN TIMBRE AND SHORT-TERM FEATURES

We developed two types of short-term features to support TMs for improving the time resolution of the timbral analysis. Therefore, we verified how much the two features correlate with the target timbre on the MIMII dataset.

We first examined the relation between AS and roughness for normal operation sounds of a slider. As shown in

TABLE 4. Classification performance evaluation of toy car.

Machine Type	case	Accuracy				F-measure				MCC			
		TMs	TMs +APF	TMs +AS	TMs +AS+AP	TMs	TMs +APF	TMs +AS	TMs +AS+AP	TMs	TMs +APF	TMs +AS	TMs +AS+AP
Toy Car	2	0.982	0.984	0.995	0.995	0.941	0.949	0.985	0.984	0.928	0.934	0.982	0.981

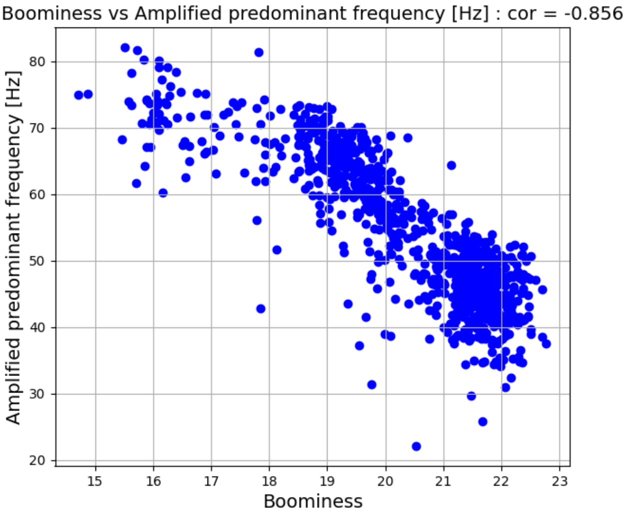


FIGURE 5. Correlation between boominess and amplified predominant frequency.

Fig. 4, AS had a strong correlation with roughness, which has a Pearson's r of 0.906. AS increases as roughness also increases. This relation matches the intention of introducing AS as a supplemental feature of roughness, as described in Section III-C2.

We also investigated the relation between APF and boominess for normal operation sounds of a fan. As shown in Fig. 5, APF had a strong correlation with boominess, which had an r of -0.856. When the predominant frequency shifted to a higher region, perception of boominess tended to weaken because boominess measures the concentration of spectral power into lower frequency region below 280 Hz (see Section II-C). Thus, this is why r is negative and APF can be used to discover anomalous sounds with rotation-frequency change.

C. EVALUATION USING A DATASET OF MINIATURE-MACHINE SOUNDS

We also evaluated the other set of machine-like operating sound data to verify the applicability of the proposed TF-ASD method. Although the MIMII dataset we used for evaluation in Section IV contained the real industrial machine sound, we used "ToyADMOS"; a dataset of miniature-machine operating sounds because we could not find any other adequate dataset except the MIMII dataset. ToyADMOS contained both normal and anomalous operation sounds of 3 kinds of miniature machines, those are toys, named Toy car, Toy conveyor, and Toy train [42].

We selected operating sound data of the Toy car whose anomalous sounds were generated by deformed gears and bent shaft as deliberate damage. Since the Toy car dataset had 4 combination types, we selected case 2 which equipped a steel bearing with a torque-tuned motor. We consider that similar damage and change of the sound in anomalous operation would occur in rotating machinery with rotor unbalance and sliding machinery with increasing friction and then assigned five TMs and two short-term features for evaluation.

Normal and abnormal sounds were recorded as 11 s-long sound files at 48 kHz sampling. To conform to the same frequency range as the MIMII dataset evaluation, all files are evaluated after being down-sampled to 16 kHz. Due to selecting the condition in which steel bearings are used, the total number of sound files is 2, 650 for normal conditions and 528 for abnormal conditions respectively.

The evaluation results are shown in Table 4. The results demonstrate that the proposed TF-ASD method can provide an excellent classification performance with accuracy of 0.995, F-measure of 0.984, and MCC of 0.981. The abnormal sound of the toy car increases in roughness due to abnormal vibration generated by deformed gear and a bent shaft. This tendency can be confirmed as the improvement of classification performance by adding AS. On the contrary, dependency on pitch tone seems very little compared with roughness. Because APF addition to TMs can improve a little in the classification performance while APF addition to a pair of TMs and AS does not improve at all in the classification performance.

D. EVALUATION USING ANOMALOUS SCORE

Since machines are well maintained regularly, they can keep operating normally most of the time. From the data-analysis perspective, this means that the amount of observation data of anomalous situations is limited compared with normal ones. Considering that the longer a machine is in operation the more likely severe malfunction will occur suddenly and will inflict severe damage to the business. Therefore, it is important to detect incipient symptoms from very few indications.

To respond to this on-site issue, an ASD method that can classify unknown anomalous sounds using prior knowledge of normal sound distribution is required. To assess the applicability of the proposed method, we verified how much each TM and short-term feature has the capability for anomalous-sound discrimination in their feature domain in terms of a probability distribution.

TABLE 5. AUC evaluation on anomalous score distribution.

Machine	ID	Boominess	Brightness	Depth	Roughness	Sharpness	AS	APF
Fan	00	0.683	0.656	0.795	0.687	0.551	0.557	0.727
	02	0.595	0.965	0.652	0.957	0.832	0.874	0.687
	04	0.791	0.802	0.763	0.842	0.637	0.699	0.620
	06	0.683	0.890	0.718	0.780	0.810	0.960	0.925
Pump	00	0.831	0.816	0.701	0.807	0.788	0.963	0.712
	02	0.602	0.831	0.748	0.724	0.856	0.625	0.682
	04	0.619	0.634	0.601	0.667	0.861	0.680	0.638
	06	0.893	0.715	0.880	0.706	0.618	0.690	0.832
Slider	00	0.980	0.985	0.968	0.925	0.993	0.997	0.675
	02	0.675	0.781	0.839	0.908	0.788	0.745	0.738
	04	0.722	0.867	0.755	0.948	0.864	0.792	0.664
	06	0.670	0.662	0.666	0.756	0.621	0.650	0.627
Valve	00	0.908	0.498	0.809	0.871	0.900	0.990	0.617
	02	0.669	0.647	0.662	0.687	0.623	0.704	0.650
	04	0.680	0.632	0.631	0.748	0.631	0.807	0.873
	06	0.653	0.731	0.641	0.676	0.726	0.674	0.607
Average on each machine								
Fan		0.688	0.828	0.732	0.817	0.707	0.773	0.740
Pump		0.736	0.749	0.733	0.726	0.781	0.739	0.716
Slider		0.762	0.824	0.807	0.884	0.817	0.796	0.676
Valve		0.727	0.627	0.686	0.745	0.720	0.794	0.687
Average in total		0.728	0.757	0.739	0.793	0.756	0.775	0.705

TABLE 6. Classification evaluation regarding the combination of timbral metrics for valve machine.

Combination	Machine	ID	Boominess	Brightness	Depth	Roughness	Sharpness	Accuracy	F-measure
#1	Valve	00	✓	✓	✓	✓	✓	1.000	1.000
		02						0.956	0.819
		04						0.972	0.852
		06						0.901	0.141
	Average in all IDs							0.957	0.703
#2	Valve	00	✓	✓	✓		✓	0.987	0.986
		02						0.924	0.737
		04						0.966	0.839
		06						0.900	0.200
	Average in all IDs							0.944	0.690
#3	Valve	06		✓	✓	✓	✓	0.897	0.093
	Average in all IDs							0.953	0.672
#4	Valve	06	✓		✓	✓	✓	0.899	0.115
	Average in all IDs							0.956	0.691
#5	Valve	06	✓	✓		✓	✓	0.892	0.000
	Average in all IDs							0.946	0.611
#6	Valve	06	✓	✓	✓	✓		0.894	0.027
	Average in all IDs							0.949	0.645

We use an anomalous score A_{score} that can measure the rareness of occurrence from the probability distribution of normal sounds [43]. Let $p(x|\text{Normal})$ be the random variable inferred from the distribution of normal sounds at value x, where "Normal" means that the random variable is obtained from the timbral metric and short-term feature vectors extracted from "normal" sounds. The A_{score} is then defined at x', which is the input value, as

$$A_{\text{score}}(x') = -\ln p(x'|\text{Normal}) \tag{22}$$

The rarer the occurrence of x', the higher the degree of anomaly, so the higher the value of the anomalous score is calculated.

In this study, we assume the probability distribution of normal sounds as a normal distribution, $N(x|\mu, \sigma^2)$.

$$N(x|\mu, \sigma^2) \triangleq \frac{1}{(2\pi\sigma^2)^{1/2}}\, exp\{-\frac{1}{2\sigma^2}\,(x-\mu)^2\} \tag{23}$$

The μ and σ^2 can be obtained by maximum likelihood estimation described below,

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)} \tag{24}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)} - \hat{\mu})^2 \tag{25}$$

Under the assumption of the probability distribution of normal sounds as a normal distribution, the anomaly score for x' is obtained by the following formula.

$$A_{\text{score}}(x') \triangleq \frac{1}{\hat{\sigma}^2}(x' - \hat{\mu})^2 = (\frac{x' - \hat{\mu}}{\hat{\sigma}})^2 \tag{26}$$

We verify the contribution of each TM and short-term feature for anomalous classification by using A_{score}.

For the validation with A_{score}, the entire normal sound is divided into a training set and a testing set, and the probability density function (PDF) is obtained from a combination of timbral metrics and short-term features extracted from the training "normal" sound set on each machine model (ID) of each machine type. Then both "normal" and "abnormal" A_{score}s are then calculated using the "normal" testing set and "abnormal" testing set based on the PDF using (26), respectively.

It is ideal that both "normal" and "abnormal" A_{score}s are sufficiently separated from each other in the A_{score} domain. To measure the discriminability between the two distributions we used the area under the receiver operating characteristics curve (AUC). If the two distributions are separated enough, the AUC is evaluated as closer to 1.0. Therefore, we utilized the AUC as a measure that represents the contribution of classification on each TM and short-term feature.

Table 5 lists the AUC evaluation results for each TM and short-term feature.

The results indicate that:

1) The AUC values of TMs were rather high from 0.728 to 0.793 on average. The dedicated timbre assignment for each machine, which is shown in the yellow box, seems to be suitable because the AUC values were also rather high from 0.627 to 0.884 on average for each machine.
2) The contribution of AS seems to be relatively high on average for all machines, as shown by the AUC values from 0.739 to 0.796. These results indicate that the AS contribution for the slider and valve seemed to be slightly higher than that for the fan and pump.
3) The AUC values of APF, from 0.676 to 0.740, are slightly lower than those of the TMs and AS. However, by comparing the results among the machine types, the AUC average value of APF for the fan was higher than that for the pump, slider, and valve. Moreover, the AUC values of APF seem to be high for the models in which those of boominess were insufficient. This indicates that APF seems to work complementarily for the timbre of boominess.

We discuss the valve ID 06 model, which has poor classification performance, and analysis the relation between the performance and the combination of TMs. Table 6 lists five TM combinations and the classification performances both in accuracy and F-measure. In this table, #1 is identical to our proposed method in which all five TMs are selected for classification. As mentioned above, all models, except the ID 06 ones, had good classification performance in terms of F-measure of over 0.852.

From the AUC evaluation on A_{score} distribution in Table 5, the AUC value of roughness in the valve ID 06 (0.676) was relatively low compared with the average value (0.745) and with those of the other three models, e.g., ID 00, ID 02, and ID 04. Therefore, we evaluated another combination of TMs, signified as #2 in Table 6, in which roughness is excluded. A slight improvement in F-measure, about 0.06, was observed under the #2 combination. From this result, we can infer that roughness does not contribute much to classification and that the exclusion of roughness might slightly improve the F-measure. By focusing on the observation that the AUC values of brightness (0.731) and sharpness (0.726) are relatively high, and those of boominess (0.653) and depth (0.641) are relatively low, as shown in Table 5, the timbre that relates to the high-frequency domain might be superior against the lower-frequency domain. However, this conjecture does not mean that boominess and depth cannot contribute to classification because the exclusion of these TMs will result in a drastic decrease in F-measure, as shown in combinations #3 and #5 in Table 6. We assume that other metrics such as "Hardness" can be used for classification, which can evaluate the change in the higher-frequency domain while estimating the lower-frequency domain.

This exploratory study showed that this combination of TMs is adequate for classifying valve machine sounds, as shown in Table 6. We will explore other machine types and model performances by considering the relation of A_{score} distribution and the machine faults from the timbre perspectives as future work.

This study also indicates that the evaluation based on the AUC on the A_{score} is useful to investigate what kind of timbre can contribute the ASD. By feeding back to the association between timbre and the noticeable difference in hearing, the proposed method has the potential to express in measurable form implicit knowledge for inspection.

VI. CONCLUSION

We proposed a timbral-feature-based anomalous sound detection (TF-ASD) method for industrial machines, which used a combination of five timbral metrics and newly developed short-term features, in which the combination is designed to follow human ''noticeable difference in hearing'' from the machine's operating sound. We uniquely investigated the association between machine faults and the noticeable difference in hearing with onomatopoeia as a mediator and then selected five timbral metrics for four types of typical machinery. We originally developed two types of short-term features to improve the time resolution of the timbral analysis, which estimate the fluctuation in sound waves to measure roughness and change in pitch in terms of harmonics to measure tone height such as boominess or sharpness. We incorporated a support vector machine as a binary classifier to perform well in high-dimensional space where the combination of TMs and our two short-term features can determine whether an observing sound is normal or anomalous.

We evaluated the proposed method using the MIMII dataset, which contains recorded sounds from four types of industrial machines and four models of each type under both ''normal'' and ''abnormal'' conditions, in terms of SVM classification performance. The results indicated that the proposed method with a tailored combination of timbral metrics for each machine type and complementary short-term features was effective in classification performance with an accuracy of 0.984 on average and the accuracy ranged from 0.971 to 0.992 in each machine type. The result also demonstrated that TF-ASD commonly provided an excellent classification for four types of machine type, while the classification accuracies were comparable to that of the recent ASD method dedicated to bearing faults detection, in which the classification accuracies ranged from 97.0 to 99.7 %.

We further verified the contribution of the combination of timbral metrics and short-term features to ASD by introducing an anomalous score. We employed AUC evaluation to measure the discriminability of each metric and feature on the anomalous score. The results indicated that the contribution of the TMs and AS for ASD was relatively high on average for almost all machine types. The results also indicated that the contribution of APF was high for fan, and relatively low for pump, slider, and valve. By introducing the anomalous score and the AUC evaluation, the contribution degree could be represented in numerical form. We also conducted an exploratory study for valve machines to adjust the SVM performance by modifying the combination of TMs with reference to the AUC value.

Through our unique analysis of causes of malfunction in industrial machines and the emitting sounds, the relation between the noticeable difference in hearing and timbre was newly derived. By developing a combination of timbral metrics and short-term timbre-related features based on the analysis, we demonstrated the effectiveness of the proposed TF-ASD with excellent classification performance under a real industrial machine's sound dataset.

The study in this paper suggests that timbre can be a powerful interpreter which can link anomalous detection with a human's noticeable difference in hearing and that a combination of five timbral metrics and our short-term features can be the primal candidates of the keys for ASD from the perspective of timbre. Through statistical analysis of the timbre-related features, various practical applications can be expected, such as an estimation of machines' faults or predictive anomalous detection from their emitting sounds. Although extensive analysis is still required, the analysis approach based on timbre has great potential to express in some measurable form the implicit knowledge of inspectors.

There are three aspects that must be addressed for future extension: 1) enhancement of the TF-ASD approach by pursuing additional features that fit human perceptual change; 2) expansion of the approach to anomalous change detection with an analysis in time series for practical use; 3) further study of the relation between the noticeable difference in hearing and timbre for sharing the implicit knowledge of inspectors.

REFERENCES

[1] *High Pressure Gas Safety: Smart Industrial Safety Action Plan*, Public-Private Sector Council on Smart Industrial Safety High Pressure Gas Safety Committee, Ministry of Economy, Trade and Industry Japan, Tokyo, Japan. 2020. Accessed on: Mar. 16, 2023. [Online]. Available: https://www.meti.go.jp/english/press/2020/pdf/0710_007a.pdf

[2] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, ''Description and discussion on DCASE2020 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring,'' in *Proc. Detection Classification Acoustic Scenes Events Workshop*, Nov. 2020, pp. 1–5.

[3] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, ''Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions,'' 2021, *arXiv.2106.04492*.

[4] S. Zhao, ''Acoustic anomaly detection based on similarity analysis,'' in *Proc. Detection Classification Acoustic Scenes Events Challenge*, 2020, pp. 1–3. Accessed: Mar. 16, 2023. [Online]. Available: https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Zhao_2_t2.pdf

[5] Y. Kawaguchi and T. Endo, ''How can we detect anomalies from subsampled audio signals?'' in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[6] Y. Kawachi, Y. Koizumi, and N. Harada, ''Complementary set variational autoencoder for supervised anomaly detection,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2366–2370.

[7] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, ''Unsupervised detection of anomalous sound based on deep learning and the Neyman–Pearson lemma,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 1, pp. 212–224, Jan. 2019.

[8] K. Li, Q. H. Nguyen, Y. Ota, and M. Unoki, ''Unsupervised anomalous sound detection for machine condition monitoring using temporal modulation features on gammatone auditory filterbank,'' in *Proc. Detection Classification Acoustic Scenes Events Challenge*, Nov. 2022, pp. 1–5.

[9] M. Hu, Z. Ji, K. Yan, Y. Guo, X. Feng, J. Gong, X. Zhao, and L. Dong, ''Detecting anomalies in time series data via a meta-feature based approach,'' *IEEE Access*, vol. 6, pp. 27760–27776, 2018.

[10] Z. Ji, Y. Wang, K. Yan, X. Xie, Y. Xiang, and J. Huang, ''A space-embedding strategy for anomaly detection in multivariate time series,'' *Expert Syst. Appl.*, vol. 206, Nov. 2022, Art. no. 117892.

[11] D. Goyal, A. Choudhary, B. S. Pabla, and S. S. Dhami, ''Support vector machines based non-contact fault diagnosis system for bearings,'' *J. Intell. Manuf.*, vol. 31, no. 5, pp. 1275–1289, Jun. 2020.

[12] R. K. Mishra, A. Choudhary, A. R. Mohanty, and S. Fatima, ''Multi-domain bearing fault diagnosis using support vector machine,'' in *Proc. IEEE 4th Int. Conf. Comput., Power Commun. Technol. (GUCON)*, Sep. 2021, pp. 1–6.

[13] K. Minemura, T. Ogawa, and T. Kobayashi, ''Acoustic feature representation based on timbre for fault detection of rotary machines,'' in *Proc. Int. Conf. Sensing, Diagnostics, Prognostics, Control (SDPC)*, Aug. 2018, pp. 302–305.

[14] T. Mian, A. Choudhary, and S. Fatima, ''An efficient diagnosis approach for bearing faults using sound quality metrics,'' *Appl. Acoust.*, vol. 195, Jun. 2022, Art. no. 108839.

[15] M. Takada, K. Tanaka, and S.-I. Iwamiya, ''Relationships between auditory impressions and onomatopoeic features for environmental sounds,'' *Acoust. Sci. Technol.*, vol. 27, no. 2, pp. 67–79, 2006.

[16] Y. Ota, S. Kura, and M. Unoki, ''Anomalous sound detection using objective metrics related to timbral attributes,'' in *Proc. 24th Int. Congr. Acoustics (ICA)*, Koria, Oct. 2022, pp. 1–7.

[17] C. J. Plack, *The Sense of Hearing*. Evanston, IL, USA: Routledge, Jun. 2018.

[18] H. Fastl and E. Zwicker, *Psycho-Acoustics: Facts and Models*, 3rd ed. Berlin, Germany: Springer, 2007.

[19] K. Jensen, ''The timbre model,'' *J. Acoust. Soc. Amer.*, vol. 112, no. 5, p. 2238, Nov. 2002.

[20] A. Pearce, T. Brookes, and R. Mason, ''Timbral attributes for sound effect library searching,'' in *Proc. AES Int. Conf. Semantic Audio*, Jun. 2017, pp. 1–8. Accessed: Mar. 16, 2023. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=18754

[21] B. C. J. Moore, *Introduction to the Psychology of Hearing*. Boston, MA, USA: Brill, 2013.

[22] P. Daniel and R. Weber, ''Psychoacoustical roughness: Implementation of an optimized model,'' *Acta Acustica United Acustica*, vol. 83, no. 1, pp. 113–123, Jan./Feb. 1997.

[23] P. N. Vassilakis, ''SRA: A web-based research tool for spectral and roughness analysis of sound signals,'' in *Proc. 4th Sound Music Comput. Conf. (SMC)*, Jan. 2007, pp. 319–325.

[24] A. Pearce, T. Brookes, and R. Mason, *Deliverable D5.2: First Prototype of Timbral Characterisation Tools for Semantically Annotating Non-Musical Content*, Audio Commons, document D5.2, Apr. 2017.

[25] S. Hatano and T. Hashimoto, ''Booming index as a measure for evaluating booming sensation,'' in *Proc. 29th Int. Congr. Exhib. Noise Eng.*, Aug. 2000, pp. 1–5.

[26] A. Pearce, T. Brookes, and R. Mason, *Deliverable D5.8: Release of Timbral Characterisation Tools for Semantically Annotating Non-Musical Content*, Audio Commons, document D5.8, Jan. 2019.

[27] E. Schubert and J. Wolfe, ''Does timbral brightness scale with frequency and spectral centroid,'' in *Acta Acustica United Acustica*, vol. 92, pp. 820–825, Jun. 2006.

[28] M. Sakamoto, ''System to quantify the impression of sounds expressed by onomatopoeias,'' *Acoust. Sci. Technol.*, vol. 41, no. 1, pp. 229–232, 2020.

[29] A. Muszynska, ''Vibrational diagnostics of rotating machinery malfunctions,'' *Int. J. Rotating Machinery*, vol. 1, nos. 3–4, pp. 266–327, 1995.

[30] (Oct. 2019). *Bearing Failure RCA: Flaking*. Worldwide Bearing Industry News. Accessed: Mar. 16, 2023. [Online]. Available: https://www.bearing-news.com/bearing-failure-rcaflaking/

[31] Bürkert Fluid Control Systems GmbH & Co KG. *What is a Solenoid Valve and How Does It Work?* Accessed: Mar. 16, 2023. [Online]. Available: https://www.burkert-usa.com/en/Company-Career/What-s-New/Press/Media/Technical-Reports/Technical-Reports-additional-topics/What-is-a-solenoid-valve-and-how-does-it-work/

[32] (Feb. 2020). *Four Common Causes of Pump Failure*. Capability Guide, in Pump Industry Magazine. Accessed: Mar. 16, 2023. [Online]. Available: https://www.pumpindustry.com.au/four-common-causes-of-pump-failure/

[33] R. M. Warren, C. J. Obusek, R. M. Farmer, and R. P. Warren, ''Auditory sequence: Confusion of patterns other than speech or music,'' *Science*, vol. 164, no. 3879, pp. 586–587, May 1969.

[34] P. Fraisse, ''Perception and estimation of time,'' *Annu. Rev. Psychol.*, vol. 35, no. 1, pp. 1–37, Jan. 1984.

[35] Aalto University. *Introduction to Speech Processing: Jitter and Shimmer 2nd Edition*. Accessed: Mar. 16, 2023. [Online]. Available: https://speechprocessingbook.aalto.fi/Representations/Jitter_and_shimmer.html

[36] M. Hirano, S. Hibi, R. Terasawa, and M. Fujiu, ''Relationship between aerodynamic, vibratory, acoustic and psychoacoustic correlates in dysphonia,'' *J. Phonetics*, vol. 14, nos. 3–4, pp. 445–456, Oct. 1986.

[37] S. Vaz Freitas, P. Melo Pestana, V. Almeida, and A. Ferreira, ''Integrating voice evaluation: Correlation between acoustic and audio-perceptual measures,'' *J. Voice*, vol. 29, no. 3, pp. 390.e1–390.e7, May 2015.

[38] E. Terhardt, ''Pitch, consonance, and harmony,'' *J. Acoust. Soc. Amer.*, vol. 55, no. 5, pp. 1061–1069, May 1974.

[39] E. Terhardt, ''Calculating virtual pitch,'' *Hearing Res.*, vol. 1, no. 2, pp. 155–182, Mar. 1979.

[40] D. J. Hermes, ''Measurement of pitch by subharmonic summation,'' *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

[41] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, ''MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection,'' 2019, *arXiv:1909.09347*.

[42] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, ''ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection,'' in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2019, pp. 313–317.

[43] S. Ntalampiras, I. Potamitis, and N. Fakotakis, ''Probabilistic novelty detection for acoustic surveillance under real-world conditions,'' *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 713–719, Aug. 2011.

YASUJI OTA (Member, IEEE) received the B.S. degree in electrical engineering from Niigata University, Japan, in 1987. He is currently pursuing the Ph.D. degree with the Japan Advanced Institute of Science and Technology (JAIST), Japan. His professional carrier started from 1987 as a Signal-Processing Researcher at Fujitsu Laboratories Ltd., and he has been an in charge of research and development engineering since 2010 for emerging industrial businesses such as mobile phones, VoIP, and the IoT with Fujitsu Ltd. He is currently a Researcher with JAIST. His current research interests include acoustic perception-related signal processing and its application development. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, and the Acoustical Society of Japan (ASJ). He received the Young Researcher's Award from IEICE, in 1995.

MASASHI UNOKI (Member, IEEE) received the M.S. and Ph.D. degrees in information science from the Japan Advanced Institute of Science and Technology (JAIST), in 1996 and 1999, respectively. He was a Visiting Researcher with the ATR Human Information Processing Laboratories, from 1999 to 2000, and the Centre for the Neural Basis of Hearing (CNBH), Department of Physiology, University of Cambridge, from 2000 to 2001. He has been with the faculty of the School of Information Science, JAIST, since 2001, where he is currently a Full Professor. His main research interests include auditory motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) Research Fellow, from 1998 to 2001. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). He received the Sato Prize from ASJ, in 1999, 2010, and 2013, for an outstanding paper; and the Yamashita Taro ''Young Researcher'' Prize from the Yamashita Taro Research Foundation, in 2005.