IMPULSIVE TIMING DETECTION BASED ON MULTI-FRAME PHASE VOTING FOR ACOUSTIC EVENT DETECTION

Sakiko Mishima* and Reishi Kondo*
* NEC Corporation, Kawasaki, Japan
E-mail: {s.mishima, kondoh}@nec.com Tel/Fax: +81-44-431-7663

***Abstract*—This paper proposes impulsive timing detection based on multi-frame phase voting for Acoustic Event Detection (AED). Since an impulsive sound exists for only a short time, the accuracy of the event timing plays an important role in the reduction of labeling costs and in the improvement of learning efficiency. We propose here a method that detects impulsive sounds on the basis of acoustic signal processing using phase slopes, which are among characteristics peculiar to impulsive sounds. Phase slopes from multiple frames are converted to sample points and aggregated by weighted voting. Evaluation employing frame-wise F-scores under noisy environmental conditions shows an improvement of 0.22 over a conventional single frame method for speech processing.**

I. Introduction

Acoustic Event Detection (AED) has become the subject of much attention as a means of situation-understanding [1], [2]. AED attempts to automatically recognize the class and timing of target events from observed acoustic signals recorded under the conditions in which numerous types of events may occur simultaneously. For the purpose of monitoring the cities and individual facilities, an impulsive sound of a short time duration may be important as a clue to understanding the situation because impulsive sounds often correspond to hazardous events as falls of objects or people, or damage to equipment.

In recent years, AED methods based on machine learning have been proposed [3], [4], [5] for modeling acoustic characteristics that commonly appear in the events. A model learns the correspondence of observed signals and labels which indicates event-class and timing of occurrence. Mislabeled data can cause deterioration of event detection performance [6]. Events with a shorter duration sounds demand higher temporal resolution [7]. This incurs the cost of annotating large training datasets, including impulsive sounds, with high temporal resolution.

One typical approach to annotation cost reduction is investigation of weakly labeled data which are only given the event-class included in a sound clip. A number of deep learning methods have been proposed [8], [9], [10], [11]. In the modeling with deep learning, investigations on self-attention architectures [12], [13] contribute to the acquisition of class representation. An event detection model with a self-attention architecture seeks and models characteristics, which contributes to event detection, in the process of training from large datasets. Weak labels include the fuzziness in the presence-timing of the target events. In the case of a target event's existing for only a short time, such as the case of an impulsive sound, estimation of the event presence section is especially difficult.

We have become interested in exploring a new detection framework for an event observed as an impulsive sound, on the bais of the acoustic characteristics of the sound. Our simple idea for the framework is that the event-occurrence timing can be given by signal processing which does not need to use a large dataset for training. The impulsive event detection system considered here consists of a timing detection block and a classification block. The timing detection block gives the classification block a cue for a section which contains characteristics particular to the event class. The classification block efficiently learns the event representation for each class by limitation of sections in the training data. Timing detection plays an important role in the detection of an event which is observed as an impulsive sound.

In this paper, we propose impulsive timing detection based on multi-frame phase voting for AED. Phase slopes, which are among characteristics peculiar to impulsive sounds [15], are calculated frame-by-frame with overlap shift. Our proposed method emphasizes sample points by weighted voting based on phase slopes from multiple frames. Section II here in provides an overview of related impulsive sound detection methods in the speech processing field and difficulties in incorporating technology in the field into AED Section III presents the proposed impulsive timing detection for AED. Section IV presents out experimental setup and give evaluation results. Finally, section V offers concluding remarks and considers directions for future work.

II. Related Work

Impulsive sounds have been detected as noise in order to suppress them for speech quality improvement [16], [17], [18], [19], [20]. Detection methods for noise suppression can be roughly divided into two types: output time-domain and feature-domain. Time-domain approaches [16], [18] exhibit detection results sample-by-sample. Frequency-domain approaches [19], [20] output the presence of impulsive sounds for individual frequencies at given frames. Frequency-domain approaches may often satisfy time resolution requirements

for AED because numerous AED methods adopt frame-wise detection [8], [9], [10], [11].

In one example of a feature-domain approach, Sugiyama et al. focused on phase linearity in impulsive sounds and incorporated this characteristic as phase slope in a detection algorithm [15], [21]. The phase slope of an input noisy signal is compared with an ideal phase slope obtained from an estimated peak point $\hat{p}$. Our proposed method has been inspired by this phase-based detection approach.

The relationship between a peak point and a phase slope can be explained using time-domain input signal $x(n)$ which has a pulse with a magnitude of a at a certain sample point p. The input signal $x(n)$ is converted to the frequency-domain signal $X(k)$ according to following formula:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j \cdot 2\pi k n / N}, \tag{1}$$

where N is frame sample size and k is frequency. Assuming that the input signal $x(n)$ is 0 for $n \neq p$, $X(k)$ can be simplified as

$$X(k) = |X(k)| e^{j\theta(k)} = a e^{\frac{-2\pi k p}{N}}, \tag{2}$$

$$\theta(k) = \frac{-2\pi k p}{N}, \tag{3}$$

where $\theta(k)$ represents the phase and j is equal to $\sqrt{-1}$. The phase slope is obtained by differentiating the phase $\theta(k)$ with frequency k as follows:

$$\frac{d\theta(k)}{dk} = \frac{-2\pi p}{N}. \tag{4}$$

Equation (4) shows that the slope is uniquely determined from the pulse position p. The differentiation can be approximated by the phase difference $\Delta\theta(k)$ in the neighboring frequency bins.

$$\Delta\theta(k) = \theta(k) - \theta(k-1) = \frac{-2\pi p}{N}. \tag{5}$$

To obtain the phase difference $\Delta\theta(k)$, a rotation vector $\bar{X}_{rot}(k)$ as given in (6) has been investigated for the purpose of avoiding phase wrapping problems [22].

$$\bar{X}_{rot}(k) = \bar{X}(k) \cdot \bar{X}^*(k-1) = e^{j\{\theta(k) - \theta(k-1)\}}, \tag{6}$$

$$\bar{X}(k) = \frac{X(k)}{|X(k)|} = e^{j\theta(k)}, \tag{7}$$

where * represents a complex conjugate. From (5) and (6), the phase difference $\Delta\theta(k)$ as shown in (6) is obtained by

$$\Delta\theta(k) = \tan^{-1} \frac{Im\{\bar{X}_{rot}(k)\}}{Re\{\bar{X}_{rot}(k)\}}. \tag{8}$$

In the method proposed by Sugiyama et al, linearity index $LI_\theta(k)$ is calculated as in (9) and compared with a threshold that is close to 0.

$$LI_\theta(k) = \Delta\theta(k) - \frac{-2\pi\hat{p}}{N}. \tag{9}$$

The impulsive sample point $\hat{p}$ is estimated from the largest magnitude sample [15] or magnitude-weighted average of phase differences [21]. However, it is difficult to apply the method to AED because event sounds, which have a variety of frequency characteristics and sound pressure levels, mask the target impulsive sound, and this results in deterioration in estimation performance.

III. PROPOSED METHOD

We propose impulsive timing detection based on multi-frame phase voting for AED. Our proposed method estimates impulsive sample points using weighted phase slopes calculated from multiple frames. One key feature of proposed method is that phase slopes can be converted to sample points in individual frames. Multi-frame voting makes a contribution to robust detection by emphasizing the presence of impulsive sample points. Figure 1 shows an overview of the proposed method.

An input signal is analyzed using overlap shift windowing with frame shift width N/D, where N is frame size and D is the number of overlaps. Equation (8) applies to the frequency transformed version of $x_m(n)$, which should be denoted as $X_m(k)$ according to the previous equation. Local appearance score $s_m(n)$ is calculated from phase slopes $\Delta\theta_m(k)$ as given in Eq. (10, 11) and stored in a storage space.

$$s_m(n) = \sum_{k=0}^{K-1} u_m(n,k), \tag{10}$$

$$u_m(n,k) = \begin{cases} 1, & if\ n = \left\lceil \Delta\theta_m(k) \dfrac{N}{2\pi} \right\rceil, \\ 0, & otherwise. \end{cases} \tag{11}$$

The storage space temporarily stores the local appearance scores of the last $D-1$ frames.

Local appearance scores calculated from current and past frames are utilized for weighted voting. Impulsive score series $y_m(n)$, which is a result of the voting, is calculated by weighting and adding while still satisfying the same sample positional relationship.

$$y_m(n) = \sum_{d=0}^{D-1} w_{m-d} \cdot s_{m-d} \left(n - \frac{N}{D} d \right), \tag{12}$$

$$w_{m-d} = \frac{1}{N} \sum_{n=0}^{N-1} \left\{ s_{m-d}(n) - \bar{s}_{m-d} \right\}, \tag{13}$$

where $\bar{s}_{m-d}$ is averaged local appearance score in frame $(m-d)$. The weighting by in-frame variance w_{m-d} emphasizes the local appearance score in any frame that includes an impulsive sample point. Equation (12) shows that D is related to total sample number for voting.

To estimate the impulsive point, impulsive score series $y_m(n)$ is fitted to Gaussian distribution $f(n;\mu,\sigma)$ by minimizing an error function $J(\mu,\sigma)$ given as

$$\min J(\mu,\sigma) = E\left[(f(n;\mu,\sigma) - y_m(n))^2 \right], \tag{14}$$

$$f(n;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(n-\mu)^2}{2\sigma^2} \right). \tag{15}$$

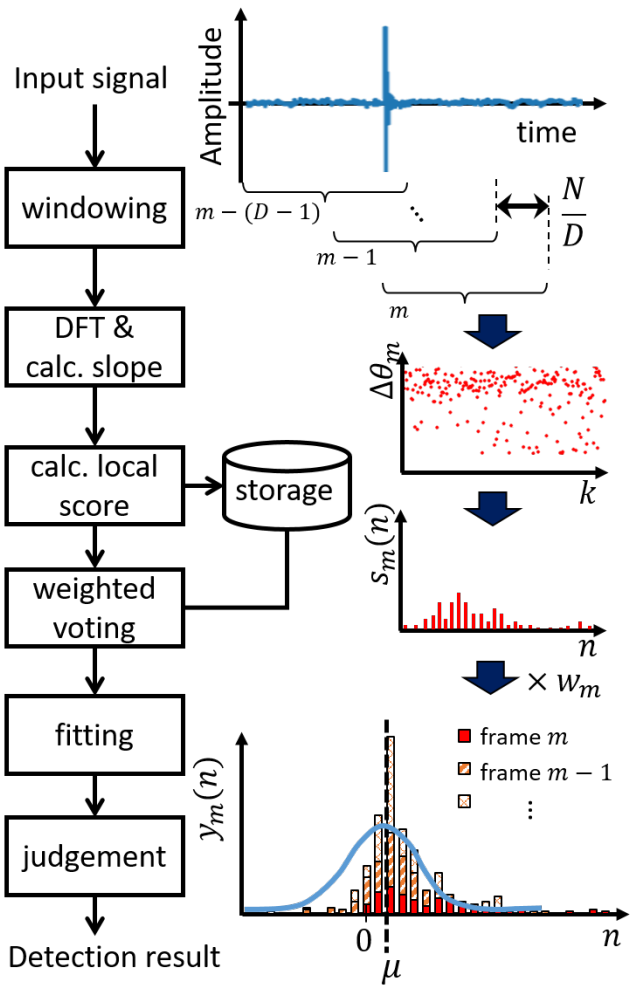


Fig. 1. Overview of the proposed method.

where μ is the mean value and σ is the standard deviation. The mean value μ of a fitted distribution is regarded as a candidate for impulsive sample point $\hat{p}$. Finally, the candidate is judged to be an impulsive sample or not on the basis of threshold α. Concentration score c_m is calculated from the sample position of candidate $\hat{p}$ and its value as

$$\text{Result} = \begin{cases} 1, & if\ c_m > \alpha\ and\ \hat{p} > 0, \\ 0, & otherwise. \end{cases} \tag{16}$$

$$c_m = \frac{f(\hat{p}; \mu, \sigma)}{\sigma}. \tag{17}$$

In the case of the result $= 1$, frame m is judged to contain an impulsive sound.

IV. EXPERIMENTS

A. Experimental condition

The timing-detection performance of the proposed method has been evaluated using synthetic data. For purposes of comparison, a conventional single frame method [15] has been applied to same data

To confirm the robustness of detection methods, sound data were synthesized by mixing clean impulsive sound signals into environmental sound signals with Peak-Signal-to-Noise Ratio (PSNR) controlled to be 10, 20, and 30 dB. Note that the PSNR is low compared to the Signal-to-Noise Ratio (SNR), as an impulsive sound is observed as instantaneous power signal. In the case of multiple peaks in a event, PSNR is calculated from the maximum peak value. Clean impulsive sound data are selected from the Real World Computing Partnership Sound Scene Database (RWCP-SSD) [23], in which the sound clips are classified into three categories in accord with the type of their sound sources. We used 2987 sound clips categorized as being of the collision-sound-source type. Five types of environmental sounds were chosen from the NOISEX-92 database [24]: white noise, pink noise, babble noise, factory noise1, and factory noise2. These sound types differ in terms of frequency characteristics, especially with reference to bandwidth where strong power conditions.

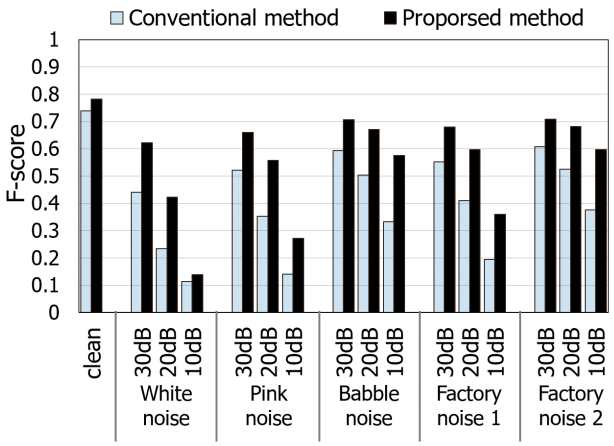
Detection performance was measured in terms of frame-wise and event-wise F-scores. F-scores were based on the total number of false negatives, true positives, and false positives [25]. Frame-wise results were calculated on the basis of frame-labels that showed a presence of peak points on a clean waveform of the frame. Peaks on waveforms were detected using a SciPy peak detection package. Event-wise results showed whether an estimated peak point existed in an event term. Event labels were obtained by means of binarization of absolute amplitudes with 70th percentiles of the sound clips.

All signals were 16 kHz sampling and applied DFT, with frame size $N = 1024$ and the number of overlaps $D = 4$. Detection threshold α was determined for each condition of the environment, using data of which 10 % had not been used in testing. With the conventional method, frame-wise detection results were obtained by thresholding for absolute average of linearity index $LI_\theta(k)$ at frequency k. The threshold used in the conventional method was obtained as it was with the proposed method.

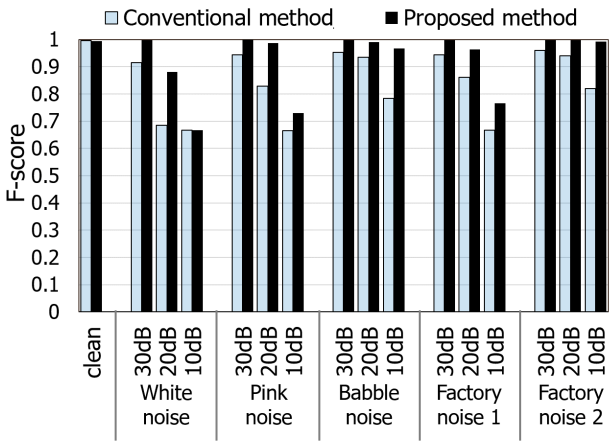
B. Experimental results

Figure 2 compares the performance of the proposed method with that of the conventional method with reference to both frame- and event-wise F-scores. The proposed method outperformed the conventional method with reference to both frame- and event-wise F-scores in noisy environments. Particularly notable is that, under a factory noise2 condition at 10 dB, the proposed method improves significantly over the conventional method by 0.22 with reference to frame-wise score and 0.18 with reference to event-wise score.

The difference between the proposed and conventional methods increased with decreasing PSNR, except under white noise conditions; while white noise had uniform power at all frequencies, the power levels of other noises differed at different frequencies. In other words, some frequencies were less affected by other event sounds under noise conditions other than those under white noise. These results show that the proposed method successfully emphasizes sample points using



(a) Frame-wise F-score



(b) Event-wise F-score

Fig. 2. Experimental results for overall conditions.

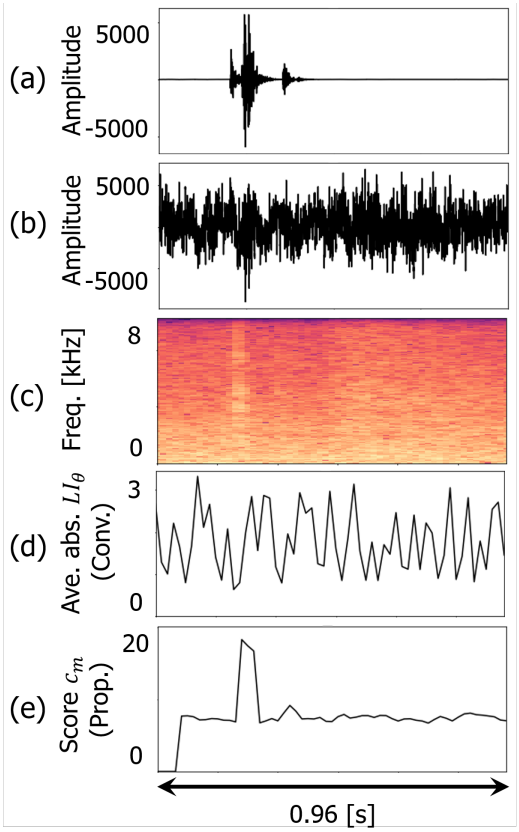


Fig. 3. Results for wood collision sounds with factory noise 1at 10 dB. (a) clean waveform, (b) noisy waveform, (c) noisy spectrogram, (d) absolute average of linearity index $LI_\theta(k)$ with conventional method (small value indicates impulse), (e) concentration score c_m with proposed method (large value indicates impulse).

the phase slopes, which some frequencies are less affected by other event sounds.

Figure 3 shows detection results for wood collision sounds with factory noise2 at 10 dB. In clean environment, the acoustic feature of impulsive sound can be clearly observed as peaks on a waveform (a). Factory noise, however, covers the features on a noisy waveform (b) and noisy spectrogram (c). The absolute average of linearity index $LI_\theta(k)$ with the conventional method fluctuates unsteadily, as can be seen in (d). On the other hands, concentration score c_m with the proposed method (e) increased in the event frame, where it can be observed as large peak on a clean waveform.

There is a gap between frame- and event-wise results because some impulsive sounds are composed of several peaks with different amplitude values, can be seen in (a). The proposed method detects predominant peaks even if it is difficult to detect all peaks in the event. This shows that the proposed method provides not only event timing but also the characteristic part of the impulsive event. Out results also indicate that the impulsive timing detection may successfully support model training with weakly labeled data with reference to training region limitation rather than self-attention modules. As mentioned in Section I, combination with classification block is indispensable for realization of impulsive event detection system because the proposed method only pays attention the feature of sound start timing. In future, combined system will be evaluated from viewpoints of detection and classification under various noise environment.

V. CONCLUSION

We have proposed here impulsive timing detection based on multi-frame phase voting for AED. Our proposed method emphasizes sample points by means of weighted voting using phase slopes, which are among characteristics peculiar to impulsive sounds. Evaluation with frame-wise F-measure under noisy environmental conditions shows performance which exceeds by 0.22 that of a conventional method for speech processing. In future, we intend to consider incorporating our timing detection into impulsive sound detection system that distinguishes detected impulsive sounds in terms of event class.

REFERENCES

[1] M. D. Plumbley T. Virtanen and D. Ellis, *Computational analysis of sound scenes and events*, Springer, 2018.

[2] D.F. Rosenthal and H. G. Okuno, *Computational auditory scene analysis*, Lawrence Erlbaum Associates Publishers, 1998.

[3] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[4] Y. Wang, J. Salamon, N. J. Bryan, and J. Pablo Bello, "Few-shot sound event detection," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 81–85, 2020.

[5] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact pf sound duration and inactive frames on sound event detection performance," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 860–864, 2021.

[6] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 21–25, 2019.

[7] X. Xia, R. Togneri, F. Sohel, and D. Huang, "Confidence based acoustic event detection," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 306–310, 2018.

[8] S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe, and M. Elhilali, "Joint acoustic and class inference for weakly supervised sound event detection," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 36–40, 2019.

[9] A. Diment and T. Virtanen, "Transfer learning of weakly labelled audio," *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (*WASPAA*), pp. 6–10, 2017.

[10] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 121–125, 2018.

[11] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[12] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Audio set classification with attention model: A probabilistic perspective," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 316–320, 2018.

[13] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Weakly-supervised sound event detection with self-attention," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 66–70, 2020.

[14] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 86–90, 2020.

[15] A. Sugiyama, R. Miyahara, and Kwangsoo Park, "Impact-noise suppression with phase-based detection," *Proc. of European Signal Processing Conference* (*EUSIPCO*), pp. 1–5, 2013.

[16] A. Kundu and S. Mitra, "A computationally efficient approach to the removal of impulse noise from digitized speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 4, pp. 571–574, 1987.

[17] S. Kohmura, A. Kawamura, and Y. Iiguni, "An efficient zero phase noise reduction method for impact noise with damped oscillation," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 892–895, 2013.

[18] C. Chandra, M. S. Moore, and S. K. Mitra, "An efficient method for the removal of impulse noise from speech and audio signals," *Proc. of IEEE International Symposium on Circuits and Systems* (*ISCAS*), vol. 4, pp. 206–208, 1998.

[19] H. Liu, R. Zhang, Y. Zhou, X. Jing, and T. Truong, "Speech denoising using transform domains in the presence of impulsive and gaussian noises," *IEEE Access*, vol. 5, pp. 21193–21203, 2017.

[20] A. Kawamura and K. Fujikura, "Impact noise suppression using speech spectral phase estimator," *IEEJ Transactions on Electronics, Information and Systems*, vol. 138, no. 11, pp. 1410–1416, 2018.

[21] A. Sugiyama and R. Miyahara, "A tapping-noise suppressor with magnitude-weighted phase-based detection for smartphones," *Proc. of IEEE International Conference on Consumer Electronics* (*ICCE*), pp. 526–527, 2014.

[22] Francois Leonard, "Phase spectrogram and frequency spectrogram as new diagnostic tools," *Mechanical Systems and Signal Processing*, vol. 21, no. 1, pp. 125 – 137, 2007.

[23] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, and Takashi Endo, "Sound scene data collection in real acoustical environments," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 225–231, 1999.

[24] Andrew Varga and Herman J.M. Steeneken, "Assessment for automatic speech recognition: II. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[25] T. Heittola A. Mesaros and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, pp. 162, 2016.