SELF-SUPERVISED MEAN OPINION SCORE PREDICTION OF PHASE-VOCODER-BASED VIRTUAL BASS SYSTEM

Jiacheng Gou[1], *Yuheng Song*[2], *Chuang Shi*[3], *Huiyong Li*[1]

[1]School of Information and Communication Engineering,
University of Electronic Science and Technology of China, China
[2]Department of Computing, Hong Kong Polytechnic University, Hong Kong
[3]Institute of Sound and Vibration Research,
University of Southampton, Southampton, United Kingdom

ABSTRACT

The virtual bass system (VBS) leverages a psychoacoustic phenomenon known as the "missing fundamental" to trick listeners into perceiving the fundamental frequency from its higher harmonics. The VBS finds common use in consumer electronic devices where miniature and flat panel loudspeakers are integrated, as they cannot reproduce satisfactory low-frequency components. The additional harmonics introduced by the VBS can lead to perceptual distortion. Therefore, evaluating the perceptual quality of the VBS necessitates subjective listening tests. Previous studies have attempted to derive objective metrics and identify combinations of model output variables to predict the perceptual quality of the VBS. However, due to the limited number of subjective test results used to obtain the combination coefficients, inconsistencies may arise in predictions. This paper proposes to adopt self-supervised deep learning models to predict the mean opinion score (MOS) of the VBS. Experiment results demonstrate a strong linear correlation between the model outputs and the human-rated MOS, indicating that a linear mapping is sufficient to convert a model output into an accurate MOS prediction.

***Index Terms*—** Virtual bass system, phase vocoder, self-supervised learning, mean opinion score prediction

I. INTRODUCTION

Small-sized loudspeakers in consumer electronic devices are often criticised for their limited bass reproduction capabilities, due to both the constraints of their physical size and cost considerations. The VBS is introduced to address this concern by leveraging a psychoacoustic phenomenon known as the "missing fundamental". It suggests that higher-order harmonics of a fundamental frequency can simulate the perceptual experience of the fundamental frequency [1]. Fig. 1 illustrates a general framework of the VBS, where the harmonic generator can be implemented using a nonlinear device (NLD) [2], [3], [4], the phase vocoder (PV) [5], or a hybrid combination of both [6], [7]. In [6], Hill and Hawksford also emphasise the importance of a delay for synchronizing the original audio with the generated harmonics.

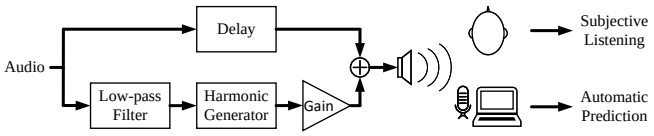

Fig. 1. Comparing subjective listening tests to automatic predictions for perceptual quality assessment of the VBS.

In addition to choosing appropriate harmonic generators, tuning the VBS parameters is crucial for achieving optimal perceptual quality [8]. For instance, most of VBS implementations involve a harmonic gain. A higher gain introduces more harmonics, but it can lead to more noticeable distortion. There is no automatic method to set this gain. Consequently, evaluating the perceptual quality of the VBS typically involves subjective listening tests that demand careful choice of method, hours of careful listening, and proof of statistical significance [5].

Simple objective scores quantifying bass enhancement and harmonic distortion were demonstrated to have a weak correlation with subjective test results [9]. To tackle this issue, Oo and Gan initially explored the *GedLee* metric, incorporating an objective model of masking [10], [11], but it failed to accurately predict the perceived distortion of the VBS. Subsequently, Oo *et al.* investigated the *Rnonlin* distortion model [12] and conducted nonlinear regression to fit the outputs of the *Rnonlin* model with subjective test results. They claimed that the *Rnonlin* model could reliably predict the perceived distortion of the NLD-based VBS [9].

On the other hand, Mu *et al.* examined the audio spectrum centroid (ASC) and the increment ratio of ASC to evaluate the unnatural timbre sharpness effect resulting from additional harmonics [13]. However, there was no reliable correlation between those metrics and the perceptual quality of

the VBS [14]. Therefore, they explored various combinations of model output variables from the ITU Recommendation ITU-R BS.1387 and developed a linear regression model using subjective test results. Their final findings suggested that steady-state and transient stimuli, as well as single instrument and polyphonic music stimuli, could be predicted by separate combinations of model output variables [15]. It should be noted that the aforementioned attempts to predict the perceptual quality of the VBS have always relied on subjective test results as prior information. This is likely to raise a suspicion of circular reasoning.

Recently, self-supervised models have demonstrated their effectiveness in automatically predicting the MOS for synthesised speech [16] and perceptual similarity for transformer noise [17]. Deep learning-based methods hold promise in learning features from extensive audio data, eliminating the need for manual design and selection of objective metrics and thereby reducing the influence of human bias. Therefore, this paper proposes a self-supervised method for predicting the perceptual quality of the VBS. This self-supervised method eliminates the need for prior subjective test results. Neural network models are trained for audio artifact detection, outputting an artifact score. Experiment results demonstrate a strong linear correlation between this artifact score and the human-rated MOS. Hence, a simple linear mapping can be employed to convert the artifact score into a reliable MOS prediction.

II. PV-BASED VBS AND SUBJECTIVE TEST

A PV-based VBS is implemented for the subjective test, where the exponential attenuation scheme is adopted as

$$W_i = exp(-\alpha \cdot i), \tag{1}$$

where W_i is the gain of the ith harmonic, and α determines the attenuation rate. Each testing audio clip in the dataset is processed by the PV-based VBS using α of 0.1, 0.3, 0.5, 0.7, and 0.9. The NLD-based VBS is not considered, for the same reason as elaborated in [15].

Two types of musical recordings: "Violin Solo" and "Vocal and Drum Mix", abbreviated as "Solo" and "Mix" for brevity, were used for the subjective test. Moreover, 4 "Solo" and 4 "Mix" clips were processed by the PV-based VBS using α of 0.1, 0.3, 0.5, 0.7, and 0.9, resulting in a total of 40 stimuli. We recruited 11 males and 10 females with normal hearing, aged between 22 and 33 years old, for the subjective test. The test was conducted in a quiet room, and participants took it individually without time constraints. They were allowed to adjust the volume of the headphone according to their listening preferences. Participants listened to every stimulus in a randomised order and rated both the perceived bass intensity and the overall sound quality using the absolute category rating scale [18].

The human-rated MOSs were then examined using one-way analysis of variance (ANOVA). The p-values of bass intensity and sound quality yielded $5e{-}5$ and $1e{-}4$ for stimuli of "Solo" and $6e{-}6$ and $1e{-}7$ for stimuli of "Mix", respectively. They were much less than 0.01, demonstrating the statistical significance [19].

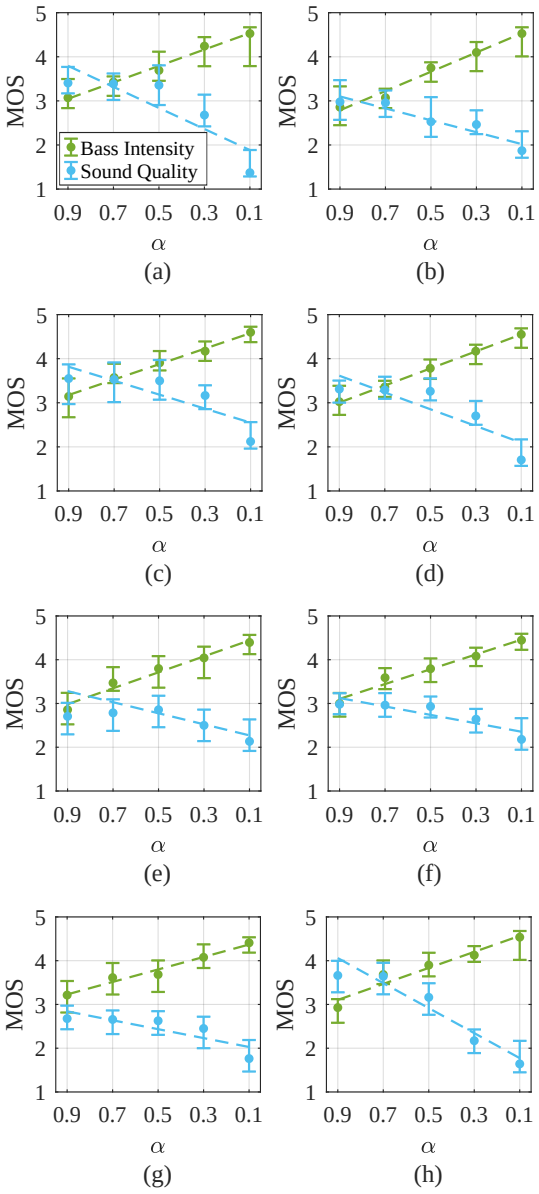

Fig. 2. Subjective test results: (a)-(d) for "Solo" Clip 1-4 and (e)-(g) for "Mix" Clip 1-4, respectively

Fig. 2 illustrates the human-rated MOSs for each original audio clips in the subjective test. The bars indicate 95% confidence intervals, the dots represent mean values, and the dashed trend lines depict the linear regressions with respect to α. It is noted that the linear correlation between α and the human-rated MOS of bass intensity exceeds 99%. In contrast, the correlation between α and the human-rated MOS of sound quality is hardly linear.

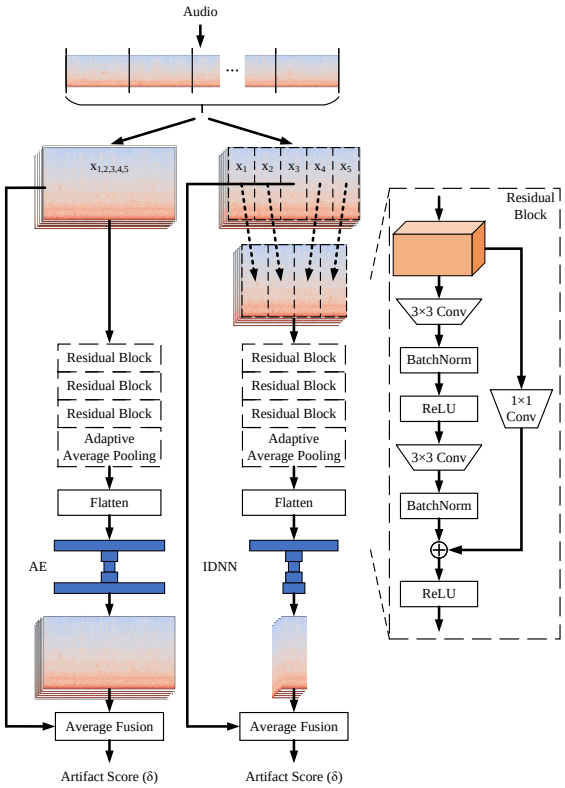

Fig. 3. Block diagrams of self-supervised models using (a) AE, (b) IDNN, and optional (c) residual blocks.

III. SELF-SUPERVISED MODELS FOR MOS PREDICTION

As depicted in Fig. 3, self-supervised models, such as the autoencoder (AE) [20] and the interpolation deep neural network (IDNN) [21], are well-suited for detecting anomalies in audio clips. In this paper, we employ both models and incorporate residual blocks to augment the extraction of localised time-frequency information. It should be noted that the network architecture design is beyond the scope of this paper.

Table I and Table II present the detailed architectures of the AE and IDNN models, both equipped with the residual blocks. Initially, an audio clip undergoes transformation into a log-mel spectrogram and is subsequently partitioned into a series of five-frame segments. In the case of AE, a five-frame output is reconstructed based on each five-frame input segment. The training loss is designed to optimise the accuracy of this reconstruction, which is expressed as

$$L_{\mathrm{AE}} = \left\| \mathbf{x}_{1,2,3,4,5} - AE\left(\mathbf{x}_{1,2,3,4,5}\right) \right\|_2^2, \tag{2}$$

where $\mathbf{x}_{1,2,3,4,5}$ represents the five-frame input segment and the subscript indicates the frame number. Conversely, in the case of IDNN, the centre frame of each five-frame segment

Table I. Architecture of AE with residual blocks.

Layers	Output Dimension	Activation
Input	128×5×1	-
Residual Block	128×5×8	ReLU
Residual Block	128×5×16	ReLU
Residual Block	128×5×32	ReLU
Adaptive Average Pooling	20×1×32	-
Flatten	640	-
Fully Connected	128	ReLU
Fully Connected	96	ReLU
Fully Connected	64	ReLU
Fully Connected	96	ReLU
Fully Connected (Output)	640	-

Table II. Architecture of IDNN with residual blocks.

Layers	Output Dimension	Activation
Input	128×4×1	-
Residual Block	128×4×8	ReLU
Residual Block	128×4×16	ReLU
Residual Block	128×4×32	ReLU
Adaptive Average Pooling	16×1×32	-
Flatten	512	-
Fully Connected	128	ReLU
Fully Connected	96	ReLU
Fully Connected	64	ReLU
Fully Connected	96	ReLU
Fully Connected (Output)	128	-

is extracted and, as the name implies, predicted through interpolation using the remaining four frames. The training loss is tailored to enhance prediction accuracy, specifically by minimising

$$L_{\mathrm{IDNN}} = \left\| \mathbf{x}_3 - IDNN\left(\mathbf{x}_{1,2,4,5}\right) \right\|_2^2. \tag{3}$$

After a model is fully trained, the loss values are averaged across the entire set of five-frame segments to generate the artifact score.

Training of the AE and IDNN models requires only a number of "normal" audio clips, commonly referred to as original audio clips in the context of VBS studies. Thus, a dataset is created, comprising 1800 training audio clips, 200 validation audio clips, and 78 testing audio clips. None of them have been presented in the subjective test. Each audio clip has a duration of 10 seconds, is monaural, and sampled at 22.05 Hz. The dataset includes two equally represented types of "Solo" and "Mix".

More specifically, with a window size of 1024 and a 50% overlap, each 10-second audio clip consists of 430 frames, further divided into 86 consecutive five-frame segments. The number of mel bands used is 128. Training in PyTorch is carried out using the adaptive moment estimation (Adam) optimiser on a single NVIDIA GeForce RTX 2080Ti graphics card, employing 300 training epochs, a batch size of 8196, and a learning rate of 0.002.

Four self-supervised models, including the AE with and without residual blocks, and the IDNN with and without residual blocks, are first compared by their AUCs. AUC

Table III. AUCs of four self-supervised models tested with two types of musical recordings and five harmonic gains.

Model	Type	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$
AE	Solo	95.99%	90.66%	84.09%	77.71%	70.41%
	Mix	89.68%	82.45%	75.02%	69.03%	64.56%
+ Residual Blocks	Solo	99.93%	99.67%	99.15%	97.50%	95.40%
	Mix	97.11%	96.19%	94.35%	92.11%	88.95%
IDNN	Solo	99.93%	99.41%	97.44%	93.29%	86.98%
	Mix	88.30%	82.64%	76.33%	70.74%	67.52%
+ Residual Blocks	Solo	100.00%	100.00%	100.00%	99.74%	99.15%
	Mix	98.16%	96.98%	95.60%	94.15%	91.98%

stands for the area under the receiver operating characteristic (ROC) curve when the self-supervised model is used to discriminate whether VBS processing has been applied. The testing results are listed in Table III. A model with a high AUC can be interpreted as having a high standard for sound quality, since it has been trained to identify even trivial changes as artifacts.

IV. MOS PREDICTION RESULTS

Thereafter, the four self-supervised models are examined for their abilities to predict the MOS of sound quality. It is worth emphasising that the self-supervised models have been trained exclusively on "normal" audio clips in the aforementioned dataset, and they are independently from the subjective test.

In Table IV, the linear correlations between the model outputs and the human-rated MOS for sound quality are detailed for each original audio clip and for every group of clips with the same musical type in the subjective test. The model outputs exhibit stronger linear correlations with the human-rated MOS than the trend lines in Fig. 2. The overall score indicates the linear correlation for a group of clips, which is not a simple average of the linear correlations for individual clips. The IDNN with residual blocks demonstrates the highest consistency across different clips, which aligns with the observation that it also achieves the highest AUCs in Table III.

However, the self-supervised model outputs exhibit distinct dynamic ranges as compared to that of MOS. A linear mapping is therefore considered, where the mapping matrix $\boldsymbol{P}$ is expressed as

$$\boldsymbol{P} = \begin{pmatrix} \bar{\delta}_{audio} & 1 \\ \bar{\delta}_{noise} & 1 \end{pmatrix}^{-1} \times \begin{pmatrix} \sigma_{max} \\ \sigma_{min} \end{pmatrix}. \tag{4}$$

In (4), $\bar{\delta}_{audio}$ and $\bar{\delta}_{noise}$ represent the averaged audio artifact scores of a range of training audio clips and pure noise clips, respectively. They are treated as the lower and upper bounds of the artifact score and are associated with the MOS range from $\sigma_{min} = 1$ to $\sigma_{max} = 5$. These settings ensure that the linear mapping is independent from the subjective test. Hence, a model output δ can be converted to a model-predicted MOS by $(\delta, 1) \times \boldsymbol{P}$.

Table IV. Linear correlations between self-supervised model outputs and human-rated MOSs.

Model	Type	Clip 1	Clip 2	Clip 3	Clip 4	Overall
Trend Line (α)	Solo	81.28%	84.56%	79.20%	71.56%	76.42%
	Mix	79.51%	91.61%	76.84%	82.43%	86.85%
AE	Solo	94.69%	88.84%	**97.68%**	**90.25%**	91.26%
	Mix	96.84%	99.44%	88.03%	95.55%	94.94%
+ Residual Blocks	Solo	94.45%	88.37%	97.28%	89.86%	91.60%
	Mix	96.68%	99.59%	88.95%	**99.28%**	97.21%
IDNN	Solo	91.75%	**89.69%**	97.15%	85.73%	90.23%
	Mix	91.13%	99.15%	86.14%	95.54%	92.70%
+ Residual Blocks	Solo	**95.85%**	86.03%	95.75%	89.99%	**93.60%**
	Mix	**97.81%**	**99.89%**	**98.12%**	98.08%	**99.24%**

Table V. MAEs between model-predicted MOSs and human-rated MOSs.

Model	Type	Clip 1	Clip 2	Clip 3	Clip 4	Overall
Trend Line (α)	Solo	0.361	0.121	0.278	0.277	0.162
	Mix	0.134	0.127	0.172	0.361	0.130
AE	Solo	0.162	0.072	0.153	0.118	0.092
	Mix	**0.070**	0.073	0.121	0.265	0.087
+ Residual Blocks	Solo	0.155	0.075	0.148	0.112	0.089
	Mix	0.112	0.057	0.073	0.365	0.052
IDNN	Solo	0.181	**0.071**	0.166	0.133	0.098
	Mix	0.071	0.091	0.139	**0.229**	0.103
+ Residual Blocks	Solo	**0.111**	0.084	**0.118**	**0.078**	**0.074**
	Mix	0.077	**0.026**	**0.049**	0.305	**0.025**

Table V shows the mean absolute error (MAE) between the model-predicted MOS and the human-rated MOS. The mean squared error is not presented because it closely aligns with the trend observed in the linear correlation in Table IV. Both Tables V and IV demonstrate that the IDNN with residual blocks is able to provide a reliable MOS prediction.

V. CONCLUSIONS

This paper proposes using self-supervised models to predict the MOS of PV-based VBS. The self-supervised learning strategy uses a dataset of original audio clips without VBS processing, eliminating the need for prior subjective test results and thereby reducing inconsistencies caused by human bias. These self-supervised models are specifically designed for audio artifact detection, and their outputs are artifact scores, which have a different dynamic range compared to the MOS. Therefore, we conducted experiments to show a strong linear correlation between the artifact scores and the human-rated MOS for the sound quality of the PV-based VBS. Among the four self-supervised models investigated, the IDNN with residual blocks leads to the highest AUC, and accordingly, its output shows the strongest and most consistent linear correlation with the human-rated MOS. Moreover, it is also demonstrated that a linear mapping is sufficient to convert the artifact score into an accurate MOS prediction.

VI. REFERENCES

[1] D. Ben-Tzur and M. Colloms, "The effect of maxxbass psychoacoustic bass enhancement on loudspeaker design," presented at the 106th Audio Eng. Soc. Conv., Munich, Germany, 1999, Paper 4892.

[2] N. Oo and W. S. Gan, "Harmonic and intermodulation analysis of nonlinear devices used in virtual bass systems," presented at the 124th Audio Eng. Soc. Conv., Amsterdam, The Netherlands, 2008, Paper 7403.

[3] R. Giampiccolo, A. Bernardini, and A. Sarti, "A time-domain virtual bass enhancement circuital model for real-time music applications," in *Proc. 2022 IEEE Int. Workshop Multimedia Signal Process.*, Shanghai, China, 2022, pp. 1–5.

[4] R. Giampiccolo, A. I. Mezza, A. Bernardini, and A. Sarti, "Virtual bass enhancement via music demixing," *IEEE Signal Process. Lett.*, vol. 30, pp. 908–912, Jul. 2023.

[5] M. R. Bai and W.-C. Lin, "Synthesis and implementation of virtual bass system with a phase-vocoder approach," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1077–1091, Nov. 2006.

[6] A. J. Hill and M. O. J. Hawksford, "A hybrid virtual bass system for optimized steady-state and transient performance," in *Proc. 2010 Comput. Sci. Electron. Eng. Conf.*, Colchester, UK, 2010, pp. 1–6.

[7] H. Mu, W. S. Gan, and E. L. Tan, "A psychoacoustic bass enhancement system with improved transient and steady-state performance," in *Proc. 2012 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, 2012, pp. 141–144.

[8] C. Shi, H. Mu, and W. S. Gan, "A psychoacoustical preprocessing technique for virtual bass enhancement of the parametric loudspeaker," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, Canada, 2013, pp. 31–35.

[9] N. Oo and W. S. Gan, "Analytical and perceptual evaluation of nonlinear devices for virtual bass system," presented at the 128th Audio Eng. Soc. Conv., London, UK, 2010, Paper 8108.

[10] E. R. Geddes and L. W. Lee, "Auditory perception of nonlinear distortion - theory," presented at the 115th Audio Eng. Soc. Conv., New York, NY, USA, 2003, Paper 5890.

[11] L. W. Lee and E. R. Geddes, "Auditory perception of nonlinear distortion," presented at the 115th Audio Eng. Soc. Conv., New York, NY, USA, 2003, Paper 5891.

[12] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, "Predicting the perceived quality of nonlinearly distorted music and speech signals," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 699–711, Jul./Aug. 2004.

[13] H. Mu, W. S. Gan, and E. L. Tan, "A timbre matching approach to enhance audio quality of psychoacoustic bass enhancement system," in *Proc. 2013 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, Canada, 2013, pp. 36–40.

[14] H. Mu and W. S. Gan, "Perceptual quality improvement and assessment for virtual bass systems," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 900–913, Jul. 2015.

[15] H. Mu, W. S. Gan, and E. L. Tan, "An objective analysis method for perceptual quality of a virtual bass system," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 5, pp. 840–850, Mar. 2015.

[16] W. C. Huang, E. Cooper, Y. Tsao, H. M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2022," in *Proc. 2022 Annu. Conf. Int. Speech Commun. Assoc.*, Incheon, Korea, 2022, pp. 4536–4540.

[17] C. Shi, M. Huang, C. Liu, and H. Li, "Active noise control with selective perceptual equalization to shape the residual sound," *Appl. Acoust.*, vol. 208, Jun. 2023, Art. no. 109376.

[18] *Methods for subjective determination of transmission quality*, Rec. ITU-T P.800, International Telecommunications Union, Geneva, Switzerland, Aug. 1996.

[19] D. Freedman, R. Pisani, and R. Purves, *Statistics*, 4th ed. New York, NY, USA: W. W. Norton & Company, 2007.

[20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[21] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech, Signal Process.*, Barcelona, Spain, 2020, pp. 271–275.