Received 13 February 2026
Revised 26 May 2026
Accepted 29 June 2026

Sound effect synthesis from language-independent mimicry: a case study on explosion sounds

Riki Takizawa
Department of Frontier Informatics, Kyoto Sangyo University, Kyoto, Japan, and Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Shigeyuki Hirai
Faculty of Information Science and Engineering, Kyoto Sangyo University, Kyoto, Japan

Asako Kanezaki
School of Computing, Institute of Science Tokyo, Tokyo, Japan, and

Hitoshi Suda
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Abstract

Existing vocal-imitation-based methods for environmental sound or sound effect synthesis typically use root-mean-square envelopes, text prompts or discrete codes, and do not directly learn spectral correspondences between imitation and target sounds. This study uses vocal mimicry, nonlinguistic imitation produced by the human articulatory system, as a control signal for sound effect synthesis. This study proposes PronounSE, a method that directly learns spectral correspondences between vocal mimicry and sound effects to generate outputs whose nuances follow the mimicry input. To address data scarcity, this study constructs a new paired vocal-mimicry data set focused on explosion sounds as a case study, since this category exhibits diverse acoustic variations and is relatively amenable to vocal mimicry due to its predominantly aperiodic, noise-like characteristics. In addition to naturalness and audio quality, this study evaluates mimicry-driven synthesis with two task-specific criteria: fidelity to the target sound and reflection of vocal mimicry nuances. These four criteria support both objective and subjective evaluation. Results show strong objective-subjective correlations for fidelity and confirm that PronounSE effectively reflects mimicry nuances. Results also indicate that fidelity is strongly affected by mimicry skill and speaker-specific characteristics, highlighting the importance of

Funding: This work was supported by AIST KAKUSEI project (FY2024). This research was partially supported by the AIST policy-based budget project "R&D on Generative AI Foundation Models for the Physical Domain" and JST SPRING (Japan Grant Number JPMJSP2157).

APSIPA Transactions on Signal and Information Processing
Vol. 15 No. 1, 2026
pp. 464-488
Emerald Publishing Limited
2048-7703
DOI 10.1108/ATSIP-02-2026-0010

evaluating fidelity and nuance reflection as distinct criteria in vocal-mimicry-based sound effect synthesis.

Keywords Sound effect, Environmental sound

Paper type Research paper

1. Introduction

Sound effects – including everyday environmental sounds, fantastical or exaggerated sounds, and user interface operation sounds – are intentionally added to enrich the user experience of various multimedia content. In visual media such as games, animation and film, a wide variety of sound effects are selected for each scene. In professional production settings, sound effects are created under the supervision of a sound director using multiple approaches: selecting audio assets from large commercial sound effect libraries or in-house libraries, performing Foley recording or field recording (Viers, 2008; Ament, 2009; Tierno, 2020), designing sounds (Sonnenschein, 2001) (conceptual/expressive sound design; including synthesis techniques with hardware or software synthesizers), generating sounds with commercially available procedural audio software (Sinclair, 2001) such as Wwise [1], Reformer Pro [2] and GameSynth [3], and sound editing and mixing (Holman, 2015; Yewdall, 2012). All of these methods demand specialized knowledge, practical experience and creative judgment.

Deep learning techniques based on large-scale data have made it increasingly easy to produce high-quality visual content (e.g. with models such as Dream Machine [4], Sora 2 [5] and Veo 3 [6]). In parallel with this, techniques for synthesizing environmental sounds, beyond speech or music, have become an active area of research. For example, the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge included dedicated environmental sound synthesis tasks: "Foley Sound Synthesis" (Choi *et al.*, 2001) in 2023 and "Sound Scene Synthesis" (Lagrange *et al.*, 2025) in 2024. Existing environmental sound synthesis approaches generate sounds from a range of input modalities, such as phoneme sequences (Okamoto *et al.*, 2021), acoustic event labels (Okamoto *et al.*, 2021), text prompts (Liu *et al.*, 2023; Liu *et al.*, 2024; Evans *et al.*, 2024; Yang *et al.*, 2023), video prompts (Luo *et al.*, 2023; Comunità *et al.*, 2024; Du *et al.*, 2023; Chen *et al.*, 2025) or audio signals (Evans *et al.*, 2024; Okamoto *et al.*, 2024a; Chung *et al.*, 2024; García *et al.*, 2025). However, these approaches often struggle to capture subtle variations in timbre and dynamics, and video-based approaches generally fail to account for sound events that occur outside the video frame. In addition, most existing methods primarily focus on everyday environmental sounds and animal calls, and the capability of synthesizing fictional sound effects is not sufficiently explored.

In this study, vocal mimicry refers to a language-independent vocal expression that captures fine-grained acoustic characteristics of sounds and cannot be adequately represented by onomatopoeia or phoneme sequences. In professional sound production, vocal mimicry of sound effects is commonly used to communicate intended sounds between collaborators (Lemaitre *et al.*, 2013; Mehrabi *et al.*, 2017; Lemaitre *et al.*, 2011). Inspired by the fact that both specialists and nonspecialists can typically render sounds they have heard or imagined through such mimicry (Hunt, 1923; Caren *et al.*, 2024), this paper proposes PronounSE, a method that synthesizes sound effects directly using vocal mimicry as a nonlinguistic control input for sound synthesis. As shown in Figure 1, PronounSE leverages the human articulatory system, incorporates a Transformer encoder-decoder network (Vaswani *et al.*, 2017) and uses a waveform vocoder (neural vocoder) to synthesize sound effects.

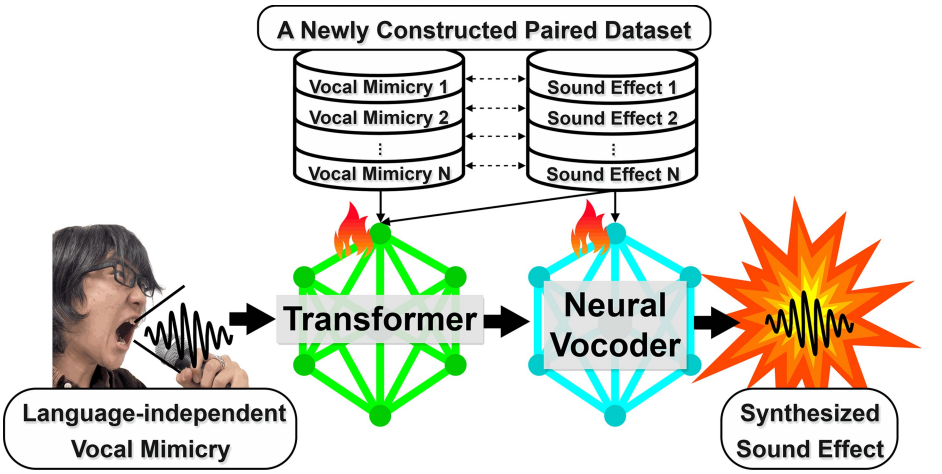

Figure 1. Schematic image of PronounSE. The method synthesizes sound effects from vocal mimicry using Transformer (Vaswani *et al.*, 2017) and a neural vocoder

Since PronounSE learns correspondences between vocal mimicry and sound effects, it requires a data set in which each sound effect is paired with a corresponding instance of vocal mimicry. To meet this requirement, we constructed a new high-quality, language-independent vocal mimicry data set. We therefore treat explosion sounds as a case study for examining the feasibility of sound effect synthesis from language-independent vocal mimicry. In the construction, we especially focus on explosion sounds because they are diverse and relatively easy to mimic vocally. Training PronounSE on this data set enables controllable synthesis of explosion sounds that reflect subtle vocal mimicry nuances and fine-grained variations in sound characteristics.

To evaluate the effectiveness of PronounSE, we conducted a comparative study with two other existing methods (Evans *et al.*, 2024; Chung *et al.*, 2024) that accept vocal mimicry as input. We set four criteria for subjective and objective evaluation: fidelity to the target sound, reflection of the vocal mimicry nuance, naturalness as an explosion sound and quality of the audio source. Accordingly, this paper should be regarded as a case study on explosion sounds rather than a full validation across all sound-effect categories, and broader validation is left for future work.

In the remainder of this paper, Section 2 describes existing sound synthesis approaches that accept vocal mimicry as input, as well as existing data sets that pair environmental sounds with vocal mimicry. Sections 3 and 4 then describe the proposed method, PronounSE, and the newly constructed data set. Sections 5 and 6 present the experimental setups and the results on the evaluation of PronounSE. Section 7 discusses the results and limitations of the method, and Section 8 concludes the paper.

The novelty of this work lies not in introducing a fundamentally new neural architecture, but in formulating language-independent vocal mimicry as a direct control signal for sound effect synthesis, together with a paired data set and an evaluation framework tailored to this task. The main contributions of this case study on explosion sound synthesis from language-independent vocal mimicry are summarized as follows:

- We collected 600 vocal mimicry samples from six speakers in their twenties for 100 commercial explosion sound samples and released the vocal mimicry subset as an evaluation data set.

- The proposed approach, which learns spectral correspondences between vocal mimicry and sound effects, can effectively reflect the nuances of vocal mimicry in synthesized sound.
- The paper introduces an evaluation scheme for nuance reflection that decomposes the evaluation into multiple aspects, enabling analysis of which parts of the synthesized sound are influenced by vocal mimicry and to what extent.
- The experimental results show that high scores for nuance reflection can coexist with low fidelity scores, indicating that fidelity depends on the skill of the vocal mimicry and highlighting the necessity of evaluating both fidelity and nuance reflection as distinct criteria.
- The results also demonstrate a strong correlation between objective scores based on both Fréchet Audio Distance (FAD) (Kilgour *et al.*, 2019) and cosine similarity and subjective fidelity ratings, supporting the effectiveness of FAD and cosine similarity for automatic evaluation of fidelity.

2. Related work

2.1 Sound synthesis techniques using vocal mimicry input

Stable Audio 2.0 (Evans *et al.*, 2024) is an audio synthesis model trained on a large-scale data set of music and sound effects. In addition to a mandatory text prompt, it can take reference audio signals such as instrumental sounds or vocal mimicry, and allows users to control the conditioning strength and the duration of the output while synthesizing sounds at a 44100 Hz sampling rate. Its training data are predominantly musical, and it is not specifically designed for sound effect synthesis.

Voice-to-Foley (Okamoto *et al.*, 2024b) is a technique that synthesizes environmental sounds from reference audio signals together with acoustic event labels. The model is trained to reconstruct Mel-spectrograms of environmental sounds from discrete vocal mimicry units obtained by applying the *k*-means algorithm to features extracted with a pretrained encoder (Niizumi *et al.*, 2022). The vocal mimicry in the ESC-50-Voice data set (Okamoto *et al.*, 2024a), which is used to train the model, consists of onomatopoeic utterances with clear linguistic characteristics rather than nonlinguistic vocal mimicry.

T-Foley (Chung *et al.*, 2024) is an environmental sound synthesis method based on acoustic event labels and the root-mean-square (RMS) envelope of the audio signal, which encodes temporal information such as dynamics and timing. During training, the model is conditioned on the RMS values computed from reference environmental sounds, along with their acoustic event labels. Therefore, paired data between vocal mimicry and environmental sounds are not required when training, and vocal mimicry is used only at synthesis time. In this way, T-Foley utilizes RMS envelopes to enforce temporal alignment between vocal mimicry and environmental sounds. Hence, the method does not learn spectral correspondences between them.

Sketch2Sound (García *et al.*, 2025) is an audio synthesis method that controls both the category and the temporal nuances of the generated sound by conditioning not only on a text prompt but also on three time-varying control signals: loudness, brightness and pitch. As with T-Foley, the method does not learn spectral correspondences between vocal mimicry and sound effects.

Krotos Reformer Pro is a procedural-audio tool for professional sound designers that synthesizes new sound effects by blending assets from sound effect libraries, conditioned on inputs such as vocal mimicry and MIDI signals. A wide range of parameters, including

blending behavior, can be adjusted to allow flexible sound creation. However, effective use of the tool requires users to manually select appropriate source assets and tune a complex set of blending and control parameters, which demands substantial expertise and training.

Beyond vocal-mimicry-driven sound synthesis itself, human involvement has also been explored in other audio-generation-related tasks. For example, Wang *et al.* (2024) introduced a human-in-the-loop framework for masked-speech enhancement, where a quality predictor learned from human subjective judgments was incorporated into the optimization loop (Wang *et al.*, 2024). Although their task and the role of human involvement differ from ours, this study is relevant as an example of integrating human evaluation into audio generation systems.

2.2 Publicly available vocal mimicry data sets

VocalSketch Data Set (Cartwright and Pardo, 2015a) was created to systematically investigate how well various types of sounds can be vocally imitated for use as inputs to audio software (Cartwright and Pardo, 2015b). The data set comprises 240 reference sounds, drawn from subsets of everyday environmental sounds, acoustic instruments, commercial synthesizers and custom synthesized sounds, and provides 4,429 vocal mimicry samples recorded by approximately 200 workers on Amazon Mechanical Turk [7]. The data set includes nonenvironmental content, such as instrument and synthesizer sounds, and provides only a small number of reference sounds. In addition, the vocal recordings may contain lip noise, breath noise and background noise due to their crowdsourced nature. Therefore, the data set does not meet the requirements of this study, which necessitates high-quality audio for sound effect synthesis.

Vocal Imitation Set (Kim and Pardo, 2018) is a vocal mimicry data set designed to support query-by-vocal-imitation and to provide baselines for acoustic classification (Bongjun *et al.*, 2018). The data set contains 302 environmental sound samples (302 classes, one sample per class) collected from Freesound [8], along with 5,601 vocal mimicry recordings contributed by 455 workers on Amazon Mechanical Turk. However, since only a single environmental sound sample is available for each class, modeling intra-class diversity is not possible using the data set alone. In addition, since the samples are derived from Freesound, a community-driven platform where anyone can upload sounds, the sample quality is not sufficiently controlled. Moreover, as with the VocalSketch Data Set, the vocal mimicry samples are collected via crowdsourcing. Hence, the data set does not meet the requirements of this study either.

ESC-50-Voice (Okamoto *et al.*, 2024b) is a data set of onomatopeic utterances constructed for training Voice-to-Foley models (Okamoto *et al.*, 2024b). The data set provides 1,240 environmental sounds from 31 classes of ESC-50[9] (Piczak, 2015), together with 9,920 onomatopoeic recordings produced by eight speakers. However, 40 reference sounds per class is too few for flexible sound effect synthesis, and the quality of the reference sounds is also limited. Moreover, the data set consists of onomatopoeic utterances that follow the Japanese phonology; nevertheless, language-independent vocal expressions are required in this study.

3. Proposed method: PronounSE

3.1 Concept

In this study, we propose PronounSE, a sound effect synthesis method based on vocal mimicry that imitates the sounds themselves, independent of language or culture, as illustrated in Figure 1. The method aims to control over subtle acoustic nuances that are difficult to capture using symbolic representations. The type of vocal mimicry considered in

this work allows intuitive expression and conveys prosodic and phonetic information that cannot be represented through linguistic symbols such as phoneme sequences and text prompts.

Since the proposed method enables immediate synthesis of sound effects from vocal mimicry, it can be incorporated into a human-in-the-loop creative workflow, as illustrated in Figure 2. In this workflow, users can iteratively input vocal mimicry and refine the output to approach the desired sound. The method will enable collaborative production of sound effects through interaction between humans and computers. This notion of human-in-the-loop differs from that in Wang *et al.* (2024), where human subjective judgments are incorporated into the optimization loop, whereas our setting emphasizes iterative creative interaction through repeated vocal mimicry and listening.

3.2 Model architecture

Figure 3 illustrates the overview of PronounSE. The inference process of PronounSE consists of the following three steps: a preprocessing step that converts the vocal mimicry waveform into a Mel-spectrogram, a conversion step from the Mel-spectrogram of the vocal mimicry into one of a target sound effect using a Transformer (Vaswani *et al.*, 2017), and a generation step using HiFi-GAN (Kong *et al.*, 2020a), which synthesizes a waveform from the converted Mel-spectrogram of the sound effect.

During training, we use paired Mel-spectrograms of vocal mimicry and the corresponding sound effects. Let T denote the number of time frames and N the number of Mel-frequency bins. Linear-layer-based Prenets first map a Mel-spectrogram of vocal mimicry $\mathbf{X} \in \mathbb{R}^{T \times N}$ and that of sound effects $\mathbf{Y} \in \mathbb{R}^{T \times N}$ to high-dimensional frequency embeddings $\mathbf{E}^{(x)}$, $\mathbf{E}^{(y)} \in \mathbb{R}^{T \times H}$ as:

$$\mathbf{E}^{(x)} = \mathrm{Prenet}_x(\mathbf{X}),\ \mathbf{E}^{(y)} = \mathrm{Prenet}_y(\mathbf{Y}), \tag{1}$$

where H denotes the hidden dimensionality, and the Prenets project the input Mel-spectrograms into higher-dimensional representations that facilitate subsequent sequence modeling by the Transformer. The embeddings $\mathbf{E}^{(x)}$ and $\mathbf{E}^{(y)}$ are then provided to a three-layer Transformer encoder and a three-layer Transformer decoder, respectively. This encoder-decoder structure is used to flexibly model temporal and spectral correspondences between vocal mimicry and the target sound effect. The decoder attends to the encoder

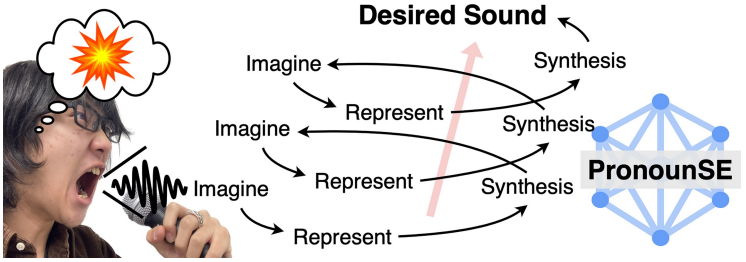

Figure 2. Sound creation flow with PronounSE. This enables a human-in-the-loop sound design workflow, in which users iteratively perform mimicking and synthesizing audio to gradually approach the desired sound

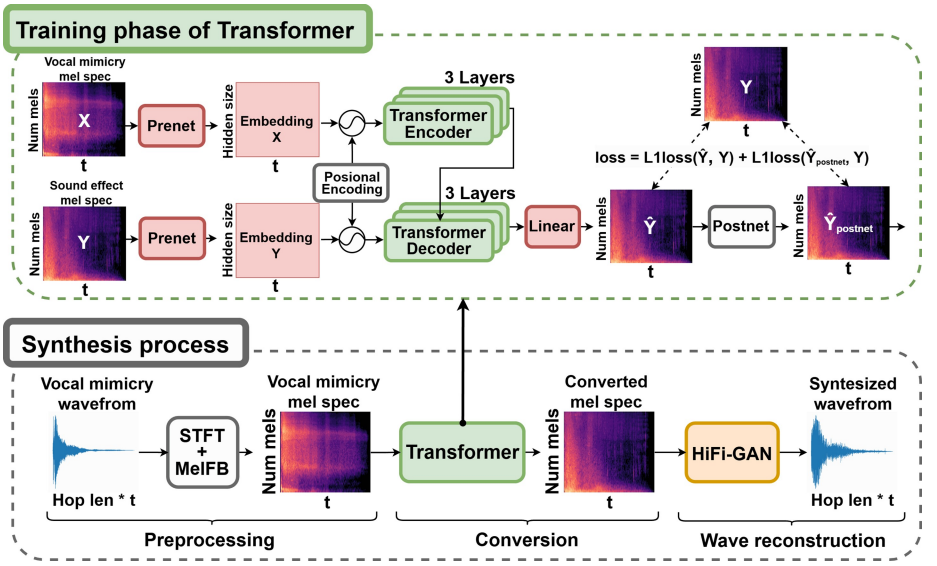

Figure 3. Architecture of PronounSE. This figure illustrates the inference-time synthesis flow of PronounSE and the training process of Transformer, which is trained to map Mel-spectrograms of vocal mimicry to those of corresponding sound effects

output and its own input embeddings. The resulting decoder output is passed through a feed-forward network that projects the features back to the original Mel-spectrogram dimension:

$$\hat{\mathbf{Y}} = \mathrm{FeedForward}\big(\mathrm{Decoder}\big(\mathbf{E}^{(y)}, \mathrm{Encoder}\big(\mathbf{E}^{(x)}\big)\big)\big), \tag{2}$$

where $\hat{\mathbf{Y}} \in \mathbb{R}^{T \times N}$ represents the predicted Mel-spectrogram features. Finally, a Postnet based on a residual correction network refines the predicted features to produce the final Mel-spectrogram $\hat{\mathbf{Y}}_{\mathrm{postnet}} \in \mathbb{R}^{T \times N}$:

$$\hat{\mathbf{Y}}_{\mathrm{postnet}} = \mathrm{Postnet}\big(\hat{\mathbf{Y}}\big). \tag{3}$$

The Postnet is introduced to refine coarse Mel-spectrogram predictions by modeling residual details, following common practice in neural speech synthesis (Shen *et al.*, 2018; Li *et al.*, 2019).

The training objective $\mathcal{L}$ is defined as a weighted sum of $\updownarrow_1$ losses applied to both the inputs and outputs of the Postnet:

$$\mathcal{L} = \alpha \|\hat{\mathbf{Y}} - \mathbf{Y}\|_1 + \beta \|\hat{\mathbf{Y}}_{\mathrm{postnet}} - \mathbf{Y}\|_1. \tag{4}$$

In all experiments shown in this paper, both α and β are set to 1. Through this training procedure, the Transformer learns to model the correspondence between acoustic features of vocal mimicry and those of sound effects. We adopt this simple $\updownarrow_1$-based objective for stable reconstruction learning; its limitations and the effect of alternative loss designs are discussed in Section 7.

During inference, the Mel-spectrogram of vocal mimicry is fed into the trained Transformer, and the resulting Mel-spectrogram is converted into a waveform using HiFi-GAN. During training, the decoder is conditioned on the ground-truth target sequence via teacher forcing, whereas during inference it uses its own previous predictions in an autoregressive manner. Specifically, the decoder generates the output Mel-spectrogram sequentially, one frame at a time, by feeding back previously predicted frames to produce the next frame. This train-inference mismatch may introduce exposure bias, which we discuss further in Section 7.

The proposed method directly learns the correspondence between the Mel-spectrograms of vocal mimicry and sound effects using paired data without relying on any prompt or acoustic event label, which differs from the existing methods described in Section 2.1.

4. Data set construction for vocal mimicry and sound effects

4.1 Required data set for this study

To train PronounSE, paired data consisting of sound effects and their corresponding vocal mimicry is required. However, as described in Section 2.2, the reference sounds in existing data sets (Okamoto *et al.*, 2024b; Cartwright and Pardo, 2015a; Kim and Pardo, 2018) are often collected from online videos or community-driven sound-sharing platforms, and many of them contain background noise or reverberation, or are recorded at low sampling rates. Although these data sets include multiple categories of environmental sounds, the number of reference samples per category is limited, and the data sets primarily focus on everyday environmental sounds. Moreover, the corresponding vocal mimicry samples are frequently collected through crowdsourcing, resulting in low-quality recordings that are often constrained by specific linguistic systems.

The goal of this study is to support the production of sound in media content, targeting sound effects commonly used in science fiction and other non-everyday contexts. High-quality sound effect synthesis is pursued by constructing a new data set that contains high-quality reference sounds and language-independent vocal mimicry capable of capturing rich acoustic nuances, rather than relying on existing data sets.

4.2 Our newly constructed data set

Since this study aims to control timbre in a language-independent manner, we constructed a data set focused on a single sound category, especially explosion sounds, as the target. Explosion sounds, while based on real acoustic events, are difficult to record or produce in reality and are in high demand in audiovisual content. Moreover, they exhibit a wide range of variations and are relatively easy for humans to imitate vocally, making them well-suited as the subject of this study. Accordingly, in this study, explosion sounds are treated as a case study for examining the feasibility of sound effect synthesis from language-independent vocal mimicry.

For the training data set, we collected 1,748 types of explosion sounds from free and commercial sound libraries. These include a variety of sounds such as stylized anime-like explosions, gunfire, dynamite blasts and destruction sounds. All collected explosion sounds are up to 12 s in length. We recorded 8,117 vocal mimicry samples from three male speakers in their 20s (4,016 samples, 2,851 samples and 1,250 samples, respectively), resulting in a data set of 8,117 explosion-mimicry pairs.

For the evaluation data set, we collected 100 explosion sounds from a separate commercial sound library [10]. Then, 600 vocal mimicry samples were recorded by six speakers (both male and female) in their 20s, identified as m-01 to m-05 and f-01, where the

first characters indicate their gender. These speakers were not included in the training data set, and each speaker mimicked all 100 explosion sounds. The training data set is speaker-imbalanced, which may bias the model toward speaker-specific vocal patterns. To mitigate trivial memorization in evaluation, we use six speakers who were not included in the training data set; however, the imbalance in the training data remains a limitation of the present study. To support reproducibility, the evaluation data set is publicly available on Zenodo [11].

All recordings were conducted in the same soundproof room using different devices: a ZOOM H4n [12] for the training data set, recorded at 44100 Hz, and a ZOOM *H4* Essential [13] for the evaluation data set, recorded at 48000 Hz.

This study involved human participants and was approved by the ethics review board of Kyoto Sangyo University (Approval No. 0250). Informed consent was obtained from all participants involved in the vocal mimicry recordings.

5. Experimental setup for comparative evaluation of audio-to-audio methods

5.1 Designing evaluation metrics for audio-to-audio

In the field of speech synthesis, as a subjective evaluation metric, mean opinion scores (MOS), which are average ratings typically of a five- or seven-point Likert scale, are widely used to assess the naturalness and intelligibility of synthesized speech. In addition, automatic MOS prediction methods have also been proposed (Patton *et al.*, 2016; Saeki and Takamichi, 2024). In the field of environmental sound synthesis, some studies (Okamoto *et al.*, 2021, 2024a; Chung *et al.*, 2024) have adopted MOS evaluations to assess audio quality and naturalness as well. However, this study focuses on sound effects, which include both fictitious sounds that do not occur in the physical world and stylized variations of real environmental sounds. The explosion sounds used in the training of PronounSE also include anime-style, exaggerated explosions and artificially created sound effects. Therefore, it is not necessarily appropriate to evaluate the synthesized sounds solely in terms of naturalness or audio quality. Accordingly, we propose an alternative evaluation approach based on a different perspective.

This study adopts the view that, in audio-to-audio synthesis techniques, including PronounSE, it is critical whether the synthesized sound both accurately reproduces the intended reference sound and maintains plausibility based on the input features. Therefore, in addition to evaluating the naturalness of the synthesized sound, we introduce two new evaluation perspectives: (1) fidelity to the target sound and (2) reflection of vocal mimicry nuances. For fidelity to the target sound, we conduct both objective and subjective evaluations. The objective evaluation adopts Fréchet Audio Distance (FAD) (Kilgour *et al.*, 2019) and cosine similarity based on PANNs (Kong *et al.*, 2020b) embeddings. Notably, the FAD metric based on PANNs was also employed in the environmental sound synthesis tasks of the DCASE Challenge (Choi *et al.*, 2001; Lagrange *et al.*, 2025). In addition, a subjective evaluation is conducted using the MOS of perceived fidelity (*Fidelity MOS*). For reflection of vocal mimicry nuances, we focus on the key characteristics that define the impression of explosion sounds: the attack (onset) and release (decay) components. We evaluate (1) the plausibility of the attack characteristics in the synthesized sound compared to the vocal mimicry (*Attack MOS*), (2) the plausibility of the release characteristics (*Release MOS*), and (3) the overall impression (*Overall MOS*). In addition to these, we also conduct subjective evaluations of the naturalness (*Naturalness MOS*) and audio quality (*Quality MOS*) of the synthesized sounds as perceived by human listeners.

For the subjective evaluations, we selected 10 reference sounds from the 100 explosion sounds in the evaluation data set so that different explosion categories were represented. The listening tests targeted general crowd workers rather than participants with specific musical

or audio-related expertise. No additional worker screening or post hoc score filtering was applied, and all submitted ratings were included in the analysis. No explicit headphone check was conducted. Instead, workers were instructed to perform the listening task in as quiet an environment as possible and to use headphones or earphones, although actual compliance was not verified.

5.2 Synthesis methods

Using the 600 vocal mimicry samples in the evaluation data set as input, explosion sounds were synthesized using three methods: PronounSE (Ours), T-Foley (TF) (Chung *et al*., 2024) and Stable Audio 2.0 (SA2) (Evans *et al*., 2024). We did not include Sketch2Sound (García *et al*., 2025) in the comparative experiment because, to the best of our knowledge, no open-source implementation is publicly available, which prevented reproducible evaluation under our setting. PronounSE is configured using a Transformer with 256 Mel bands and an embedding dimension of 512, trained on the constructed training data set for 20,000 epochs with a batch size of 64, together with a HiFi-GAN model trained on explosion sounds within the same data set. The implementation of PronounSE and the pretrained models used in this experiment are publicly available on GitHub [14].

TF is based on the official implementation [15] and trained for 5,000 epochs with a batch size of 8 using 1,748 explosion sounds from the PronounSE training data set. Both PronounSE and TF are trained using audio sampled at 22050 Hz. SA is evaluated using the synthesis engine provided on the official website [16]. For SA2, the text prompt was "Real Explosion Sound," and all input intensities were set to the default value (75%). The resulting audio was downsampled to 22050 Hz to match the sampling rate of the synthesized sounds of PronounSE and TF.

For reproducibility, we report the hardware configuration and inference efficiency of PronounSE. Real-time factor (RTF) was measured using 100 vocal mimicry samples on a single NVIDIA *H200* GPU. The average inference time was 1.06 s (SD: 0.54 s), and the average RTF was 0.317 (SD: 0.033), indicating that PronounSE can generate audio at about 3.2× faster than real time. In this measurement, the inference time included Mel-spectrogram generation by the Transformer and waveform generation by HiFi-GAN, while Mel-spectrogram conversion of the input audio, file I/O and visualization was excluded.

6. Evaluation results of audio-to-audio

6.1 Qualitative evaluation of synthesized sounds

Figure 4 shows representative Mel-spectrograms of three types of audio: the reference sounds, six vocal mimicry samples from six speakers and the synthesized outputs from the three methods [17].

TF often produces synthesized sounds that decay immediately after the attack, failing to reflect intensity nuances. In addition, outputs for m-02, m-03, m-05 and f-01 contain noise below 100 Hz, with m-02 showing noise across the entire frequency range, resulting in low sound quality. First, one possible reason why intensity nuances are not properly reflected is that the RMS envelope of the explosion sound differs substantially from the RMS envelope of the vocal mimicry. As a result, the explosion-like dynamics may not be accurately reproduced. In addition, in regions where the vocal mimicry has low amplitude, the corresponding RMS envelope is likely to cause the synthesized sound to decay too rapidly.

In the results of PronounSE, the timbre of the synthesized attack subtly changes in accordance with the consonant and vowel characteristics of the vocal mimicry's attack phase. On the other hand, harmonic artifacts (appearing as horizontal stripe noise) that are not characteristic of explosion sounds are observed during the decay phase in the frequency

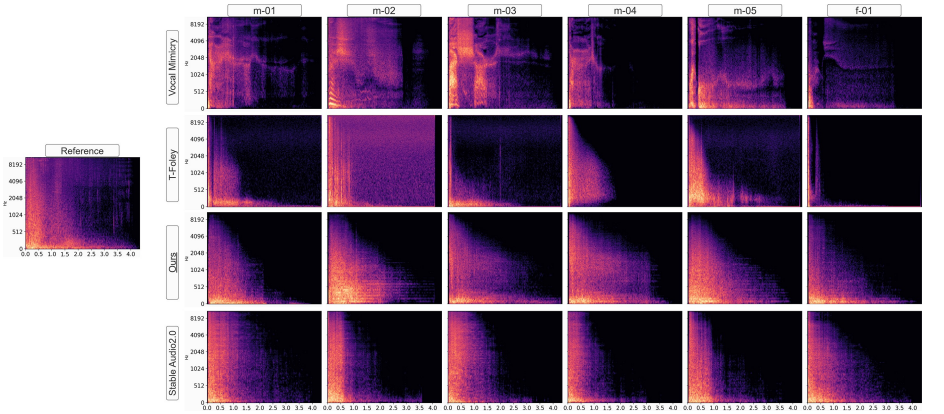

Figure 4. Examples of the reference sound, vocal mimicry from six speakers and sounds synthesized using PronounSE, T-Foley (Chung *et al.*, 2024) and Stable Audio 2.0 (Evans *et al.*, 2024). The synthesized results are shown in the order of T-Foley, PronounSE and Stable Audio 2.0

range below 2 kHz for the results of m-02 and m-05. The possible causes of these artifacts are discussed further in Section 7, including the train-inference mismatch and the limitations of the current loss design.

Similar to PronounSE, the results of SA2 exhibit synthesis that reflects the consonants, vowels and intensity of the vocal mimicry. However, the overall sound tends to resemble destructive sounds, such as glass breaking, rather than having the timbre typically associated with explosions, resulting in a different sound character from PronounSE.

6.2 Quantitative evaluation of synthesized sounds

In this section, we present the quantitative evaluation results of the synthesized sounds, including both objective and subjective evaluations. We did not compute an explicit inter-rater agreement statistic in the present study. Instead, to provide a qualitative view of rating consistency and variability across listeners, we report the score distributions for each MOS as stacked bar charts in the following subsections.

6.2.1 Fidelity to the target sound. To calculate the FAD scores, we used the 100 reference explosion sounds from the evaluation data set as the evaluation set, and the 600 synthesized samples generated by each model as the test set. This reference set corresponds to all explosion sounds in our evaluation data set. Although this setup enables consistent comparison within the present benchmark, we acknowledge that a larger reference set could provide more stable distributional estimation for FAD. The cosine similarity scores are calculated between the embedding vectors of each reference explosion sound and its corresponding synthesized sounds. The mean and standard deviation were then computed for each model.

To evaluate the fidelity to the target sound subjectively, a listening test was conducted using ten explosion sounds from the evaluation data set. A total of 350 crowd workers participated in the experiment, each being assigned one of the ten explosion sounds. They evaluated the similarity between the assigned explosion sound and the corresponding synthesized sounds (18 combinations of six vocal mimicry samples and three synthesis

methods) on a five-point scale: 1 (completely different), 2 (different), 3 (neutral), 4 (similar) and 5 (very similar). Based on these evaluations, the Fidelity MOS was calculated.

Table 1 presents the results of the objective evaluation based on FAD and cosine similarity, and Figure 5 and Table 2 show the results of the subjective evaluation of the fidelity to the target sound. As shown in Table 1, PronounSE achieved lower FAD scores and higher cosine similarity than the other methods. In the subjective fidelity evaluation results shown in Figure 5 and Table 2, SA2 achieved the highest Fidelity MOS. However, based on the results of the three-group test (Friedman test) and post hoc analysis (Nemenyi test), no significant differences were observed between SA2 and PronounSE, except for the cases of m-02, m-05, and the whole set. On the other hand, the Fidelity MOS of TF was lower than that of PronounSE and SA2, with a higher proportion of scores rated as 1 (red), indicating lower fidelity. Therefore, in terms of fidelity, PronounSE demonstrated superior performance in objective evaluation and achieved performance comparable to SA2 in subjective evaluation.

To investigate the relation between the objective and subjective scores, we calculated the correlation coefficients between the objective evaluation metrics (FAD scores and cosine similarity) and the subjective Fidelity MOS. Table 3 shows the results. A significant negative correlation was observed between Fidelity MOS and FAD, with a Spearman correlation coefficient of $-0.573\,(p = 0.0129)$ and a Pearson correlation coefficient of $-0.937\,\left(p = 1.08 \times 10^{-8}\right)$, indicating that lower FAD scores tend to correspond to higher subjective fidelity ratings. Similarly, a positive correlation was observed between Fidelity MOS and cosine similarity, with a Spearman correlation coefficient of $0.475\,(p = 0.0465)$ and a Pearson correlation coefficient of $0.898\,\left(p = 4.27 \times 10^{-7}\right)$, suggesting that higher cosine similarity tends to be associated with higher subjective fidelity ratings. From these observation, FAD and cosine similarity are considered sufficient to adequately evaluate the fidelity of the synthesized sounds to the reference sounds.

6.2.2 Reflection of vocal mimicry nuances. Similar to the subjective evaluation of fidelity, a listening test was conducted via crowdsourcing with 350 participants. Each participant was assigned one of the ten explosion sounds and, after listening to the corresponding vocal mimicry samples from six speakers and the synthesized sounds, evaluated the reflection of nuances in the synthesized sounds in terms of three aspects: attack, release and overall impression. The evaluation was performed on a five-point scale: 1 (not reflected at all), 2 (not reflected), 3 (neutral), 4 (reflected) and 5 (strongly reflected). Based on these ratings, the Attack MOS, Release MOS and Overall MOS were computed.

Table 4 shows the subjective evaluation results for nuance reflection in the synthesized sounds with respect to the attack, release and overall impression of the vocal mimicry. For Attack MOS, PronounSE achieved the highest MOS only for speakers m-01 and m-03, while SA2 obtained the highest MOS for the other speakers. For Release MOS and Overall MOS, PronounSE achieved the highest MOS for all speakers. In addition, PronounSE outperformed TF in all conditions.

Figure 6 shows the score distributions for each MOS. Across all three MOS measures, the bars corresponding to scores 1–3 (red, orange and yellow) for PronounSE and SA2 are mostly shorter than those for TF. In addition, for Release MOS and Overall MOS, PronounSE has fewer responses with score 1–3 than both TF and SA2.

The results suggest that, compared with TF, PronounSE can better reflect the attack and release nuances of the vocal mimicry in the synthesized sounds. Although SA2 generally achieved higher Attack MOS than PronounSE, PronounSE obtained the highest Release MOS and Overall MOS, indicating that PronounSE surpasses SA2 in terms of reflection of vocal mimicry nuances. On the other hand, despite PronounSE achieving nuance reflection

Table 1. FAD scores and average cosine similarities based on PANNs embeddings between the set of reference explosion sounds and the synthesized sound set from PronounSE, T-Foley and Stable Audio 2.0. The ± indicates the standard deviation

	FAD scores by PANNs ↓			Cosine similarity ± SD ↑		
Speaker ID	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)
m-01	54.02	*17.85*	21.35	0.73 ± 0.07	*0.85 ± 0.08*	0.83 ± 0.06
m-02	54.47	*17.05*	20.94	0.74 ± 0.07	*0.86 ± 0.06*	0.84 ± 0.06
m-03	48.85	*17.28*	20.95	0.75 ± 0.07	*0.86 ± 0.07*	0.84 ± 0.06
m-04	48.52	*16.65*	21.46	0.75 ± 0.08	*0.85 ± 0.07*	0.84 ± 0.06
m-05	47.89	*19.29*	19.75	0.75 ± 0.08	*0.86 ± 0.07*	0.85 ± 0.07
f-01	56.94	*19.28*	22.01	0.72 ± 0.07	*0.85 ± 0.07*	*0.85 ± 0.06*
Whole	55.93	*14.09*	21.15	0.74 ± 0.07	*0.85 ± 0.07*	0.84 ± 0.06

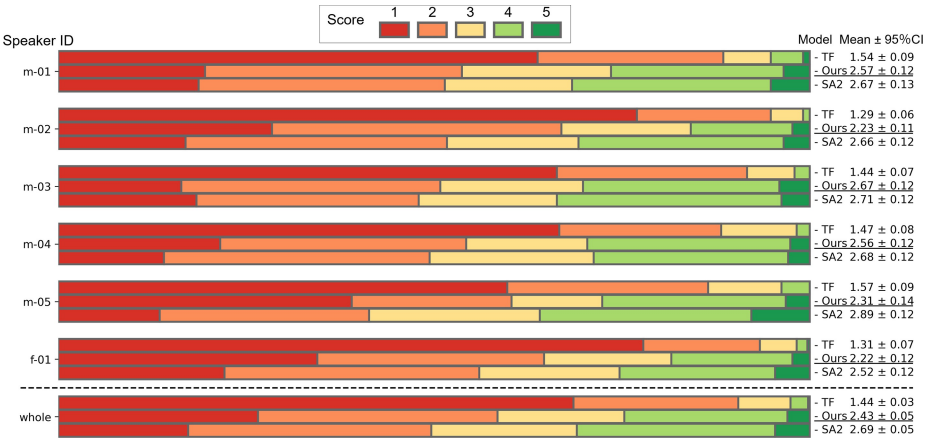

Figure 5. Score distributions in subjective evaluations of fidelity. For each speaker, the frequency distribution of similarity scores between the reference sounds and the synthesized sounds from each model is shown as a stacked bar chart. The corresponding Fidelity MOS (mean ± 95% confidence interval) is displayed next to each bar

Table 2. Fidelity MOS ↑ for each speaker. The ± indicates the 95% confidence intervals for the scores

Speaker ID	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)
m-01	1.54 ± 0.09	2.57 ± 0.12	*2.67 ± 0.13*
m-02	1.29 ± 0.06	2.23 ± 0.11	*2.66 ± 0.12*
m-03	1.44 ± 0.07	2.67 ± 0.12	*2.71 ± 0.12*
m-04	1.47 ± 0.08	2.56 ± 0.12	*2.68 ± 0.12*
m-05	1.57 ± 0.09	2.31 ± 0.14	*2.89 ± 0.12*
f-01	1.31 ± 0.07	2.22 ± 0.12	*2.52 ± 0.12*
Whole	1.44 ± 0.03	2.43 ± 0.05	*2.69 ± 0.05*

Table 3. Correlation coefficients between Fidelity MOS and FAD scores / mean cosine similarity (with *p*-values)

Type of correlation	FAD scores	Cosine similarity
Spearman	$-0.573\ (p = 0.0129)$	$0.475\ (p = 0.0465)$
Pearson	$-0.937\ (p = 1.08 \times 10^{-8})$	$0.898\ (p = 4.27 \times 10^{-7})$

MOS around 3 overall, its Fidelity MOS were around 2.5 and slightly lower than those of SA2, suggesting that fidelity to the target sound depends on how accurately speakers can mimic the reference sounds.

6.2.3 Naturalness and audio quality of the synthesized sounds. Subjective evaluations of the naturalness and sound quality of the synthesized sounds were also conducted via crowdsourcing with 350 participants. Each participant was randomly assigned one of the ten

Table 4. Attack MOS, release MOS and overall MOS for each speaker. The ± indicates the 95% confidence intervals for the scores

	Attack MOS ↑			Release MOS ↑			Overall MOS ↑		
Speaker ID	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)
m-01	2.95 ± 0.11	*3.13 ± 0.12*	3.08 ± 0.12	2.87 ± 0.11	*3.03 ± 0.12*	2.98 ± 0.12	2.80 ± 0.11	*3.04 ± 0.12*	2.94 ± 0.12
m-02	2.65 ± 0.12	2.87 ± 0.13	*2.93 ± 0.12*	2.68 ± 0.12	*3.05 ± 0.12*	2.93 ± 0.12	2.45 ± 0.12	*2.92 ± 0.12*	2.85 ± 0.11
m-03	2.55 ± 0.12	*2.92 ± 0.12*	2.90 ± 0.12	2.58 ± 0.12	*2.86 ± 0.12*	2.80 ± 0.11	2.50 ± 0.11	*2.81 ± 0.11*	2.76 ± 0.11
m-04	2.76 ± 0.12	3.03 ± 0.12	*3.07 ± 0.12*	2.81 ± 0.12	*3.12 ± 0.11*	2.93 ± 0.11	2.71 ± 0.12	*3.03 ± 0.11*	2.95 ± 0.12
m-05	2.64 ± 0.12	2.74 ± 0.13	*2.78 ± 0.11*	2.63 ± 0.11	*2.83 ± 0.12*	2.75 ± 0.11	2.61 ± 0.11	*2.69 ± 0.12*	2.63 ± 0.10
f-01	2.84 ± 0.12	2.85 ± 0.12	*2.86 ± 0.13*	2.76 ± 0.12	*2.92 ± 0.12*	2.71 ± 0.12	2.79 ± 0.12	*2.81 ± 0.11*	2.68 ± 0.11
Whole	2.73 ± 0.05	2.92 ± 0.05	*2.94 ± 0.05*	2.72 ± 0.05	*2.97 ± 0.05*	2.85 ± 0.05	2.64 ± 0.05	*2.88 ± 0.05*	2.80 ± 0.05

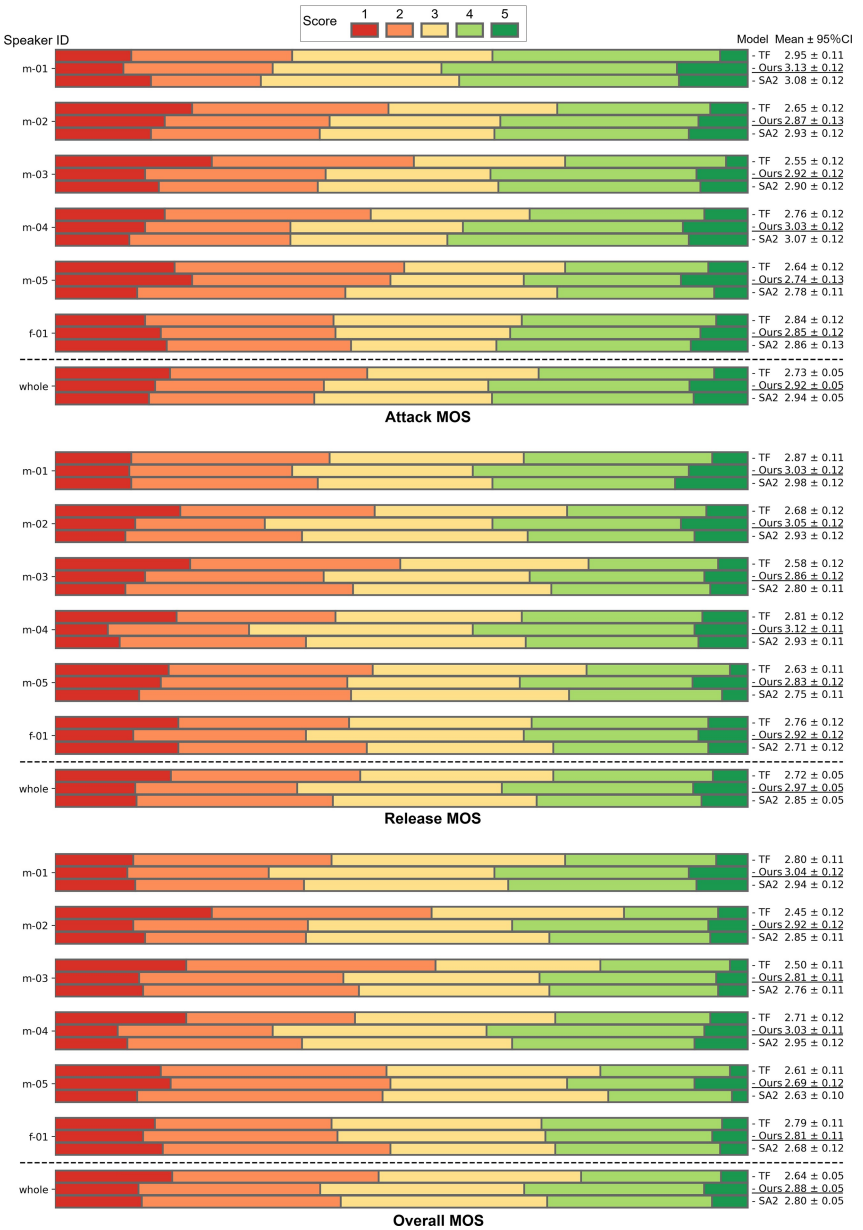

Figure 6. Score distributions in subjective evaluations of nuance reflection for attack, release and overall impression (from top to bottom). For each speaker, the frequency distribution of nuance reflection scores for the synthesized sounds is shown as a stacked bar chart. The corresponding MOS (mean ± 95% confidence interval) is displayed next to each bar

explosion sounds and asked to listen to both the assigned explosion sound (ground truth) and the corresponding synthesized sounds. They then rated the naturalness and quality of the synthesized sounds as explosion sounds on a five-point scale (Naturalness: 1. very unnatural, 2. unnatural, 3. neutral, 4. natural, 5. very natural; Quality: 1. very poor, 2. poor, 3. neutral, 4. good, 5. very good). Based on these ratings, the Naturalness MOS and Quality MOS for each method were calculated.

Table 5 shows the subjective evaluation results for the naturalness and audio quality of the synthesized sounds. For naturalness, PronounSE achieved the highest MOS only for speaker m-01, while SA2 obtained the highest MOS for all other speakers. In terms of audio quality, SA2 achieved the highest MOS across all speakers.

Figure 7 shows the score distributions for each MOS. Except for speaker m-01, the bars corresponding to scores 1–3 (red, orange and yellow) become shorter in the order TF, PronounSE and SA2, indicating that PronounSE and SA2 yield substantially fewer low scores than TF.

Taken together, the two MOS measures and their score distributions suggest that SA2 performs best in terms of naturalness as an explosion sound and overall audio quality. For PronounSE, as described in Section 6.1, harmonic noise artifacts are frequently observed in the synthesized sounds; this likely increases the number of low ratings (scores 1 and 2, shown in red and orange) for naturalness and audio quality compared with SA2, resulting in lower MOS. For TF, the frequent occurrence of low-frequency noise below 100 Hz and broadband noise, as discussed in Section 6.1, is likewise considered the main reason for its low MOS in naturalness and audio quality.

6.3 Summary of evaluation results

In all subjective evaluations, PronounSE achieved higher MOS than TF, and the subjective results for reflection of vocal mimicry nuances showed that PronounSE outperformed SA2 in this aspect. These findings indicate that, for nuance reflection, a method that learns correspondences between spectral feature representations of vocal mimicry and sound effects is effective. On the other hand, subjective evaluations revealed that, for PronounSE, the naturalness and audio quality are degraded by noise with a harmonic structure that does not correspond to the acoustic characteristics of explosion sounds. This is considered to be caused by the limited amount of audio data available to train the Transformer and the neural vocoder.

In the fidelity evaluation for explosion sounds, clear correlations were observed between objective and subjective scores, demonstrating the effectiveness of automatic fidelity assessment using FAD and cosine similarity.

In the subjective evaluation of fidelity, PronounSE achieved slightly lower MOS than SA2, whereas in the subjective evaluation of nuance reflection, PronounSE obtained the highest MOS. These results suggest that, even when the synthesized sound reflects the nuances of the vocal mimicry well, high fidelity to the reference sound is not always guaranteed. This is likely because fidelity depends not only on the model itself but also on how accurately the speaker can vocally mimic the reference sound. Therefore, evaluating only fidelity in vocal-mimicry-based audio synthesis does not adequately capture the controllability of the model. Consequently, this study highlights the need to evaluate nuance reflection and fidelity separately.

7. Discussion

In the subjective evaluations, PronounSE consistently outperformed T-Foley and showed better performance than Stable Audio 2.0 in terms of reflection of vocal mimicry nuances. On the other hand, cases were observed in which harmonic noise artifacts inconsistent with

Table 5. Naturalness MOS and quality MOS for each speaker. The ± indicates the 95% confidence intervals for the scores

		Naturalness MOS ↑			Quality MOS ↑	
Ground truth		3.70 ± 0.11			3.70 ± 0.10	
Speaker ID	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)	TF (Chung *et al.*, 2024)	Ours	SA2 (Evans *et al.*, 2024)
m-01	2.86 ± 0.12	*3.50 ± 0.11*	3.29 ± 0.11	2.73 ± 0.12	3.36 ± 0.11	*3.36 ± 0.10*
m-02	2.56 ± 0.11	3.09 ± 0.12	*3.42 ± 0.10*	2.47 ± 0.11	3.16 ± 0.11	*3.45 ± 0.11*
m-03	2.82 ± 0.12	3.46 ± 0.11	*3.48 ± 0.10*	2.55 ± 0.11	3.32 ± 0.10	*3.52 ± 0.10*
m-04	2.95 ± 0.12	3.55 ± 0.11	*3.57 ± 0.10*	2.83 ± 0.12	3.39 ± 0.10	*3.48 ± 0.10*
m-05	2.69 ± 0.11	2.93 ± 0.12	*3.72 ± 0.10*	2.49 ± 0.11	2.95 ± 0.11	*3.57 ± 0.10*
f-01	2.76 ± 0.12	3.14 ± 0.12	*3.53 ± 0.11*	2.69 ± 0.11	2.99 ± 0.11	*3.45 ± 0.10*
Whole	2.77 ± 0.05	3.28 ± 0.05	*3.50 ± 0.04*	2.62 ± 0.05	3.19 ± 0.04	*3.47 ± 0.04*

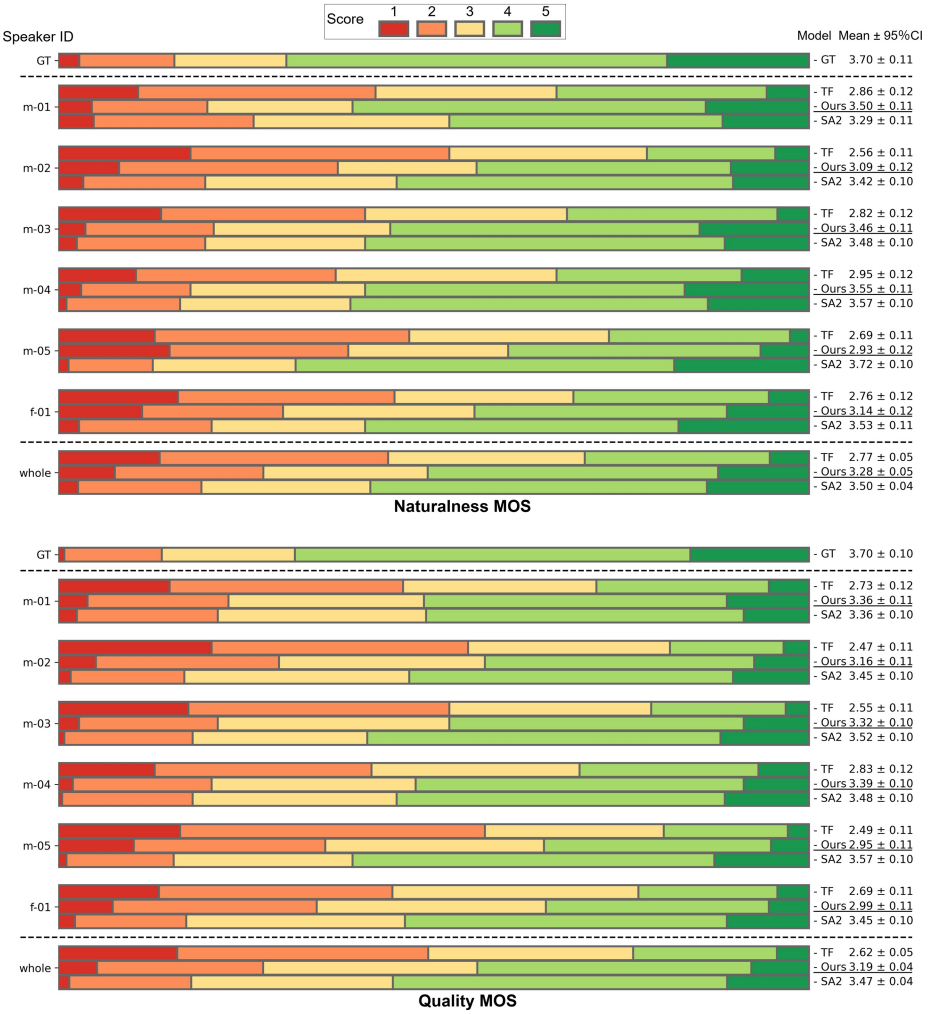

Figure 7. Score distributions in subjective evaluations on the naturalness and sound quality of the synthesized explosion sounds (top: Naturalness MOS, bottom: Quality MOS). For m-01, the subjective evaluation results for Ground Truth are shown above. For each speaker, the frequency distribution of subjective scores is presented as a stacked bar chart. The corresponding MOS (mean ± 95% confidence interval) is displayed next to each bar

the acoustic characteristics of explosion sounds were generated by PronounSE, degrading the naturalness and audio quality of the synthesized sounds. In the fidelity evaluation, clear correlations were found between objective measures (FAD and cosine similarity) and subjective ratings, supporting the validity of these metrics for automatic fidelity assessment.

Trade-off between fidelity and nuance reflection: Stable Audio 2.0 achieved higher fidelity to the target sound than PronounSE, whereas PronounSE received higher scores for

reflection of vocal mimicry nuances. This contrast suggests that agreement with the reference sound and following the intent encoded in the input cannot always be maximized at the same time. PronounSE readily transfers the temporal structure and timbral variations of the vocal mimicry to the output, but the degree of agreement with the reference sound largely depends on how accurately each speaker can mimic it. In other words, even when the synthesized sound reflects the nuances of the vocal mimicry well, high fidelity to the reference sound is not always guaranteed. This is likely because reproducing the reference sound itself depends strongly on how faithfully the input vocal mimicry captures the reference. Overall, these results indicate that, in vocal-mimicry-based audio synthesis, fidelity and nuance reflection should be evaluated as separate criteria.

Speaker diversity and mimicry-skill dependence: A related limitation is the restricted speaker diversity of the current training data set. The training data consist of three male speakers, and the numbers of vocal mimicry samples are not balanced across speakers. In practice, vocal mimicry styles can vary substantially across individuals, for example, in fundamental-frequency characteristics and articulation style, as well as in whether the imitation is more percussive or more speech-like. Such variability may affect how well the model generalizes across speakers and may partly increase its dependence on the speaker's mimicry skill. Constructing a data set that captures broader speaker diversity will therefore be an important direction for reducing this dependency in future work.

Harmonic noise artifacts in PronounSE: The synthesized explosions sometimes contained pitched, harmonic noise that did not match typical explosion characteristics, which degraded perceived naturalness and audio quality. In a diagnostic experiment, ground-truth explosion Mel-spectrograms resynthesized with the trained HiFi-GAN did not exhibit such noise. This suggests that the artifacts mainly stem from the Transformer, likely due to over-generation of periodic structure in the predicted Mel-spectrograms rather than from the vocoder itself. A plausible cause is the limited size and diversity of the explosion training data, which may underrepresent inharmonic noise events and bias the model toward harmonic patterns. Another possible factor is the mismatch between teacher-forced training and autoregressive inference, which may introduce exposure bias and accumulate prediction errors during sequential generation. In addition, the current simple $\updownarrow_1$-based objective may be insufficient to discourage undesired harmonic structure in the predicted Mel-spectrograms. Possible remedies therefore include enlarging and diversifying the training set, refining loss functions and regularization to better handle inharmonic components, mitigating the train-inference mismatch, and introducing quantitative checks on generated Mel-spectrograms, such as harmonicity measures or spectral-slope statistics, to monitor and constrain their quality.

Relationship between objective metrics and subjective evaluations: Correlations observed between subjective ratings and both FAD and cosine similarity in the fidelity evaluation suggest that these objective metrics are reasonably valid as automatic measures of similarity to the reference sound. To use subjective evaluations more efficiently, a key future task is to design objective metrics that more directly capture nuance reflection. Possible candidates include measures of agreement between the temporal (energy) envelopes of the vocal mimicry and the synthesized sound (e.g. correlation or Dynamic Time Warping distance), and measures of how closely the Mel-spectrogram patterns of the two signals match. Similarity in an acoustic embedding space between vocal mimicry and synthesized sounds is another promising option. Introducing such metrics would extend the current framework from automatic evaluation of fidelity to a scheme that also quantifies how well input nuances are reflected in the synthesized sounds.

A further limitation of this study is that our experiments are restricted to explosion sounds. Thus, the present work should be regarded as a case study on sound effect synthesis

from language-independent vocal mimicry, and the applicability of PronounSE to other sound effect categories remains an important topic for future work.

8. Conclusion

This study proposed PronounSE, a sound-effect synthesis method that generates sound effects from vocal mimicry imitating a reference sound and can reflect nuances contained in the input. Focusing on explosion sounds, we construct a training data set for PronounSE and prepare an evaluation data set suitable for both subjective and objective evaluations, thereby establishing a reproducible evaluation environment. Subjective evaluation results showed that PronounSE consistently outperformed T-Foley and achieved higher ratings than Stable Audio 2.0 in terms of reflection of vocal mimicry nuances, while Stable Audio 2.0 yielded superior scores for fidelity to the target sound (agreement between reference and synthesized sounds), suggesting a potential trade-off between fidelity and nuance reflection. In addition, strong correlations were observed between objective measures (FAD and cosine similarity) and subjective ratings in the fidelity evaluation, supporting the validity of automatic fidelity assessment based on these metrics.

Future work includes conducting user studies with sound designers to examine the practical usefulness of PronounSE in real production workflows. Beyond the quality, fidelity and nuance reflection of synthesized sounds, such studies should consider additional axes, such as usability, time, the number of iterations required to achieve intended nuances and the impact on the creative process. Production-oriented tasks combined with qualitative feedback (e.g. open-ended comments and interviews) will be important for clarifying PronounSE's practical utility and for identifying design issues, such as the difficulty of vocal input or desired forms of system feedback.

Accordingly, the present work should be regarded as a case study on explosion sounds rather than a full validation across sound-effect categories. Nevertheless, it is essential to verify the applicability of PronounSE to other sound-effect categories (e.g. impacts, footsteps, machinery and scraping sounds) to clarify its practical scope. In addition, there remains a large room to improve synthesis quality, such as reducing harmonic noise, by expanding and diversifying the training data and refining loss functions and regularization. Finally, developing objective metrics that more directly capture nuance reflection is required, for example, by measuring temporal-envelope agreement or spectral consistency between vocal mimicry and synthesized sounds.

Acknowledgements

The authors used an AI-based language tool to assist with limited aspects of English wording and manuscript editing. All scientific content, analysis, interpretations and final revisions were reviewed and approved by the authors, who take full responsibility for the manuscript.

Notes

[1.] www.audiokinetic.com/en/wwise/overview/

[2.] www.krotosaudio.com/reformer-pro/

[3.] https://tsugi-studio.com/web/en/products-gamesynth.html

[4.] https://lumalabs.ai/dream-machine

[5.] https://sora.chatgpt.com/

[6.] https://gemini.google/us/overview/video-generation

[7.] www.mturk.com/

[8.] https://freesound.org/

[9.] An environmental sound data set collected from Freesound, consisting of 2,000 five-second audio clips (40 samples for each of 50 classes).

[10.] https://cine-tools.com/explosions/. This library provides cinematic, heavily designed explosion sound effects intended for film/game production, including destruction- and debris-like elements.

[11.] https://doi.org/10.5281/zenodo.20283840

[12.] www.zoom.co.jp/ja/products/handy-recorder/h4nsp-handy-recorder

[13.] https://zoomcorp.com/en/us/handheld-recorders/handheld-recorders/h4essential/

[14.] https://github.com/Jinmaro/PronounSE

[15.] https://github.com/YoonjinXD/T-foley

[16.] https://stableaudio.com/

[17.] These audio samples are available at https://jinmaro.github.io/PronounSE_demo/

References

Ament, V.T. (2009), *The Foley Grail: The Art of Performing Sound for Film*, Routledge, New York, NY.

Bongjun, K., Madhav, G., Bryan, P. and Zhiyao, D. (2018), "Vocal imitation set: a dataset of vocally imitated sound events using the AudioSet ontology", *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop, DCASE2018*.

Caren, M., Chandra, K., Tenenbaum, J., Ragan-Kelley, J. and Ma, K. (2024), "Sketching with your voice: 'non-phonorealistic', Rendering of sounds via vocal imitation", *SIGGRAPH Asia 2024 Conference Papers. SA '24, Association for Computing Machinery, Tokyo, Japan*, ISBN: 9798400711312, doi: 10.1145/3680528.3687679.

Cartwright, M. and Pardo, B. (2015a), *VocalSketch Data Set v1.0.4*, Zenodo, doi: 10.5281/zenodo.13862.

Cartwright, M. and Pardo, B. (2015b), "VocalSketch: vocally imitating audio concepts", *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, Association for Computing Machinery, Seoul, Republic of Korea*, pp. 43-46, ISBN: 9781450331456, doi: 10.1145/2702123.2702387.

Chen, Z., Seetharaman, P., Russell, B., Nieto, O., Bourgin, D., Owens, A. and Salamon, J. (2025), "Video-guided foley sound generation with multimodal controls", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Choi, K., Im, J., Heller, L.M., McFee, B., Imoto, K., Okamoto, Y., Lagrange, M. and Takamichi, S. (2001), "Foley sound synthesis at the DCASE 2023 challenge", *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023), Tampere, Finland*, pp. 16-20.

Chung, Y., Lee, J. and Nam, J. (2024), "T-FOLEY: a controllable waveform-domain diffusion model for temporal-event-guided Foley sound synthesis", *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*.

Comunità, M., Gramaccioni, R.F., Postolache, E., Rodolà, E., Comminiello, D. and Reiss, J.D. (2024), "Syncfusion: multimodal onset-synchronized video-to-audio Foley synthesis", *ICASSP 2024-2024*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 936-940, doi: 10.1109/ICASSP48485.2024.10447063.

Du, Y., Chen, Z., Salamon, J., Russell, B. and Owens, A. (2023), "Conditional generation of audio from video via Foley analogies", *Conference on Computer Vision and Pattern Recognition 2023*.

Evans, Z., Parker, J., Carr, C., Zukowski, Z., Taylor, J. and Pons, J. (2024), "Long-form music generation with latent diffusion", International Society for Music Information Retrieval Conference.

García, H.F., Nieto, O., Salamon, J., Pardo, B. and Seetharaman, P. (2025), "Sketch2Sound: controllable audio generation via time-varying signals and sonic imitations", *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, doi: 10.1109/ICASSP49660.2025.10888184.

Holman, T. (2015), *Sound for Film and Television*, Routledge, doi: 10.4324/9780240814322.

Hunt, R. (1923), "The phonetics of bird-sound", *The Condor*, Vol. 25 No. 6, pp. 202-208, doi: 10.2307/1362681, ISSN1938-5129, eprint: available at: https://academic.oup.com/condor/articlepdf/25/6/202/28133987/condor0202.pdf

Kilgour, K., Zuluaga, M., Roblek, D. and Sharifi, M. (2019), "Fréchet audio distance: a reference-free metric for evaluating music enhancement algorithms", *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2350-2354, doi: 10.21437/Interspeech.2019-2219, available at: www.isca-archive.org/interspeech_2019/kilgour19_interspeech.pdf

Kim, B. and Pardo, B. (2018), *Vocal Imitation Set v1.1.3: Thousands of Vocal Imitations of Hundreds of Sounds from the AudioSet Ontology*, Zenodo, doi: 10.5281/zenodo.1340763.

Kong, J., Kim, J. and Bae, J. (2020a), "HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis", *Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc, Vancouver, BC, Canada*, ISBN: 9781713829546.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W. and Plumbley, M.D. (2020b), "PANNs: large-scale pretrained audio neural networks for audio pattern recognition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 28, pp. 2880-2894, doi: 10.1109/TASLP.2020.3030497, available at: https://arxiv.org/abs/1912.10211

Lagrange, M., Lee, J., Tailleur, M., Heller, L.M., Choi, K., McFee, B., Imoto, K. and Okamoto, Y. (2025), "Sound scene synthesis at the DCASE 2024 challenge", *arXiv preprint arXiv:2501.08587*.

Lemaitre, G., Dessein, A., Susini, P. and Aura, K. (2011), "Vocal imitations and the identification of sound events", *Ecological Psychology*, Vol. 23 No. 4, pp. 267-307, doi: 10.1080/10407413.2011.617225.

Lemaitre, G., Rocchesso, D., Susini, P., Lambourg, C. and Boussard, P. (2013), "Using vocal imitations for sound design", *10th International Symposium on Computer Music Multidisciplinary Research. cote interne IRCAM: Lemaitre13b, Marseille, France*, available at: https://hal.science/hal-01107170

Li, N., Liu, S., Liu, Y., Zhao, S. and Liu, M. (2019), "Neural speech synthesis with transformer network", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33 No. 1, pp. 6706-6713, doi: 10.1609/aaai.v33i01.33016706.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W. and Plumbley, M.D. (2023), "AudioLDM: text-to-audio generation with latent diffusion models", *Proceedings of the International Conference on Machine Learning*, pp. 21450-21474.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y. and Plumbley, M.D. (2024), "AudioLDM 2: learning holistic audio generation with self-supervised pretraining", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp. 2871-2883, doi: 10.1109/TASLP.2024.3399607.

Luo, S., Yan, C., Hu, C. and Zhao, H. (2023), "Diff-Foley: synchronized video-to-audio synthesis with latent diffusion models", in Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M. and Levine, S. (Eds), *Advances in Neural Information Processing Systems*, Curran Associates, Inc, Vol. 36, pp. 48855-48876.

Mehrabi, A., Dixon, S. and Sandler, M.B. (2017), "Vocal imitation of synthesised sounds varying in pitch, loudness and spectral centroid", *The Journal of the Acoustical Society of America*, Vol. 141 No. 2, pp. 783-796, doi: 10.1121/1.4974825, ISSN0001-4966, available at: https://pubs.aip.org/asa/jasa/article-pdf/141/2/783/15322216/783_1_online.pdf

Niizumi, D., Takeuchi, D., Ohishi, Y., Harada, N. and Kashino, K. (2022), "BYOL for audio: exploring pre-trained general-purpose audio representations", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31 No. November, pp. 137-151, doi: 10.1109/TASLP.2022.3221007, ISSN2329-9290.

Okamoto, Y., Imoto, K., Takamichi, S., Nagase, R., Fukumori, T. and Yamashita, Y. (2024a), "Environmental sound synthesis from vocal imitations and sound event labels", *ICASSP 2024 – 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 411-415, doi: 10.1109/ICASSP48485.2024.10446906.

Okamoto, Y., Imoto, K., Takamichi, S., Nagase, R., Fukumori, T. and Yamashita, Y. (2024b), *ESC-50-Voice: Dataset of Vocal Imitation for Environmental Sound in ESC-50*, Zenodo, doi: 10.5281/zenodo.11385662.

Okamoto, Y., Imoto, K., Takamichi, S., Yamanishi, R., Fukumori, T. and Yamashita, Y. (2021), "Onoma-to-wave: environmental sound synthesis from onomatopoeic words", *APSIPA Transactions on Signal and Information Processing*, Vol. 11 No. 1, p. e13, doi: 10.1017/ATSIP.2022.13, doi available at: https://arxiv.org/abs/2102.05872

Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K., Saurous, R.A. and Sculley, D. (2016), "AutoMOS: learning a non-intrusive assessor of naturalness-of-speech", NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop, available at: https://arxiv.org/abs/1611.09207

Piczak, K.J. (2015), "ESC: dataset for environmental sound classification", *Proceedings of the 23rd Annual ACM Conference on Multimedia, ACM Press, Brisbane, Australia*, pp. 1015-1018, ISBN: 978-1-4503-3459-4, doi: 10.1145/2733373.2806390, available at: http://dl.acm.org/citation.cfm?doid2806390

Saeki, T. and Takamichi, S. (2024), "UTMOS: a state-of-the-art mean opinion score prediction system developed by UTokyo SaruLab", *The Journal of the Acoustical Society of Japan*, Vol. 80 No. 7, pp. 401-408, doi: 10.20697/jasj.80.7_401.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R.J., Saurous, R.A., Agiomyrgiannakis, Y. and Wu, Y. (2018), "Natural TTS synthesis by conditioning wave net on mel spectrogram predictions", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

Sinclair, J.L. (2001), *Principles of Game Audio and Sound Design: Sound Design and Audio Implementation for Interactive and Immersive Media*, Focal Press, New York, doi: 10.4324/9781315184432.

Sonnenschein, D. (2001), *Sound Design: The Expressive Power of Music, Voice, and Sound Effects in Cinema*, Michael Wiese Productions, Studio City, CA.

Tierno, M. (2020), *Location and Postproduction Sound for Low-Budget Filmmakers*, Routledge, London, doi: 10.4324/9780429331305.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*, Curran Associates Inc, Long Beach, CA, pp. 6000-6010, ISBN: 9781510860964.

Viers, R. (2008), *The Sound Effects Bible: How to Create and Record Hollywood Style Sound Effects*, Michael Wiese Productions, Studio City, CA.

Wang, S.S., Chen, J.Y., Bai, B.R., Fang, S.H. and Tsao, Y. (2024), "Unsupervised face-masked speech enhancement using generative adversarial networks with human-in-the-loop assessment

metrics", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp. 3826-3837, doi: 10.1109/TASLP.2024.3426996.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y. and Yu, D. (2023), "Diffsound: discrete diffusion model for text-to-sound generation", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31 No. April, pp. 1720-1733, doi: 10.1109/TASLP.2023.3268730, ISSN2329-9290.

Yewdall, D.L. (2012), Routledge, doi: 10.4324/9780240815237.

Corresponding author
Riki Takizawa can be contacted at: taikiriki1216@gmail.com