IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 December 2023, accepted 26 January 2024, date of publication 1 February 2024, date of current version 9 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3360885

METHODS

JVNV: A Corpus of Japanese Emotional Speech With Verbal Content and Nonverbal Expressions

DETAI XIN[1], (Member, IEEE), JUNFENG JIANG[1], SHINNOSUKE TAKAMICHI[1], (Member, IEEE), YUKI SAITO[1], (Member, IEEE), AKIKO AIZAWA[2], (Member, IEEE), AND HIROSHI SARUWATARI[1], (Member, IEEE)

[1]Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113-8656, Japan
[2]National Institute of Informatics, Chiyoda, Tokyo 101-8430, Japan

Corresponding author: Detai Xin (detai_xin@ipc.i.u-tokyo.ac.jp)

This work was supported in part by the Japan Science and Technology Agency (JST) Support for Pioneering Research Initiated by the Next Generation (SPRING) under Grant JPMJSP2108; in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant JP23KJ0828, Grant 21H04900, Grant 22H03639, and Grant 21H05054; in part by the Moonshot Research and Development under Grant JPMJPS2011; and in part by JST Fusion Oriented Research for disruptive Science and Technology (FOREST) under Grant JPMJFR226V.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Graduate School of Information Science and Engineering, University of Tokyo, Japan.

ABSTRACT We present the JVNV, a **J**apanese emotional speech corpus with **v**erbal content and **n**on**v**erbal vocalizations whose scripts are generated by a large-scale language model. Existing emotional speech corpora lack not only proper emotional scripts but also nonverbal vocalizations (NVs) that are essential expressions in spoken language to express emotions. We propose an automatic script generation method to produce emotional scripts by providing seed words with sentiment polarity and phrases of nonverbal vocalizations to ChatGPT using prompt engineering. We select 514 scripts with balanced phoneme coverage from the generated candidate scripts with the assistance of emotion confidence scores and language fluency scores. Experimental results show that JVNV has better phoneme coverage and emotion recognizability than previous Japanese emotional speech corpora. We then benchmark JVNV on emotional text-to-speech synthesis using discrete codes to represent NVs. The results demonstrate that there still exists a gap between the performance of synthesizing read-aloud speech and emotional speech, and adding NVs in the speech makes the task even harder, which brings new challenges for this task and makes JVNV a valuable resource for relevant works in the future. To our best knowledge, JVNV is the first speech corpus that generates scripts automatically using large language models.

INDEX TERMS Corpus design, emotional speech corpus, Japanese corpus, large-scale language model, nonverbal expression, nonverbal vocalization.

I. INTRODUCTION

Nonverbal expressions, such as vocal, facial, and gestural expressions [1], play an important role in human communication [2], [3]. In human speech, nonverbal expressions are called nonverbal vocalizations (NVs), which refer to vocalizations containing no linguistic information like laughter, sobs, and screams [4]. These expressions are relatively casual and usually used in spoken language [5]. One of the most important functions of NVs is conveying affects [6], [7]. Especially, emotional NVs, also called affect bursts, widely exist in many cultures [8]. Though NVs were ignored by most previous research on emotional speech [9], recent works have shown the possibility of applying NVs in many conventional speech-processing tasks, including speech emotion recognition [10] and expressive

The associate editor coordinating the review of this manuscript and approving it for publication was Lin Wang.

TABLE 1. A comparison between previous emotional corpora and JVNV. Here "partial" means only a part of the utterances of the corpus includes nonverbal expressions.

Corpus	Phoneme-balanced	Emotional scripts	Nonverbal expressions	Balanced emotion labels
JTES [18]	✗	✓	✗	✓
EmoV-DB [16]	✓	✗	Partial	✓
STUDIES [19]	✓	✓	✗	✓
OGVC [20]	✓	✓	Partial	✗
JVNV (this work)	✓	✓	✓	✓

speech synthesis [11], [12], [13]. However, the number of open-sourced emotional speech corpora with NVs is quite limited. Most existing work used in-house datasets to conduct experiments [12], [14] or even purchased a commercial corpus with limited size [15] for their experiments. As a result, further progress on NVs is impeded by a serious low-resource problem. We attribute this problem to two reasons. Firstly, it is difficult to obtain proper scripts for emotional speech corpus. Adigwe et al. [16] reused neutral scripts from an existing speech corpus [17] and asked the speakers to utter them with different emotions. Such a method is efficient because it can utilize existing annotations of the original scripts and skip to consider important factors for corpus design like phoneme balance. However, the neutral scripts make it difficult for speakers to utter them naturally with designated emotions. Takeishi et al. [18] collected emotional scripts from social media but failed to guarantee the phoneme balance property of the proposed corpus. Saito et al. [19] even employed workers to manually write emotional scripts, which is costly and difficult to scale up. Secondly, things become more troublesome when making emotional scripts with NV phrases. Adigwe et al. [16] tried to accomplish it by encouraging speakers to add NVs when they uttered the scripts. However, since this is not compulsory, not all utterances of the proposed corpus contain NVs. Arimoto et al. [20] collected a spontaneous speech corpus from online game chats. Though it is possible to find nonverbal expressions in this corpus, annotating the position and emotion labels for all NVs is prohibitive.

In this paper, to solve the above problems, we propose a corpus construction method assisted by large language models (LLMs) to generate emotional scripts with nonverbal expressions. The proposed method first samples candidate seed words from a Japanese sentiment polarity dictionary [21] and NV phrases from a previous Japanese NVs corpus [22] to form prompts to generate emotional scripts with an LLM. To further improve the generation quality, we leverage the ability of in-context learning of LLMs by adding handwritten exemplars in the prompts. Then, we select high-quality scripts from the generated scripts with the help of a pretrained emotion classifier and a pretrained language model. Based on the proposed method, we construct JVNV, an emotional Japanese speech corpus uttered by professional speakers with Verbal content and Non-Verbal expressions. JVNV consists of about four hours of emotional speech data covering six basic emotions [23] from four native speakers. Every utterance has at least one NV phrase. We also annotate the duration of the NV phrases in each utterance. JVNV is large enough to support relevant tasks like emotional speech synthesis. Besides, since the phrases and duration of NVs are provided explicitly, JVNV is more suitable for further research on NVs compared to existing corpora with NVs. In the experiments, we first technically validate the effectiveness of the proposed corpus construction method from the aspects of phoneme coverage and emotion recognizability. We then benchmark JVNV on emotional text-to-speech (TTS) synthesis using discrete codes obtained from a self-supervised learning (SSL) model to represent NVs. The results demonstrate that there still exists a gap between the performance of synthesizing read-aloud speech and emotional speech, and adding NVs in the speech even makes the task harder, which brings new challenges for this task and makes JVNV a valuable resource for relevant tasks in the future.

The contributions of this work can be summarized as follows:

- We propose a corpus construction method for emotional speech with NVs using LLMs for script generation, which is to our best knowledge the very first try of making scripts for a speech corpus with LLMs.
- We construct JVNV, a phoneme-balanced Japanese emotional speech corpus with both verbal and nonverbal expressions. JVNV is expected to further advance relevant works with NVs in the future.
- We conduct comprehensive objective and subjective experiments to validate the effectiveness of JVNV.
- We benchmark JVNV on emotional TTS and show the challenges of synthesizing emotional speech with both verbal and nonverbal expressions.

We release JVNV at our project page.[1]

The rest of the paper is organized as follows. Section II introduces related work. We give a detailed explanation of the proposed script generation method in Section III and describe how JVNV is recorded in Section IV. In Section V, we validate the effectiveness of JVNV from phoneme coverage and emotion recognizability. In Section VI, we benchmark JVNV on emotional TTS. Section VII describes potential

[1] https://sites.google.com/site/shinnosuketakamichi/research-topics/jvnv_corpus

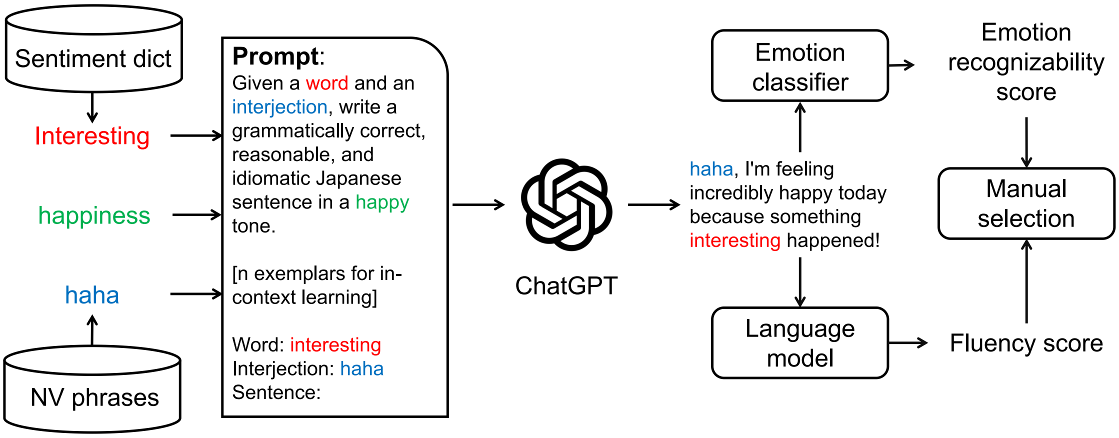

Sentiment dict
Interesting
happiness
haha
NV phrases
Prompt:
Given a word and an interjection, write a grammatically correct, reasonable, and idiomatic Japanese sentence in a happy tone.
[n exemplars for in-context learning]
Word: interesting
Interjection: haha
Sentence:
ChatGPT
haha, I'm feeling incredibly happy today because something interesting happened!
Emotion classifier
Emotion recognizability score
Manual selection
Language model
Fluency score

FIGURE 1. Overview of the proposed emotional script generation method with NV phrases. Here we use happiness as the emotion, interesting as the seed word, and haha as the NV phrase. Note we use the word "interjection" to replace NV so that ChatGPT can understand this concept.

social impacts of JVNV. Finally, we conclude the paper in Section VIII.

II. RELATED WORK

A. SCRIPT SELECTION FOR CONSTRUCTING SPEECH SYNTHESIS CORPORA

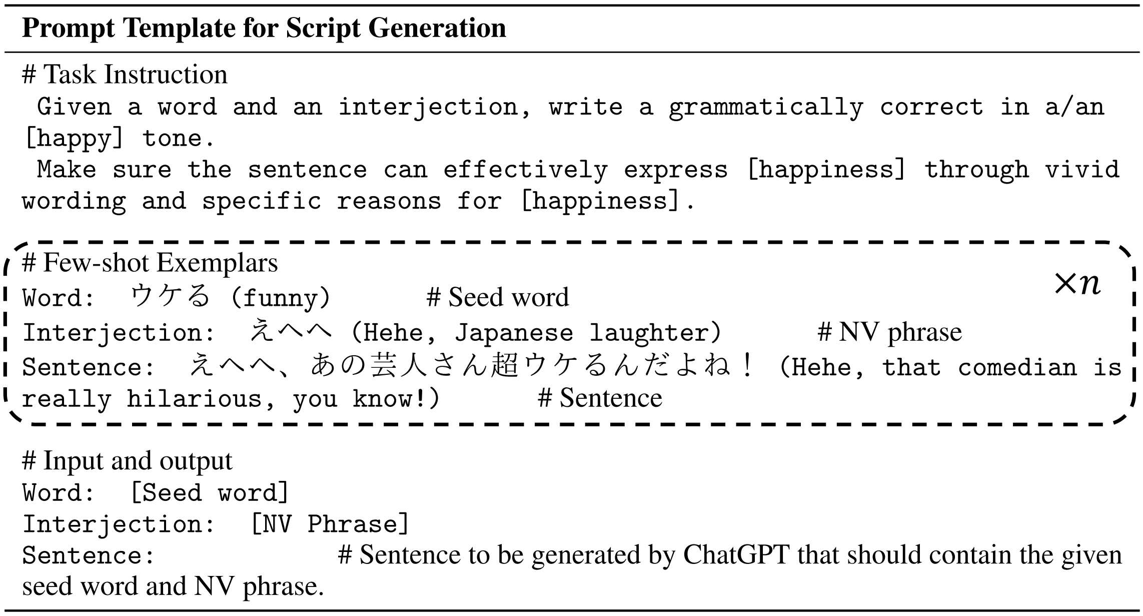
Since speech data that is suitable for training a TTS system has to be clean and include as little noise as possible, instead of directly collecting dirty speech data from the internet, common read-aloud speech corpora for TTS [24], [25] were constructed by employing speakers to utter designated scripts prepared before the recording in an anechoic chamber. With a limited budget, the scripts have to cover more phonemes to ensure the generalization ability of the TTS systems trained on the corpus, which stimulates a lot of script selection methods to select proper scripts from an existing script set. Such methods either regarded the selection process as a set covering problem to cover more phoneme combinations [26], [27] or tried to match the phoneme distribution of the selected scripts with a desired distribution [28], [29]. Recently, Nose et al. [30] proposed a selection criterion called extended entropy to measure the phonetic and prosodic coverage of the selected scripts.

To design an emotional speech corpus, one has to not only ensure the emotion recognizability of the selected scripts, i.e., the scripts should express the target emotions, but also consider the phonemic properties of the scripts like phoneme coverage. However, it is difficult to accomplish these two goals simultaneously. Takeishi et al. [18] collected emotional scripts from Twitter but had narrow phoneme coverage compared to other read-aloud corpora since the number of collected scripts was not enough to utilize a script selection method, and the scripts themselves were short and simple. Adigwe et al. [16] utilized phoneme-balanced scripts of an existing read-aloud corpus and asked the speakers to utter them emotionally to construct an emotional corpus, but the neutral scripts usually failed to express the desired emotions. Saito et al. [19] employed workers to write emotion-recognizable scripts, but the number of scripts is limited due to the high employment fee.

In this work, instead of adopting existing scripts, we propose to utilize a powerful LLM to generate emotional scripts by prompt engineering. LLMs have demonstrated impressive performance in natural language generation tasks [31], [32] and can generate scripts cheaply. We propose to use seed words with proper emotional polarity in the prompts. The seed words designate the topics of the generated scripts and thus can improve the diversity of the scripts to ensure sufficient phoneme coverage. To make sure the scripts can cover all Japanese phonemes, we use specific seed words that contain rare phonemes for generation. Finally, we propose to compute an emotion recognizability score and a language fluency score for each script using a pretrained emotion classifier and a pretrained language model. The final script set of JVNV is selected with the assistance of the computed scores to ensure the scripts have high quality.

B. EMOTIONAL SPEECH CORPUS WITH NONVERBAL EXPRESSIONS

As mentioned, the number of emotional corpora with nonverbal expressions is limited. Emotional speech corpora can be roughly categorized into two types: acted and spontaneous [33]. In acted corpora, scripts and a recording instruction are provided to the speakers before the recording, which makes it easy to control label balance and almost does not need extra manual annotation. Therefore, the difficulty of constructing an acted corpus with NVs lies in making emotional scripts with nonverbal expressions. Adigwe et al. [16] dealt with this difficulty by encouraging speakers to utter NVs during recording, but their method cannot ensure every utterance has NVs, let alone control the content of NVs. On the other hand, spontaneous corpora have fewer constraints on speakers. Usually, the speakers can speak any content in any style. Since NVs are casual expressions, it is easy to incorporate NVs in a spontaneous corpus. However, these corpora require more effort to do annotation and



Prompt Template for Script Generation

Task Instruction
Given a word and an interjection, write a grammatically correct in a/an [happy] tone.
Make sure the sentence can effectively express [happiness] through vivid wording and specific reasons for [happiness].

Few-shot Exemplars ×n
Word: ウケる (funny) # Seed word
Interjection: えへへ (Hehe, Japanese laughter) # NV phrase
Sentence: えへへ、あの芸人さん超ウケるんだよね！ (Hehe, that comedian is really hilarious, you know!) # Sentence

Input and output
Word: [Seed word]
Interjection: [NV Phrase]
Sentence: # Sentence to be generated by ChatGPT that should contain the given seed word and NV phrase.

FIGURE 2. Prompt template for script generation. We use happiness as an example. Texts embraced by [] are replaced by proper content during script generation. Texts starting with # are comments. We use $n = 3$ demonstrations during the script generation. We provide one of them as an example. The English translation is also attached.

usually have a low controllability on emotion distribution. Arimoto et al. [20] proposed OGVC, an emotional corpus with many nonverbal expressions collected from online game chats. However, the emotion labels of the utterances were biased to common emotions like happiness, and only laughter was annotated in this corpus.

JVNV is an acted emotional corpus, and we solve the problem of making scripts by prompting the LLM to generate scripts with designated NV phrases. Compared to spontaneous corpora, the scripts of JVNV have a balanced emotion distribution. Furthermore, every utterance of JVNV has at least one NV, and we also provide the duration information of the NVs in each utterance for future research. We compare previous corpora and JVNV in Table 1.

III. PROPOSED CORPUS CONSTRUCTION METHOD

We aim to construct a Japanese emotional corpus with NVs covering six basic emotions [23]: anger, disgust, fear, happiness, sadness, and surprise. The core idea is to create proper prompts for controllable script generation. The overview of the proposed script generation method is illustrated in Fig. 1. To generate a script expressing an emotion with a nonverbal expression, we first sample a seed word from a Japanese sentiment polarity dictionary [21] and an NV phrase in the JNV corpus [22] of the emotion. The seed word is used as a topic word to improve the diversity of the scripts and must be included in the generated scripts. Formally, given the emotion e, the seed word w, and the NV phrase v_e of e, we aim to model and find scripts s to maximize the following probability distribution:

$$p(s \mid e, v_e, w). \tag{1}$$

Since there is no available dataset, it is almost impossible to directly learn this distribution using supervised learning. Fortunately, recent work has shown that LLMs have an impressive ability of zero-shot text generation with prompt engineering [31], [32], which enables us to find proper scripts without training a new model. In prompt engineering, a text prompt describing the task requirements is fed to the LLM, and since the LLM can interpret natural language, it will generate the text satisfying the requirements, if the prompt is properly designed. We thus prompt an LLM to generate an emotional script that not only includes the sampled seed word and the NV phrase but also expresses the designated emotion. The prompt is constructed by filling the emotion, the seed word, and the NV phrase to a predefined prompt template as shown in Fig. 1. In this way, assume $f(\cdot)$ is a function mapping the template into the prompt given e, v_e, and w, i.e. prompt $= f(e, v_e, w)$, we convert Eq. 1 into:

$$p(s \mid f(e, v_e, w)), \tag{2}$$

which can be easily learned by LLMs without training. Especially, we use ChatGPT as the LLM because of the high quality of its generated texts and its moderate price. Finally, we select high-quality scripts from the generated scripts with the assistance of a pretrained emotion classifier and a pretrained language model. In the following sections, we introduce the proposed corpus construction method step by step.

A. SESSIONS

Following Trouvain and Truong [5], we define NVs as vocalizations uttered by humans that are difficult or even not able to be transcribed into orthographical forms. Therefore, NVs include not only nonverbal expressions like laughter but also common interjections like ''Oh''. Though there exists a common set of NV phrases in Japanese, people usually have their own unique NV phrases to express certain emotions [22]. Therefore, we design to create two kinds of scripts for two different sessions in the corpus: regular session and phrase-free session. In the regular session, we designate a certain NV to utter in each script. In the phrase-free session, we do not include NV phrases in the scripts but ask the speakers to utter NVs by themselves with no restriction on the phonetic content. This approach can ensure the generality of the phrases while maintaining the personality of each speaker.

B. PHRASES OF NVS

For phrases used for the regular session, we adopt NV phrases collected in JNV corpus [22], a corpus of Japanese NVs covering six basic emotions. JNV collects NV phrases by large-scale crowd-sourcing and thus can cover a wide range of phrases used in daily conversations. We choose phrases that have high emotion recognizability in JNV (phrases with recognition accuracy larger than 66.7% as described in the original paper [22]), which results in 11/7/8/16/7/19 phrases for anger/disgust/fear/happiness/sadness/surprise, respectively. Readers are recommended to refer to JNV[2] to get detailed information on the phrases.

C. SEED WORDS WITH SENTIMENT POLARITY

With the impressive ability of text generation of ChatGPT, it is possible to ask ChatGPT to generate scripts by simply prompting it with a phrase and an emotion, i.e. maximizing:

$$p(s \mid f(e, v_e)), \tag{3}$$

where the seed word w is removed from the prompt. However, in our preliminary experiments, the generated scripts suffered from semantic overlapping. Their sentence structures and vocabularies also lacked diversity, making this naive method inappropriate for generating proper scripts for a speech corpus. Moreover, we also tried to find proper scripts from existing sentiment analysis corpora and insert NV phrases into the original emotional sentences. But this method is also infeasible because the selected NV phrase is not necessarily suitable for a certain sentence, producing unnatural and incoherent scripts.

Therefore, we propose to generate scripts by adding a seed word together into the <phrase, emotion> pair in the prompt to designate the topic of the generated script. We use a Japanese sentiment polarity dictionary that includes words with positive, negative, and neutral polarities [21]. The polarity is properly selected according to the emotion. In this way, the proposed method can generate coherent and diverse scripts with NV phrases, and avoid the semantic overlapping problem. Furthermore, it is intuitively more natural and easier to make scripts using words with a proper polarity rather than random words. For example, we may randomly select a negative word like ''bad'' to create a happy script, which is unsuitable.

In practice, we first filter inappropriate words (e.g. toxic words) from the dictionary[3] using an existing word list.[4] We sample negative words for the script generation of anger, disgust, fear, and sadness, while positive words are sampled for generating happy scripts. For the emotion of surprise, although it is possible to sample all kinds of words, we sample only neutral words to avoid the possible confusing patterns between surprise and other emotions (e.g., happiness or fear) [7], [22]. The detailed generation algorithm is described in the next section.

D. SCRIPT GENERATION WITH PROMPT ENGINEERING

With the emotions, the phrases, and the seed words prepared, we prompt ChatGPT to generate scripts. The prompt template is shown in Fig. 2. It consists of three parts: a task instruction, n exemplars for few-shot in-context learning, and a place-holder for the script to be generated. The first sentence of the instruction describes the basic requirements including the information about the target emotion. The second sentence stresses that the generated script should effectively express the target emotion. We also use expressions like ''vivid wording'' and ''specific reasons'' to serve as conditions for text generation to make the scripts more diverse and reasonable. Note that, we use ''interjection'' in the prompt since we found that ChatGPT could not understand the meaning of NVs well in our preliminary experiments.

In addition to the instruction, we also add n exemplars in the prompt because it has been shown that the performance of LLMs like ChatGPT becomes better by providing some exemplars in the prompt [34], which is called in-context learning. This is quite intuitive since the exemplars can make the task from a zero-shot setting to a few-shot setting for LLMs. In this way, Eq. 2 becomes:

$$p(s \mid f(e, v_e, w, \{a_e^i\}_{i=1}^n)), \tag{4}$$

where a_e^i is the i-th exemplar of e. As shown in Fig. 2, each exemplar comprises a seed word, an NV phrase, and a script that satisfies our requirements, which helps ChatGPT understand what we expect it to generate. In this work, we manually create $n = 3$ exemplars for each emotion with non-overlapped seed words and phrases. Finally, we fill in the sampled seed word and phrase to the same template as the exemplars and leave the script term blank for ChatGPT to generate.

We use the `gpt-3.5-turbo-0301` API from OpenAI.[5] Since ChatGPT performs better on English tasks than on Japanese tasks, we use English to compose the prompt so that it can understand our instructions better. For generating the

[2] https://sites.google.com/site/shinnosuketakamichi/research-topics/jnv_corpus

[3] https://www.cl.ecei.tohoku.ac.jp/Open_Resources-Japanese_Sentiment_Polarity_Dictionary.html

[4] https://github.com/MosasoM/inappropriate-words-ja

[5] https://openai.com/blog/chatgpt

TABLE 2. Numbers of selected scripts (regular session + phrase-free sessions) for each emotion. Noted that these two sets have no overlapping.

Set	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Total
Core	$44 + 10$	$49 + 10$	$49 + 10$	$48 + 10$	$49 + 10$	$57 + 10$	356
Extra	22	15	28	41	14	38	158

scripts of the phrase-free session, the phrase information is excluded from the prompt template. In this stage, we finally generate 13k candidate scripts. We filter out scripts that do not contain the NV phrase in the prompt.

The `gpt-3.5-turbo` model used in the proposed method costs $0.002/1k tokens, as described in https://openai.com/pricing. Generating 13k scripts cost 8 USD in total.

E. SCRIPT SELECTION

Though the generated scripts already have a high quality, we further rate them to select scripts with high quality. We consider two criteria that are important for scripts of an emotional speech corpus: emotion recognizability and fluency because speakers will find it easy to utter the script if the script is fluent and reasonable to express the emotion. To this end, we propose to use a pretrained emotion classification model and a pretrained language model to rate the scripts.

To obtain the emotion recognizability score, we train an emotion classifier by fine-tuning a Japanese RoBERTa[6] [35] using the Japanese emotion analysis corpus, WRIME [36], which covers the six emotions used in JVNV. Subsequently, the classification probability of the emotion of each script is regarded as the emotion recognizability score. To obtain the fluency score, we compute the pseudo-log-likelihood scores (PLL) [37] based on the Japanese BERT model[7] [38]. PLL can be regarded as an approximation of the perplexity of the generated script. So, it is suitable to serve as the fluency score. We also tried other models including Japanese RoBERTa and GPT2, but the Japanese BERT had the best performance in our preliminary experiments. We describe further details of the fine-tuning and the language fluency score at Appendix A-A and A-B, respectively.

We compute the two scores for all generated 13k scripts and convert the two scores to $[0, 1]$ so that they have the same scale. We then sort them descendingly by the summation of these two scores and select the top-k scripts for each <emotion, phrase> pair. All selected scripts are double-checked by us to avoid possible bad cases and ethical issues. We first select a core script set that contains 356 scripts that are required to be uttered by all speakers, where each emotion has 10 scripts without NV phrases for the phrase-free session. An extra script set with 158 scripts is also selected for the regular session in case the speakers have spare time to utter more samples. We balance the number of scripts of different emotions and phrases. The detailed information is shown in Table 2.

TABLE 3. Number of utterances for each (emotion, speaker) pair in JVNV. The total duration of each speaker/emotion is shown in the last row/column.

Emotion	F1	F2	M1	M2	$\sum$ (hrs)
Anger	54	76	54	65	0.53
Disgust	59	74	59	66	0.59
Fear	59	78	59	69	0.65
Happiness	58	90	58	74	0.75
Sadness	59	73	59	66	0.82
Surprise	67	86	67	86	0.60
$\sum$ (hrs)	0.94	1.11	0.92	0.97	3.94

F. PHONEME COVERAGE

In our preliminary experiments, we found that it was hard to cover some rare phonemes in Japanese like “デュ” (pronounced as /dyu/) by randomly choosing seed words. Therefore, we manually choose some words (less than 5) containing rare phonemes as the seed words for the phrase-free session. This process ensures the full script set covers all Japanese phonemes. In Section V-A, we show that our proposed script set has better phoneme coverage than a previous Japanese emotional corpus.

IV. JVNV CORPUS

We used the selected scripts to construct the proposed JVNV corpus. Four professional speakers (two males and two females) were employed to utter these scripts. We got the consent from all speakers before the recording. Specifically, we first commissioned a company to find potential speakers. After selecting the speakers, the company described all relevant information about this project and obtained consent from the speakers. The four speakers are all native speakers and experienced in professional voice acting.

Right before the recording, we first described our goal of collecting emotional speech with both verbal content and nonverbal expressions. We then described some key points of the recording, including the definition of NV, the definition of the two sessions, and their differences. Each speaker was required to utter the scripts with the designated emotions as many as possible within four hours. The speakers were instructed to express the designated emotions as naturally as possible. Also, they were allowed to utter NVs with more flexibility on duration and contents than the verbal parts. For example, for an NV phrase “はは (haha)”, they can utter as “ははは (hahaha)”, as long as they think it was more natural for expressing the given emotion. In the phrase-free session, we encouraged them to utter their personal NV phrases that

[6] https://huggingface.co/rinna/japanese-roberta-base

[7] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

TABLE 4. Extended phoneme entropy of each corpus. Higher entropy means better phonemic balance.

ITA	JVNV (full)	JVNV (core)	JTES
34.64	33.33	33.20	33.10

do not exist in the regular session, even if the phrases are not commonly used by other people. However, if the speakers found it difficult to find a personal phrase, they were allowed to refer to the phrases in the regular session. We paid 800, 000 JPY for the recording involving four speakers.

All utterances were recorded in an anechoic chamber to reduce possible noises and were saved as 48 kHz Wav files. All speakers uttered the core set. Two of them uttered scripts in the extra set. We show the number of utterances of each <emotion, speaker> pair in Table 3. The final corpus contains about 3.94 hours of emotional speech data. Note that due to our limited budget, the size of the corpus is limited, but it is enough for the experiments described in Section VI and is easy to scale up by employing more speakers.

In addition to the audio and scripts, we also annotate the duration information of the NV in each utterance. Such labels can be used to study nonverbal expressions or verbal content separately and support other methods like data augmentation described in Section VI.

V. TECHNICAL VALIDATION

In this section, we validate the effectiveness of the proposed corpus construction method and JVNV from phoneme coverage and emotion recognizability. We use the scripts or audio of the following existing corpora for comparison:

- **ITA**[8]: A phoneme-balanced script set designed for TTS. It comprises 324 neutral scripts and 100 emotional scripts. The scripts are manually selected to ensure phoneme coverage.
- **JTES** [18]: An emotional corpus with four emotions (anger, happiness, sadness, neutral) uttered by nonprofessional speakers, where the scripts are collected from Twitter. Each emotion has 50 scripts.
- **OGVC** [20]: A spontaneous speech corpus collected from online game chats with post-annotated emotion labels. OGVC includes ten different emotions, covering all emotions used in our JVNV.
- **STUDIES** [19]: A manually designed Japanese dialogue speech corpus with four emotions (anger, happiness, sadness, neutral) uttered by professional speakers, where the scripts are collected by employing workers to write emotional dialogues.

A. PHONEME COVERAGE

As for the phoneme coverage, we only compare JVNV with ITA and JTES since the number of scripts in OGVC (6579) and STUDIES (5311) is much larger than others, which is unfair to perform a comparison. We first compute extended entropy [30] of each script set, which is defined as the summation of the entropy of m consecutive phonemes:

$$S = \sum_{m=1}^{M} w_m S_m, \qquad S_m = -\sum_{n=1}^{N_m} p_{mn} \log_2 p_{mn}, \tag{5}$$

where M is the maximal number of consecutive phonemes. S_m and w_m are the entropy of m phonemes approximated from the script set and the weight of it, respectively. p_{mn} is the probability of the n-th combination of m consecutive phonemes. N_m is the number of all possible arrangements of m consecutive phonemes in the scripts. In this work, we set $M = 4$ and $w_m = 0.25(m = 1, 2, 3, 4)$. The result is shown in Table 4.

It shows that both the core and full sets of JVNV have better phoneme coverage than JTES, which demonstrates that the generated emotional scripts of the proposed method have better quality than the manually collected emotional scripts of JTES. Besides, the ITA corpus has the largest extended entropy, which is expected since ITA is designed for good phoneme coverage without considering emotions.

Fig. 3 shows the number of arrangements of m consecutive phonemes of each corpus. It can be seen that JVNV has a similar performance to ITA, but JTES fails to capture diverse phoneme arrangements when $m = 4$, which further shows that JVNV has better phoneme coverage than JTES.

TABLE 5. Subjective emotion recognition accuracy (%) of each corpus.

JVNV	JVNV-V	JTES	STUDIES	OGVC
94.21	87.37	80.85	62.05	54.69

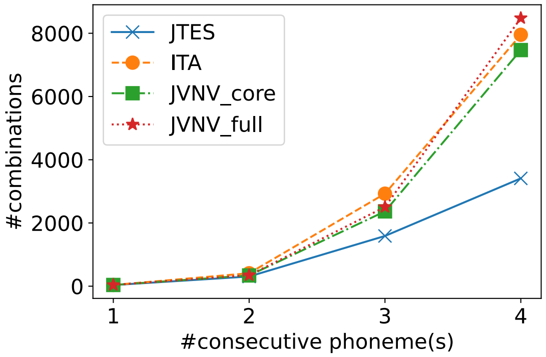


FIGURE 3. The number of unique arrangements of consecutive phoneme(s) of each corpus.

B. EMOTION RECOGNIZABILITY

As for emotion recognizability, we compare JVNV with JTES, OGVC, and STUDIES. We exclude ITA since it has no emotional utterances. For JTES and STUDIES, we exclude the neutral scripts. For OGVC, we select the utterances with the six emotions used in JVNV. Each utterance in OGVC is annotated with three labels from three annotators, and we only use those utterances whose three labels are the same,

[8] https://github.com/mmorise/ita-corpus

which only accounts for 3% of the whole corpus. To show the contributions of NVs on emotion recognizability, we also evaluate the verbal parts of JVNV by removing nonverbal parts from the utterances of JVNV denoted as "JVNV-V".

We conducted a forced choice task on a Japanese crowdsourcing platform.[9] For each corpus, we randomly picked up 60 emotion-balanced samples. Thirty workers participated in the evaluation. Each worker evaluated 30 utterances by listening to the corresponding audio and selecting the most possible expressed emotion from seven choices (the six emotions used in JVNV and an extra "None of the above" choice to avoid artificially high accuracy [39]). For further details of the evaluation, please refer to Appendix B. The result is shown in Table 5.

Firstly, we find that JVNV has much higher accuracy than JTES, OGVC, and STUDIES, which demonstrates the utterances of JVNV are expressive enough for people to recognize the emotions. Secondly, the accuracy of JVNV degrades after removing the NVs, which demonstrates the necessity of considering NVs in emotional speech. Note that, even without NVs, JVNV-V still performs better than JTES and STUDIES, whose scripts are written or designed by humans, showing the effectiveness of using prompt engineering with ChatGPT to generate proper emotional scripts. Finally, to our surprise, OGVC has the lowest emotion recognizability even if we select those utterances with high agreement from the original annotators. We assume it is because collecting speech of uncommon emotions in a specific situation (e.g., game chats in this case) is difficult. We further inspected the results and found that the recognizability of anger, fear, and disgust in OGVC is much lower than that of happiness, sadness, and surprise. This observation supports our assumption since happiness, sadness, and surprise are more common emotions appearing during game playing than the other three emotions, which demonstrates the difficulty of controlling the balance of emotion labels for a spontaneous emotional speech corpus. In summary, JVNV not only has good phoneme coverage but also has high emotion recognizability with nonverbal expressions, showing the effectiveness of the proposed corpus construction method.

VI. TTS BENCHMARK

A. THE TTS MODEL WITH MIXED REPRESENTATIONS

We benchmark JVNV on the task of emotional TTS since it is the main design goal of JVNV, and we leave other tasks like emotion recognition for future works. One of the key problems for synthesizing speech with both nonverbal expressions and verbal contents in the framework of TTS is how to represent NVs in a symbolic form. This is because NVs used in daily life usually have various phonetic contents and duration (the design choices like the phrase-free session of the proposed method also aim to simulate this fact), which makes it quite difficult to transcribe NVs like the verbal contents.

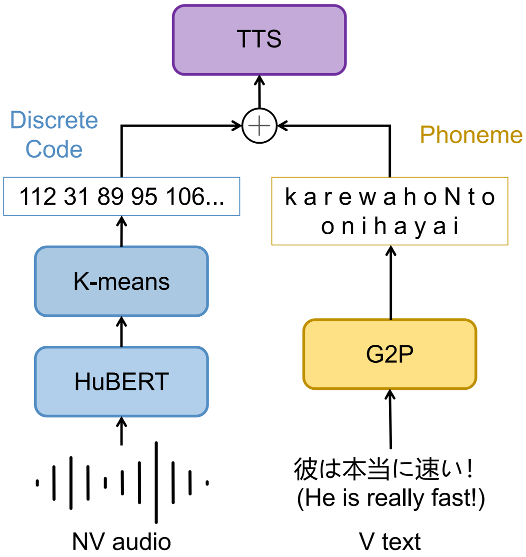


FIGURE 4. The proposed TTS method uses codes to represent NVs. Codes and phonemes are concatenated together and fed to the TTS model.

Fortunately, recent work implies that it is possible to use discrete tokens extracted by a self-supervised model to represent NVs [11], [12], [13], [40]. Following this idea, we use a TTS model that is used to synthesize laughter in a previous work [13] and adapt it to JVNV. The general architecture is illustrated in Fig. 4. Inspired by the previous work [12], [13], we use discrete codes generated by a k-means model trained on the features extracted by HuBERT [41] from the waveforms of NVs to represent NVs, and use phonemes as the representations for verbal text. The two representations are then concatenated together and fed to the TTS model. For the TTS model, we use the same architecture of Xin et al. [13] based on FastSpeech2 [42] with an additional emotion embedding table. An unsupervised alignment module is used to get the duration of codes/phonemes automatically, which is trained with connectionist temporal classification loss [43] adapted from GlowTTS [44]. Readers are recommended to refer to the original paper [13] for more details. Furthermore, we propose a data augmentation strategy, where we randomly select an NV with the same emotion to replace the original one in the utterance during training.

B. EXPERIMENTS

1) SETUP

To show the difficulties of synthesizing emotional speech, we use JVNV and a read-aloud Japanese speech corpus, JSUT [45], to train all models. JSUT is a read-aloud single-speaker Japanese speech corpus with 5, 000 utterances. The emotion labels of all utterances of JSUT are treated as neutral. In addition to the model trained by the proposed method (denoted as "NV+V" hereafter), we also trained two variations of the proposed method. The first variation, denoted as "V", is trained on JVNV-V with no NV in the

[9] https://www.lancers.jp/

TABLE 6. Performance of the models trained with different representations for NVs. For all models we use the same test set except for the V model, where we exclude all NVs in the test set. Bold indicates the best score with $p < 1e$-5.

Model	NV	MCD(↓)	F0-RMSE(↓)	MOS-Emo(↑)	MOS-JSUT(↑)
GT	✓	-	-	4.7	4.3
HiFi-GAN	✓	3.1	26.4	4.1	3.8
V	✗	5.9	49.5	2.9	3.4
NV+V	code	**6.2**	**50.9**	**2.4** (3.0 w/o NV)	**3.5**
Phoneme	phoneme	6.7	53.1	1.9	3.4

training set. In the second variation denoted as "Phoneme", we use phonemes of the phrases to represent NVs.

We mixed JSUT with JVNV, which resulted in 6, 615 utterances. We split these utterances into train/validation/test sets with 6, 447/84/84 utterances, respectively. In the test set, each of the 6 emotions used in JVNV had 12 utterances equally selected from the 4 JVNV speakers. We also excluded 12 samples from the training set of JSUT for testing. The validation set had the same structure as the test set.

All waveforms were downsampled into 22.05 kHz. The pitch information of each utterance was extracted with WORLD vocoder [46]. We used OpenJTalk[10] as the G2P (grapheme-to-phoneme) model to convert texts into phoneme sequences. We used the pretrained `hubert-base-ls960` model[11] to extract the HuBERT [41] features used in the proposed method. This model was trained on the 960-hour LibriSpeech corpus [47] with a 12-layer transformer-based architecture [41]. The outputs of the 12-th (last) layer of the model were used as the features. For k-means clustering, we used the implementation of sklearn[12] to train the model. We also used K-means++ [48] to accelerate the model initialization. The cluster number was set to 200, which means that there are 200 unique codes representing the NVs. The batch size was set to 10, 000. The k-means model converged in about 250 iterations.

We used the same architecture of the original FastSpeech2 [42]. The dimension of the speaker embedding and the emotion embedding were set to 256. For the alignment module, we used the same training strategy used in Xin et al. [13]. Readers are recommended to refer to the original paper for more details [13]. The batch size was set to 16. Adam [49] was used as the optimizer with a scheduled learning rate proposed in Vaswani et al. [50]. All TTS models converged in about 200k steps with an NVIDIA A100 GPU card in 24 hours. We used HiFi-GAN [51] to convert mel-spectrograms output by the TTS model into time-domain waveforms. Note that, during inference, we directly use the codes of NVs extracted from ground truth (GT) utterances for simplicity, but it is also possible to use another model like a language model to sample codes [13].

2) RESULTS AND DISCUSSIONS

We use both objective and subjective metrics to evaluate the performance. For objective metrics, we use mel-cepstral distortion (MCD) and F0 root mean square error (F0-RMSE) to evaluate speech quality and prosody. For subjective metrics, we conduct a standard five-scale mean opinion score (MOS) test to evaluate the naturalness of the synthesized speech from 1 (very unnatural) to 5 (very natural). For further details of the MOS test, please refer to Appendix C. For the NV+V model, we additionally remove the NV part of each synthesized utterance to evaluate the performance of NV+V on synthesizing verbal speech. We denote the MOS scores of the synthesized emotional speech and neutral speech as "MOS-Emo" and "MOS-JSUT", respectively. Note that, the 12 test samples of JSUT are only used to compute the MOS-JSUT score and are not used to obtain any objective metric.

The results are shown in Table 6. First, we can see that NV+V consistently outperforms Phoneme in all metrics, showing the effectiveness and necessity of the proposed method using discrete codes to represent NVs. During training, we also observed that the Phoneme model even could not converge. We suppose it is because phoneme is not a proper representation of NVs. Second, for all models, the MOS-Emo scores are worse than the MOS-JSUT scores, which demonstrates the difficulty of synthesizing emotional speech with various prosody patterns. In our opinion, this difficulty is derived from the multimodal property of the pitch distribution of emotional speech, and the mean square error loss used by FastSpeech2 [42] only enables the model to learn a single-modal distribution. We believe a more powerful model can solve this problem and leave this as a future work. Third, although the MOS-JSUT scores of V and NV+V are quite similar, the difference in MOS-Emo scores is too large to neglect, which can be regarded as the extent of performance degradation by adding NVs. This observation is further verified by evaluating the verbal part of the synthesized samples of NV+V, which results in a MOS-Emo score of 3.0 that is even larger than the one of V, showing the difficulty of synthesizing nonverbal expressions. By listening to the samples, we found that some NVs were still not synthesized

[10] https://github.com/r9y9/pyopenjtalk
[11] https://huggingface.co/facebook/hubert-base-ls960
[12] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html

with correct phonetic contents. We assume this is because the discrete code obtained from HuBERT is still not a perfect representation for NVs, even if it is better than the phoneme. This is quite intuitive since HuBERT is trained on speech corpora with rare NVs [41].

To sum up, we show that there still exists a big gap between the performance of normal TTS and emotional TTS, and adding NVs in emotional speech makes the task even harder. Based on our observations, we realize that the discrete code cannot fully represent NVs. Therefore, finding a proper representation for NVs seems to be a core problem for this task in the future, which further makes JVNV a valuable resource for future work in this field.

VII. POTENTIAL SOCIAL IMPACTS

JVNV should intuitively be a useful resource for all emotional speech processing tasks with NVs, including but not limited to speech emotion recognition [10], [52], emotional speech synthesis [11], and NVs detection [53]. As previous work showed that NVs can effectively improve the emotion recognizability [54] and expressiveness [55] of real/synthetic speech, we expect JVNV to be utilized to construct expressive TTS systems or robust high-accuracy SER systems in the future.

However, JVNV also brings several potential negative impacts. First, for TTS systems that can synthesize speech with NVs, since NVs make synthetic speech more realistic and difficult to distinguish from real speech for listeners, they might be used in voice phishing. Such a problem can be possibly solved by recent speech anti-spoofing technologies [56]. Second, powerful SER systems that can recognize emotions from not only verbal speech but also nonverbal signals like sighs and laughter, might be maliciously used to analyze the mental states of others, causing a severe privacy problem. Technologies for solving such a problem are usually called SER evasion, in which the original speech is perturbed to remove emotional information but preserve content information [57].

VIII. CONCLUSION

This paper first presented JVNV, a Japanese emotional speech corpus with verbal content and nonverbal expressions. The scripts of JVNV are generated by providing seed words with sentiment polarity and phrases of NVs to ChatGPT based on prompt engineering. To our best knowledge, JVNV is the first speech corpus that uses LLMs to generate scripts. We technically validated the effectiveness of the proposed corpus-design method and demonstrated that JVNV has better phoneme coverage and significantly higher emotion recognizability than previous Japanese emotional corpora. Finally, we benchmark JVNV on the emotional TTS synthesis task. We propose a method using mixed representations of discrete codes and phonemes to represent NVs and verbal content, respectively. Experimental results demonstrated that the proposed mixed representation is consistently better than the phoneme for utterances mixed with NVs and verbal content. Finally, we showed the challenges of emotional TTS with NVs compared to normal TTS. We believe JVNV can serve as a valuable resource for future work in all relevant tasks.

APPENDIX A
DETAILS OF THE PROPOSED METHOD

A. EMOTION RECOGNIZABILITY SCORES

In this section, we describe the details of the emotion recognizability scores used for assisting high-quality script selection. Specifically, we train an emotion classifier to assist in selecting scripts with strong emotions, filtering out those samples with incorrect or vague emotions. We use WRIME [36] dataset as the training data, which is a well-known Japanese emotion analysis dataset, containing 43k crowd-sourcing data. Each sentence is annotated by one writer and three readers. The annotations follow Plutchik's [58] eight-category emotion schema on a four-point intensity scale, ranging from 0 to 3. With its annotations, we define the emotion of each sentence by weighted averaging annotators' scores:

$$s = \frac{1}{2}(s^{writer} + \frac{1}{n}\sum_{k=1}^{n} s^{reader_k}), \tag{6}$$

where n is the number of reader annotators. If the emotion score is larger than 1, we regard the sentence as having at least weak emotion with respect to an emotion category. Under this experimental setting, we use a RoBERTa-based model to train an emotion classifier with Binary Cross Entropy loss.

The WRIME dataset has two versions that are WRIME-ver1 and WRIME-ver2. WRIME-ver2 is a subset of WRIME-ver1 and asks another three annotators to annotate the emotion intensity together with the sentiment polarity. We also consider the extra annotations from WRIME-ver2 making the annotations more robust. After training, our emotion classifier can calculate the probability of a given sentence conveying a certain emotion. In this way, this probability can assist us in selecting those generated scripts with correct and strong emotions efficiently.

We implemented the classifier using PyTorch [59] and also utilized PyTorch Lightning[13] to build the pipeline in our experiments. We downloaded the model weights of Japanese RoBERTa from Huggingface Transformers [60] using `rinna/japanese-roberta-base`. The Adam [49] optimizer was used to optimize model parameters with a learning rate of 1×10^{-4} and a batch size of 64. L2-regularization with weight decay of 1×10^{-5} was also applied to avoid over-fitting. We trained our supervised model with 10 epochs. We employed early stopping when the validation loss did not improve for half of the total number of epochs.

B. LANGUAGE FLUENCY SCORES

We used exactly the same algorithm described in Salazar et al. [37] to compute the language fluency scores.

[13] https://pytorch-lightning.readthedocs.io/

As for the pretrained Japanese BERT model, we used the `cl-tohoku/bert-base-japanese-v2` from Huggingface Transformers for computing. We normalized the score by the length of each sentence.

APPENDIX B
DETAILS OF THE FORCED CHOICE TASK FOR EMOTION RECOGNIZABILITY

For the forced choice task conducted to evaluate the emotion recognizability of JVNV, JVNV-V, JTES, OGVC, and STUDIES, we employed 30 workers to evaluate the $60 \times 5 = 300$ utterances. Each worker rated 30 unique utterances equally sampled from the 5 corpora. As a result, each utterance had 3 answers, and we averaged the scores of each corpus as the final results. In the evaluation, we stressed that the workers should consider not only the content of the speech but also the prosody of it to choose the emotion. We also provided the definition of each emotion to help the workers. The workers could select from the 6 emotions and an additional "None of the above" (どれでもない) choice. We paid each worker 68 JPY based on the minimum hourly wage (1, 072 JPY) of Tokyo, Japan.

APPENDIX C
DETAILS OF THE NATURALNESS MOS TEST OF THE TTS BENCKMARK

We conducted 2 MOS tests to evaluate the 72 emotional samples and the 12 neutral samples from JSUT of each model, respectively. In the MOS test for the emotional samples, we evaluated $72 \times 6 = 432$ samples from NV+V, V, Phoneme, the HiFi-GAN vocoder, the Ground-truth (GT) data, and the NV-removed synthesized emotional samples of NV+V. 48 workers joined in the evaluation; each rated 33 utterances of which 6 were dummy samples used to familiarize the workers with the task. As a result, each utterance had 3 answers, and we averaged the scores of each corpus as the final results. The answers of the dummy samples were not counted in the final results. We paid each worker 76 JPY based on the minimum hourly wage (1, 072 JPY) of Tokyo, Japan.

In the MOS test for the neutral samples from JSUT, we evaluated $12 \times 5 = 60$ samples from NV+V, V, Phoneme, the HiFi-GAN vocoder, and the GT data. 9 workers joined in the evaluation; each rated 25 utterances of which 5 were dummy samples. The post-processing procedures are the same as the previous test. We paid each worker 60 JPY.

For all subjective evaluations, we described the potential risks to the workers with a document (Japanese).[14] All workers have to understand and consent to the content of the document to join in the tests.

ACKNOWLEDGMENT

(Detai Xin and Junfeng Jiang contributed equally to this work.)

[14] http://sarulab.sakura.ne.jp/rinri_jp.pdf

REFERENCES

[1] M. Tatham and K. Morton, *Expression in Speech: Analysis and Synthesis*. London, U.K.: Oxford Univ. Press, 2004.

[2] K. R. Scherer and U. Scherer, "Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the emotion recognition index," *J. Nonverbal Behav.*, vol. 35, no. 4, pp. 305–326, Dec. 2011.

[3] J. A. Hall, S. A. Andrzejewski, and J. E. Yopchick, "Psychosocial correlates of interpersonal sensitivity: A meta-analysis," *J. Nonverbal Behav.*, vol. 33, no. 3, pp. 149–180, Sep. 2009.

[4] A. Mehrabian, *Nonverbal Communication*. Evanston, IL, USA: Routledge, 2017.

[5] J. Trouvain and K. P. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *Proc. LREC Workshop Corpora Res. Emotion Sentiment Social Signals*, 2012, pp. 36–39.

[6] K. R. Scherer, "Affect bursts," in *Emotions: Essays Emotion Theory*, vol. 161. New York, NY, USA: Psychology Press, 1994, p. 196.

[7] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing," *Behav. Res. Methods*, vol. 40, no. 2, pp. 531–539, May 2008.

[8] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 6, pp. 2408–2412, Feb. 2010.

[9] C. F. Lima, S. L. Castro, and S. K. Scott, "When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1234–1245, Dec. 2013.

[10] D. Xin, S. Takamichi, and H. Saruwatari, "Exploring the effectiveness of self-supervised learning and classifier chains in emotion recognition of nonverbal vocalizations," in *Proc. ICML Expressive Vocalizations Workshop*, 2022. [Online]. Available: https://arxiv.org/html/2207.06958

[11] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T. A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, "Textless speech emotion conversion using discrete & decomposed representations," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2022, pp. 11200–11214.

[12] H. Zhang, X. Yu, and Y. Lin, "NSV-TTS: Non-speech vocalization modeling and transfer in emotional text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[13] D. Xin, S. Takamichi, A. Morimatsu, and H. Saruwatari, "Laughter synthesis using pseudo phonetic tokens with a large-scale in-the-wild laughter corpus," in *Proc. Interspeech*, Aug. 2023, pp. 17–21.

[14] P. Tzirakis, A. Baird, J. Brooks, C. Gagne, L. Kim, M. Opara, C. Gregory, J. Metrick, G. Boseck, V. Tiruvadi, B. Schuller, D. Keltner, and A. Cowen, "Large-scale nonverbal vocalization detection using transformers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[15] H.-T. Luong and J. Yamagishi, "LaughNet: Synthesizing laughter utterances from waveform silhouettes and a single laughter example," 2021, *arXiv:2110.04946*.

[16] A. Adigwe, N. Tits, K. El Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," 2018, *arXiv:1806.09514*.

[17] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Workshop Speech Synth.*, 2004, pp. 223–224.

[18] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *Proc. Conf. Oriental Chapter Int. Committee Coordination Standardization Speech Databases Assessment Techn. (O-COCOSDA)*, Oct. 2016, pp. 16–21.

[19] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "STUDIES: Corpus of Japanese empathetic dialogue speech towards friendly voice agent," in *Proc. Interspeech*, Sep. 2022, pp. 5155–5159.

[20] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, "Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment," *Acoust. Sci. Technol.*, vol. 33, no. 6, pp. 359–369, 2012.

[21] M. Higashiyama, K. Inui, and Y. Matsumoto, "Learning sentiment of nouns from selectional preferences of verbs and adjectives," in *Proc. 14th Annu. Meeting Assoc. Natural Lang. Process.*, 2008, pp. 584–587.

[22] D. Xin, S. Takamichi, and H. Saruwatari, "JNV corpus: A corpus of Japanese nonverbal vocalizations with diverse phrases and emotions," *Speech Commun.*, vol. 156, Jan. 2024, Art. no. 103004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639323001383

[23] P. Eckman, "Universal and cultural differences in facial expression of emotion," in *Proc. Nebraska Symp. Motivat.* Lincoln, Nebraska: Univ. Nebraska Press, 1972, pp. 207–248.

[24] K. Ito and L. Johnson. (2017). *The LJ Speech Dataset*. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[25] C. Veaux, J. Yamagishi, K. MacDonald. (1028). *Superseded-CSTR VCTK Corpus: English Multi-speaker Corpus for Cstr Voice Cloning Toolkit*. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443

[26] H. Francois and O. Boeffard, "Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem," in *Proc. 7th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 2001, pp. 829–832.

[27] B. Bozkurt, O. Ozturk, and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 2003, pp. 277–280.

[28] A. Krul, G. Damnati, F. Yvon, and T. Moudenc, "Corpus design based on the Kullback–Leibler divergence for text-to-speech synthesis application," in *Proc. Interspeech*, Sep. 2006, Paper no. 1647-Wed3BuP.2.

[29] D. Cadic, C. Boidin, and C. d'Alessandro, "Towards optimal TTS corpora," in *Proc. LREC*, Malta, U.K., 2010, pp. 99–104.

[30] T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, and Y. Shiga, "Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 1107–1116, May 2017.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[32] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *Trans. Mach. Learn. Res.*, Aug. 2022.

[33] F. Chenchah and Z. Lachiri, "Speech emotion recognition in acted and spontaneous context," *Proc. Comput. Sci.*, vol. 39, pp. 139–145, Jan. 2014.

[34] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1877–1901.

[35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[36] T. Kajiwara, C. Chu, N. Takemura, Y. Nakashima, and H. Nagahara, "WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2021, pp. 2095–2104.

[37] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2699–2712. [Online]. Available: https://aclanthology.org/2020.acl-main.240

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[39] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, Apr. 2003.

[40] C.-C. Hsu, "Synthesizing personalized non-speech vocalization from discrete speech representations," 2022, *arXiv:2206.12662*.

[41] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[42] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2020. [Online]. Available: https://iclr.cc/virtual/2021/poster/2919

[43] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[44] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 8067–8077.

[45] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," 2017, *arXiv:1711.00354*.

[46] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[47] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[48] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[51] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17022–17033. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

[52] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-H. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5866–5870.

[53] J. Gillick, W. Deng, K. Ryokai, and D. Bamman, "Robust laughter detection in noisy environments," in *Proc. Interspeech*, Aug. 2021, pp. 2481–2485.

[54] A. Lausen and K. Hammerschmidt, "Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters," *Humanities Social Sci. Commun.*, vol. 7, no. 1, pp. 1–17, Jun. 2020.

[55] M. Cohn, C.-Y. Chen, and Z. Yu, "A large-scale user study of an Alexa prize chatbot: Effect of TTS dynamism on perceived quality of social dialog," in *Proc. 20th Annu. SIGdial Meeting Discourse Dialogue*, 2019, pp. 293–306.

[56] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2507–2522, 2023.

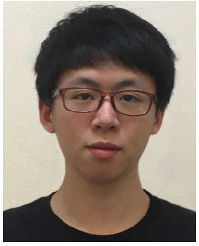
[57] B. Testa, Y. Xiao, H. Sharma, A. Gump, and A. Salekin, "Privacy against real-time speech emotion detection via acoustic adversarial evasion of machine learning," 2022, *arXiv:2211.09273*.

[58] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of Emotion*. Amsterdam, The Netherlands: Elsevier, 1980, pp. 3–33.

[59] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, vol. 32, 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf and https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[60] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Syst. Demonstrations*, 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

DETAI XIN (Member, IEEE) received the B.E. degree from Beihang University (former Beijing University of Aeronautics and Astronautics), Beijing, China, in 2019, and the M.E. degree from the Graduate School of Information Science and Technology, The University of Tokyo, Japan, in 2021, where he is currently pursuing the Ph.D. degree. He has published more than ten articles on speech synthesis and speech processing. His research interests include speech synthesis, speech processing, and deep learning. He received the IEEE SPS Tokyo Joint Chapter Student Award, in 2022.



JUNFENG JIANG received the B.S. degree from the School of Mathematics, Sun Yat-sen University, Guangzhou, China, in 2019, and the M.E. degree from the Graduate School of Information Science and Technology, The University of Tokyo, Japan, in 2022, where he is currently pursuing the Ph.D. degree, researching large language models, dialogue systems, and deep learning. He has published more than five articles in the field of natural language processing.

SHINNOSUKE TAKAMICHI (Member, IEEE) received the B.E. degree from the Nagaoka University of Technology, Nagaoka, Japan, in 2011, and the M.E. and Ph.D. degrees from the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Japan, in 2013 and 2016, respectively. He is currently a Lecturer with The University of Tokyo. He has received more than 20 paper/achievement awards, including the 2020 IEEE Signal Processing Society Young Author Best Paper Award.

YUKI SAITO (Member, IEEE) received the Ph.D. degree in information science and technology from the Graduate School of Information Science and Technology, The University of Tokyo, Japan, in 2021. His research interests include speech synthesis, voice conversion, and machine learning. He is a member of the Acoustical Society of Japan, IEEE SPS, and the Institute of Electronics, Information and Communication Engineers. He was a recipient of eight paper awards, including the 2020 IEEE SPS Young Author Best Paper Award.

AKIKO AIZAWA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from The University of Tokyo, in 1985, 1987, and 1990, respectively. She was a Visiting Researcher with the University of Illinois at Urbana–Champaign, from 1990 to 1992. Currently, she is the Vice Director General and a Professor with the National Institute of Informatics, Japan. She is also a Professor with the Graduate School of Information Science and Technology, The University of Tokyo, and the Graduate Institute for Advanced Studies (SOKENDAI). She is a Research Supervisor of the Strategic Basic Research Programs "Core Technologies for Trusted Quality AI Systems," funded by the Japan Science and Technology Agency (2020–2027). Her research interests include natural language understanding, dialogue systems, text-based content and media processing, and information retrieval. She has served as an organizer and a program committee member for related conferences and workshops. She has organized mathematical formula retrieval tasks at NTCIR-10, NTCIR-11, and NTCIR-12.

HIROSHI SARUWATARI (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Nagoya University, Nagoya, Japan, in 1991, 1993, and 2000, respectively. In 1993, he joined the SECOM IS Laboratory, Tokyo, Japan, and the Nara Institute of Science and Technology, Ikoma, Japan, in 2000. Since 2014, he has been a Professor with The University of Tokyo, Tokyo, Japan. His research interests include statistical audio signal processing, blind source separation, and speech enhancement. He has put his research into the world's first commercially available independent-component-analysis-based BSS microphone, in 2007. He was a recipient of several paper awards from IEICE, in 2001 and 2006; TAF, in 2004, 2009, 2012, and 2018; IEEE-IROS2005, in 2006; and APSIPA, in 2013 and 2018. He received the DOCOMO Mobile Science Award, in 2011; the Ichimura Award, in 2013; the Commendation for Science and Technology by the Minister of Education, in 2015; the Achievement Award from IEICE, in 2017; and the Hoko-Award, in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ. Since 2018, he has been an APSIPA Distinguished Lecturer.