# Emotional Speech with Nonverbal Vocalizations: Corpus Design, Synthesis, and Detection

*Detai Xin*

Graduate School of Information Science and Technology, The University of Tokyo, Japan.

detai_xin@ipc.i.u-tokyo.ac.jp

## Abstract

The main topic of this research proposal is emotional speech with nonverbal vocalizations. Nonverbal vocalizations (NVs) refer to short and affective vocalizations that contain rich emotional information. However, current research usually ignores this component. Therefore, this research aims to advance related research by developing new technologies of corpus design, synthesis, and detection of NVs. We first introduce the background and motivation of this research, and describe our recent progress in emotion recognition of NVs. Then, three future projects of this research are described in detail. Finally, the contributions and possible impacts of this research will be discussed.

**Index Terms**: Emotional Speech, Nonverbal Vocalizations, Emotion Recognition, Speech Synthesis

## 1. Introduction

### 1.1. Background

Emotion is a universal concept that exists across different cultures and languages [1, 2]. Human communication contains both verbal and nonverbal parts [3]. While the verbal part conveys essential linguistic information, the nonverbal part contains most emotional information, which indicates the internal intentions and mental states of speakers. Nonverbal expressions play an important role in human communication [4, 5], and can be expressed by vocal, facial, and hand expressions [6]. In human speech, the emotional nonverbal expression is called nonverbal vocalizations (NVs) or affect bursts, which refers to vocalizations containing no linguistic information like laughter, sobs, screams [7, 8]. They are relatively casual expressions and are usually not used in written languages [9]. Although the researches about NVs are in an initial stage, many works have shown the importance of NVs in emotion processing from various aspects like behavior [10], clinical treatments [11], and human development [12].

### 1.2. Research goals

As illustrated in Figure 1, this research consists of three parts: corpus design, TTS synthesis with NVs, and NVs detection.

1. **Corpus design**: Designing a Japanese emotional corpus with NVs.
2. **TTS with NVs**: Building a TTS system that can synthesize emotional speech with NVs based on the designed corpus.
3. **NVs detection**: Detecting emotional NVs in unlabeled data based on the designed corpus.

### 1.3. Motivations

The motivations of this research are discussed below.

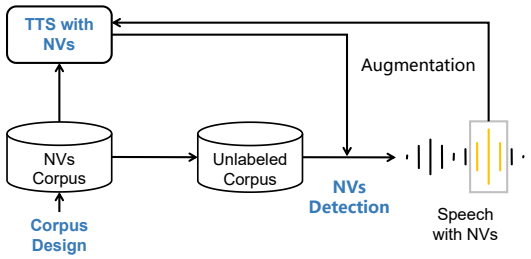


Figure 1: *Overview of this proposal. The designed corpus can be used in TTS with NVs and NVs detection. The synthesized and detected speech with NVs can then be used to augment other modules.*

First, NVs are ignored by most current research. It turns out that current work in emotional speech synthesis mainly focuses on generating emotional prosody for the verbal part of emotional speech. As mentioned in the previous section, emotional information in conversational speech is mainly contained in NVs, thus only the verbal part is not sufficient to express emotions, let alone other emotional information is contained in other modals like facial expressions and gestures [13].

Second, resources of emotional speech with NVs are insufficient. Previous NVs corpora, though with high quality and diversity, have a relatively small size (usually only 100 recordings) [8, 14, 15], since the purpose of these corpora is more of analysis than synthesis. Such corpora are not suitable for powerful data-driven machine-learning methods like deep neural networks (DNNs). Furthermore, the language of these corpora is mainly English, which makes it difficult to conduct multilingual experiments.

Third, there are many theoretical and technical challenges in synthesizing NVs. Several questions about NVs still remain unsolved. For example, is there any individuality existing in NVs? Although Pisanski et al. showed that the F0 difference exists in both verbal and nonverbal vocalizations [16], it is still unclear whether speakers have individual preferences on the phonetic tokens of NVs. This question is important for speech synthesis since the researchers in this field usually want to model the personality of each speaker.

Finally, it is helpful to develop NVs detection technologies so that large-scale unlabeled data can be utilized. Since emotional speech usually exists in spoken languages, it is important to consider the authenticity of the datasets [17]. Furthermore, the detected data can also be used to augment the TTS model to improve its performance. However, directly annotating NVs in spontaneous speech is difficult, hence it is reasonable to detect speech with NVs in unlabeled corpora.

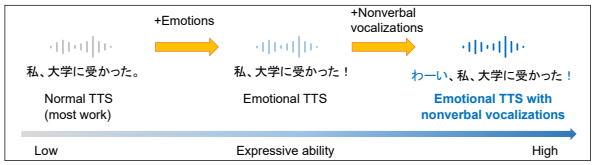


Figure 2: *The proposed emotional TTS system with NVs compared to previous work.*

## 2. Emotion Recognition of NVs

In our recent research, we explored the effectiveness of self-supervised learning (SSL) and classifier chain (CC) in emotion recognition of NVs [18]. In the proposed method, we used SSL to extract features from NVs and used a CC to model label dependency between different emotion labels. Experimental results demonstrated that SSL models trained on corpora with mainly verbal speech can be used for NVs, and it was beneficial to model the label dependency between emotion labels. Our proposed system obtained a mean concordance correlation coefficient (CCC) of $0.725$ in the validation set and $0.739$ in the test set, while the best baseline method only obtained $0.554$ in the validation set, which was in the 1st-place among all other systems.

## 3. Research plan

### 3.1. Corpus design

In this project, a Japanese emotional corpus with nonverbal vocalizations will be made.

**Methodology** Possible NVs in Japanese will first be collected by crowd-sourcing. After that, these vocalizations will be added to phoneme-balanced texts. The emotions used in the corpus will be selected based on Russell's circumplex model to a cover wide range of emotions [19].

**Evaluations** To evaluate the corpus, not only the valence and arousal but also the recognition accuracy and authenticity of each emotion will be evaluated by subjective tests to clarify the quality of the corpus [14].

**Challenges** First, NVs are highly diverse in phonetic tokens and meanings [9], so how to control the content of each NV will become a key challenge of this project. Also, since each speaker has his own style in uttering NVs, how to model and control individuality is also a challenge.

### 3.2. Emotional TTS with NVs

In this project, the corpus built in the first project will be used to train a Japanese emotional TTS system that can synthesize emotional speech with NVs. As Figure 2 illustrates, emotional TTS systems with NVs can synthesize speech with higher expressive ability compared to most previous work.

**Methodology** Two possible methods: (1) insertion [20] and (2) textless synthesis [21] can be used to synthesize NVs. These two methods and other techniques will be used in the experiments to explore the best method for synthesizing emotional speech with Japanese NVs.

**Evaluations** In the evaluations, the speech naturalness and the accuracy of emotion recognition will firstly be evaluated. Secondly, the emotional speech without nonverbal vocalizations will also be synthesized and evaluated to elucidate the effectiveness of nonverbal vocalizations.

**Challenges** The key challenge of this project is that NVs are difficult to be represented by texts. Most current speech synthesis systems rely on text input to work, but this is not applicable for NVs synthesis.

### 3.3. Emotional NVs detection

In this project, the corpus designed in the first project will be used to train a parametric model to detect NVs in unlabeled speech.

**Methodology** DNN-based methods will be used to discover speech with NVs [22, 23]. Noise-robustness, which is important for such NVs detection systems, will also be considered by training the system on real-world large-scale data [24]. Intuitively the TTS system with NVs and the NVs detection system are complementary since the synthesized speech can be used to train the detection model, and vice versa.

**Evaluations** Standard metrics like F score will be used to evaluate the system. Besides, the noise-robustness of the system will be evaluated by feeding noisy speech into the system. Finally, the benefits of data augmentation will also be evaluated in both the detection and the TTS systems.

**Challenges** The key challenge of this project is improving robustness of the proposed method. Since NVs are highly diverse in phonetic tokens, unseen NVs might have a bad influence on the robustness of the developed system.

## 4. Discuss

### 4.1. Main Contributions

The main contributions of this research are:

- Advancing research on NVs by making a Japanese NVs corpus.
- Improving the expressive ability of current emotional TTS systems by synthesizing emotional verbal speech and NVs together.
- Developing emotional NVs detection technologies to stimulate unsupervised learning methods in this field.
- Exploring and clarifying both technical and theoretical problems in this field.

### 4.2. Possible impacts

This research can have several impacts on both industries and academia. Possible impacts are listed as follows:

- **Human-computer interaction (HCI)**: Incorporating emotional functions is a popular and common topic in HCI and human-robot interaction (HRI) [25, 26]. It includes two tasks: (1) recognizing users' emotions [27] and (2) generating emotional expressions [28, 29]. In this research, not only the NVs detection technologies can improve the recognition accuracy but also the TTS system with NVs can improve the expressive ability of the agents in HCI or HRI.
- **Speech emotion recognition (SER)**: SER aims to recognize emotions from a given speech. Current state-of-the-art methods have low accuracy on popular datasets without NVs even using DNNs [30, 31]. Previous work has shown that adding NVs in speech can substantially improve recognition accuracy [32], thus the proposed corpus might benefit this task.
- **Other fields**: NVs are also studied in other fields like psychology [33], behavior [10], and clinical treatment [11]. The proposed synthesis system may benefit these fields like previous work [34].

**Acknowledgements:** This work was supported by JST SPRING, Grant Number JPMJSP2108.

# 5. References


[1] P. Eckman, "Universal and cultural differences in facial expression of emotion," in *Nebraska symposium on motivation*. University of Nebraska Press Lincoln, 1972.

[2] P. E. Ekman and R. J. Davidson, *The nature of emotion: Fundamental questions*. Oxford University Press, 1994.

[3] A. Mehrabian, *Nonverbal communication*. Routledge, 2017.

[4] 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," *Journal of Nonverbal Behavior*, vol. 35, no. 4, pp. 305–326, 2011.

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

[6] M. Tatham and K. Morton, *Expression in speech: analysis and synthesis*. Oxford University Press on Demand, 2004.

[7] K. R. Scherer, "Affect bursts," *Emotions: Essays on emotion theory*, vol. 161, p. 196, 1994.

[8] 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," *Behavior research methods*, vol. 40, no. 2, pp. 531–539, 2008.

[9] J. Trouvain and K. P. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *Proceedings of the LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals*. Citeseer, 2012, pp. 36–39.

[10] P. E. Bestelmeyer, J. Rouger, L. M. DeBruine, and P. Belin, "Auditory adaptation in vocal affect perception," *Cognition*, vol. 117, no. 2, pp. 217–223, 2010.

[11] D. Dellacherie, D. Hasboun, M. Baulac, P. Belin, and S. Samson, "Impaired recognition of fear in voices and reduced anxiety after unilateral temporal lobe resection," *Neuropsychologia*, vol. 49, no. 4, pp. 618–629, 2011.

[12] E. M. Hunter, L. H. Phillips, and S. E. MacPherson, "Effects of age on cross-modal emotion perception." *Psychology and Aging*, vol. 25, no. 4, p. 779, 2010.

[13] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

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

[15] N. Holz *et al.*, "The variably intense vocalizations of affect and emotion (vivae) corpus prompts new perspective on nonspeech perception." *Emotion*, 2022.

[16] K. Pisanski, J. Raine, and D. Reby, "Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries," *Royal Society open science*, vol. 7, no. 2, p. 191642, 2020.

[17] V. Aubergé, N. Audibert, and A. Rilliard, "Why and how to control the authentic emotional speech corpora," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[18] X. Detai, T. Shinnosuke, and S. Hiroshi, "Exploring the effectiveness of self-supervised learning and classifier chains in emotion recognition of nonverbal vocalizations," in *ICML Expressive Vocalizations Workshop*, 2022.

[19] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, 1980.

[20] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *The 10th ISCA Speech Synthesis Workshop*, 2019.

[21] 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 decomposed and discrete representations," *arXiv preprint arXiv:2111.07402*, 2021.

[22] V. Nallanthighal and H. Strik, "Deep sensing of breathing signal during conversational speech," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association, Interspeech*. Graz, Austria:[Sn], 2019, pp. 4110–4114.

[23] S. Condron, G. Clarke, A. Klementiev, D. Morse-Kopp, J. Parry, and D. Palaz, "Non-verbal vocalisation and laughter detection using sequence-to-sequence models and multi-label training," *Proc. Interspeech*, pp. 2506–2510, 2021.

[24] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.

[25] C. Breazeal and R. Brooks, "Robot emotion: A functional perspective," *Who needs emotions*, pp. 271–310, 2005.

[26] R. Beale and C. Peter, "The role of affect and emotion in hci," in *Affect and emotion in human-computer interaction*. Springer, 2008, pp. 1–11.

[27] S. Brave and C. Nass, "Emotion in human-computer interaction," *Human-computer interaction fundamentals*, vol. 20094635, no. 53-68, p. 4, 2009.

[28] L. Bechade, G. D. Duplessis, and L. Devillers, "Empirical study of humor support in social human-robot interaction," in *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 2016, pp. 305–316.

[29] N. Mirnig, G. Stollnberger, M. Giuliani, and M. Tscheligi, "Elements of humor: How humans perceive verbal and non-verbal aspects of humorous robot behavior," in *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 2017, pp. 211–212.

[30] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "Copypaste: An augmentation method for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6324–6328.

[31] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

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

[33] M. Schröder, "Experimental study of affect bursts," *Speech communication*, vol. 40, no. 1-2, pp. 99–116, 2003.

[34] A. Anikin, "Soundgen: an open-source tool for synthesizing nonverbal vocalizations," *Behavior research methods*, vol. 51, no. 2, pp. 778–792, 2019.