Interpretable Control for Emotional Text-to-Speech System toward Development of Sympathetic Educational-Support Robots

Jingyi Feng*, Tomohiro Yoshikawa† and Tomoki Toda‡
* Nagoya University, Nagoya, Japan
E-mail: feng.jingyi@g.sp.m.is.nagoya-u.ac.jp
† Suzuka University of Medical Science, Suzuka, Japan
E-mail: yoshi@suzuka-u.ac.jp
‡ Nagoya University, Nagoya, Japan
E-mail: tomoki@icts.nagoya-u.ac.jp

***Abstract*—With increasing aging, sympathetic educational-support robots (sympathetic robots) that can assist young learners have been attracting the attention of researchers. However, the visually-driven interactions (emotional body motions and facial expressions of robots) that have been studied don't provide sufficiently good interactivity and may distract learners. In this paper, we develop an emotional text-to-speech (TTS) synthesis system to be implemented for sympathetic robots. As a speech system oriented to be equipped with sympathetic robots that can speak and express their own emotions by voice, the control of variable emotional expression in the synthesized speech during interaction needs to be fully considered. Towards the development of sympathetic robots providing sufficiently good interactivity, we propose an emotional TTS system architecture using both a global style tokens (GSTs) module and a set of arousal-valence tokens to flexibly control the emotional expression of synthesized speech by two interpretable annotations, categorical and dimensional, respectively. The experimental results demonstrate that our model can flexibly control the emotional expression of the synthesized speech and can satisfy the demand of the application to sympathetic robots.**

Index Terms: educational-support robots, speech synthesis, prosody control, human-robot interaction

I. Introduction

In recent years, educational support robots that can support learners have attracted increasing attention. A problem with these educational support robots is that learners find the robots' behavior too monotonous, which makes collaborative learning with robots rather boring [1]. To address this problem, previous research has proposed a sympathy expressions method [2], where the robot expresses emotions similar to those of the learner's emotions, thus resonating with the learner. The sympathy expressions method (SEM) is based on Russell's circumplex model (also called emotional space [3]) for expressing emotion, which is based on some of the learner's behaviors during the learning process (e.g. the number of correct answers or the time taken to answer the questions.). The educational-support robots equipped with the SEM are called sympathetic educational support robots, abbreviated as sympathetic robots in this paper.

With sympathetic robots, emotional expressions based on body motions and facial expressions were discussed [4] in conventional studies. However, the robot's body motions and facial expressions are visually driven interactions. The learner needs to take their eyes off the tasks they are focusing on to complete the interaction process. This may lead to interruptions in the learning process and distract the learner's attention. And based on past research in human-robot interaction, speech is also an effective form of emotional expression for robots. In the case of learning support robots, auditory-driven interaction can provide a sense of companionship while ensuring the continuity of the learning process. Therefore, we implement a speech synthesis system for the development of sympathetic robots.

The current development of the neural speech synthesis model has been accompanied by a large amount of research focusing on high-quality speech synthesis. This allows interactive devices to use high-quality speech to complete the interaction process. To achieve more fluid interaction scenarios, emotional speech synthesis systems are also widely used in the process of human-robot interaction [5]. In these studies, interactive devices have achieved not only communication of content but also speech-based communication of emotion, which enables human-robot interaction closer to natural communication. In our study, we focus on educational-support robots that extend the function of sympathy expression based on emotional expression. Thus, we attempt to use synthesized speech to achieve the functions of content delivery, emotion expression, and sympathy expression in the interaction process.

In this paper, towards the development of flexible emotion control required for sympathetic function, we implement a speech synthesis system based on the need.

1) Emotional control of synthesized speech using emotional labels (categorical annotations).
2) Emotional change of synthesized speech using arousal-valence values (dimensional annotations) based on emotion space.
3) High-quality emotional speech synthesis by using a

high-performance speech synthesis framework Fastspeech2 [6] as a backbone.

II. Relate work

A. Sympathy Expression Method and Sympathetic Robots

The sympathy expression method (SEM) is an approach proposed to alleviate the monotony of the robot for the learner during an interaction, which is a method of making the robot capable of sympathy with the learner based on Russel's circumplex model [3]. In Fig. 1, there are 12 categories of emotions, and each emotion has strong and weak aspects in the conventional SEM. During the emotional expressive processing, the emotions are expressed using positive ($\vec{A}$) and negative ($\vec{B}$) vectors based on a judgment (e.g. fast or slow writing time, correct or incorrect answers to questions). Equipped with the SEM, the sympathetic educational-support robots, called sympathetic robots, are constructed.

In conventional research, two emotional expression methods of sympathetic robots through facial expressions and body motions were studied. Twenty-four (12 emotional categories and each emotion have 2 levels "strong" and "weak") facial expressions and body motions are designed manually in each study and are used for the emotional expression of the sympathetic robot. The sympathetic robot can express its emotion (e.g. Fig. 2) and sympathy to the learner, with the SEM and some kinds of emotional expression methods in the learning interaction processing.

However, while visually-driven interaction somewhat alleviates the monotony of robot companionship, there is a risk of interrupting the learning process of the learner. Therefore, our study uses a more natural method, speech interaction, for semantic expression, emotional expression and sympathetic expression. The use of auditory-driven interaction allows the learner to live with a sufficient sense of companionship while reducing learner gaze shift so that can focus more on the current learning process. In addition, the auditory-driven interaction method is the basis for further implementation of audiovisual integration of interaction of sympathetic robots.

B. Emotional Control Text-to-Speech

With the development of end-to-end models of deep neural networks, such as Tacotron2 [7], it is possible to obtain high-quality natural synthesized speech. Fastspeech [8] and Fastspeech2 [6] are the non-autoregressive TTS models which improve training speed and show fast synthesis speed in the inference. These neural TTS models generate a mel-spectrogram from the text and then convert the mel-spectrogram into a speech waveform using vocoders such as Griffin-Lim [9] and HiFiGAN [10]. However, the emotional expression of the synthesized speech of these models depends on the training dataset but has little control over the emotional expression of each sentence.

To address this issue, the process of the emotional control module is added to the model of speech synthesis. The GSTs module [11] has been jointly trained with Tacotron, which is a set of style tokens containing acoustic variations. The module

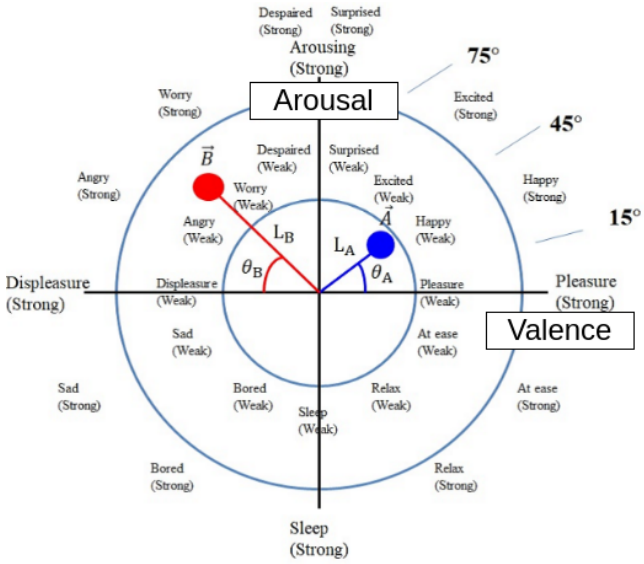


Fig. 1: *The sympathy expression method is based on emotional space (Russell's circumplex model). [2]*

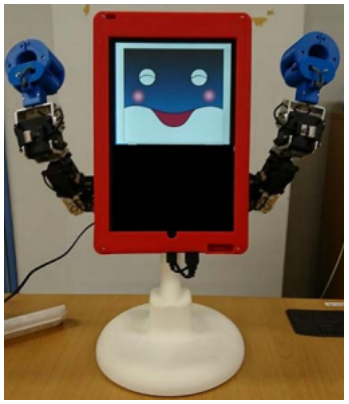


Fig. 2: *An educational-support robot express "happy" by the combination of facial expression and body motion.*

takes a reference speech as input and outputs the corresponding style embedding as an input feature to the TTS model. In Sivaprasad et al. [12] an interpretable emotional expression intensity control method was proposed. In this study, arousal-valence values are input into a prosody control block, which are used as weights of learnable tokens (arousal tokens and valence tokens). Two token learned arousal-valence-related features during the training and the block input arousal-valence values can change during the inference to change the emotional expression intensity of synthesized speech, achieving an interpretable emotional control of emotional speech synthesis by using the emotional space-based arousal valence.

Toward the robotic study, the demand for smooth human-robot interaction should be considered based on conventional speech synthesis systems. The single module of emotion expression control in past studies does not accomplish both emotional expression and sympathy expression (emotion change) of the sympathetic robots. Thus we attempt to combine the emotion category control and emotion expression intensity control modules to implement an emotional speech synthesis

model to meet the diverse emotional expression demand of sympathetic robots.

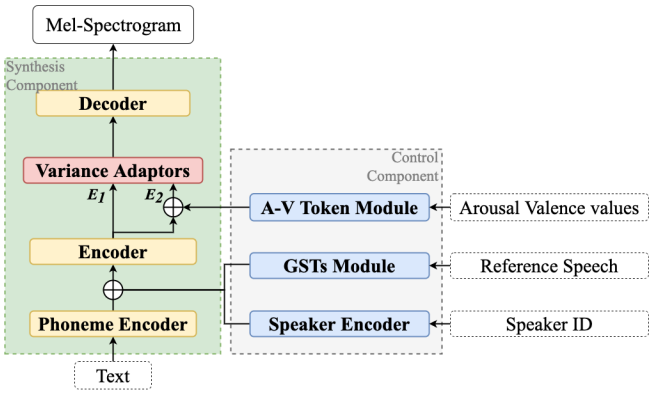
III. Proposed Method

Considering the past research, the functionalities of a speech synthesis system for sympathetic robots are clarified, (1) synthesize emotional speech with a specified emotional category, and (2) synthesize emotional speech with the controllable intensity (”strong” and ”weak” levels) of emotional expression. Due to the limited and small number of emotional category labels usually available in the corpus for training emotional speech synthesis models, it is difficult to achieve flexible and diverse emotion variations in synthesized speech using only discrete emotion labels for training.

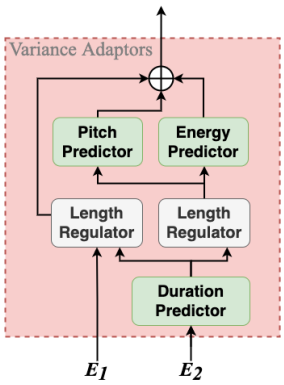
To address this problem, not only the categorical but also dimensional annotations of the emotion are used in the proposed model training. The arousal and valence values are used as the dimensional annotations in our study. Arousal is the level of autonomic activation that an event creates, and ranges from calm (or low) to excited (or high) and valence is the level of pleasantness that an event generates and is defined along a continuum from negative to positive. They are the vertical and horizontal axes of Russel’s circumplex model (emotional space circle, in Fig. 1) respectively, which can be used as a measurement of SEM. All arousal and valence values are provided by the 5-point self-assessment method [13], which directly measures the valence and arousal associated with a person’s affective reaction to a wide variety of stimuli (in this study the stimuli is speech). In the proposed model, the emotional categories information controls the specified emotional category of synthesized speech and the arousal-valence value controls the emotional expression intensity of synthesized speech instead of ”strong” and ”weak” levels’ labels.

The proposed model (shown in Fig. 3) consists of two main components, a synthesis component and a control component. Fastspeech2 [8] is used as the synthesis component, which is a non-autoregressive text to mel-spectrogram generation model that ensures high-quality synthesized mel-spectrogram synthesis. On the other hand, for the control component, as mentioned earlier, the emotional expression and sympathetic expression of sympathetic robots are by using the arousal-valence value based on emotional space. Also based on the explicit need for emotional expression, emotion categories are introduced into the speech emotion control process. Furthermore, with the multi-speaker dataset being used for training, the speaker identifies control also considered in the proposed model. These three modules are the control component, GSTs module [11] and arousal-valence (A-V) tokens module [12], and a speaker encoder implemented in the model structure. They can embed emotion category information, arousal-valence values information (which is related to the intensity of emotional expression), and speaker information for speech synthesis, respectively. The *E1* and *E2* embeddings sequences in Fig. 3 are considered as the input of the variance adaptors module, which determines the pitch, duration and energy and impacts the emotional expression of the synthesized speech. By the emotional categories control and the arousal-valence values (interpretable in emotion space) control, the proposed structure achieves interpretable emotional control.

There are two stages in training. In stage 1, a high-quality normal speech dataset is used to train the pretrained Fastspeech2 model for high-quality speech synthesis. In stage 2, an emotional speech dataset is used for learning the emotional expression features. To address insufficient quality and the limited size of the emotional speech dataset, the pre-trained Fastspeech2 parameters are input into the proposed structure and the phone encoder, encoder and decoder are frozen during the stage 2 training, to ensure high-quality synthesis.



(a) *the overview of model structure*



(b) *The variance adaptors*

Fig. 3: *The proposed emotional TTS model architecture*

IV. Experiments

A. Experimental conditions

We used LibriTTS corpus [14] as a high-quality normal speech dataset and IEMOCAP corpus [15] as an emotional speech dataset in the training. Four emotions (neutral, angry, sad, happy) speech samples from the 10 speakers (5 male speakers and 5 female speakers) in the IEMOCAP corpus were used for training. The arousal-valence values were provided for each sample in the IEMOCAP corpus by a 5-point self-assessment method [13] also used. The ”Fastspeech2+GSTs” and the ”Fastspeech2+A-V tokens” were trained as the baseline

TABLE I: The MOS with 95% confidence intervals

Naturalness	**MOS**
Fastspeech2	**3.61 ± 0.11**
Proposed model	3.58 ± 0.08
Emotional Category Assess	**Accuracy in %**
Fastspeech2+GSTs	62.5%
Proposed model	**64.6%**
Expressive Intensity Assess	**Accuracy in %**
Fastspeech2+A-V tokens	80.6%
Proposed model	**84.3%**

models, without the arousal-valence tokens module and GSTs module in the control component, respectively.

For the evaluation of the proposed model, three assessments, (1) naturalness, (2) emotional category control, and (3) expressive intensity control, were conducted. In the naturalness evaluation, the pre-trained Fastspeech2 is used as the baseline to compare voice quality with the proposed model. Forty samples were synthesized for the experiment from the baseline and the proposal, respectively, by texts which are from the IEMOCAP corpus unseen in training. In the emotional category control assessment, the emotional TTS model "Fastspeech+GSTs" is used as a baseline. Twelve samples of each emotional category, in total, 48 samples were synthesized from each model (baseline and proposal) by texts and emotional categories input data from the IEMOCAP corpus unseen in training. The subjects were asked to listen to synthesized speech samples and select the emotion category corresponding to this sample. In the expressive intensity control assessment, the "Fastspeech2+A-V tokens" is considered as the baseline. Two synthesized samples were taken as a group synthesized by related synthesis conditions. One input for inference is called "original input", which is a sample from the IEMOCAP corpus unseen in training. And the other input is called "manual input", gets by modifying the arousal-valence value of the "original input" to a selected suitable arousal-valence value based on the emotional space. Different arousal-valence values will determine the 'weak' and 'strong' expressions of synthesized speech. Eight groups of each emotional category (without 'neutral'), in total, 48 samples of each model were used in the experiment. The subjects were asked to select the samples with stronger expressions in the samples of different intensities ('weak' and 'strong').

Thirteen subjects participated in each evaluation experiment. HiFi-GAN [10] was used as a vocoder for converting mel-spectrogram to waveform in this experiment.

B. Results

The results of three experiments are in Table I.

Naturalness Evaluation: To evaluate the naturalness of the synthesized speech, the subject is asked to make quality judgments about the naturalness using the Mean Opinion Score (MOS). The MOS on 1 to 5 (1 is "completely unnatural" and 5 is "completely natural") is used as a measure in the subjective test of naturalness. As the results in Table I, we find that our model performs similarly in the naturalness of the synthesized speech compared to an original Fastspeech2.

Emotional Category Control Assess: In this assessment, the emotional similarity of synthesized speech is evaluated. The listeners were asked to select the emotion category that was closest to the emotion expressed in the heard synthesized speech sample. Also, subjects were asked to judge the emotional category without considering the textual content contained in the synthesized speech samples. The accuracy of the selected results is used as the measure. The results indicate that the emotional expression categories of the synthesized speech of the proposed model and the baseline model are recognizable and the proposed model is better than the baseline model.

Expressive Intensity Control Assess: In the expressive intensity assessment, subjects were asked to distinguish the differences in the intensity of emotional expression in a sample group (including a "strong" sample and a "weak" sample). The accuracy of the "strong" sample being selected is used as the measure. The result shows that the different intensity expression control by arousal-valence values change is feasible. And the proposed model performed better than the baseline in this assessment.

V. Conclusions and Future Work

Our work implements an emotional speech synthesis model for the sympathetic educational-support robots. The Fastspeech2 model is used as the backbone which ensures the synthesis speed and the quality of the synthesized speech. The GSTs module and arousal-valence tokens module provide the encoding of emotional categorical attributes and dimensional attributes, respectively, to achieve emotional control during the synthesis. The discrete emotion categorical labels and arousal-valence values based on interpretable emotion space are used to control the emotional expression of synthesized speech. The evaluated results show that the proposed model can meet the requirements for applying a sympathetic educational-support robot. The proposed model can control both category and intensity of emotional expression, which achieve a flexible method of emotional expression for the sympathetic educational-support robot during the interaction.

In future research, we will evaluate the performance of a sympathetic robot equipped with the proposed speech synthesis system in a real interaction scenario. Moreover, improving the interpretable expressive performance of the emotional synthesis speech system is one of our future research purposes.

Acknowledgment

This work was supported by JST CREST Grant Number JPMJCR19A3, JSPS KAKENHI Grant Number JP21H05054, and JST SPRING Grant Number JPMJSP2125.

References

[1] F. Jimenez, M. Kanoh, T. Yoshikawa, and T. Furuhashi, "Effect of collaborative learning with robot that prompts constructive interaction," in *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2014, pp. 2983–2988.

[2] F. Jimenez, T. Yoshikawa, T. Furuhashi, and M. Kanoh, "A proposal of model of emotional expressions for robot learning with human," in *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Istanbul, Turkey: IEEE, Aug. 2015, pp. 1–5.

[3] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980, place: US Publisher: American Psychological Association.

[4] Y. Tanizaki, F. Jimenez, T. Yoshikawa, and T. Furuhashi, "Impression investigation of educational support robots using sympathy expression method by body movement and facial expression," in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, 2018, pp. 1254–1258.

[5] J. Crumpton and C. L. Bethel, "A survey of using vocal prosody to convey emotion in robot speech," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 271–285, 2016.

[6] 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 *International Conference on Learning Representations*, 2021.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4779–4783, iSSN: 2379-190X.

[8] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019.

[9] D. W. Griffin, Jae, S. Lim, and S. Member, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoustics, Speech, Signal Process*, pp. 236–242, 1984.

[10] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033.

[11] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Jul. 2018, pp. 5180–5189, iSSN: 2640-3498.

[12] S. Sivaprasad, S. Kosgi, and V. Gandhi, "Emotional Prosody Control for Speech Generation," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 4653–4657.

[13] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[14] H. Zen, V.-T. Dang, R. A. J. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *ArXiv*, vol. abs/1904.02882, 2019.

[15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.