Thu-P-OS-10-3

STUDIES: Corpus of Japanese Empathetic Dialogue Speech Towards Friendly Voice Agent

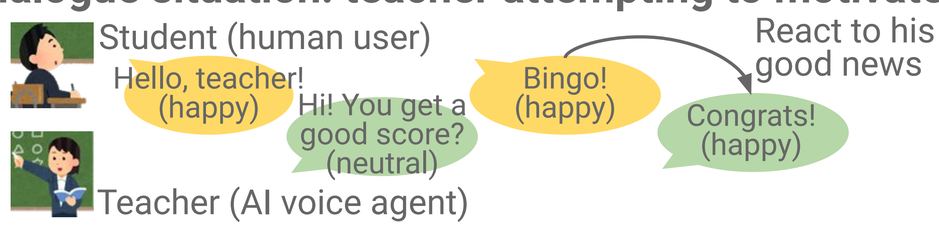
Yuki Saito[1], Yuto Nishimura[1], Shinnosuke Takamichi[1], Kentaro Tachibana[2], and Hiroshi Saruwatari[1]
(1: The University of Tokyo, Japan, 2: LINE Corp.)

Synopsis: corpus for empathetic dialogue speech synthesis

- **What is "empathy"?**
  - Definition: active attempt to get inside the interlocutor [1]
    - e.g., doctor–patient [2], teacher–student [3]
    - Essentials: both text contents [4] and speech (e.g., prosody) [5]
  - Different from "sympathy"
    - Passive attempt to share the speaker's feeling

- **Empathetic dialogue speech synthesis**
  - For AI voice agent that empathizes with humans
  - Corpus for this purpose: uncultivated so far…

- **Dialogue situation: teacher attempting to motivate students**



Get it now!
http://sython.org/Corpus/STUDIES

- **STUDIES: Japanese empathetic dialogue speech synthesis**
  - Professional speakers, emotion labels, and short/long dialogues
  - Speech synthesis experiment in this poster
  - Opensourced for research purpose only (the above QR code!)

Methodology for corpus construction

Dialogue scenario

- **Two issues of chat-oriented dialogue with specific persona [6]**
  a. Inconsistent persona of the agent [7]
  b. Lack of long-term memory (i.e., forgetting chat history) [8]

- **Predetermining the agent's persona**

  a. Female in her early twenties
  b. Tokyo dialect
  c. Cheerful and friendly personality
  d. etc. (see our paper for the detail)

- **Two types of dialogue durations**
  - Long (10~20 turns): challenging
  - Short (4 turns): easier than "Long"
    - User → Agent → User → Agent

Crowdsourcing chat dialogue lines

- **Instruction to crowdworkers**
  a. Situation: teacher-student dialogue
  b. Agent's persona: see left.
  c. Dialogue durations: 4 or 10~20 turns
  d. etc. (see our paper for the detail)

- **Crowdworkers' actions**
  - Making teacher-student dialogue lines
  - Labeling the fictional characters' emotion

- **Hiring crowdworkers**
  a. Macrotask crowdsourcing for Long dialogues
     i. Hiring crowdworkers on the basis of their skills
     ii. Diffucult task for the crowdworkers
  b. Microtask crowdsourcing for Short dialogues
     i. Hiring crowdworkers without approval
     ii. Simple task for the crowdworkers

Voice recording

- **Prior to recording**
  a. Manual revision of dialogue lines
     i. Unnatural in grammar, syntax, and/or dialogue
     ii. Inconsistent persona

- **Voice recording**
  - Three professional speakers
    - Teacher and two students
  - Speaker-wise recording in studio
    - Due to the COVID-19 pademic
  - Recording two subsets
    - Dialogue lines (Long and Short)
    - Phoneme-balance sentences with emotion labels (ITA [9])
      - Neutral, Happy, Angry, and Sad
      - Coutermeasure for data imbalance

Corpus analysis and speech synthesis experiments

Corpus analysis

- **Corpus specification**
  - #utterances and #hours in each subset

Speaker	Long	Short	ITA	Total (hours)
Teacher	1,201	1,440	724	3,365 (5.0)
Male student	609	720	N/A	1,329 (1.5)
Female student	621	720	N/A	1,341 (1.7)

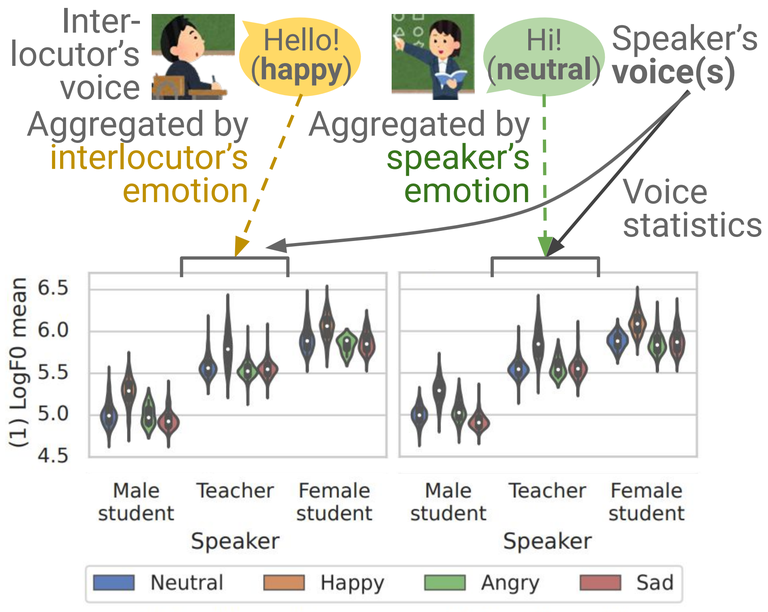
  - #utterances for each emotion label

Speaker	Neutral	Happy	Sad	Angry
Teacher	2,220	715	312	118
Male student	351	381	329	268
Female student	300	417	340	284

- 8 hrs, 6,000 utts in total
  - Sufficient for neural TTS
- Larger than existing relevant corpora (e.g., UUDB [10])

- The size is well balanced.
  - Except teacher's "angry"
- Dialogue in which the teacher gets angry is rare.
  - "Empathy" ≠ act to share angry emotion

- **F0 statistics distributions**



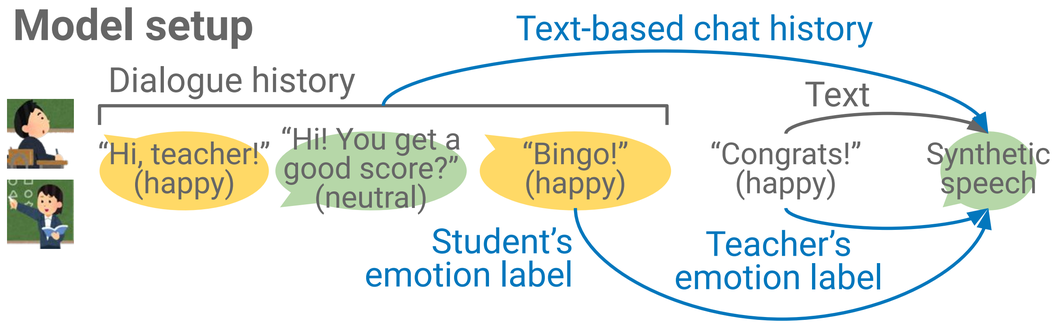
- Statistics of speaker's F0
  - Aggregated by two ways
  - (see our paper for other analysis results)

- Similar distributions between different aggregations
  - Possibility to synthesize speech using not only the speaker's own emotion but also interlocutor's one

- Speech samples: 👉👉

Speech synthesis experiments

- **Model setup**



Acoustic model for TTS	FastSpeech 2 [11]
Chat history encoder	BERT-based [12]
Emotion label format	One-hot vector → lookup emb.
Training/Evaluation data	3,600/210 utterances (teacher)

- **Subjective evaluation (MOS on naturalness)**
  - Utterance level (i.e., teacher's synthetic speech only)
  - Dialogue level (i.e., student's natural speech and teacher's synthetic speech)

Emotion label	Text history	Utt-level	Dial-level
(baseline)		3.42	3.20
Teacher		3.43	**3.38**
Student		3.38	3.29
	✓	3.32	**3.37**
Teacher	✓	3.38	3.34
Student	✓	3.43	**3.41**

MOSs for reconst. speech: 4.08 and 4.33

- **Cases better than baseline (ordinary TTS)**
  - Teacher's emotion
  - **Text history**
  - **Student's emotion & text history**
- Importance of **dialogue context** (history and situation) in empathetic dialogue speech synth.

Reference

[1] M. H. Davis et al., Routedge, 1996.
[2] M. Hojat et al., Journal of Health and Human Services Administration, 2009.
[3] C. Warren et al., The Urban Review, 2015.
[4] H. Rashkin et al., ACL, 2019.
[5] C. Regenbogen et al., Cognition and Emotion, 2012.
[6] S. Zhang et al., ACL, 2018.
[7] J. Li et al., ACL, 2016.
[8] O. Vinyals et al., ICML, 2015.
[9] J. Koguchi et al., 2021.
[10] H. Mori et al., Speech Communication, 2011.
[11] Y. Ren et al., ICLR, 2021.
[12] H. Guo et al., SLT, 2021.

Saruwatari-Koyama Lab., The University of Tokyo, Japan.

Interspeech 2022, Korea

This is the joint work between The University of Tokyo & LINE Corp. Japan.