Thu-P5.21 (#1098)

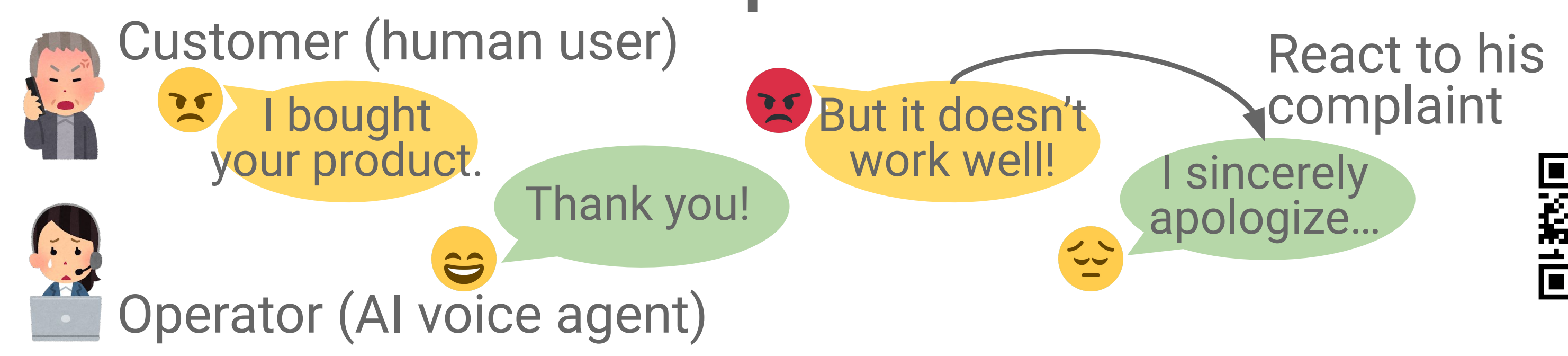
CALLS: Japanese Empathetic Dialogue Speech Corpus of Complaint Handling and Attentive Listening in Customer Center

Yuki Saito[1], Eiji Iimori[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 in polite dialogue situation

- **Task: E**mpathetic **D**ialogue **S**peech **S**ynthesis **(EDSS)**
  - Towards an AI voice agent that **empathizes** with humans
    - Control the prosody of synthetic speech considering the interlocutor's mental state (e.g., happy → high pitch) & **situation**
- **Existing corpus for EDSS: STUDIES[1]**
  - Situation: chat betw. a teacher & student in a cram school[2] (informal & intensely expressive speaking style)
  - Limitation: only 1 situation & 8 hours empathetic dialogues…
    - **How can we construct a corpus for EDSS in a different situation?**

- **New situation for EDSS: operator & customer in customer center**



🔥 **Get it now!**
https://sytho n.org/Corpus/ STUDIES-2

- **CALLS:** corpus of **C**omplaint handling & **A**ttentive **L**istening **L**ines **S**peech
  - Covering negative/positive feedbacks from customers
  - The same speaker as the STUDIES teacher → **multi-domain EDSS**
  - Opensourced for research purpose only (scan the above QR code!)

Methodology for corpus construction

Dialogue scenario

- **Difficulties in recording actual customer-center dialogues**
  - Privacy preservation for speakers
    → Collecting **simulated** dialogue lines by **crowdsourcing**
  - Limited bandwidth of phone calls
    → Recording the simulated dialogues by a professional voice actor **in studio**
- **Other settings**
  - The agent's persona
    - Female in her early twenties
    - Tokyo dialect
    - Gentle tone of voice, etc.
  - Two dialogue subsets
    - Complaint handling (2 ~ 12 turns)
    - Attentive listening (4 turns)

Instructions for crowdworkers

- **For the complaint handling subset…**
  - Use (1) **seed situation** & (2) **user's proposal** w/ (3) **metadata** when writing dialogue lines.
    1. Text data describing an anonymous user's complaint about a specific service or product
    2. An user's opinion to deal with a complaint
    3. User's age, gender, job, and locale

    → Selected from the **FKC corpus**[3]
  - Anonymize the name of a particular company or product if the situation included it.
- **For the attentive listening subset…**
  - Write short dialogue lines where a customer & operator are **talking happily** on a phone call.
  - Don't include the name of a specific service or product in the dialogue lines.
    → Similar to "Short" dialogue in STUDIES[1]

Voice recording

- **Prior to the recording**
  - Screening the obtained dialogue lines
    - Corrected unnatural sentence in grammar and/or syntax
    - Removed outcomes from spam workers (e.g., wrote only "a") or ones who didn't follow our instructions
- **Recording the agent's voices**
  - Speaker: **the same as the STUDIES teacher** (a female voice actor)
    - We didn't record the customers' voices because the recording is unrealistic.
  - Device: a unidirectional microphone
  - Period: 10 days (3 hours per day)
    - Complaint handling: 6 days
    - Attentive listening: 4 days

Corpus analysis and EDSS experiments

Corpus analysis

- **Corpus specification**
  - # of utterances for each emotion label
    - (complaint handling + attentive listening)

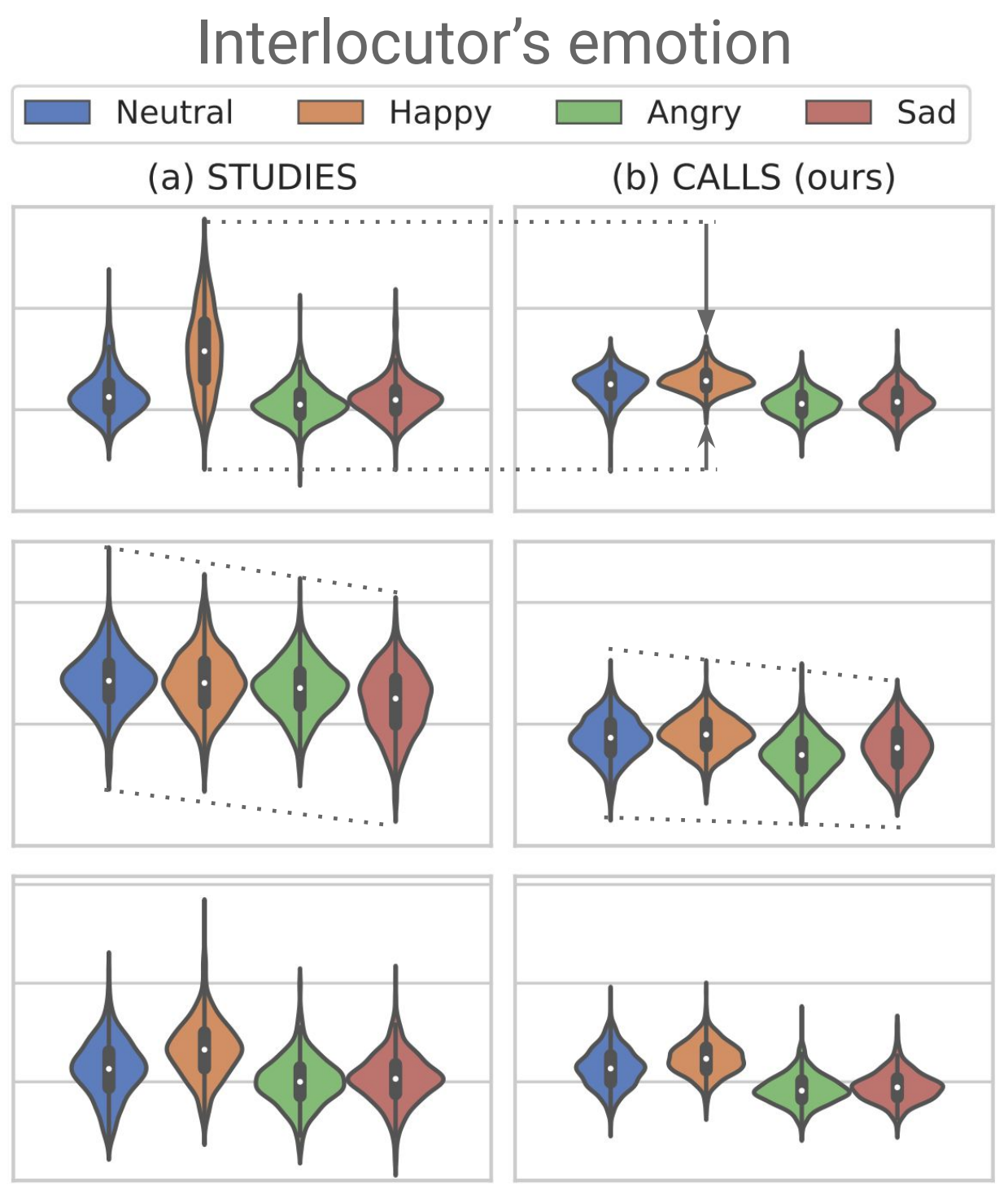
Spkr.	Neutral	Happy	Sad	Angry	Total
Operator	414 + 243	719 + 950	939 + 7	**0** + 0	3,232 (6.5h)
Customer	760 + 389	144 + 790	263 + 21	945 + 0	3,312 (N/A)

- **Comparison with existing corpora**

Corpus	Dialogue type	Open-sourced?	Dur. [h]	# Spkr.	Emotion labeled?
Hiraoka+[4]	Persuasive	No	5.7	22	No
Kawahara+[5]	Attentive listening	No	2.3	8	No
STUDIES[2]	Empathetic (casual)	Yes	8.0	3	Yes
CALLS (ours)	Empathetic (formal)	Yes	6.5	1	Yes

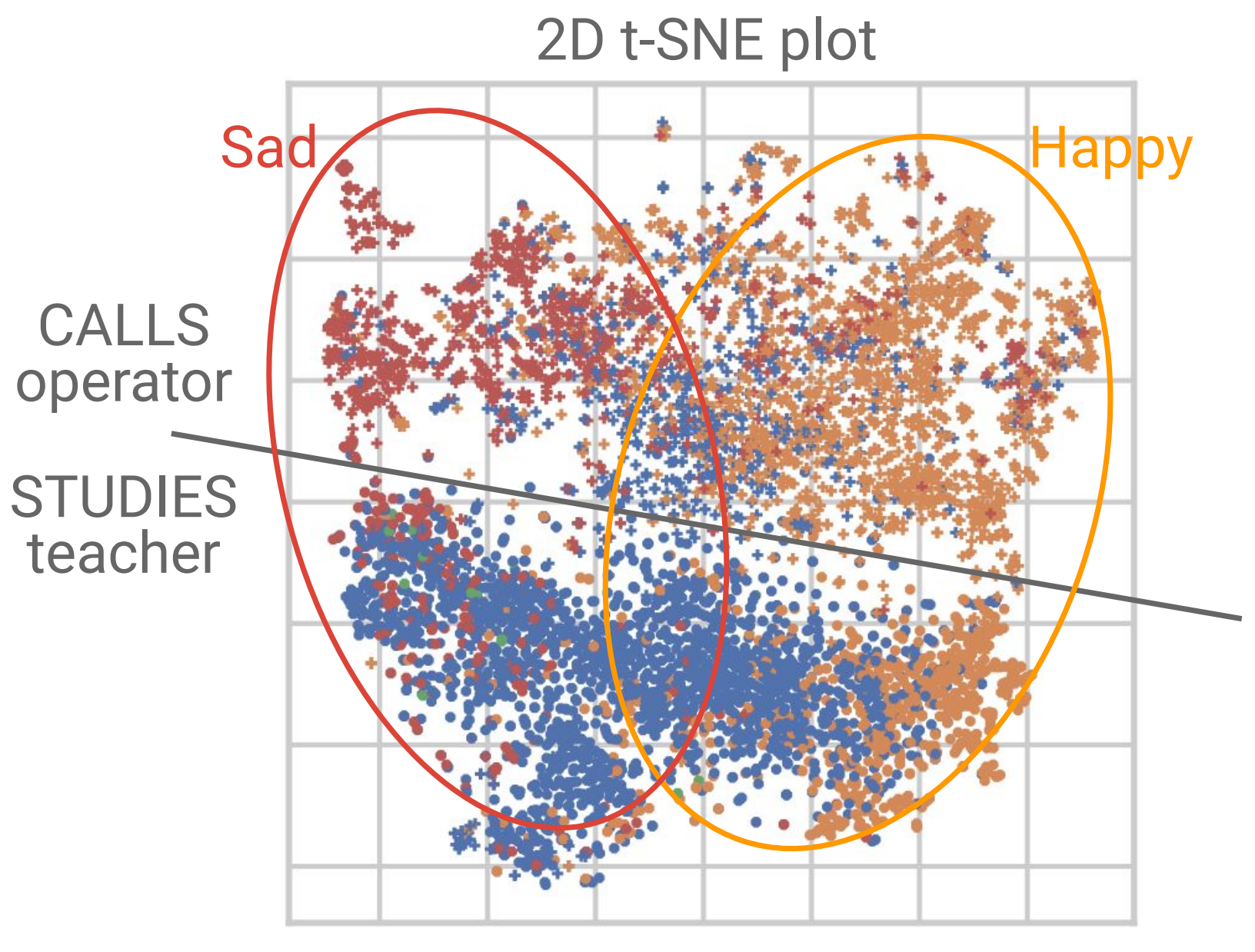
- 6,544 utterances in total
  - 6.5h of agent's empathetic voices w/ **formal** styles
- Data imbalance in emotion
  - 0 angry voices by the operator
    → Unfavorable in the customer center situation
- New domain for EDSS
  - Combined with STUDIES
    → 10h of multi-domain empathetic speech corpus
  - Related to persuasion & counselor's attentive listening

- **Prosody feature statistics & sentence embedding visualization**



Less expressive in "happy" emotion

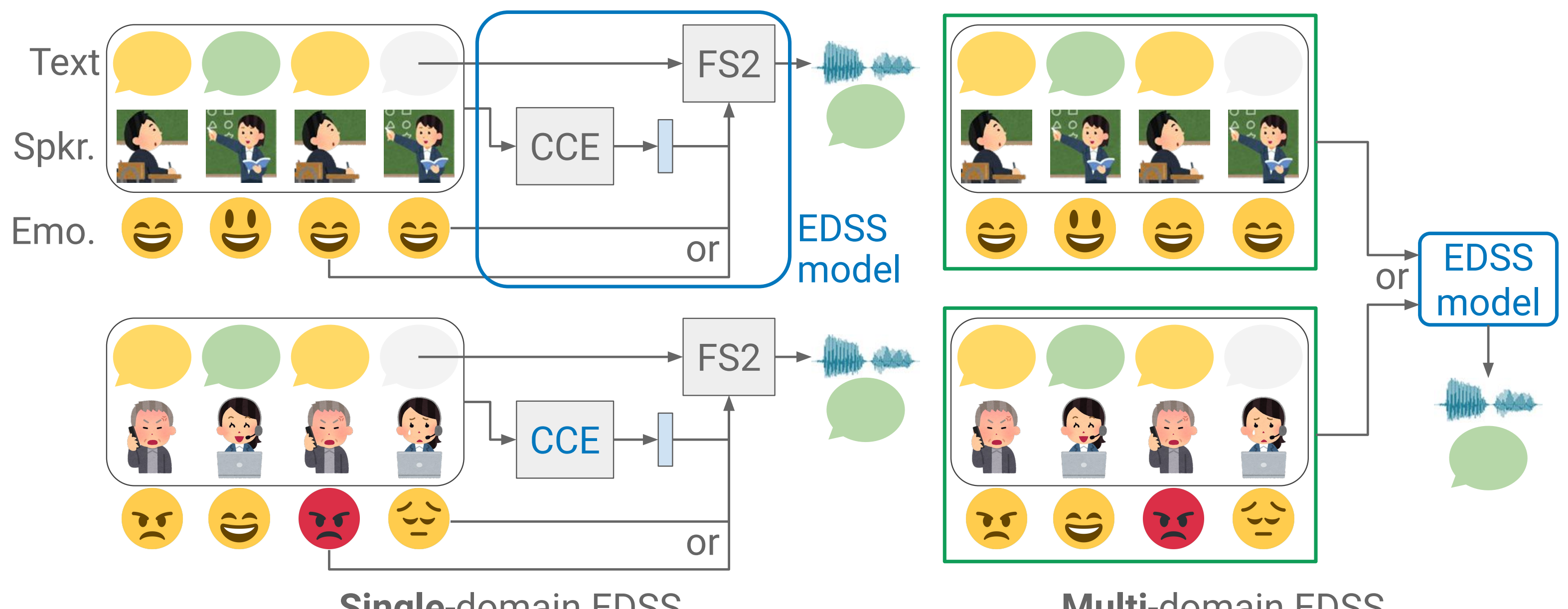
Generally lower stds of logF0 & Energy



CALLS covers a different range than STUDIES!

EDSS experiments

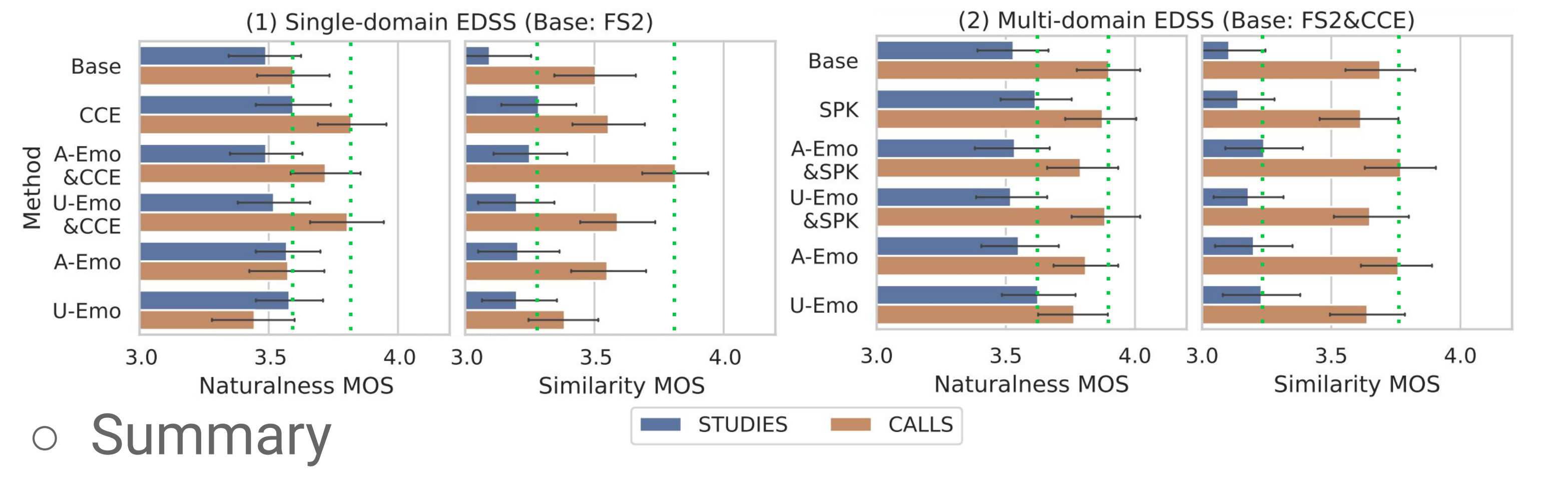
- **Experimental setup: single-/multi-domain EDSS**
  - Acoustic model: FastSpeech 2 (FS2)[6] conditioned by:
    - Emotion label of { agent, user } (A-Emo, U-Emo)
    - Chat history embedded by Conversational Context Encoder[7] (CCE)
    - Speaker ID (SPK: only available for multi-domain EDSS)



Single-domain EDSS
(separately trained 2 EDSS models)

Multi-domain EDSS
(1 EDSS model trained w/ 2 corpora)

- **Subjective evaluation: 2 MOS tests** (w/ 400 listeners)
  - Criteria: naturalness & speaking-style similarity



  - Summary
    - **CALLS operator's voices can be trained more easily.**
    - **CCE is effective, but the domain gap should be considered.**

Reference

[1] Y. Saito et al., INTERSPEECH, 2022.
[2] C. Warren et al., The Urban Review, 2015.
[3] K. Mitsuzawa et al., LREC NIEUW Workshop, 2016.
[4] T. Hiraoka et al., Speech Communication, 2016.
[5] T. Kawahara et al., IWSDS, 2015.
[6] Y. Ren et al., ICLR, 2021.
[7] H. Guo et al., SLT, 2021.