国際会議報告
INTERSPEECH2022

河野 誠也

理化学研究所ガーディアンロボットプロジェクト（GRP）
知識獲得・対話研究チーム 特別研究員

ガーディアンロボット
プロジェクト
Guardian Robot Project

RIKEN

E-mail: Seiya.kawano@riken.jp
Web: https://kwnsiy.github.io/index.html

INTERSPEECHの概要

- ISCA（International Speech Communication Association）主催の国際会議
  - 対象分野：音声情報処理全般
  - ICASSPと並ぶ音声処理分野のトップ会議

- 今年の開催形態
  - ハイブリット開催（2022/9/18~9/22）
  - 韓国仁川（現地）&オンライン(Gather)
  - 約7割の参加者が現地参加

- 採択率・投稿件数
  - 2022：52.4%（1121/2140件）
  - 2021：48.4%（963/1990）

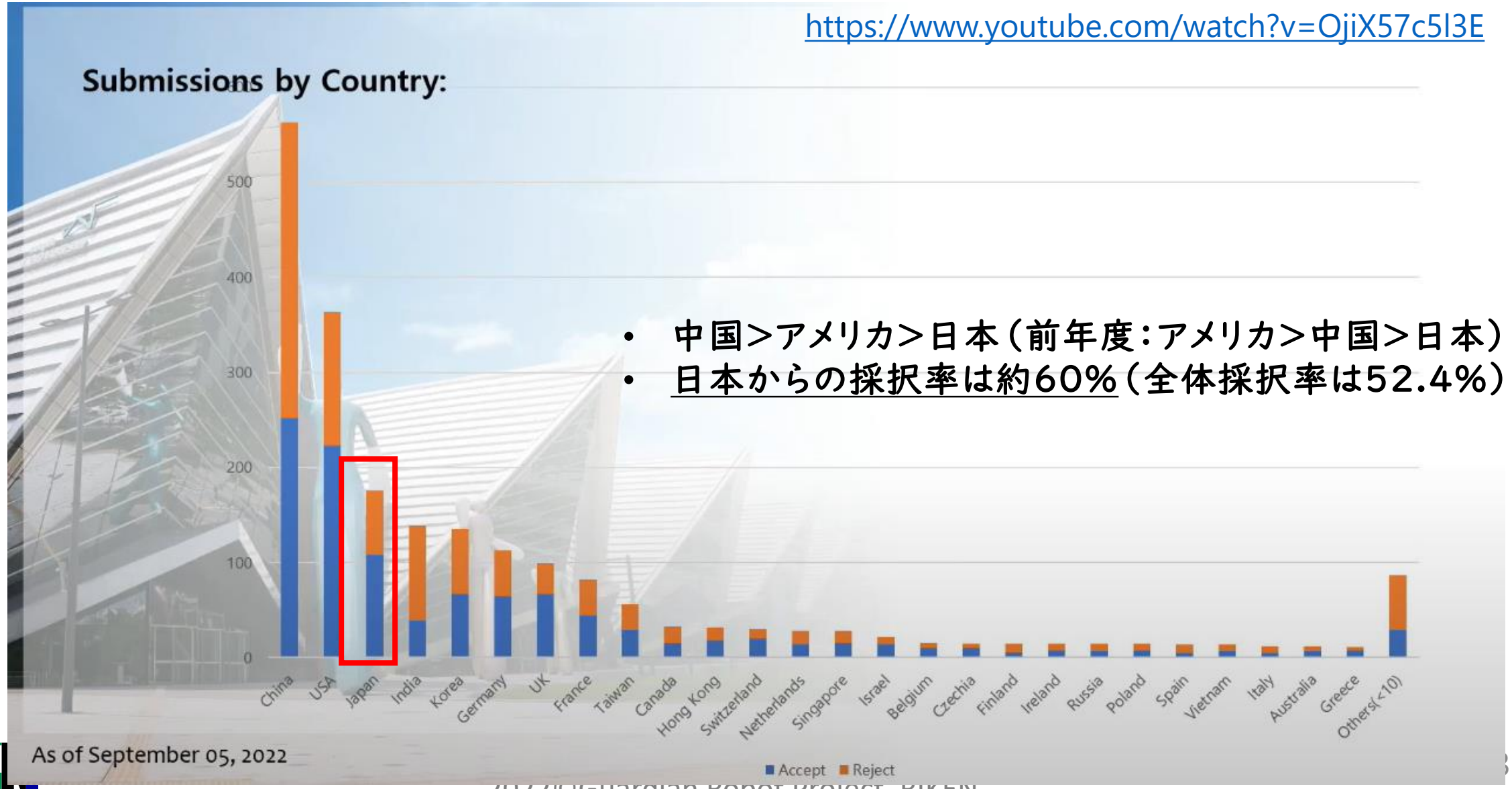
投稿件数
過去最多！

	INTERSPEECH (ICSLP) Jeju 2004	INTERSPEECH Incheon 2022
number of attendance	854	2311 in-person: 1571 remote: 740
papers presented	772	1102
number of countries	36	56

https://www.youtube.com/watch?v=OjiX57c5l3E

国別投稿・発表件数

https://www.youtube.com/watch?v=OjiX57c5l3E



- 中国＞アメリカ＞日本（前年度：アメリカ＞中国＞日本）
- 日本からの採択率は約60%（全体採択率は52.4%）

分野別投稿件数と傾向

- 音声認識・合成，音声符号化，音声強調に関連する投稿が年々増加傾向
  （全体の投稿件数も増加）

- SLUD研究会との関連トピック
  - 音声対話/会話システム
  - 音声翻訳・情報検索・言語資源
  - パラ言語分析
  - 音声学・音韻論
  - 音声知覚・言語産出・言語学習

  → 全体に占める割合：22%
  （投稿数は例年と比較して横ばい）

	'22	'21	'20	'19	'18	'17
1. Speech Perception, Production and Acquisition	121	91	148	161	126	212
2. Phonetics, Phonology, and Prosody	83	112	71	103	77	164
3. Analysis of Paralinguistics in Speech and Language	122	116	157	140	115	159
4. Speaker and Language Identification	191	204	186	173	131	172
5. Analysis of Speech and Audio Signals	313	245	254	212	172	140
6. Speech Coding and Enhancement	206	189	193	122	84	81
7. Speech Synthesis and Spoken Language Generation	315	259	246	173	98	130
8. Speech Recognition Signal Processing, Acoustic Modeling, Robustness, and Adaptation	299	247	265	252	175	171
9. Speech Recognition - Architecture, Search, and Linguistic Components	121	124	97	71	50	57
10. Speech Recognition – Tech's and Systems for new Apps	105	85	126	80	65	57
11. Spoken dialog systems and conversational analysis	102	99	85	110	50	106
12. Spoken Language Processing: Translation, Information Retrieval, Summarization, Resources and Evaluation	129	121	93	97	54	119
13. Speech, voice, and hearing disorders	118	82	0	0	0	0
Special Sessions:	265	303	161	182		
TOTALS:	2490	2277	2138	1855	1379	1568
	(2140)	(1990)				

https://www.youtube.com/watch?v=OjiX57c5l3E

対話系の研究に関して

■ モダリティ
- 音声が基本（純粋なNLPは少ない）
- マルチモーダル

■ アプローチ
- 対話（現象，スタイル）の分析
- 言語理解，感情認識，データセット構築
- 対話のリアルタイム性（ターンテイキング，音声認識，音声合成）の考慮

- 言語（テキスト）生成・対話制御に関する研究は少ない

チュートリアルセッション

■ 会議初日に8つのチュートリアルが実施

■ 午前
- Leaning from Weak Labels
- Reinforcement Learning and Bandits for Speech and Language Processing
- Self-supervised Representation Learning for Speech Processing
- Speech enhancement for cochlear implants: From psychoacoustics to machine learning

■ 午後
- Deep Spoken Keyword Spotting Hybrid
- Personalized Speech Enhancement: Data- and Resource-Efficient Machine Learning
- A Speech Brain for Everything: State of the PyTorch Ecosystem for Speech Technologies
- Neural Speech Synthesis

基調講演

- From Semantics to Self-supervised Learning for Speech and Beyond (Lin-shan Lee, National Taiwan University)

- David V.S. Goliath: the Art of Leaderboarding in the Era of Extreme-Scale Neural Models
  (Yejin Choi, University of Washington and AI2)

- Blurring the line between human and computer-generated speech: opportunities and challenges
  (Rupal Patel, Voice & Accessibility)

- Representations and Geometry for Multimodal Learning (Daniel Dongyuel Lee, Samsung and Tisch University)

特別セッション

■ 13の特別セッションが採択

■ Tue-SS-OS-5-5: Speaking Styles and Interaction Styles

- 対話における**発話/インタラクションのスタイル**について扱った特別セッション

➢ Text-driven **Emotional Style** Control and Cross-speaker **Style Transfer** in Neural TTS
➢ Comparison of Models for Detecting **Off-Putting Speaking Styles**
➢ Strategies for a developing Conversational **Speech Dataset** for Text-To-Speech Synthesis
➢ **Multimodal Persuasive Dialogue Corpus** using Teleoperated Android → 報告者らの研究グループによる発表
➢ Deep CNN-based Inductive Transfer Learning for **Sarcasm** Detection in Speech
➢ End-to-End Text-to-Speech Based on **Latent Representation of Speaking Styles** Using Spontaneous Dialogue
➢ Attention-based conditioning methods using variable frame rate for **style-robust speaker verification**
➢ Learning from **human perception** to improve automatic speaker verification in style-mismatched conditions
➢ Exploring audio-based **stylistic variation** in podcasts

Best Paper Awards

■ 3本の論文が選出（12件の論文がNominee）

1. Transfer Learning Framework for Low-Resource Text-to-Speech using a Large-Scale Unlabeled Speech Corpus
   - 低リソースTTSのための転移学習フレームワークの提案

2. Tree-constrained Pointer Generator with Graph Neural Network Encodings for Contextual Speech Recognition
   - 発話者の文脈知識（バイアス単語）を音声認識モデルに反映

3. Investigating perception of spoken dialogue acceptability through surprisal
   - LMによるサプライザル推定と人間による対話受容性判断の関連

個人的に気になった論文（1）

- タイトル：Investigating perception of spoken dialogue acceptability through surprisal（Wallbridge et al.）

- サプライザル（Surprisal）とは？
  - 人間の漸進的な文処理のモデル化に用いられる概念
  - 人間は先読みしながら文章を読む
    - 予想通りの情報が出現→サプライズが低
    - 予想と異なる情報が出現→サプライズが大→処理コストの増大
  - 言語モデルによるサプライザルの推定：$S(u_n) = -\log_2 p(u_n|u_{<n}).$

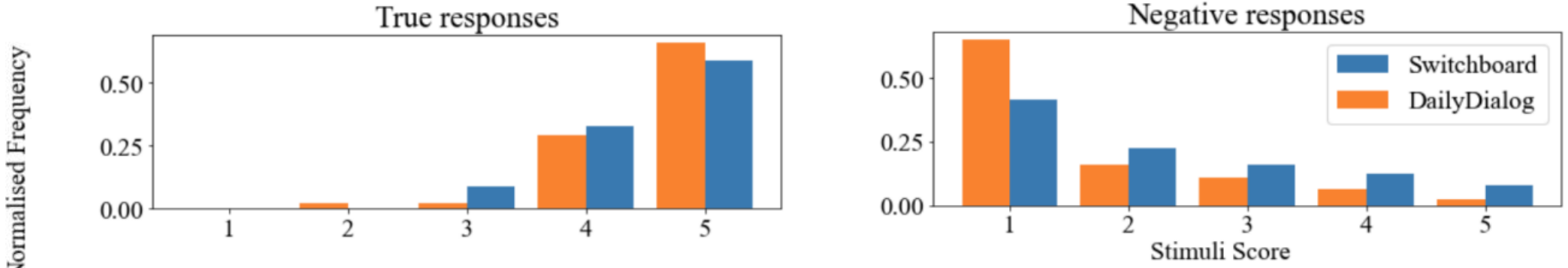
→人間の文書の読み取り時間や文法的妥当性の判断といった知覚特性と相関

  - モノローグデータ以外でのサプライザルと知覚の関係については証拠がほとんどない（一方で，人間の言語使用の形式は通常インタラクティブ）

個人的に気になった論文（2）

■ 人手で書かれた対話と口頭の対話におけるサプライザルと知覚特性の関連を調査

■ 対話受容性判断タスクの提案
- 対話文脈cを参加者（人間）に提示し、その後に続くターンr の発話の妥当性を判定
- 5段階評価（“Very Unlikely” – “Very Likely”）



書かれた対話と口頭の対話の両方で、人間は次発話の妥当性について正確な判断が可能
口頭の対話では否定的な刺激に対してより分散したスコア分布

個人的に気になった論文（3）

■ 言語モデル（GPT）によって推定されたサプライザルと人間の対話受容性の関連

- 負の相関関係
- 発話内トークンの予測結果の累積に基づいたグローバル指標よりも特定のトークンに着目したローカルな指標で最も大きな負の相関（モノローグベースのタスクに基づいた従来研究と相違）
- グローバルとローカル指標を組み合わせることで人間の対話受容性の予測に寄与

Table 1: *Surprisal measure definitions: r and c are response and context word sequences resp.*

$$S_{total}(\mathbf{r}|\mathbf{c}) = \sum_{n=1}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{mean}(\mathbf{r}|\mathbf{c}) = \frac{1}{N}\sum_{n=1}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{relative}(\mathbf{r}|\mathbf{c}) = S_{mean}(\mathbf{r}|\mathbf{c}) - S_{mean}(\mathbf{r})$$

$$S_{max}(\mathbf{r}|\mathbf{c}) = \max[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{var}(\mathbf{r}|\mathbf{c}) = \frac{1}{N-1}\sum_{n=2}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c}) - S(r_{n-1}|\mathbf{r}_{<n-1}, \mathbf{c})]^2$$

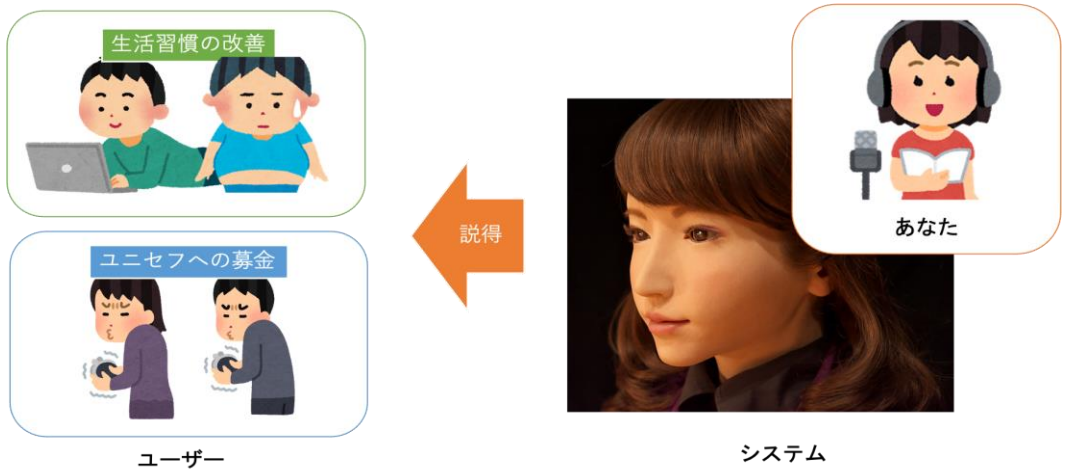
Table 2: *Correlation between surprisal and median judgement scores in DailyDialog*

Surprisal	DailyDialog		Switchboard	
	ρ	p-val	ρ	p-val
Total	-0.341	0.001	-0.273	0.006
Mean	-0.350	<0.001	-0.299	0.003
Relative	-0.360	<0.001	-0.262	0.009
Max	-0.407	<0.001	-0.400	<0.001
Variance	-0.295	0.003	-0.217	0.038

報告者による発表

■ タイトル: Multimodal Persuasive Dialogue Corpus using Teleoperated Android

- 説得対話システム構築のためのマルチモーダルコーパスをWoZ法で構築
- 60人の被験者, 三つの説得ドメイン（運動習慣改善, ネット依存改善, 寄付の促し）
- 被験者に対する詳細な事前・事後調査
  - 説得の前後に実施
  - 意識, パーソナリティ, ロボットの印象
- 説得の効果（実際の行動変容）の追跡調査
  - 対話の二週間後に実施
- 対話に対する言語・非言語情報のアノテーション
  - 対話行為ラベル
  - Action Unit
  - 感情情報, etc.



■ 説得の成功/失敗の予測可能性の分析 → 言語・非言語の両方の情報を考慮したモデルで最高精度

INTERSPEECH2023

- 開催地：アイルランド，ダブリン（2023年8月20日-8月24日）
- 論文投稿〆切期限：2023年3月1日
- 論文更新期限：2023年3月8日
- 採択通知：2023年5月15日

https://www.interspeech2023.org/important-dates-deadlines/