TAPAS
Training Adapted Personalised Affective Social Skills with Cultural Virtual Agents

Multimodal Voice Activity Prediction:
Turn-taking Events Detection in Expert-Novice Conversation

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura
Nara Institute of Science and Technology, Japan

Code

HAI 2023 – Conference in Human-Agent Interaction '23
4 – 7 December, Gothenburg, Sweden

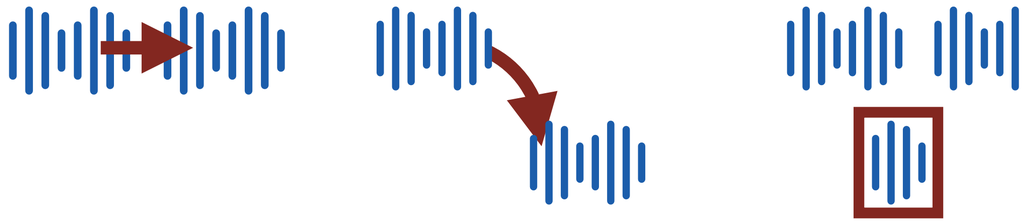
JST CREST Japan Science and Technology Agency
Symbiotic Interaction
ANR AGENCE NATIONALE DE LA RECHERCHE

Background

Turn-taking is ···

the process of communicating in a conversation
in which the speaker and the listener take turns speaking

Turn-hold　　Turn-shift　　Back-channel (BC)



... various turn-taking

Challenges for Virtual Agents

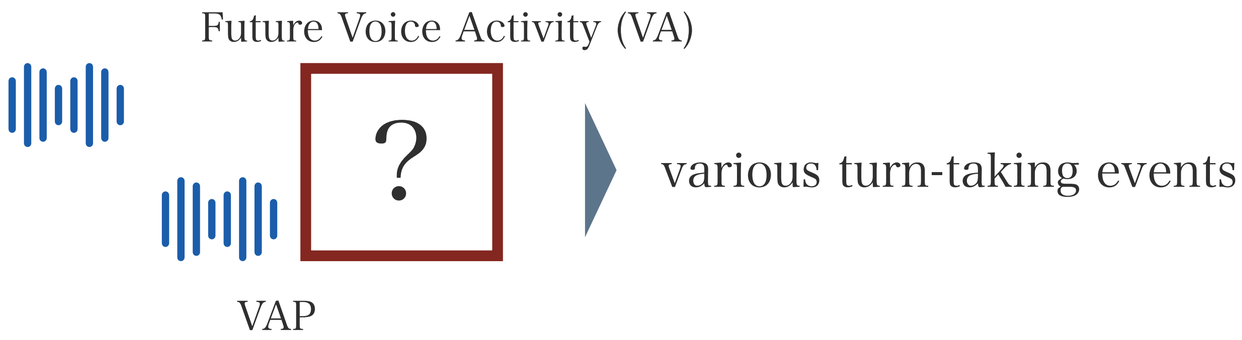
Improving virtual agent usage satisfaction ...
need human-like turn-taking performance [Toshihiko+, 2009]

→ Predictive modeling to determine turn-taking

Related Work

Previous Models

Unified approach to various turn-taking events
with a **voice activity prediction (VAP) model** [Skantze, 2017]



- Proposed architecture to process audio and linguistic features [Roddy+, 2018]
- Define turn-taking events that extend to and evaluate Transformer models [Ekstedt+, 2022a]
- Investigate the impact of prosody on models [Ekstedt+, 2022b]

In This Study

Research Question

Humans make predictions based on three features (audio, linguistic, non-verbal)

→ Non-verbal features have not been verified in VAP model

Are non-verbal features useful in voice activity prediction models?

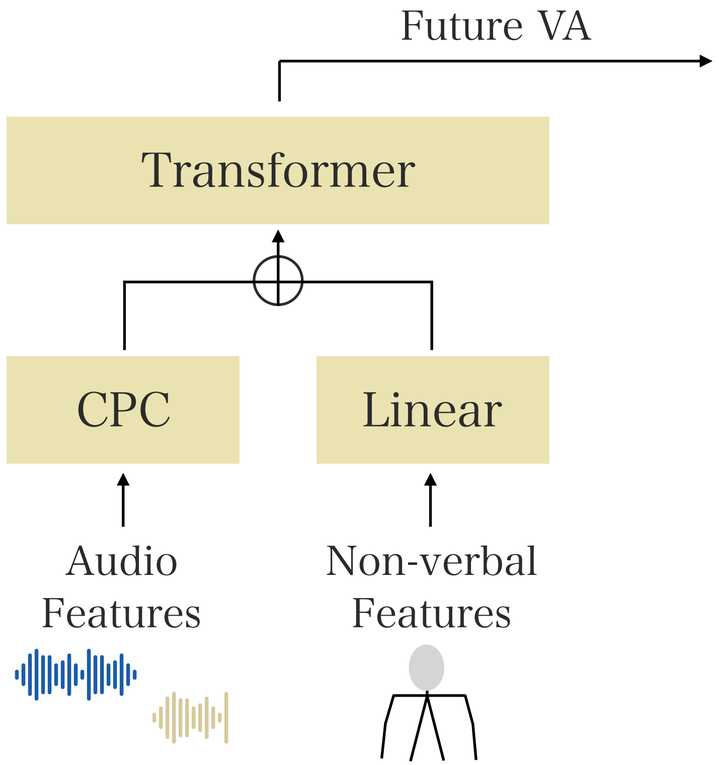
Experiments

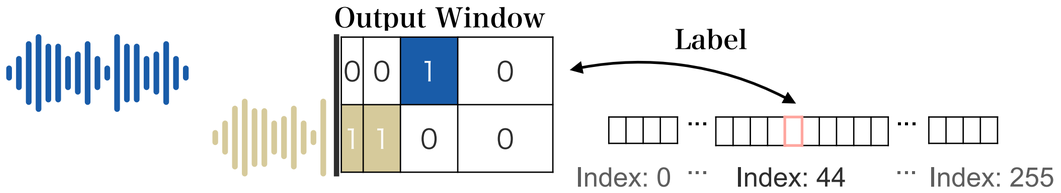
Verify three conditions

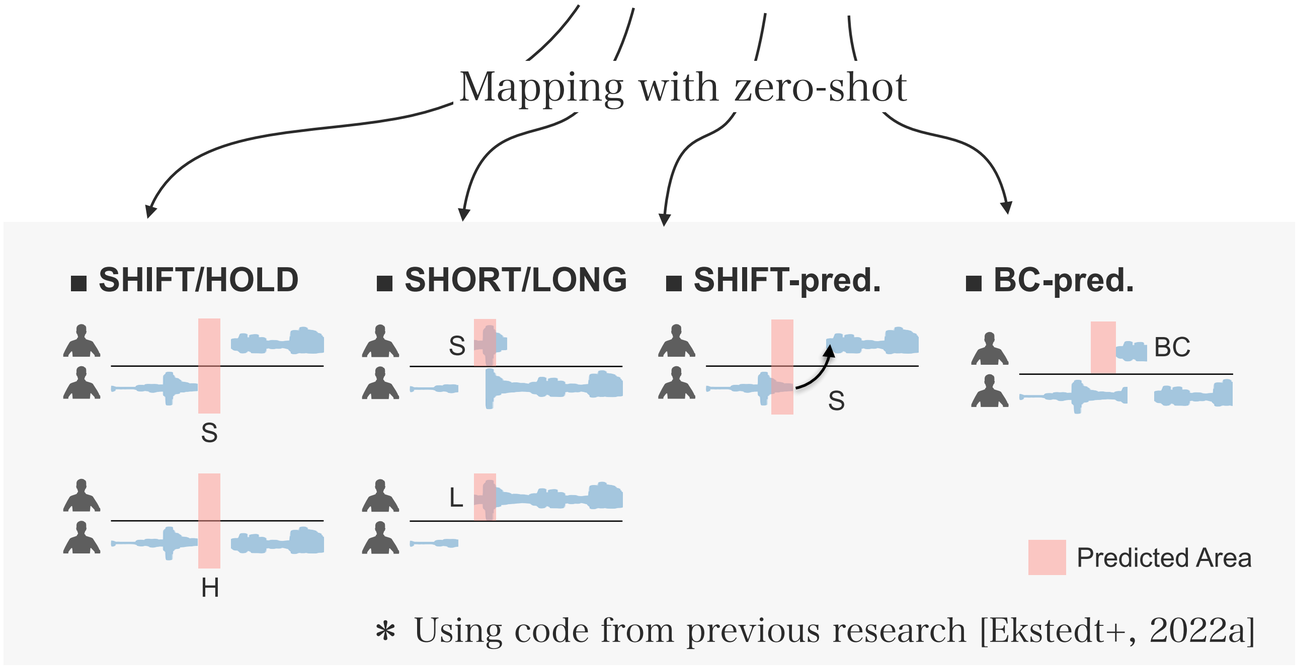
(1) Audio Features Only
(2) Non-verbal Features Only
(3) Audio + Non-verbal Features

Multimodal Model

Evaluating Turn-taking Events From VAP







∗ Code is available

∗ Using code from previous research [Ekstedt+, 2022a]

Input Features

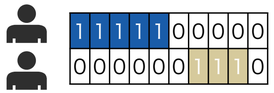
Audio Features

Audio Waveform

- Mixing two speakers' audio waveforms

Voice Activity Frame (VA-frame)

- VA expressed as active 1 and inactive 0



Voice Activity History (VA-history)

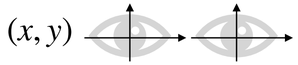
- Percentage of VA in the particular sections

$$VA_{history}(section) = \frac{VA_1 \ active \ time}{VA_2 \ active \ time}$$

Non-verbal Features

Gaze Direction

- Direction vector of line of sight



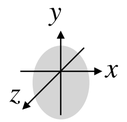
Action Units

- Eyebrows, eyelids, mouth, and facial expressions

Head Pose

- Head movement in 3 dimensions

$$head_{x,y,z}(t) = \sqrt{\left((x,y,z)_t - (x,y,z)_{t-1}\right)^2}$$

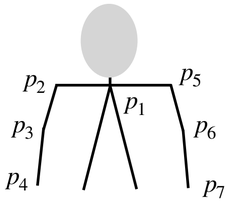


※ Difference from one previous frame

Articular Points

- Capture gestures from joint coordinates

$$Articular^n(t) = \sqrt{\left(Articular_t^n - Articular_{t-1}^n\right)^2}$$



※ Difference from one previous frame

Data

Corpus

NoXi Database

- A pair of people through a screen
- Expert shares topics with Novice
- Multilingual database (English, German, French, ...)
- Dialogue similar to an agent system
- Record audio and video

Expert

Novice

similar to an agent system

Splitting of Training Data

	Number of Sessions	Duration [h]	S/H	S/L	S-pred.	BC-pred.
Train	54	16.87	1015/11757	3382/1015	1015	3375
Validation	12	3.35	246/2447	595/246	246	592
Test	18	5.29	354/3659	1007/354	354	1005

(1) Audio Features

Audio Features Only

Audio Features			Turn-taking Events (F1 Score)			
Audio Waveform	VA-frame	VA-history	S/H	S/L	S-pred.	BC-pred.
✔			0.734	0.629	0.657	0.650
✔	✔		0.885	0.804	0.704	**0.691**
✔	✔	✔	**0.888**	**0.811**	**0.725**	0.687

+ VA-frame

- Improved scores on all turn-taking events
- Facilitated speaker identification of mixed audio waveforms

+ VA-history

- Improved S-pred. scores by +2.1 points ($p < 0.025$)
- Speech ratios affect turn shift

(2) Non-verbal Features

Non-verbal Features Only

Non-verbal Features				Turn-taking Events (F1 Score)			
Gaze	AU	Head	Articular	S/H	S/L	S-pred.	BC-pred.
✔				0.814	0.625	0.656	0.530
	✔			0.837	**0.649**	0.676	**0.591**
		✔		**0.856**	0.643	**0.697**	0.428
			✔	**0.856**	0.631	0.692	0.467

- Achieves 70 points for S-pred. and 59 points for BC-pred.
- Outperforms results with audio waveform-only input
- Non-verbal features have a significant impact on turn-taking

- Possibility that action units in particular affect BC-pred. scores

Audio + Non-verbal Features

Audio Features and Non-verbal Features

Audio Features			Non-verbal Features				Turn-taking Events (F1 Score)			
Audio Waveform	VA-frame	VA-history	Gaze	AU	Head	Articular	S/H	S/L	S-pred.	BC-pred.
✔	✔	✔	✔				0.889	0.810	0.738	0.688
✔	✔	✔		✔			**0.897**	0.812	**0.744**	**0.707**
✔	✔	✔			✔		0.892	0.814	0.739	0.693
✔	✔	✔				✔	0.891	**0.816**	0.735	0.691

Scores significantly different compared to all audio features only

S/H and S/L

No significant difference

Due to high baseline F1 scores and a the bias in the number of tasks

S-pred.

Significant difference when adding gaze direction or action units

BC-pred.

Significant difference when adding action units

Result

All Audio Features and All Audio and Non-verbal Features

	Audio Features, mean of F1 score (SD)	Audio + Non-verbal Features, mean of F1 score	Cohen's *d*	*P* value
S/H	0.888 (0.006)	0.890 (0.009)	0.144	0.782
S/L	0.811 (0.012)	0.810 (0.013)	0.035	0.950
S-pred.	0.725 (0.011)	0.748 (0.012)	1.985	0.000
BC-pred.	0.687 (0.013)	0.705 (0.005)	1.900	0.001

- Finally, +2.3 points improvement for S-pred. and +1.8 points for BC-pred.
- Non-verbal features are useful in VAP model

Summary

Are non-verbal features useful in VAP models?

(1) Audio Features Only

- VA-frame is essential for speaker identification in mixed audio waveform
- Speech ratio affects turn-shift and improves +2.1 points

(2) Non-verbal Features Only

- 70 points for turn-shift and 59 points for back-channel
- Non-verbal features have a significant impact on turn-taking

(3) Audio + Non-verbal Features

- +2.3 improvement for turn-shift and +1.8 for back-channel
- Action units are particularly effective in predicting turn-shift and back-channel

Code

Appendix