Speech Aware Dialogue Management

Professor M. Inés Torres, Universidad del País Vasco (UPV/EHU), Bilbao, Spain

---

RTTH Fall School - Jaca
November, 2023

Outline

- Context of the talk
- Part I: Emotional Information in the Speech
- Part II: Speech aware policy
- Conclusions, questions, opinions, thoughts……..

Speech Aware Dialog Systems Technology Challenge

Vast majority of research in dialog systems are performed on written dialog corpora (MultiVOZ, SGD, MetalWOZ). As a result, most dialog systems are trained on written text but are deployed as spoken interfaces, leading to a mismatch between training and testing setup—-Google says

Multi- Domain Wizard-of-Oz dataset (MultiWOZ), a fully-labeled collection of human-human written conversations spanning over multiple domains and topics.

Metric	DSTC2	SFX	WOZ2.0	FRAMES	KVRET	M2M	MultiWOZ
# Dialogues	1,612	1,006	600	1,369	2,425	1,500	**8,438**
Total # turns	23,354	12,396	4,472	19,986	12,732	14,796	**115,424**
Total # tokens	199,431	108,975	50,264	251,867	102,077	121,977	**1,520,970**
Avg. turns per dialogue	14.49	12.32	7.45	**14.60**	5.25	9.86	13.68
Avg. tokens per turn	8.54	8.79	11.24	12.60	8.02	8.24	**13.18**
Total unique tokens	986	1,473	2,142	12,043	2,842	1,008	**24,071**
# Slots	8	14	4	**61**	13	14	25
# Values	212	1847	99	3871	1363	138	**4510**

- Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Jeffrey Zhao, Ye Jia, Wei Han , Yuan Cao and Aramys Miranda (2023), "Speech Aware Dialog System Technology Challenge (DSTC11)", INTERSPEECH pp 668-4672, doi=10.21437/Interspeech.2023-1037}

DSTC11: Speech Aware Dialog Systems Technology Challenge

MultiWoz benchmark + spoken user utterances = four different datasets

Audio + ASR transcription

Dataset	**Description**
Human-Verbatim	Verbatim, spoken user utterances
Human-Paraphrased	Paraphrased, spoken user utterances (Figure 1)
Human-Noisy	Human-Verbatim with synthetic background noise [10]
TTS-Verbatim	Verbatim user utterances generated using a TTS model

Spoken task-oriented dialogue system, which is robust to various spoken language phenomena such as disfluencies, speech repair and background noise.

Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Jeffrey Zhao, Ye Jia, Wei Han , Yuan Cao and Aramys Miranda (2023), "Speech Aware Dialog System Technology Challenge (DSTC11)", INTERSPEECH pp 668-4672, doi=10.21437/Interspeech.2023-1037}

DSTC11: Speech Aware Dialog Systems Technology Challenge

Speaker	Utterance
User	I need to book a hotel in the east that has 4 stars.
	Hey I want to book a hotel in the east that has 4 stars
System	I can help you with that. What is your price range?
User	That doesn't matter as long as it has free wifi and parking.
	Hmmm aaa that does not matter as long as I get free wifi and parking
System	If you'd like something cheap, I recommend the Allenbell. For something moderately priced, I would recommend the Warkworth House.
User	Could you book the Wartworth for one night, 1 person?
	I see aaa could you book the wartworth for one night aaa 1 person
System	What day will you be staying?
User	Friday and Can you book it for me and get a reference number ?
	Oh that is friday aa and can you book it for me aa also get a reference number
System	Booking was successful. Reference number is : BMUKPTG6. Can I help you with anything else today?
User	I am looking to book a train that is leaving from Cambridge to Bishops Stortford on Friday.
	I am also looking to book a train that is leaving from cambridge to bishops strotford oh no actually it is bishops stortford on friday
System	There are a number of trains leaving throughout the day. What time would you like to travel?
User	I want to get there by 19:45 at the latest.
	I want to get there by quarter to 8 at the latest
System	Okay! The latest train you can take leaves at 17:29, and arrives by 18:07. Would you like for me to book that for you?
User	Yes please. I also need the travel time, departure time, and price.
	Yes please ummm I also need the travel time um departure time and the price
System	Reference number is : UIFV8FAS. The price is 10.1 GBP and the trip will take about 38 minutes. May I be of any other assistance?
User	Yes. Sorry, but suddenly my plans changed. Can you change the Wartworth booking to Monday for 3 people and 4 nights?
	Yes oh sorry but suddenly my plans changed actually can you change the wartworth booking to monday for 3 people and 4 nights
System	I have made that change and your reference number is YF86GE4J
User	Thank you very much, goodbye.
	Thank you very much umm goodbye
System	You're welcome. Have a nice day!

DSTC11: Speech Aware Dialog Systems Technology Challenge

How much degradation do systems suffer when **switching from written to ASR transcripts**?

A: Comparison between **Human-Verbatim** and state-of-the-art **text-based** models built on MultiWOZ dataset.

How much degradation do systems suffer due to **spontaneous speech phenomena** (e.g., paraphrasing, speech repairs)?

A: Comparison between **Human-Paraphrased** and **Human-Verbatim.**

How much degradation do systems suffer when the audio contains **undirected speech** and **background noise**? Noise will be superimposed in audio domain to degrade the ASR performance by a measured amount (e.g., 20% WER degradation).

A: Comparison between **Human-Verbatim** and **Human-Noisy.**

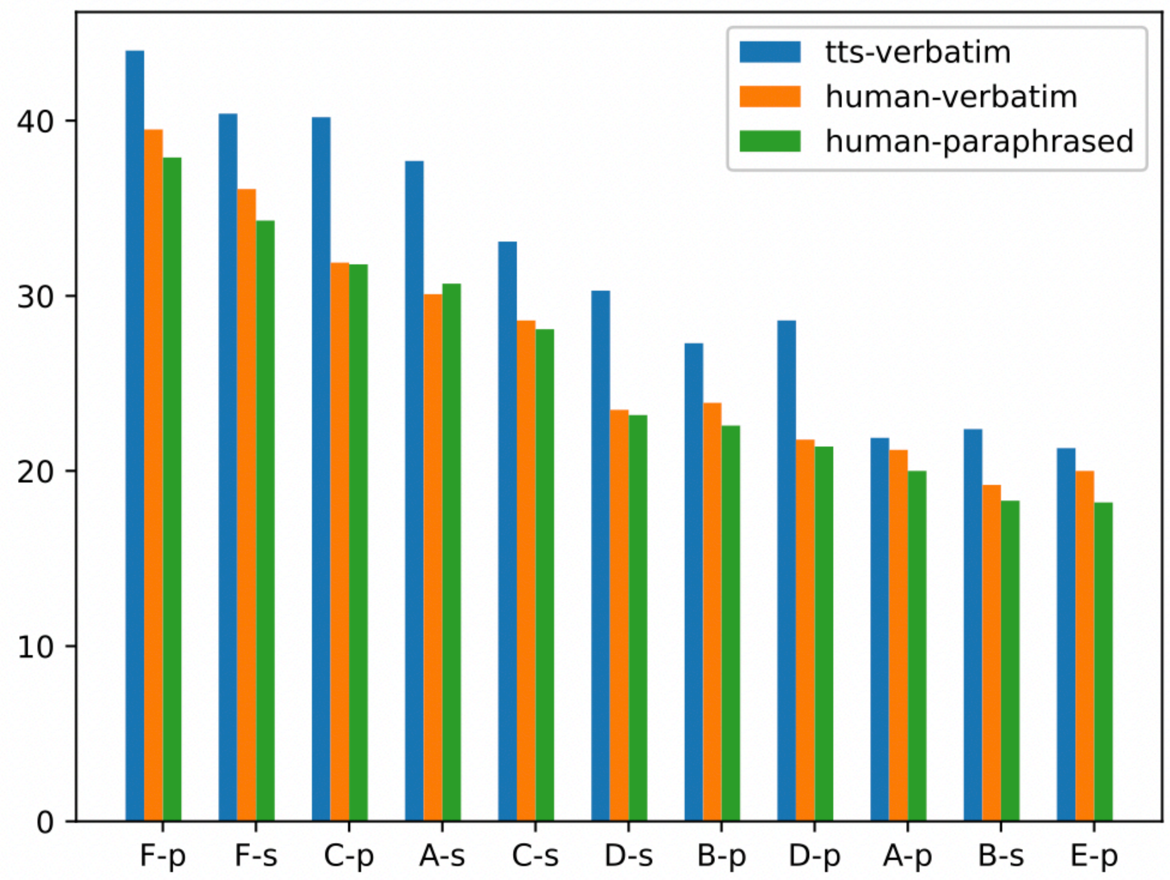
Can degradation in human responses be emulated using synthetic speech via TTS? If the user utterances synthesised with TTS model can characterise the system degradation as well as crowd-sourced utterances, then this would unblock future research in this area.

A: Comparison between **Human-Verbatim** and **TTS- Verbatim**

Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Jeffrey Zhao, Ye Jia, Wei Han , Yuan Cao and Aramys Miranda (2023), "Speech Aware Dialog System Technology Challenge (DSTC11)", INTERSPEECH pp 668-4672, doi=10.21437/Interspeech.2023-1037}

DSTC11: Speech Aware Dialog Systems Technology Challenge

Joint Goal Accuracy (JGA) defined as the fraction of turns where the ground-truth dialogue state exactly matches the prediction.



"…….Similarly, we hope the release of the audio and the audio encoders outputs will allow researchers to evaluate the power of joint audio-text encoders on dialog tasks."

Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Jeffrey Zhao, Ye Jia, Wei Han , Yuan Cao and Aramys Miranda (2023), "Speech Aware Dialog System Technology Challenge (DSTC11)", INTERSPEECH pp 668-4672, doi=10.21437/Interspeech.2023-1037}

Context in Human Conversations

What speech and audio analysis can provide?

- Speaker identification, language recognition, first or second Language, accent recognition, emotion recognition, some soft mental diseases features such as depression or anxiety, etc
- Paralinguistic information resulting in speaking styles: overlappings, pauses, spontaneous speech events, etc
- Genre, age, drugs, mood, noise, environment, etc etc
- Audio acoustic events: laugh, cough, different types of noise,….etc

Outline

- Context of the talk
- Part I: Emotional Information in the Speech
- Part II: Speech aware policy
- Conclusions, questions, opinions, thoughts……..

Emotional information in Speech

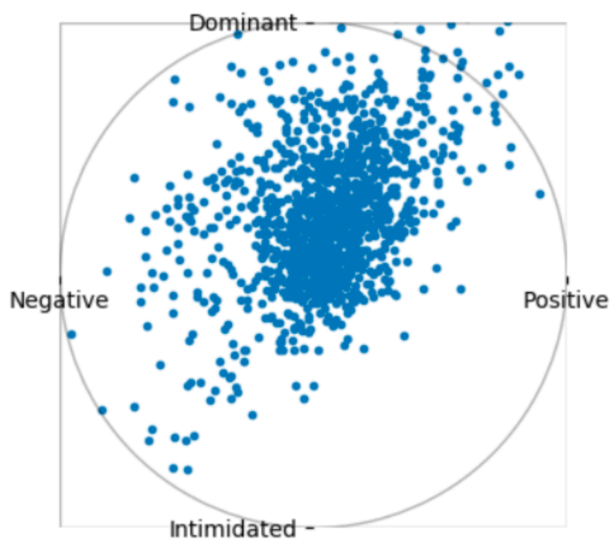
Emotions are not so extreme in real tasks.

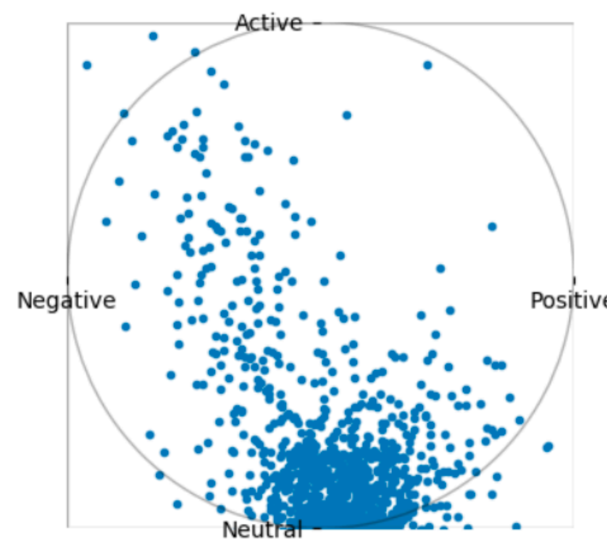
The human interaction with machines do not promote their expression…

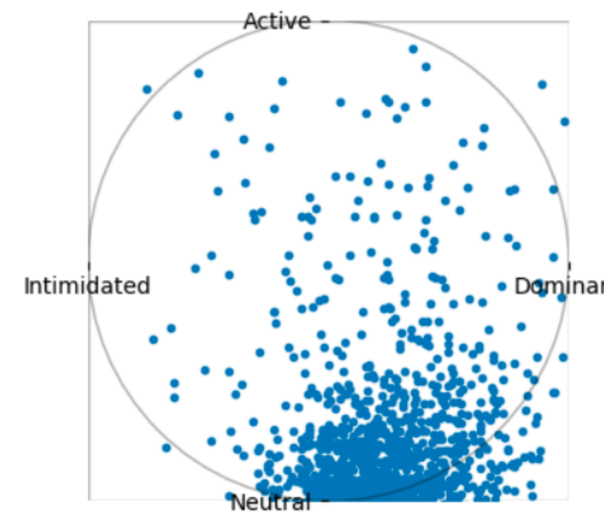
Mood and/or emotion identification vs facial expressions….gaze.

Annotation difficulty: very subjective task, speech or content?,

Emotional information in Speech







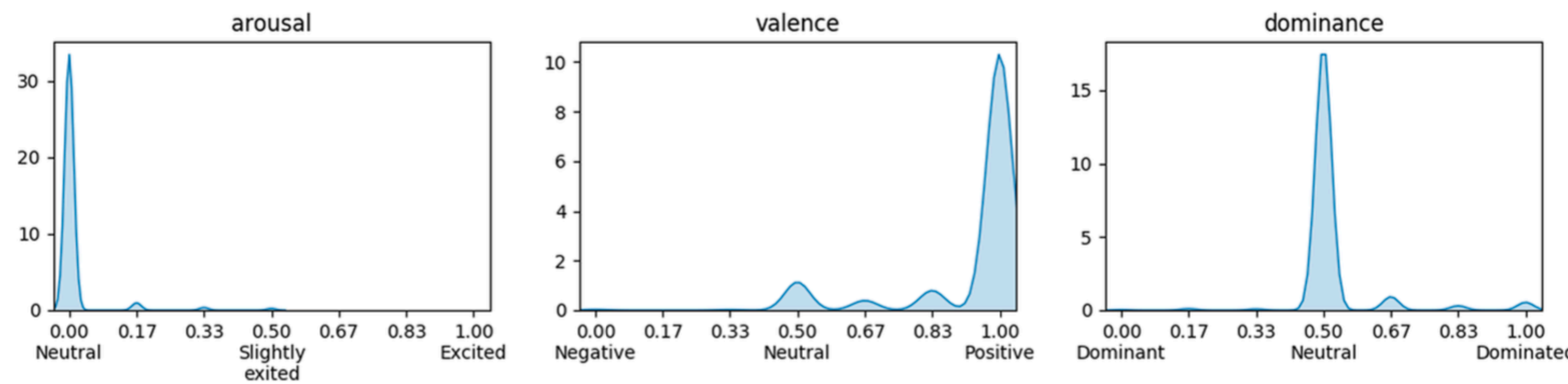
National Project: UPM, UPV, UZ and UPV/EHU

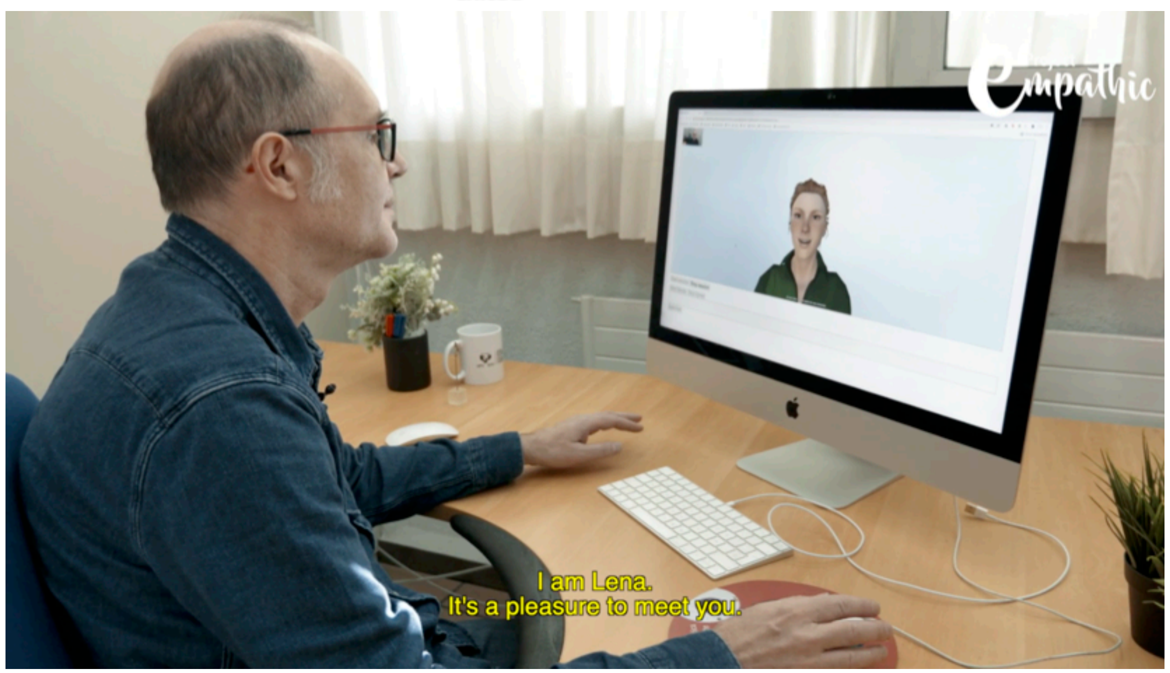
La Sexta Noche, political debates: annotation analysis

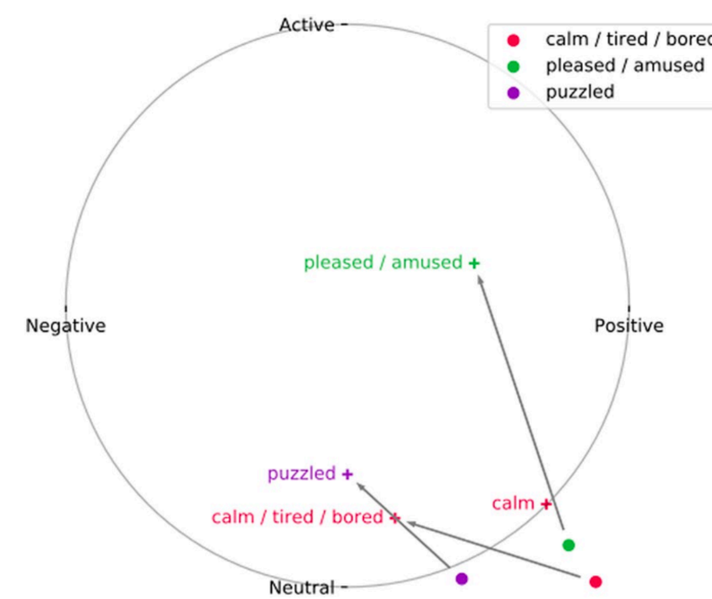
Human Human Interaction

Mikel deVelasco-Vázquez, Raquel Justo, and M. Inés Torres (2022) "Automatic Identification of Emotional Information in Spanish TV Debates and Human-Machine Interactions" *Applied Sciences*, Vol 12, number 4, 1902 https://doi.org/10.3390/app12041902

Emotional information in Speech







H2020 EMPATHIC PROJECT: annotation analysis

Human Machine Interaction

Target population: >65, no pathologies, independent life

R. Justo, L. Ben Letaifa, C. Palmero, E. Gonzalez-Fraile, A. Torp Johansen, A. Vázquez, G. Cordasco, S. Schlögl, B. Fernandez Ruanova, M. Rodrigues da Silva, S. Escalera, M. deVelasco, J. Tenorio-Laranga, A. Esposito, M. Kornes and M. Inés Torres (2021): "Analysis of the Interaction between Elderly People and a Simulated Virtual Coach." *Journal of Ambient Intelligence and Humanized Computing*, Vol 11, N 12 pp 6125-6140 https://doi.org/10.1007/s12652-020-01983-3

SPEECH INTERACTIVE RESEARCH GROUP

MENHIR MCSA RISE

eman ta zabal zazu
Universidad del País Vasco | Euskal Herriko Unibertsitatea

WEMBWS

Template to be filled during the interaction

	CLIENT ANSWER				
Statements	**None of the time**	**Rarely**	**Some of the time**	**Often**	**All the time**
I'm feeling optimistic about the future					
I'm feeling useful					
I'm feeling relaxed					
I'm feeling interested in other people					
I've energy to spare					
I'm dealing with problems well					
I'm thinking clearly					
I'm feeling good about myself					
I'm feeling close to other people					
I'm feeling confident					
I'm able to make up my own mind about things					
I'm feeling loved					
I'm interested in new things					
I'm feeling cheerful					

CONSELLOR PERCEPTION	
Arousal	**Valence**
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p
a++ a+ a0	n sn sp p

Final report

Final report
Participant ID:

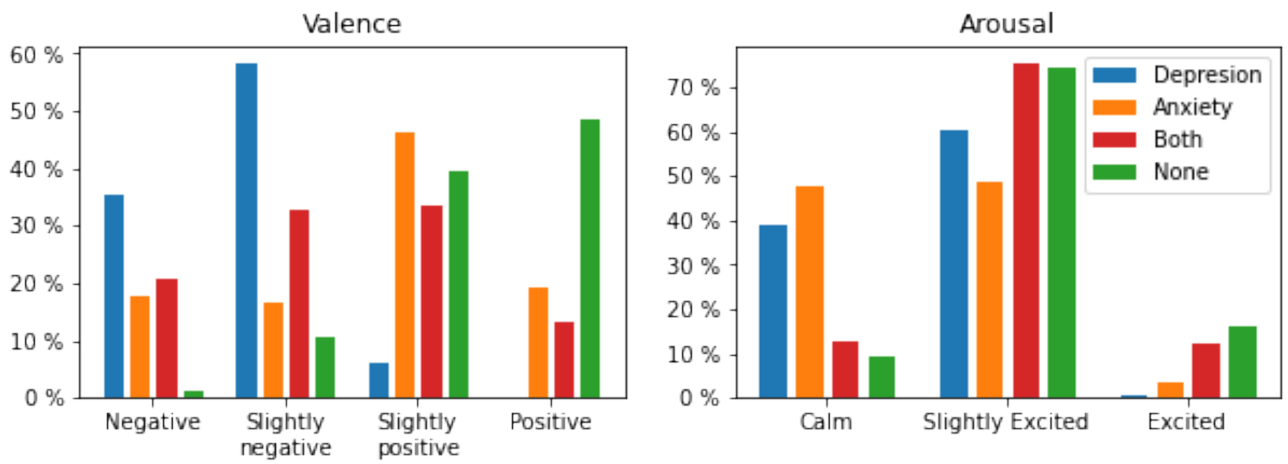
		yes	no
Before each conversation starts			
Has been informed about their right to withdraw from the study at all times without providing a reason			
Has signed the informed consent			
Has filled the questionnaire DASS21 **(only for UU-CG group)**			
Belongs to the **AMH_MG subset** to be recorded once a year during the project			
To be filled after the conversational interaction phase			

Answer to introductory question	How are you feeling now?	*One word/phrase/sentence summarizing client answer*
Main interaction (once each interaction has finished) and before the participant reads the final text)	**Level of verbosity** of the client during the interaction (according to the counsellor perception)	*One word/phrase/sentence summarizing the counselor perception of the client verbosity: Low ,Medium High, other……*
	The participant has provided **personal information** to be anonymized	yes / no
To be filled after the reading phase		
Final reading	**Level of comfort of** the client during the reading (according to the counsellor perception)	*One word/phrase/sentence summarizing the counselor perception of the client comfort with the situation: Low, Medium High, other……*
Final remarks	*Free text summarizing any client data or counselor observation.*	

Emotion, depression and anxiety in Speech

Prediction of Anxiety vs. Depression

	GeMAPS	HuBERT
Anxiety	0.64	**0.71**
Depression	0.53	**0.70**

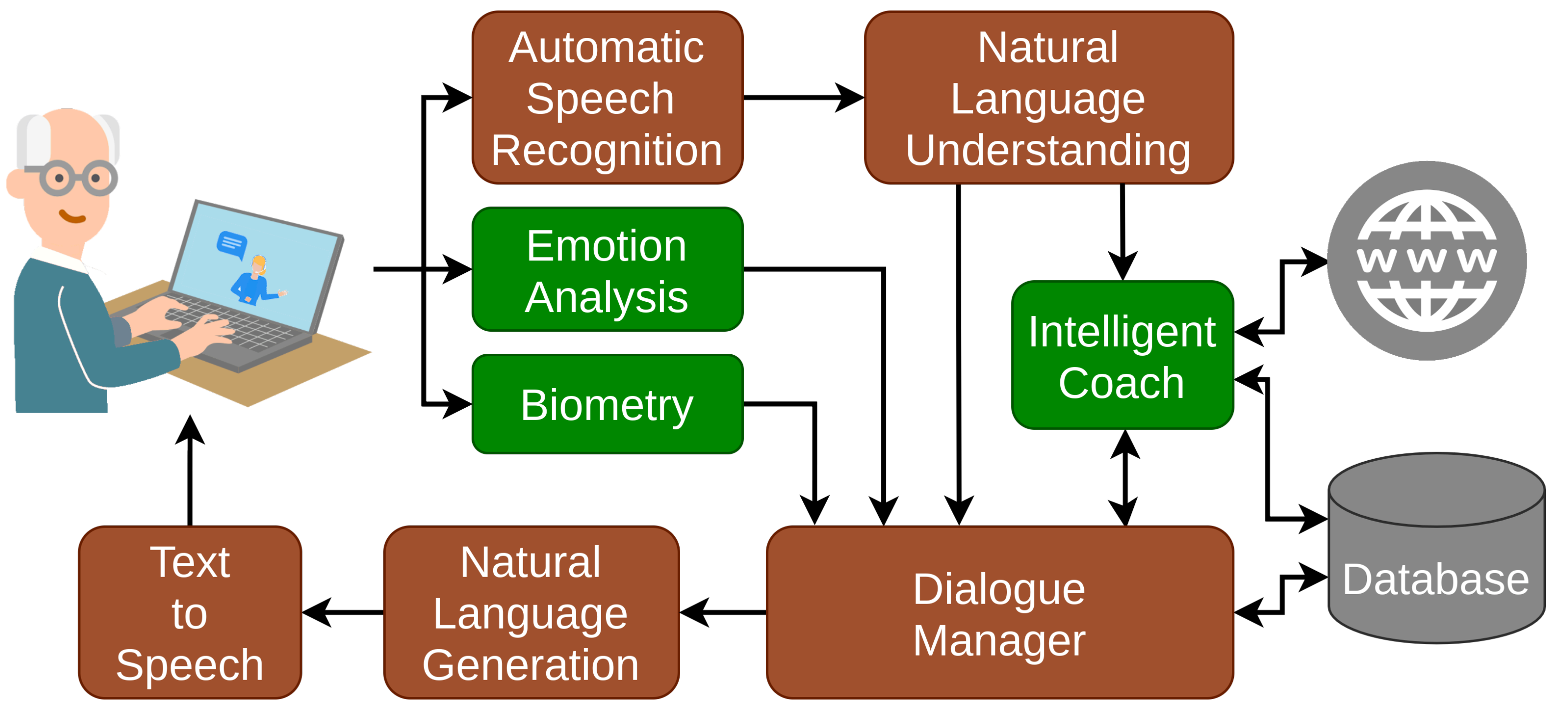


	GeMAPS	HuBERT
Valence	0.35	0.46
Arousal	0.41	**0.57**

What speech and audio analysis can provide?

- Speaker identification, language recognition, first or second Language, accent recognition, emotion recognition, some soft mental diseases features such as depression or anxiety, etc
- Paralinguistic information resulting in speaking styles: overlappings, pauses, spontaneous speech events, etc
- Genre, age, drugs, mood, noise, environment, etc etc
- Audio acoustic events: laugh, cough, different types of noise,….etc

How to Manage Emotions in Human - Machine Interaction?

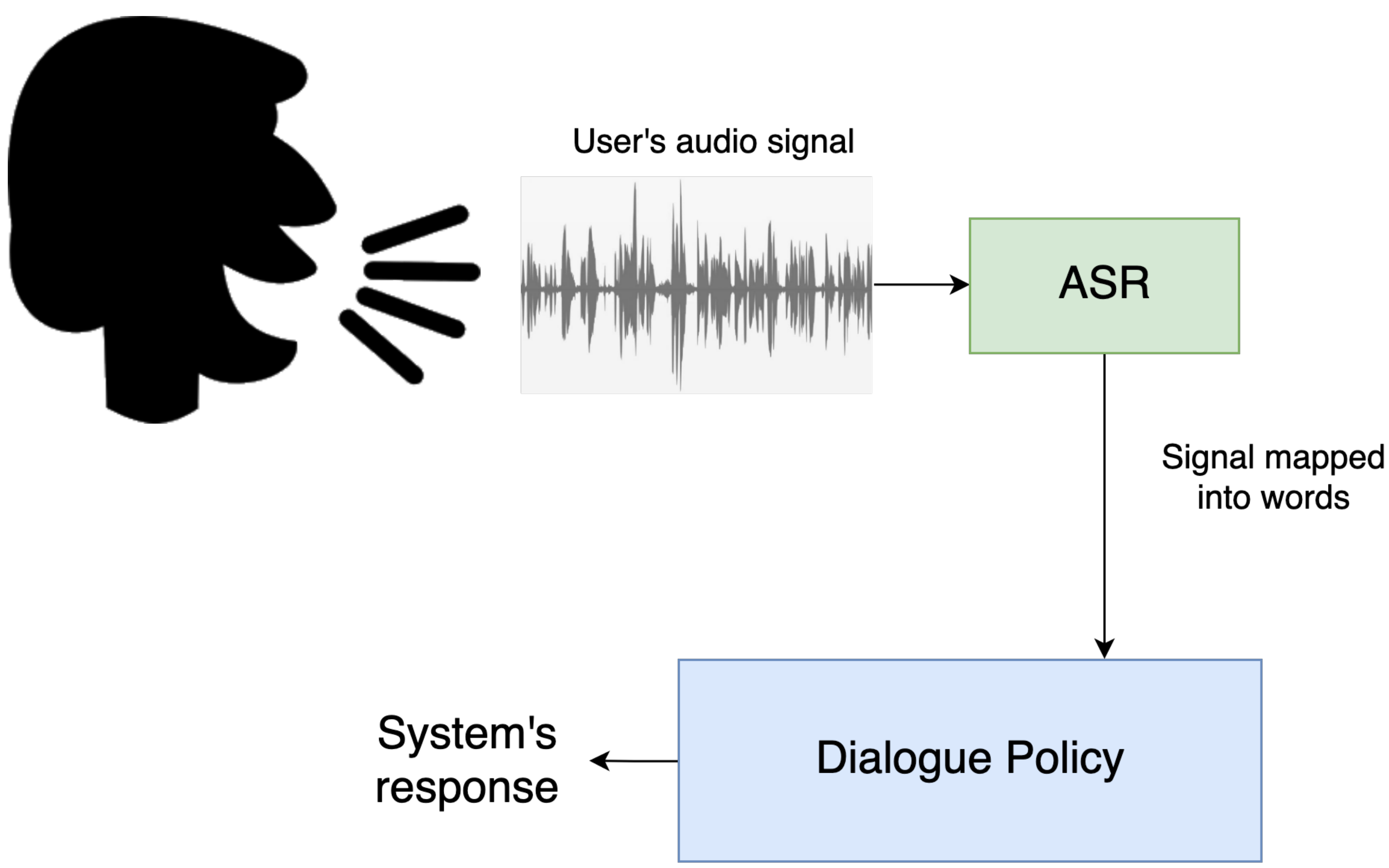


- What for:
  - A better understanding of the user profile and behaviour (NLU).
  - Assist the DM to take decisions
- The NLG and TTS have to create language and voices according to the task but mainly current according to the speaker input.
  - Simple rules.
  - Implement a behavioural model.

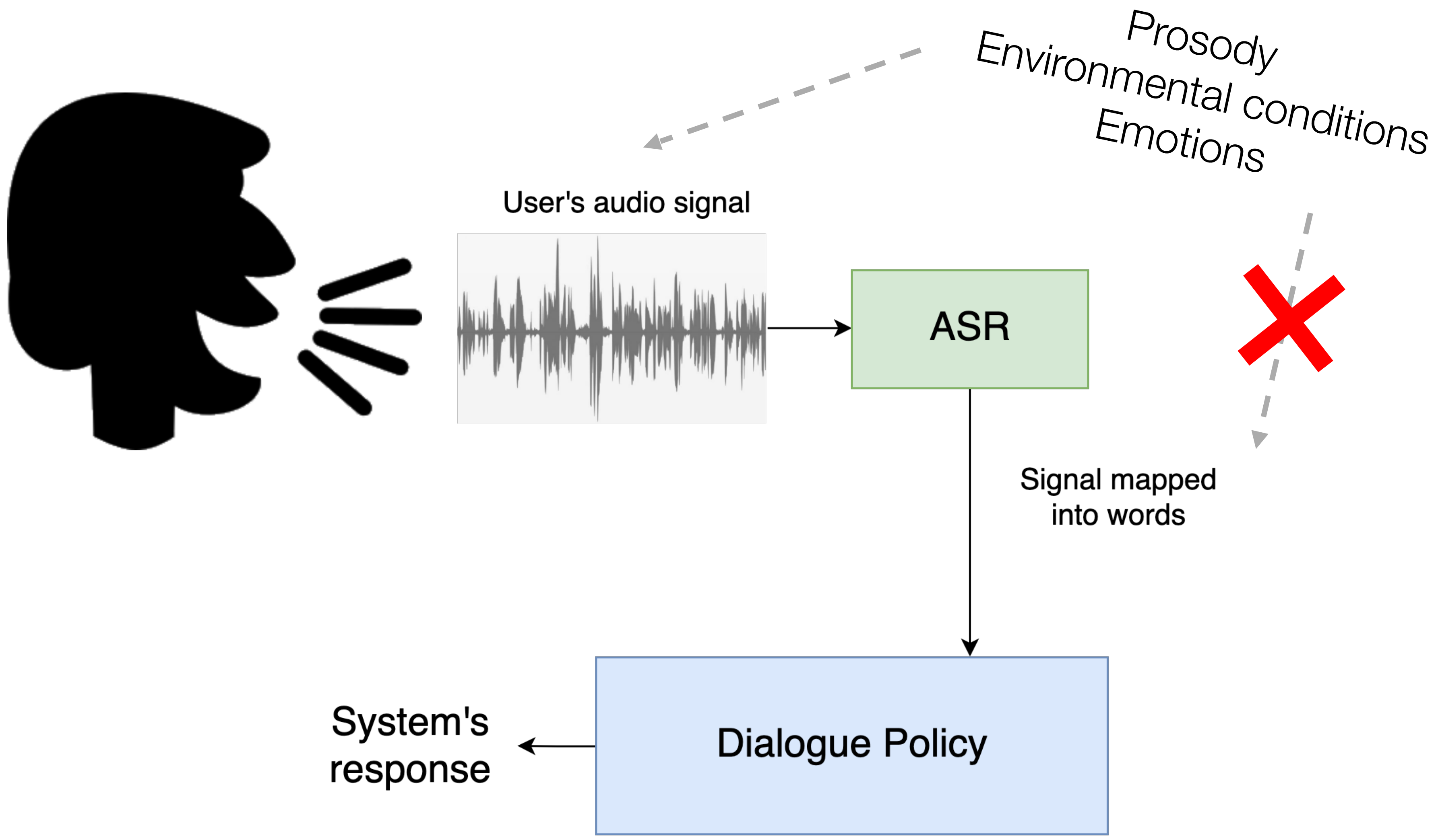
Outline

- Context of the talk
- Part I: Emotional Information in the Speech
- Part II: Speech aware policy
- Conclusions, questions, opinions, thoughts……..

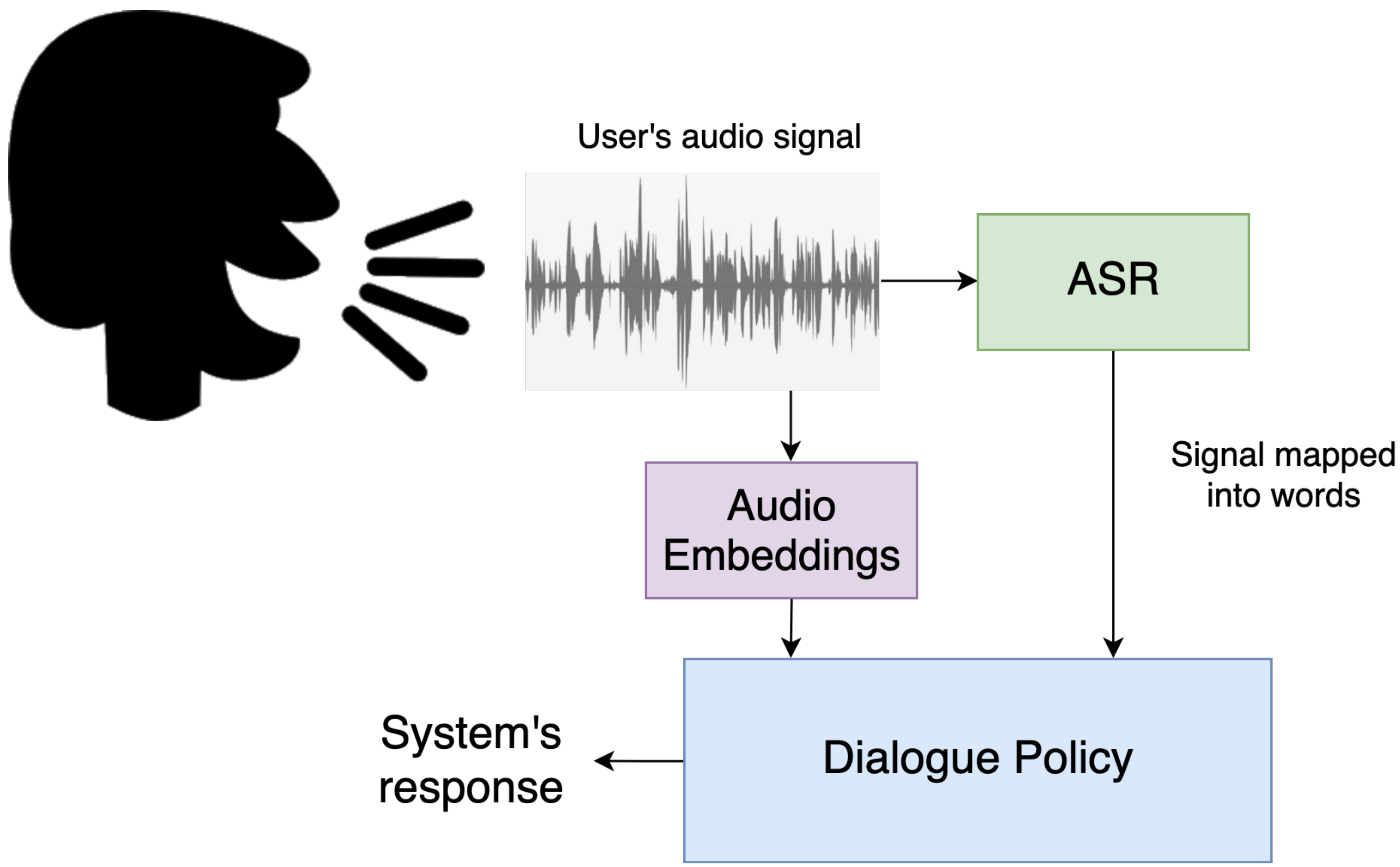
Motivation


User's audio signal
ASR
Signal mapped into words
Dialogue Policy
System's response

Motivation


Prosody
Environmental conditions
Emotions
User's audio signal
ASR
Signal mapped into words
Dialogue Policy
System's response

Our main proposal



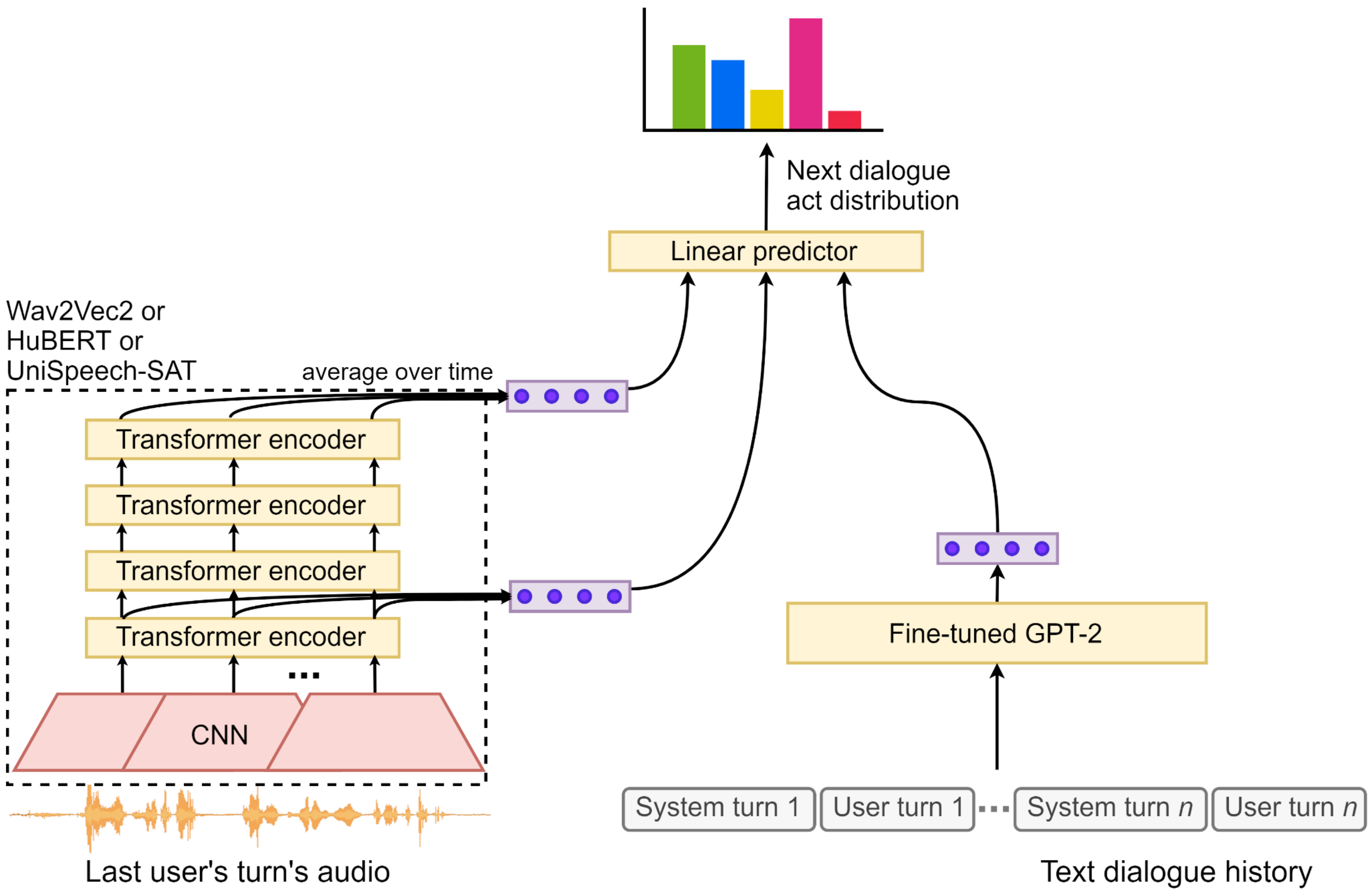
Our main proposal

- How audio embeddings can be used to include this kind of information in dialogue policies, and yield better dialogues policies
- A transformer-based DM capable of processing both the text dialogue history and the audio signal of the last user's turn.
- We compare it against a version of itself that does not explicitly process audio, in a variety of conditions and with different learning algorithms.
- We also compare three audio embedding models (Wav2Vec2 (Baevski et al., 2020), HuBERT (Hsu et al., 2021) and UniSpeech-SAT (Chen et al., 2021)) and two different methodologies to extract the speech representations from the user turns.
- Automatic metrics, human evaluation and manual inspection in the DSTC2 dataset are in favour of our hypothesis: audio embeddings help to learn better dialogue policies.

Our main proposal

Two transformer networks in parallel:

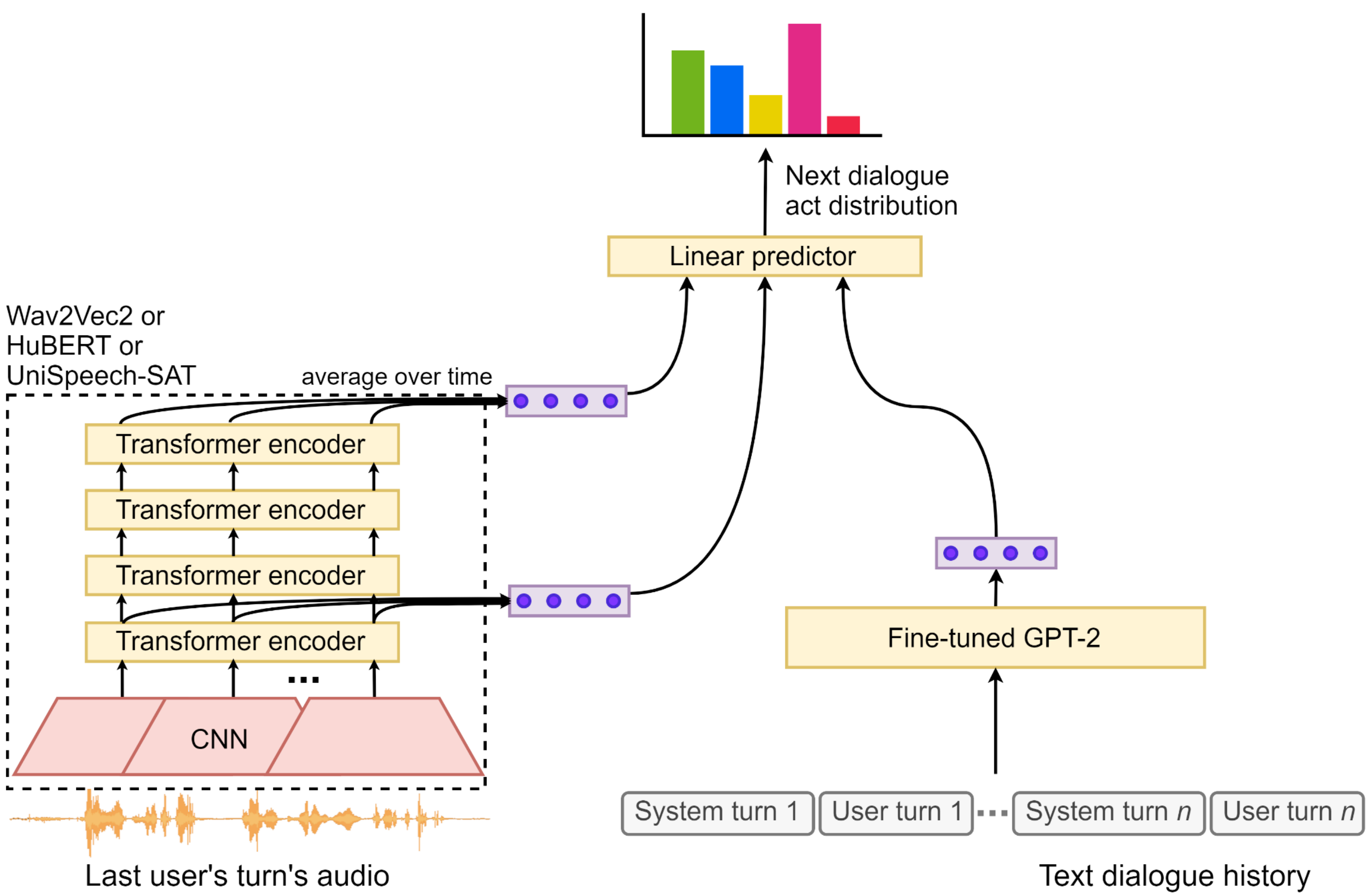
- A pretrained GPT-2 network for the text history. Each turn in the dialogue history is represented as raw text, i.e. no dialogue acts or named entities are used as input to the policies—to keep our approach as simple as possible.
- Wav2Vec2, HuBERT or UniSpeech-SAT network for the last user's turn's audio.
- A last linear layer to predict the next dialogue act.



Our main proposal

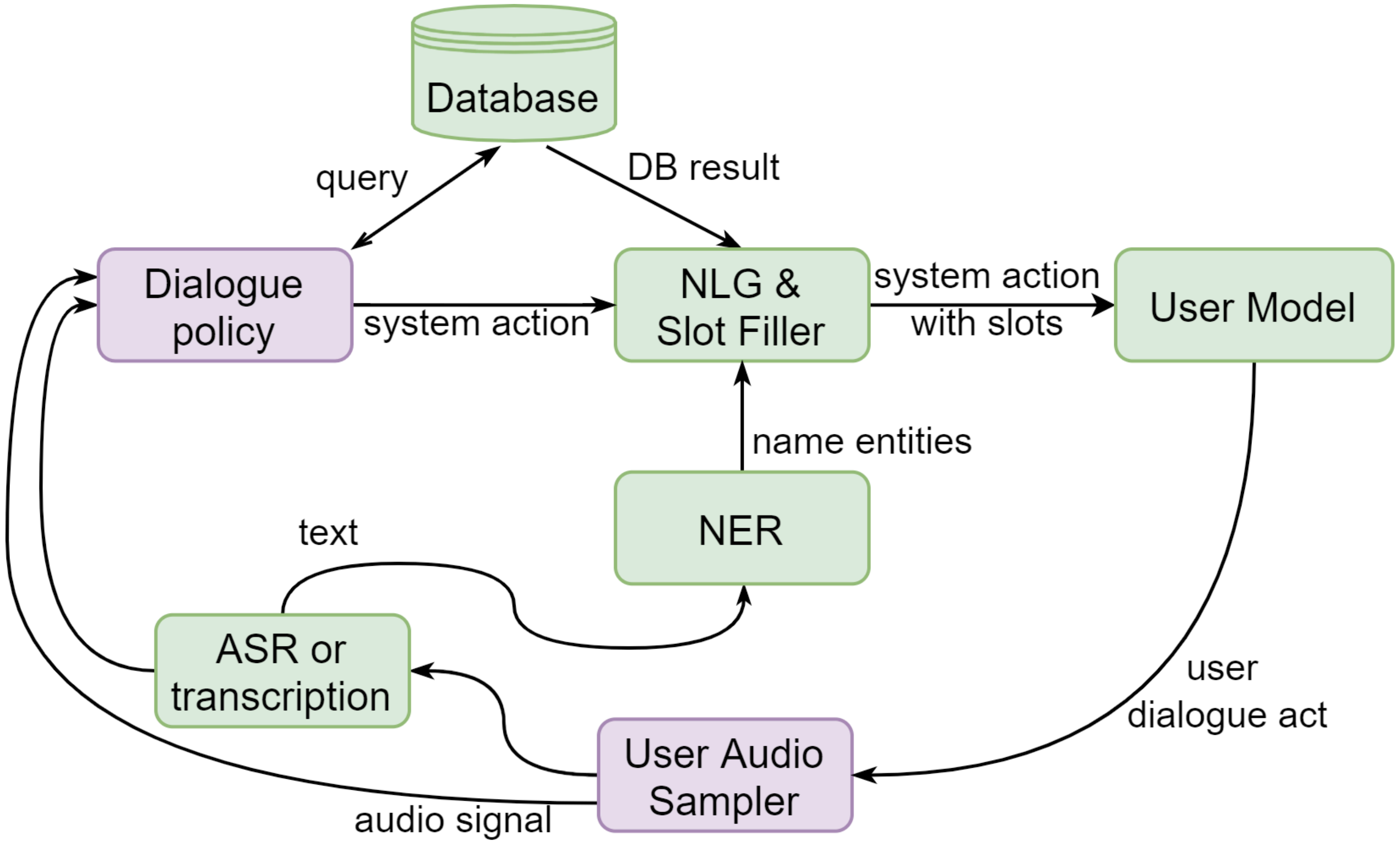
Compare dialogue policies that use the audio processing part vs. one that does not in different conditions. Mainly:

- Manual transcription or 2 ASRs of different qualities as the text input.
- Training them via Supervised Learning and two Reinforcement Learning algorithms: REINFORCE and Actor-Critic.



Experimental framework

- Corpus: DSTC2 (restaurant search).
- 3235 dialogues: 1612 for training, 1623 for test.
- DB, NLG, NER & slotfiller taken from DeepPavlov.
- UM based on bi-automata.
- ASR1: based on Wav2Vec2 | ASR2: Vosk toolkit



Experimental framework

- 3 evaluation metrics:
  - User Request Score (URS). This score indicates whether the system answers to the user in focus. This score is high if the system provides a phone number after the user has requested it.
  - System Offered Valid Venue (SOVV). This score indicates how correct the system informs are. It is the ratio of the system informs that satisfy the constraints of the user over the total informs.
  - Can't Help Score (CHS). This score (also between 0 and 1) indicates how correct the system informs are. It is the ratio of the system informs that satisfy the constraints of the user over the total informs.
  -

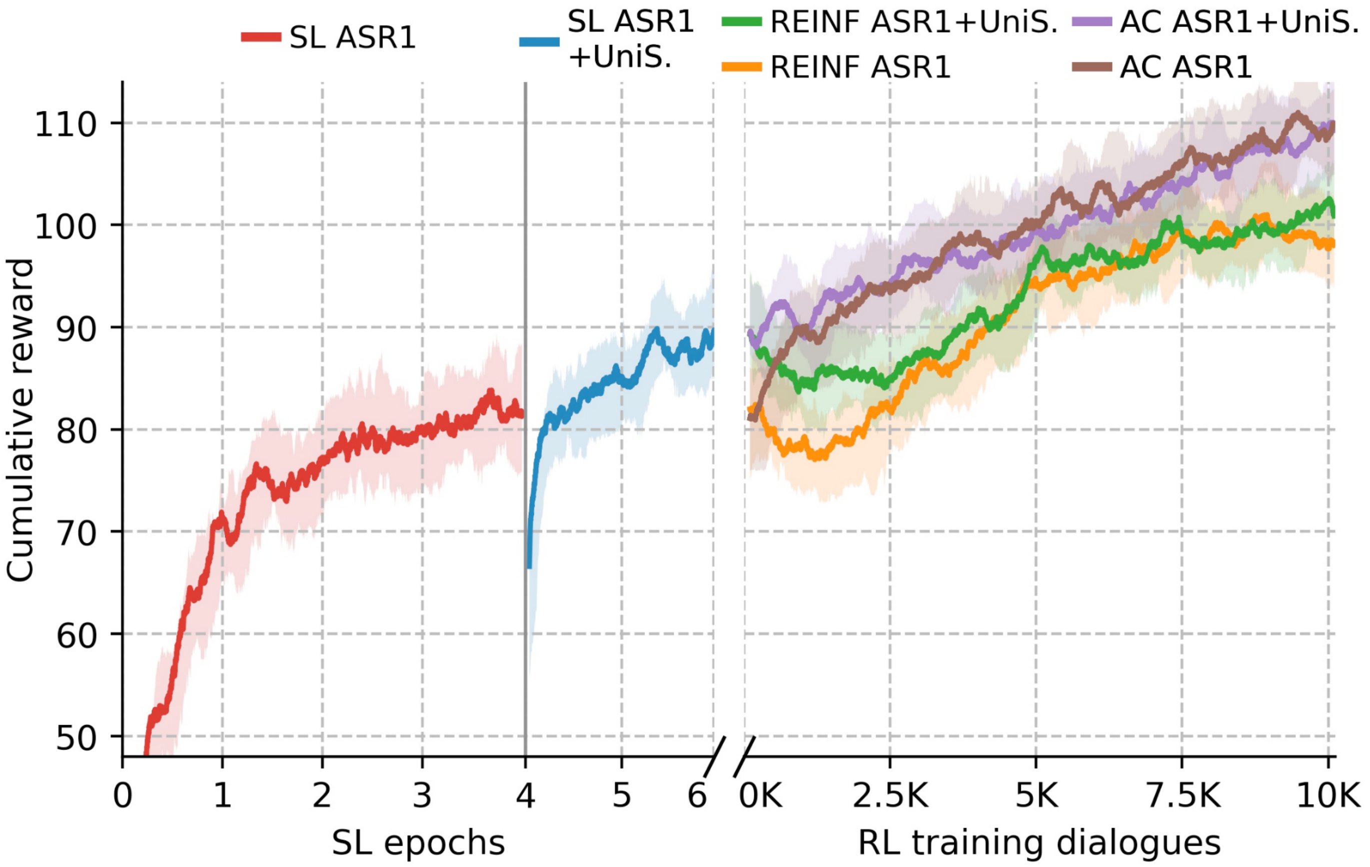
$$\text{Evaluation score} = 0.2 \cdot \text{URS} + 0.4 \cdot \text{SOVV} + 0.4 \cdot \text{CHS}.$$

AVERAGED EVALUATION METRICS USING THE TEST UM AFTER SUPERVISED LEARNING (SL), REINFORCE AND ACTOR-CRITIC, WITH DIFFERENT TEXT INPUTS AND AUDIO EMBEDDING MODELS. THE POLICIES WITH RESULTS IN PURPLE WERE PART OF THE HUMAN EVALUATION.

	SL				REINFORCE				Actor-Critic			
	Text	+W2V2	+UniS.	+HuBERT	Text	+W2V2	+UniS.	+HuBERT	Text	+W2V2	+UniS.	+HuBERT
Evaluation score												
ASR 1	0.771	0.790*	0.792*	**0.795***	0.792	0.796	**0.805***	0.796	0.818	0.820	0.822	**0.823**
ASR 2	0.934	0.935	0.932	**0.937**	0.916	0.918	**0.920**	0.911	0.927	0.930	0.931	**0.934***
TRS	0.940	**0.951***	0.947*	0.948*	0.928	0.931	**0.932**	0.928	0.947	**0.953***	0.953	0.950
Cumulative reward												
ASR 1	83.2	89.5*	90.4*	**91.4***	97.3	100.5*	**101.4***	98.3	**109.2**	105.7	106.1	108.1
ASR 2	137.8	140.3*	141.1*	**141.6***	138.9	**141.0**	139.7	139.5	**169.1**	165.3	168.6	168.3
TRS	135.4	147.0*	**147.4***	146.1*	142.9	144.4	**146.3***	144.0	**183.9**	182.9	182.1	180.6
User Request Score (URS)												
ASR 1	0.945	0.962*	**0.975***	0.969*	0.958	0.964	**0.971***	0.967*	0.987	**0.988**	0.987	0.988
ASR 2	0.984	0.988*	**0.991***	0.991*	0.982	0.984	**0.988***	0.979	0.991	0.993	0.993	**0.993**
TRS	0.974	0.986*	**0.989***	0.987*	0.975	0.978	**0.981**	0.979	0.991	0.992	**0.992**	0.992
System Offered Valid Venue (SOVV)												
ASR 1	0.750	0.766*	0.762*	**0.768***	0.773	0.774	**0.783**	0.775	0.791	0.793	0.795	0.796
ASR 2	0.917	0.920	0.912	**0.921**	0.894	0.901	**0.902**	0.896	0.909	0.911	0.913	**0.917***
TRS	0.880	**0.938***	0.932*	0.935*	0.912	**0.919**	0.919	0.918	0.936	**0.943**	0.942	0.938
Can't Help Score (CHS)												
ASR 1	0.668	0.701*	**0.721***	0.703*	0.629	0.643	**0.651**	0.629	0.674	0.703*	**0.713***	0.695
ASR 2	0.967	**0.968**	0.965	0.966	**0.922**	0.906	0.905	0.895	0.940	0.942	**0.954***	0.950
TRS	**0.989**	0.988	0.978	0.983	**0.925**	0.915	0.915	0.887	0.943	0.963*	**0.965***	0.959*
SOVV-CHS combined score												
ASR 1	0.747	0.766*	0.766*	**0.768***	0.758	0.761	**0.770***	0.760	0.779	0.788	**0.791**	0.789
ASR 2	0.925	**0.928**	0.921	0.928	0.899	0.902	**0.903**	0.897	0.913	0.918	0.921	**0.924***
TRS	0.900	**0.945***	0.939*	0.942*	0.914	0.917	**0.917**	0.912	0.934	0.946	**0.964***	0.941

Some results – learning curves

- Audio embeddings help the most with the noisier ASR, and with Supervised Learning.
- RL policies learn to deal with uncertainty better and thus they improve less.



Some results – human evaluation

Table 5.5.: Questionnaire used by judges in the human evaluation.

Q1: The system offered a restaurant satisfying the user constraints, or correctly informed that there were no such restaurants.
- Yes. (1)
- No. (0)

Q2: The system provided the information the user was looking for (phone number, post code, address...).
- Yes. (1)
- Partially. (0.5)
- No. (0)
- None–if there are no user requests.

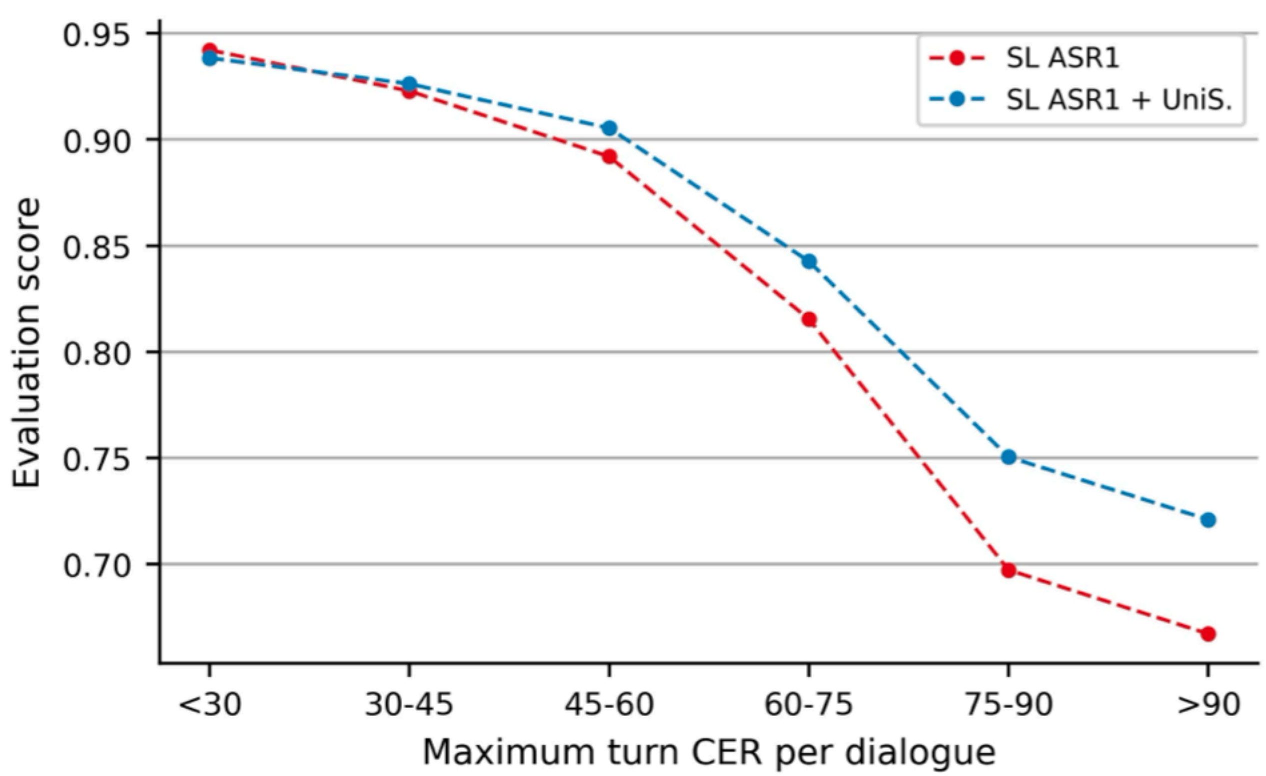
Q3: The conversation felt natural.
- Strongly agree. (1)
- Agree. (0.75)
- Neither agree nor disagree. (0.5)
- Disagree. (0.25)
- Strongly disagree. (0)

Table 5.6.: Human evaluation results.

#	Algo.	Input	**Q1**	**Q2**	**Q3**	**Avg.**
1	SL	ASR1	0.656	0.848	0.535	0.629
2	SL	ASR1+UniS.	**0.760***	**0.902***	**0.601***	**0.716***
3	REINF	ASR1	0.730	0.892	**0.637**	0.716
4	REINF	ASR1+UniS.	**0.762**	**0.901**	0.632	**0.721**
5	AC	ASR1	0.761	**0.919**	**0.605**	0.718
6	AC	ASR1+UniS.	**0.789**	0.907	0.585	**0.719**

Some results

- We generated and analysed a number of simulated dialogues with policies she the text processing part.
- Many times dialog strategies develop similarly whether the use policies with or without speech representations



Example 1: better user understanding

Context	Only ASR 1, SL policy continuation	ASR 1 + UniSpeech-SAT, SL policy continuation
S: Hello, welcome to the Cambridge restaurant system. You can ask for restaurants by area, price range or food type. How may I help you? U: looking for a restaurant in the east part of town S: What kind of food would you like? U: chinese (ASR 1: try me) S: <api_call> S: sitar tandoori is a nice place in the east of town and the prices are expensive. U: can i have the address of the venue (ASR 1: and add the atters of devinon) S: Sure, sitar tandoori is on 43 high street cherry hinton cherry hinton. U: can i get the phone number (ASR 1: figet the faw number) S: The phone number of sitar tandoori is 01223 249955. U: [noisy] whats the post code (ASR 1: hirst)	S: <api_call> S: pipasha restaurant is a nice place in the east of town and the prices are expensive. U: [noise] (ASR 1:) S: Sure, pipasha restaurant is on newmarket road fen ditton. U: post code (ASR 1: i was going) S: pipasha restaurant is a great restaurant. U: and for good luck what is the post code (ASR 1: and for good luck what is the postcod) S: The post code of pipasha restaurant is c.b 5, 8 p.a.	S: The post code of sitar tandoori is c.b 1, 9 h.x.

Example 2: better noise/not understanding awareness

Context

S: Hello, welcome to the Cambridge restaurant system. You can ask for restaurants by area, price range or food type. How may I help you?
U: [sluttering] yes indian food
(ASR 1: yes in the edfu)

Only ASR 1, REINFORCE policy continuation

S: <api_call>
S: hk fusion serves chinese food
U: indian food
(ASR 1: indico)
S: Sure, hk fusion is on 21 burleigh street city centre.
U: indian
S: <api_call>
S: curry garden serves indian food
U: phone number
(ASR 1: falg in)
S: Sure, curry garden is on 106 regent street city centre.
U: phone number.
(ASR 1: bo number)
S: The phone number of curry garden is 01223 302330.

ASR 1 + UniSpeech-SAT, REINFORCE policy continuation

S: Sorry I can't hear you.
U: indian
S: <api_call>
S: india house serves indian food
U: is there anything else
S: <api_call>
S: cocum serves indian food
U: can i have the address and phone number
(ASR 1: can at the actebundefore number)
S: the phone number of cocum is 01223 366668.
U: address
(ASR 1: aga)
S: Sure, cocum is on 71 castle street city centre.

Asier López Zorrilla, María Inés Torres and Heriberto Cuayáhuitil (2021): “Audio Embeddings Help to learn better Dialogue Policies” IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena de Indias, Colombia, December 13-17 (Virtual)

Asier López-Zorrilla, M. Inés Torres and Heriberto Cuayáhuitl (2022): “Audio embedding-aware dialogue policy learning” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Vol 31, pp 525-538 DOI: 10.1109/TASLP.2022.3225658

Speech Aware Dialogue Management

Conclusions, questions, opinions, thoughts……..

RTTH Fall School - Jaca
November, 2023