Towards Learning to Speak and Hear Through Multi-agent Communication over a Continuous Acoustic Channel

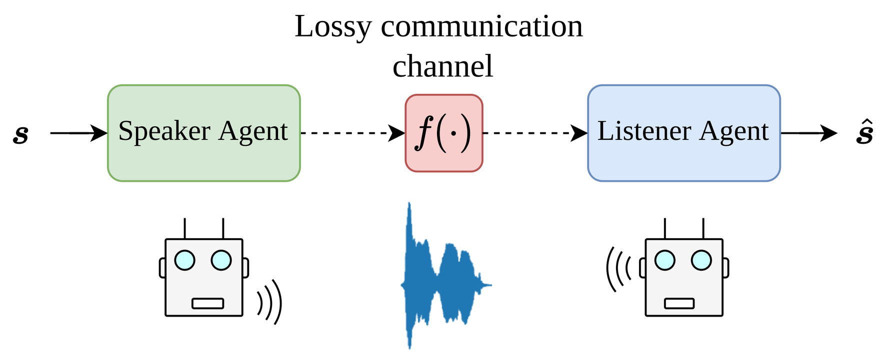
Kevin Eloff[1,2*] Arnu Pretorious[2] Herman A. Engelbrecht[1] Herman Kamper[1]

[1]MediaLab, E&E Engineering, Stellenbosch University, South Africa

[2]InstaDeep, Cape Town, South Africa

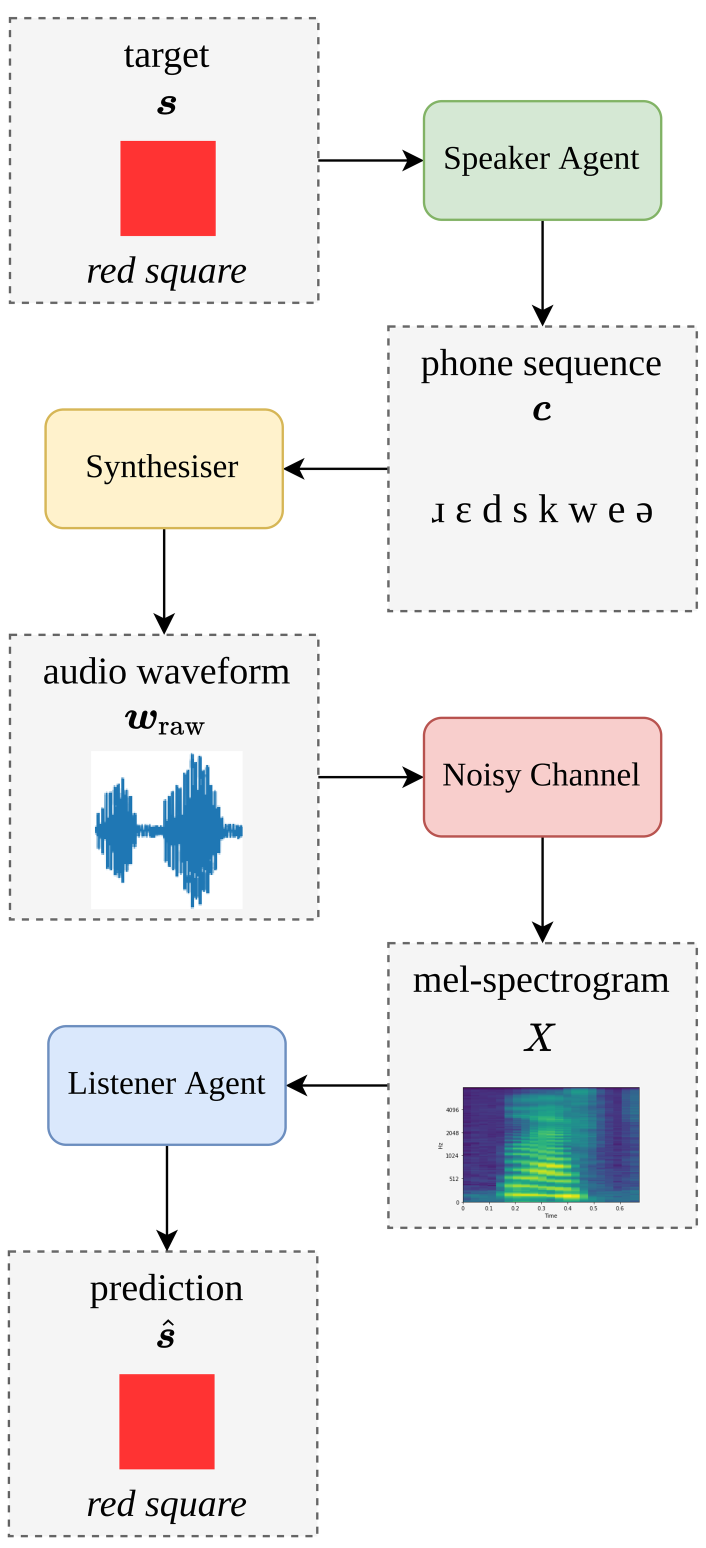
Background

- Multi-agent reinforcement learning has proven effective for investigating emergent communication.
- However, most studies focus on communication with discrete symbols.
- Humans learn language over a continuous channel and language evolved through spoken communication.
- Are we able to observe emergent language between agents with a continuous communication channel?
- We provide a platform to study emergent continuous signalling in order to see how it relates to human language acquisition and evolution.
- We propose a messaging environment where a Speaker agent needs to convey a set of attributes to a Listener over a noisy acoustic channel.



Environment

- Let $\boldsymbol{s}$ represent a set of attribute values the Speaker must communicate to a Listener agent.
- Taking these attributes as input, the Speaker produces a waveform as output, which passes over a lossy acoustic channel.
- The agents must develop a common communication protocol such that $\boldsymbol{s} = \hat{\boldsymbol{s}}$.



Speaker agent

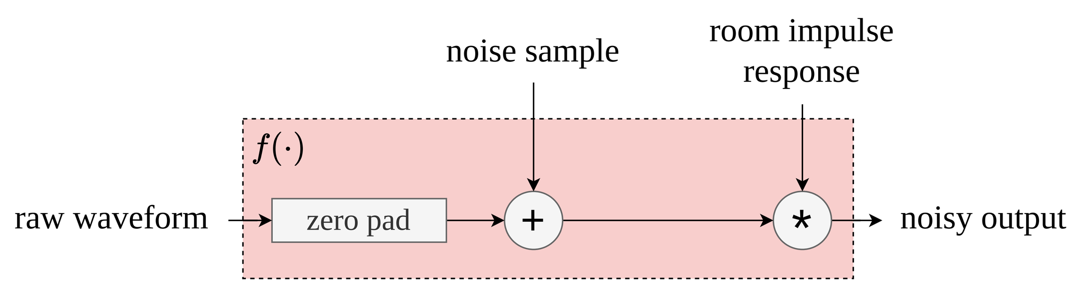
- Speaker agent generates a phone sequence $\boldsymbol{c}$ given $\boldsymbol{s}$.
- GRU-based sequence generation model.
- Speaker is able to generate arbitrary length sequences, up to a maximum length.

Listener agent

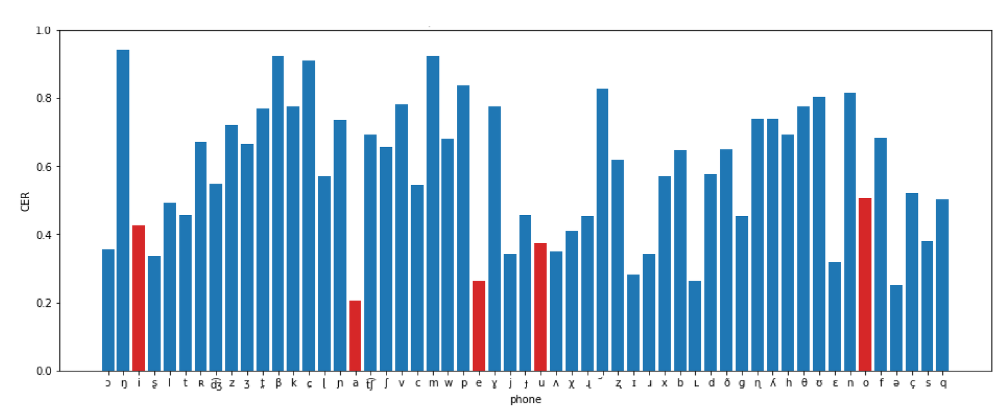
- **End-to-end:** The Listener agent produces $\hat{\boldsymbol{s}}$ directly from the mel-spectrogram X.
  - CNN-GRU based architecture.
  - No intermediary steps from X to $\hat{\boldsymbol{s}}$.
- **Phone recogniser:** A static pre-trained phone recogniser combined with a discrete Listener.
  - First extract a phone sequence from X, which is consumed by a discrete GRU-based Listener agent.
  - Phone recogniser trained to 6.42% CER.

Realistic communication channel

- Implementation of channel function $f(\cdot)$:



- The channel samples background noise and a room impulse response in each pass.
- CER per phone in the evaluation channel:



Different approaches in noisy environments

- Per attribute accuracy of various models:

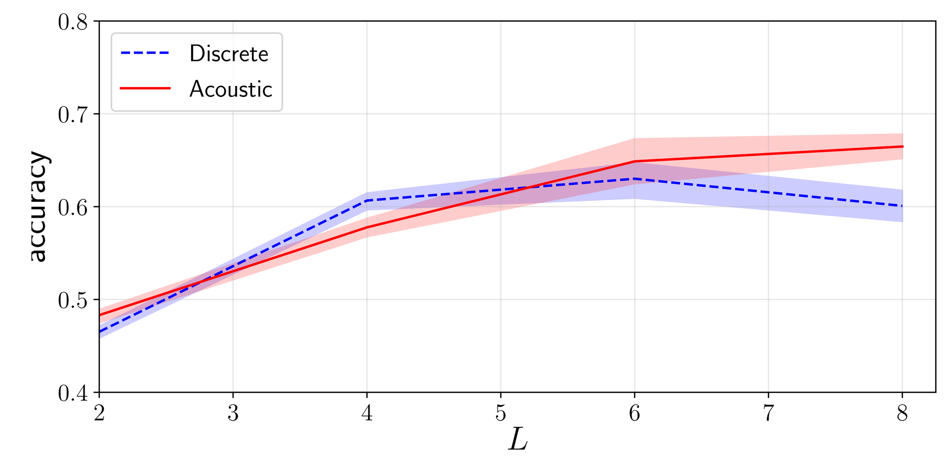
Model	*Training rooms*	*Eval. rooms*
Discrete baseline	0.621	0.612
Acoustic end-to-end	0.611	0.566
Acoustic* end-to-end	**0.973**	**0.958**
Acoustic + phone recogniser	0.539	0.535
Acoustic* + phone recogniser	0.609	0.591

- The discrete baseline is first trained in the discrete task, and then used with a phone mapping and phone recogniser during evaluation.
- Acoustic* uses the discrete baseline for pretraining.

Increasing sequence length

			No background noise			10dB SNR background noise		
Model	L	*no room*	*training rooms*	*meeting*	*stairway*	*training room*	*meeting*	*stairway*
Discrete	5	0.977	0.720	0.742	0.519	0.621	0.675	0.549
Acoustic	5	0.863	0.679	0.680	0.500	0.609	0.650	0.532
Discrete	8	**1.000**	0.714	**0.789**	0.559	0.575	0.660	0.544
Acoustic	8	0.867	**0.758**	0.759	**0.601**	**0.668**	**0.707**	**0.619**

- Per-attribute accuracy as a function of maximum phone length L:



Emergent redundancy

- Samples of phones produced by the **Acoustic Speaker:**

$\boldsymbol{s}_2$ \ $\boldsymbol{s}_1$	0	1	2	3	4
0	oua u**aoao**	uoo**aoao**a	oouo**aoao**	**ao**uuu**ao**o	uuea**ao**aa
1	eaoaaaaa	eo**aoaoao**	oo**aoao**aa	**ao**o**ao**o**ao**	eeaa**ao**aa
2	uuouuuau	uuuuouuu	ouuouuua	auuuuuuo	uuuuuuua
3	uouua**ao**a	uouua**ao**a	ouou o**ao**o	auouuuua	uuue**aoao**
4	uuue**aoao**	uoo**aoao**o	oouo**ao**oa	auoouo**ao**	ue**aoaoao**

- Samples of phones produced by the **Discrete Speaker:**

$\boldsymbol{s}_2$ \ $\boldsymbol{s}_1$	0	1	2	3	4
0	ouiaeaao	uoeaioia	oouea iao	auaoieau	ueeaiaeo
1	eoeauaie	eoeaeioe	ooeeaioa	aeaoeaui	eeeaeioe
2	uoiuauei	uouiuaoe	ououiaou	auuaouiu	uueauiai
3	uoiaeeou	uouaeiai	oouea uia	auaouiea	ueuaeaii
4	euaoiaee	euoaioia	oouea aio	auaoiaea	eueaeaio

- **Acoustic Speaker:** 2.477 repeated phones per utterance and 2.810 unique phones per utterance.
- **Discrete Speaker:** 1.258 repeated phones per utterance and 3.978 unique phones per utterance.

Conclusion

- We have laid the foundation for answering the larger question of whether we can observe emergent language over continuous acoustic channel trained through RL.
- We allow our agents to generate unique audio waveforms.
  - Speaker uses discrete units, could consider continuous articulation in future.
- We observe that the acoustic Speaker learns redundancy which improves Listener coherency.
  - An example of emergent linguistic behaviour that is not modelled in a purely discrete setting.
- **Future:** Multi-round communication games between two or more agents.

*Work done during an internship at InstaDeep, Cape Town, South Africa.

k.eloff@ml.sun.ac.za