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

Kevin Eloff\textsuperscript{1,2*} \hspace{1em} Arnu Pretorious\textsuperscript{2} \hspace{1em} Herman A. Engelbrecht\textsuperscript{1} \hspace{1em} Herman Kamper\textsuperscript{1}

\textsuperscript{1}MediaLab, E&E Engineering, Stellenbosch University, South Africa  \hspace{1em} \textsuperscript{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 a 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 \( s = \hat{s} \).

Speaker agent
- Speaker agent generates a phone sequence \( c \) given \( s \).
- 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 \( \hat{s} \) directly from the mel-spectrogram \( X \).
  - CNN-GRU based architecture.
  - No intermediary steps from \( X \) to \( \hat{s} \).
- 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:

<table>
<thead>
<tr>
<th>Model</th>
<th>Training error</th>
<th>Eval. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete baseline</td>
<td>0.823</td>
<td>0.822</td>
</tr>
<tr>
<td>Acoustic real-end</td>
<td>0.811</td>
<td>0.819</td>
</tr>
<tr>
<td>Acoustic* real-end</td>
<td>0.973</td>
<td>0.958</td>
</tr>
<tr>
<td>Acoustic* = phone recogniser</td>
<td>0.530</td>
<td>0.535</td>
</tr>
<tr>
<td>Acoustic* = phone recogniser</td>
<td>0.809</td>
<td>0.561</td>
</tr>
</tbody>
</table>

  - 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.

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