ArticulationGAN: Unsupervised Modelling of Articulatory Learning

Gašper Beguš\textsuperscript{1}, Alan Zhou\textsuperscript{2}, Peter Wu\textsuperscript{1}, and Gopala Anumanchipalli\textsuperscript{1}

\textsuperscript{1}University of California, Berkeley, \textsuperscript{2}Johns Hopkins University
\{begus, peterw1, gopala\}@berkeley.edu, azhou23@jhu.edu

\textbf{Introduction}

- Most generative speech synthesis models are trained to directly generate waveforms or spectral data
- Humans, however, produce speech by performing articulatory gestures
- Can a deep neural network learn to produce speech with human-like articulatory gestures given only an unsupervised training objective?

\textbf{Model Architecture}

- Three subnetworks in a GAN framework
  - An Articulatory Generator that takes in random noise and generates synthetic electromagentic articulography (EMA) data to pass to a physical model
  - A pre-trained Physical Model that transforms articulatory gestures from the Generator into a speech waveform
  - A Discriminator that receives the outputs from the articulatory model or real speech data and produces a realism store
- During training, we freeze the physical model, and update the generator and discriminator according to a WGAN-GP training objective
- We train the model on 8 words from TIMIT, and compare the Articulatory Generator’s outputs with real EMA data

\textbf{Training Data}

- The physical model was trained on the MNGU\textsubscript{30} dataset, consisting of articulatory data from one male British English speaker
- The rest of the model was trained on 8 words sliced from TIMIT (\textit{ask, dark, year, water, wash, rag, oily, and greasy})

\textbf{Model Performance}

- A trained phonetician was hired to transcribe speech outputs of tested models and annotate them as \textit{Intelligible, Unintelligible}. Intelligible outputs were further annotated as \textit{Innovative} if they did not appear in the training data
- Overall, while ArticulationGAN was less intelligible than WaveGAN, its intelligible outputs were much more innovative

\textbf{Smoothing}

- As the articulatory generator is not penalized for producing extremely fast movements, we smooth the outputs using LOESS smoothing

\textbf{Quantitative Comparison}

\begin{table}[h]
\begin{tabular}{lcccc}
& wash & fast & tongue tip & tongue body & lower lip & upper lip & lower incisor & tongue dorsum \\
\hline
WaveGAN & 174 (87%) & 26 (13%) & 87 (50%) & 70 (37%) & 30 (17%) & 70 (37%) & 70 (37%) & 70 (37%) \\
ArticulationGAN & 143 (72%) & 57 (29%) & 110 (77%) & 70 (37%) & 30 (17%) & 70 (37%) & 70 (37%) & 70 (37%) \\
\end{tabular}
\end{table}

- We see similar gestures between real and generated EMA
  - For wash (left), tongue gestures are extremely similar
  - For fast (right), we see almost identical patterns for gestures in tongue tip and lower lip, and high correlations elsewhere

\textbf{Conclusions}

- Our model is able to generate human-like articulatory gestures in a fully unsupervised setting
- While our model is somewhat less intelligible than a traditional mode, it also produces a much higher proportion of innovative intelligible outputs
- We argue that this model is not only a more cognitively plausible model of how humans learn to produce speech, but also potentially useful for creating more realistic speech synthesis technologies

\textbf{References}

- See paper.