ArticulationGAN: Unsupervised Modelling of Articulatory Learning

Gašper Beguš[1], Alan Zhou[2], Peter Wu[1], and Gopala Anumanchipalli[1]

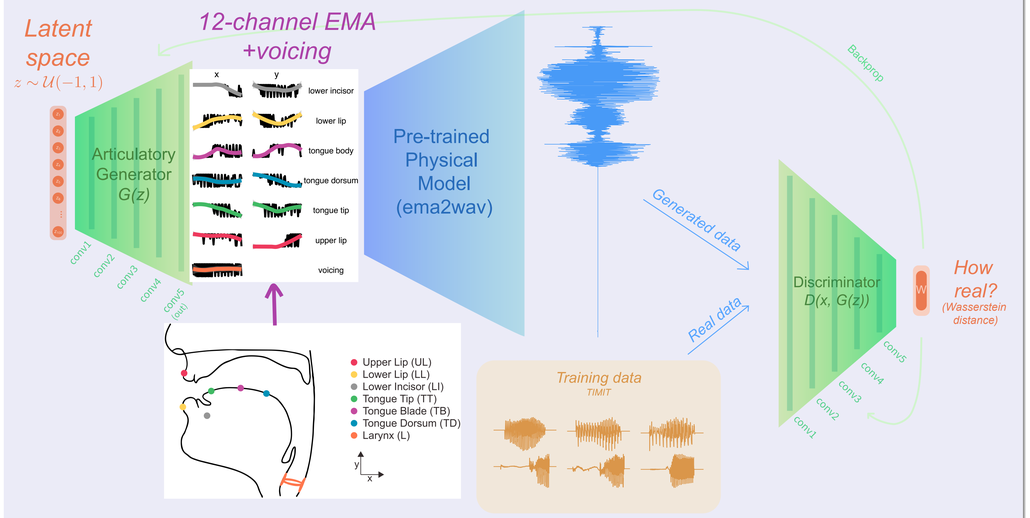
[1]University of California, Berkeley, [2]Johns Hopkins University
{begus, peterw1, gopala}@berkeley.edu, azhou23@jhu.edu

5406
ICASSP 2023
4 - 10 JUNE, RHODES ISLAND, GREECE

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 a unsupervised training objective?

Model Architecture

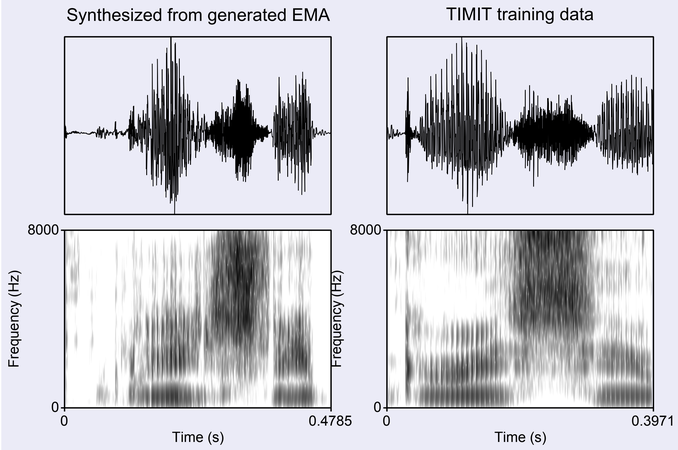


- Three subnetworks in a GAN framework
  - An **Articulatory Generator** that takes in random noise and generates synthetic electromagentic articulography (EMA) data to pass to a physical model
  - A *pre-trained* **Physical Model** that transforms articulatory gestures from the Generator into a speech waveform
  - A **Discriminator** that receives the outputs from the articulatory model or real speech data and produces a *realness 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

Training Data

- The physical model was trained on the MNGU0 dataset, consisting of articulatory data from one male British English speaker
- The rest of the model was trained on 8 words sliced from TIMIT (*ask, dark, year, water, wash, rag, oily,* and *greasy*)

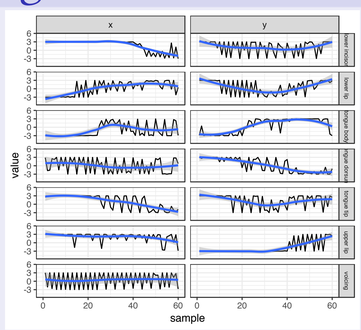
Model Performance



Model	Intelligible	Unintelligible	Innovative
WaveGAN	174 (87%)	26 (13%)	87 (50%)
ArticulationGAN	143 (72%)	57 (29%)	110 (77%)

- A trained phonetician was hired to transcribe speech outputs of tested models and annotate them as *Intelligible*, *Unintelligible*. Intelligible outputs were further annotated as *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

Smoothing



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

Quantitative Comparison

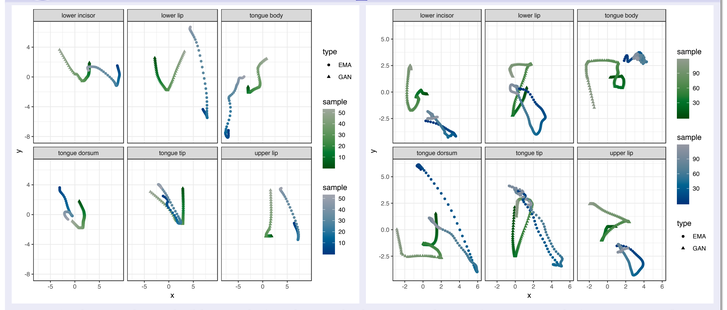


Figure: Real EMA channels (blue circles) and smoothed, generated EMA (green triangles) in 2D space for output transcribed as *wash* (left) and *fast* (right).

	wash		fast	
Place	x	y	x	y
tongue tip	0.70	0.90	0.99	0.96
tongue body	0.94	0.91	0.32	0.79
lower lip	-0.52	0.70	0.85	0.94
upper lip	0.51	0.90	0.64	0.43
lower incisor	0.87	0.66	0.31	0.72
tongue dorsum	0.41	0.91	0.24	0.89

Table: Pearson's product-moment correlation (r) for *wash* and *fast* after DTW alignment of two time series.

- 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

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 potentually useful for creating more realistic speech synthesis technologies

References:
See paper.

Manuscript:
`arxiv.org/pdf/2210.15173.pdf`