USC Viterbi School of Engineering
Ming Hsieh Department of Electrical and Computer Engineering

Acoustic-to-Articulatory Inversion for Dysarthric Speech: Are Pre-Trained Self-Supervised Representations Favorable?

Sarthak Kumar Maharana, Krishna Kamal Adidam, Shoumik Nandi, Ajitesh Srivastava

Introduction

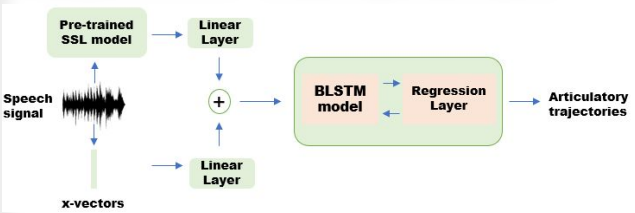
- **Dysarthria**: Speech disorder causing a decline in speech clarity that affects the articulators like the lips, jaw, tongue, velum, etc.
- Estimating articulatory movements from acoustic embeddings, called **acoustic-to-articulatory inversion** (AAI), has a one-to-many mapping and is non-linear.
- **Challenge:** The collection of acoustic- articulatory data is very tedious due to sensor fall-off.
- **Objective:** Perform AAI on dysarthric speech by studying the effectiveness of pre-trained SSL features, over the standard MFFCs.
- **Findings:** Pre-trained SSL models like wav2vec, APC, and DeCoAR, tend to capture the complex dysarthric articulatory trajectories well.

Dataset

- **TORGO Dataset:** 4 speakers - 2 healthy controls (MC01, MC04) and 2 patients (F03, F04) with **complete** parallel acoustic-articulatory data.
- **Electromagnetic Articulography (EMA):** Articulatory movements of 6 articulators, at 100 Hz, in the X and Y directions.
- Sensor coils were attached to the tip, middle, and back of the tongue, as well as the jaw and lips.
- **Features:** 12-dim articulatory features + 6-dim velocity + 6-dim acceleration = 24-dim features.
- **Data:** An average of 7535 and 3066 utterances from healthy controls and patients, respectively.

Modeling

- **Inputs:** Concatenation of pre-trained SSL acoustic features + x-vectors.
- **Model:** BLSTM + Linear regression layers.



Experimental Setup

- Sampling rate: 100 Hz for articulatory features.
- **Pre-trained SSL models: wav2vec, APC, NPC, DeCoAR, TERA, Mockingjay, vq_wav2vec.**
- **Baseline feature:** 39-dim MFCCs (25 ms window, 10 ms shift).
- **Training schemes: Seen** subject conditions- Subject-specific, Pooled, and Fine-tuned. **Unseen** subject conditions - Leave-one-person-out method.
- **Evaluation metric**: Pearson Correlation Coefficient (CC) between the ground-truth and predicted articulatory trajectories.

Results

Seen evaluation (averaged CC (std) across all articulators, sentences and folds)

Features	Subject-specific		Pooled		Fine-tuned	
	Healthy Controls	Patients	Healthy Controls	Patients	Healthy Controls	Patients
MFCCs	0.7627 (0.0578)	0.5534 (0.1374)	0.7493 (0.048)	0.5436 (0.1194)	0.7629 (0.0572)	0.5808 (0.1200)
wav2vec	0.7648 (0.0561)	0.5591 (0.1283)	0.756 (0.0437)	0.5808 (0.1106)	0.7649 (0.0554)	0.593 (0.1216)
APC	0.7544 (0.0596)	0.5438 (0.1265)	0.7481 (0.0458)	0.5717 (0.1159)	0.7642 (0.0551)	0.5867 (0.1224)
NPC	0.7561 (0.059)	0.5441 (0.1443)	0.7501 (0.0468)	0.5421 (0.1116)	0.7592 (0.0594)	0.5563 (0.1357)
DeCoAR	**0.7699 (0.0540)**	**0.5832 (0.1344)**	**0.7628 (0.0445)**	**0.5928 (0.1187)**	**0.7767 (0.0562)**	**0.6073 (0.1312)**
TERA	0.7481 (0.0613)	0.5451 (0.1322)	0.7515 (0.0492)	0.5562 (0.1242)	0.7657 (0.0587)	0.5702 (0.1348)
Mockingjay	0.7291 (0.0624)	0.5121 (0.1437)	0.7289 (0.0499)	0.5287 (0.118)	0.7428 (0.0595)	0.547 (0.1326)
vq_wav2vec	0.7192 (0.0683)	0.5299 (0.1337)	0.7209 (0.0559)	0.5631 (0.1066)	0.7361 (0.0631)	0.5824 (0.1154)

Unseen evaluation (averaged CC (std) across all articulators, sentences and folds)

Features	Unseen subjects			
	F03	F04	MC01	MC04
MFCCs	0.4201 (0.1342)	0.4505 (0.187)	0.4439 (0.0861)	0.568 (0.1059)
wav2vec	0.4422 (0.1362)	**0.5126 (0.2003)**	**0.5134 (0.0879)**	0.5781 (0.1179)
APC	0.4284 (0.1489)	0.4502 (0.1826)	0.4879 (0.0875)	0.5608 (0.1111)
NPC	0.4255 (0.1427)	0.4257 (0.2106)	0.4852 (0.0859)	0.5458 (0.1237)
DeCoAR	**0.4659 (0.1414)**	0.4953 (0.1931)	0.5102 (0.0863)	**0.5938 (0.1022)**
TERA	0.4437 (0.1418)	0.4471 (0.191)	0.4901 (0.0937)	0.5504 (0.1358)
Mockingjay	0.4182 (0.1329)	0.4139 (0.1807)	0.4503 (0.0923)	0.5031 (0.1407)
vq_wav2vec	0.4341 (0.1432)	0.4823 (0.193)	0.4836 (0.0994)	0.5874 (0.1149)

Conclusion

- Pre-trained SSL features, as acoustic features, are effective for dysarthric AAI.
- With minimal training data for dysarthric AAI, DeCoAR outperforms MFCCs.
- **Future work:** Study effects of SSL models on low-resourced language-mismatched dysarthric corpus.

maharana@usc.edu