Enhancing Speech Articulation Analysis Using a Geometric Transformation of The X-Ray Microbeam Dataset

Ahmed Adel Attia[1] , Mark Tiede[2] , Carol Y. Espy-Wilson[1]

[1]University of Maryland College Park, MD, [2]Haskins Laboratories, New Haven, CT

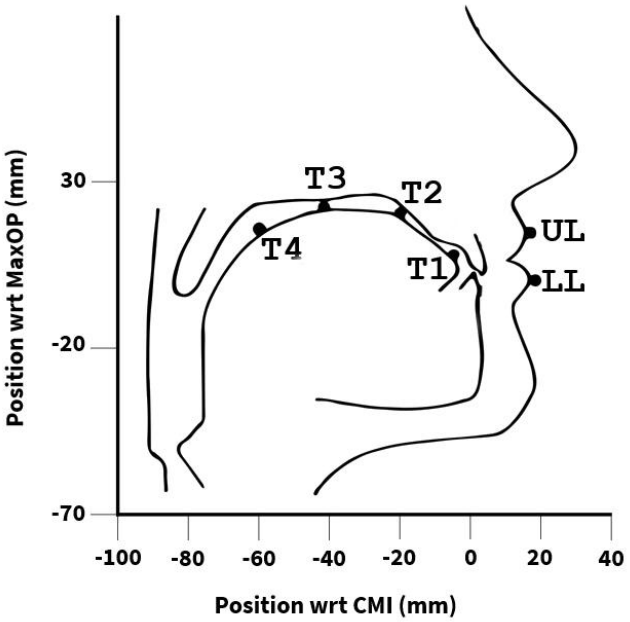
aadel@umd.edu, mark.tiede@yale.edu, espy@umd.edu

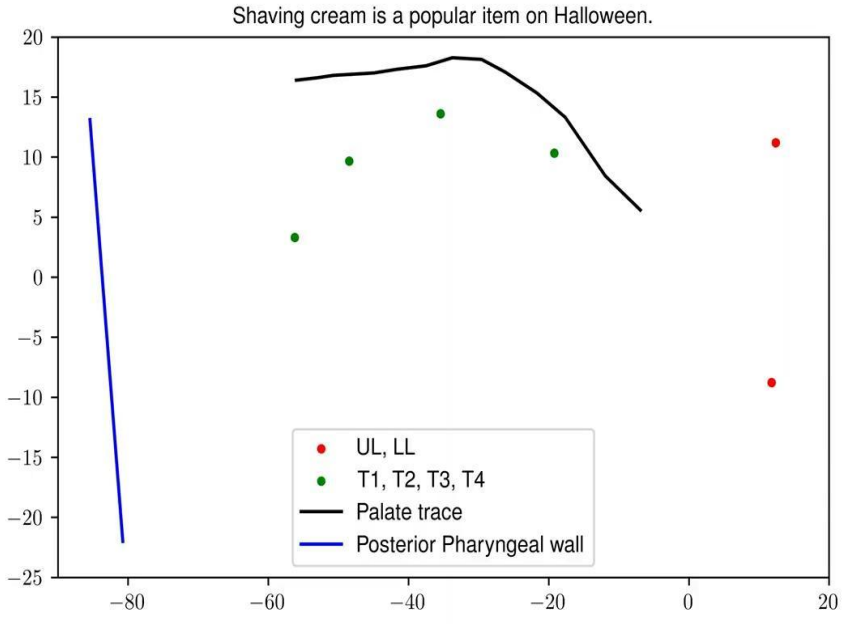
INTERSPEECH 2023

A. James Clark School of Engineering

Institute for Systems Research

X-ray Microbeam (XRMB) Dataset





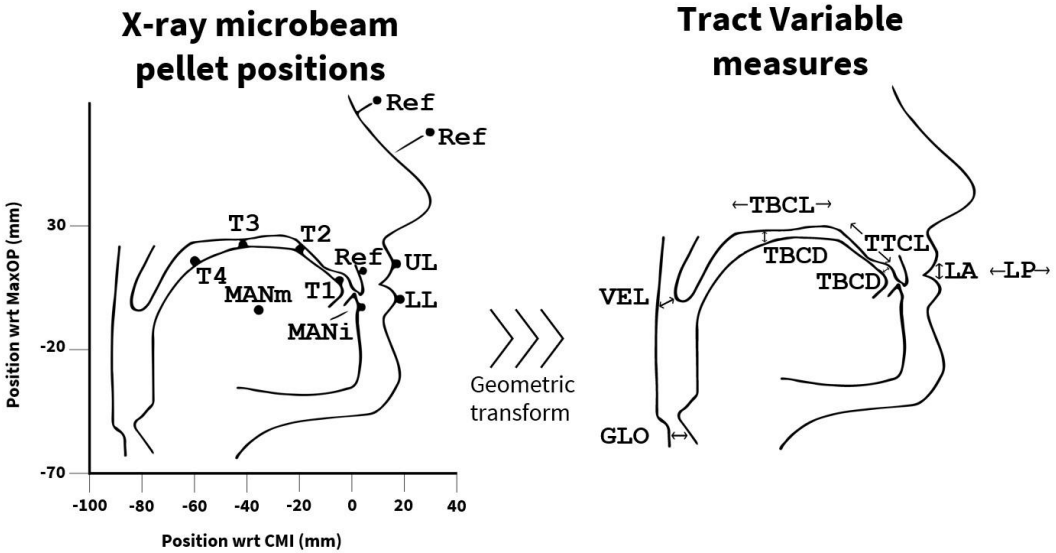
Applications of Articulatory Data

- Mental health assessment and diagnosis (Sirwadena 2021, Seneviratne 2020)
- Automatic speech recognition (Mitra 2011, 2018)
- Speech synthesis (Ling, 2013)
- Speech therapy (Hueber, 2013)
- Many more…

Challenges in Analyzing Articulatory Data

- Variability in speaker anatomy and pellet placement makes accurate analysis challenging.
- Pellet positioning in the X-Y plane is closely linked to the speaker's anatomy, resulting in significant variability among speakers for the same sound.
- Quantifying vocal tract shape is best achieved by measuring the location and degree of constrictions, called Tract Variables (TVs), which are relative measures.

X-ray Microbeam (XRMB) Dataset



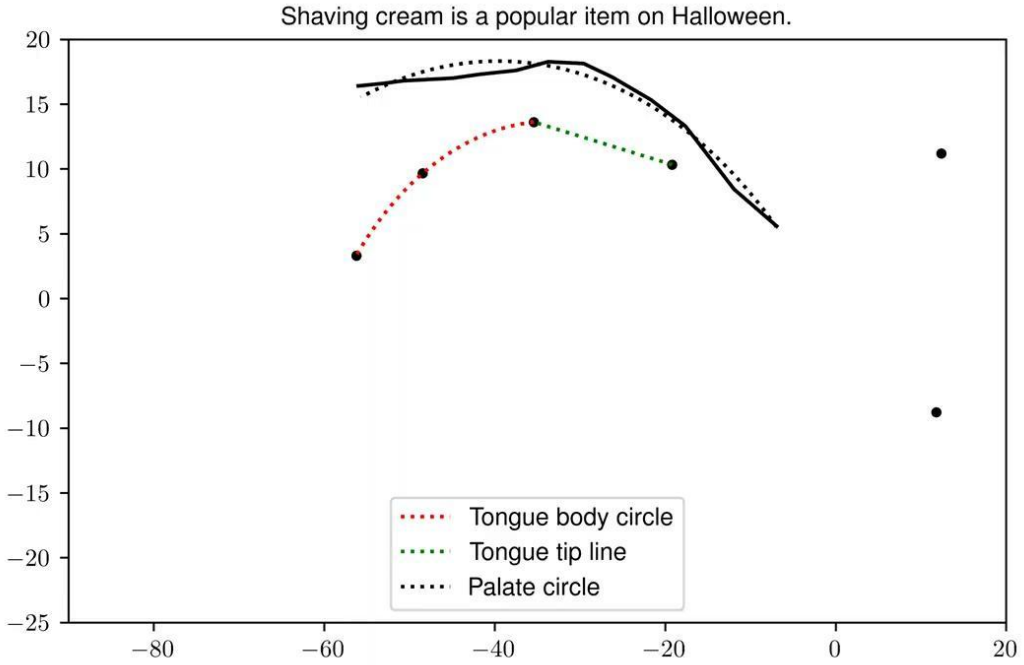
Absolute positions of articulators *dependent* on speaker dimensions

Relative measures of constrictions *independent* of speaker dimensions

Tract Variables for XRMB Dataset

- Sivaraman et al.[1] proposed a geometric transformation to obtain TVs from the XRMB Pellet Trajectories (PTs).
- The hard palate was approximated as a large circle using curve fitting through the palate trace.
- The tongue body was approximated as a smaller circle.
- The tongue tip was modeled separately using the segment T2-T1.

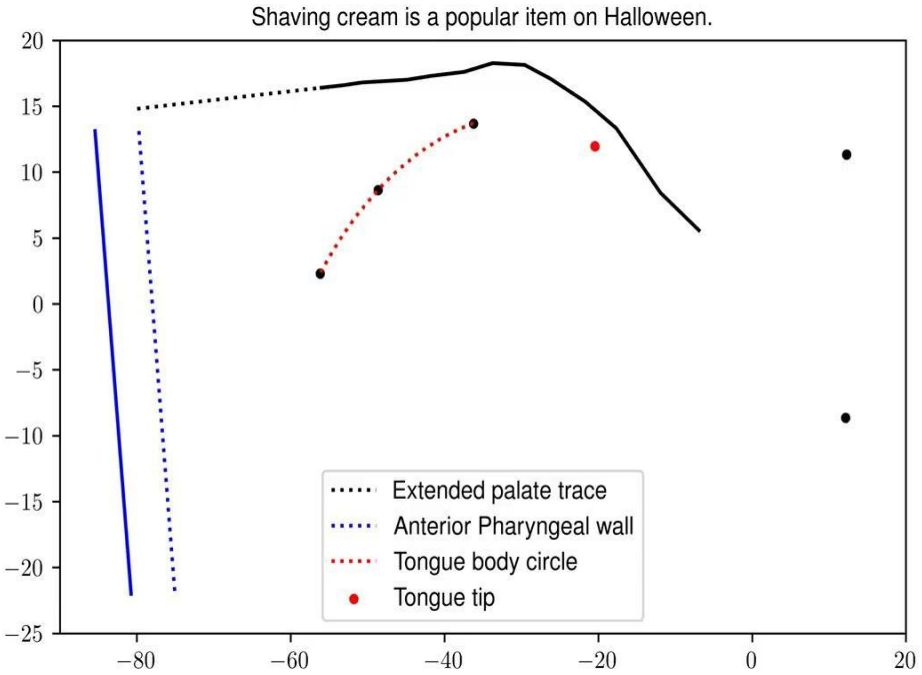
Tract Variables for XRMB Dataset



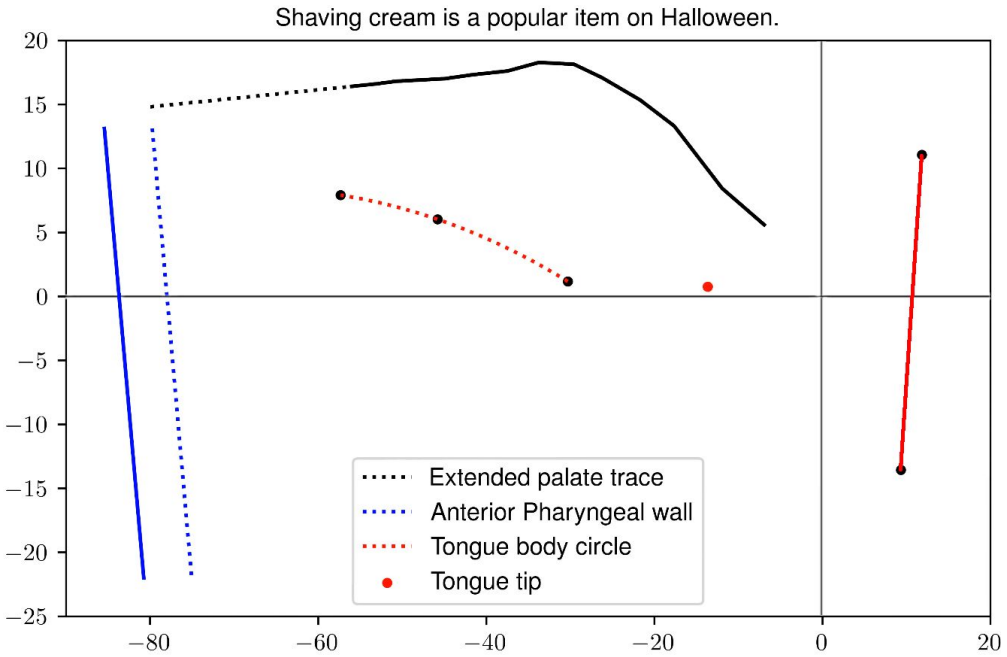
Drawbacks of Current TVs

- Palatal trace does not cover the soft palate.
- Circular arc model of the palate trace does not represent the actual shape of the palate.
- T2 is too posterior to accurately model the tongue tip.
- No formulae indicated for the transformation, which limits reproducibility and applications on new data.
- Consequently, these transformations couldn't be applied to the reconstructed recordings of the XRMB dataset[2].

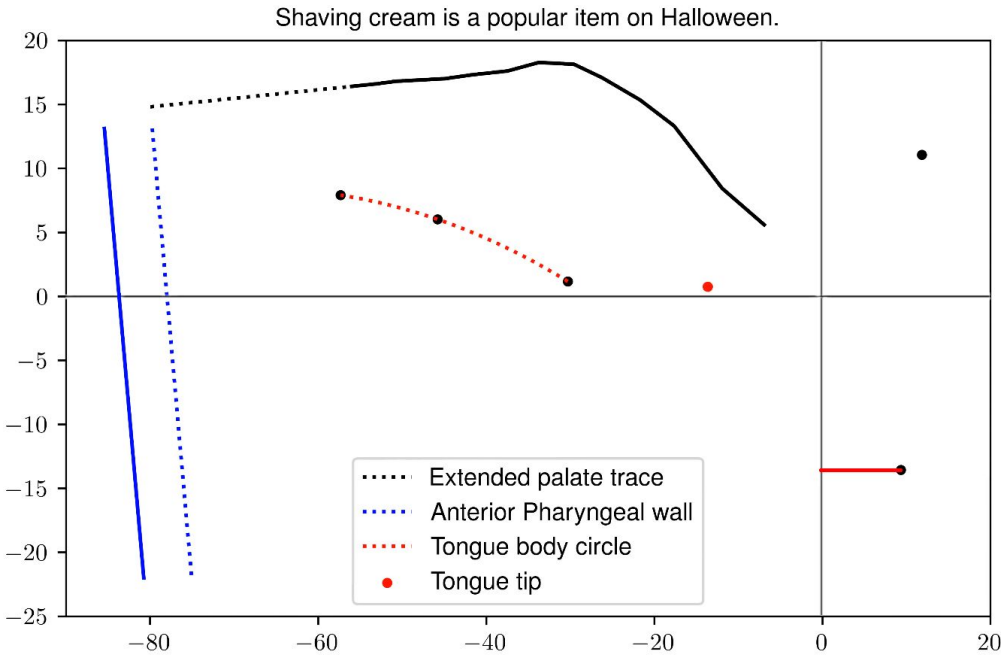
Proposed Model



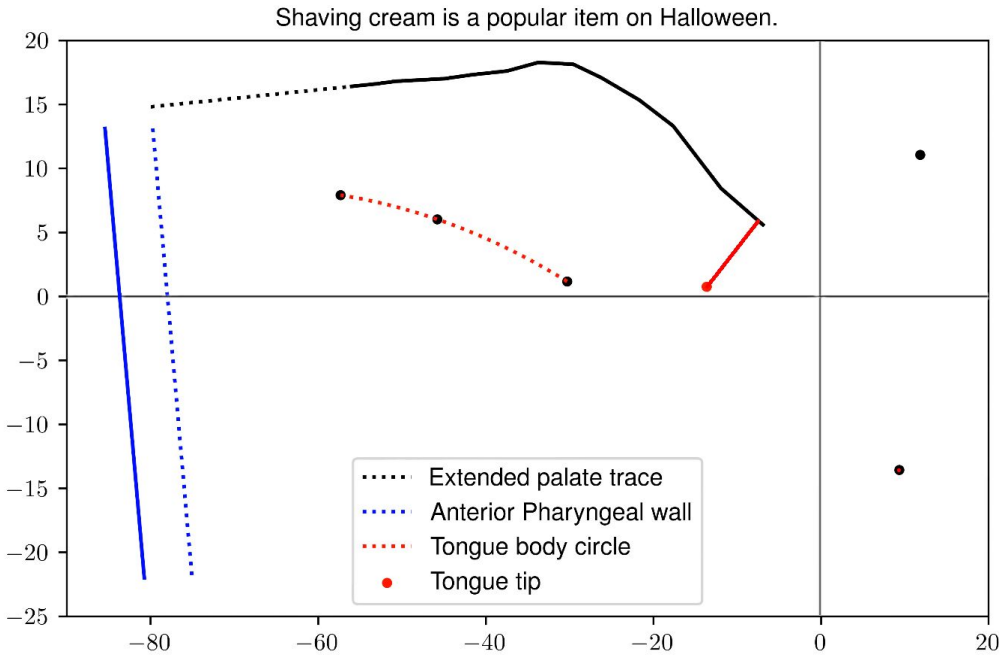
Lip aperture



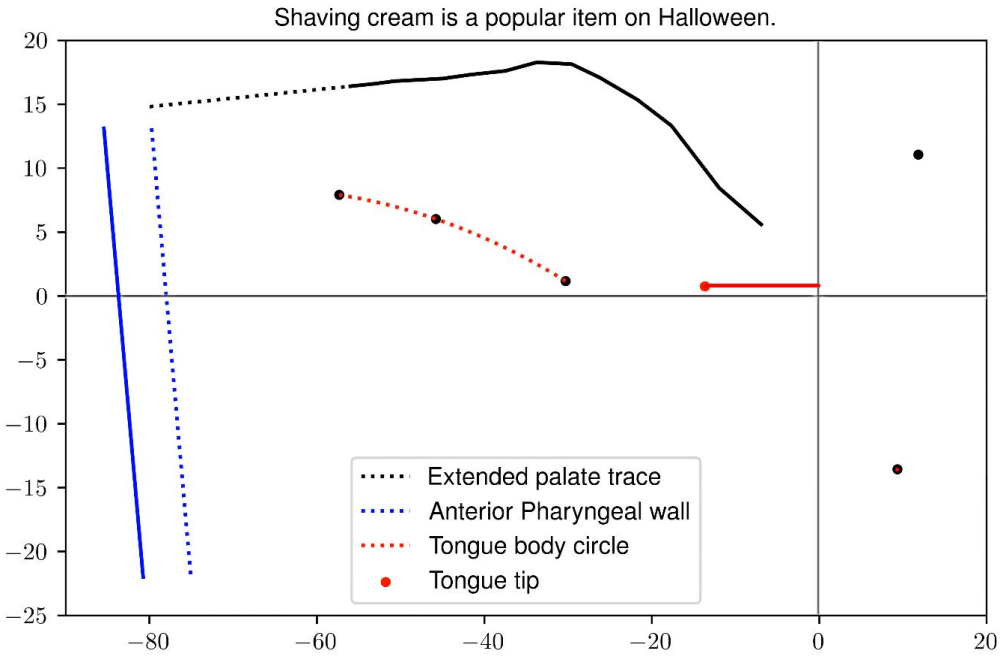
Lip protrusion



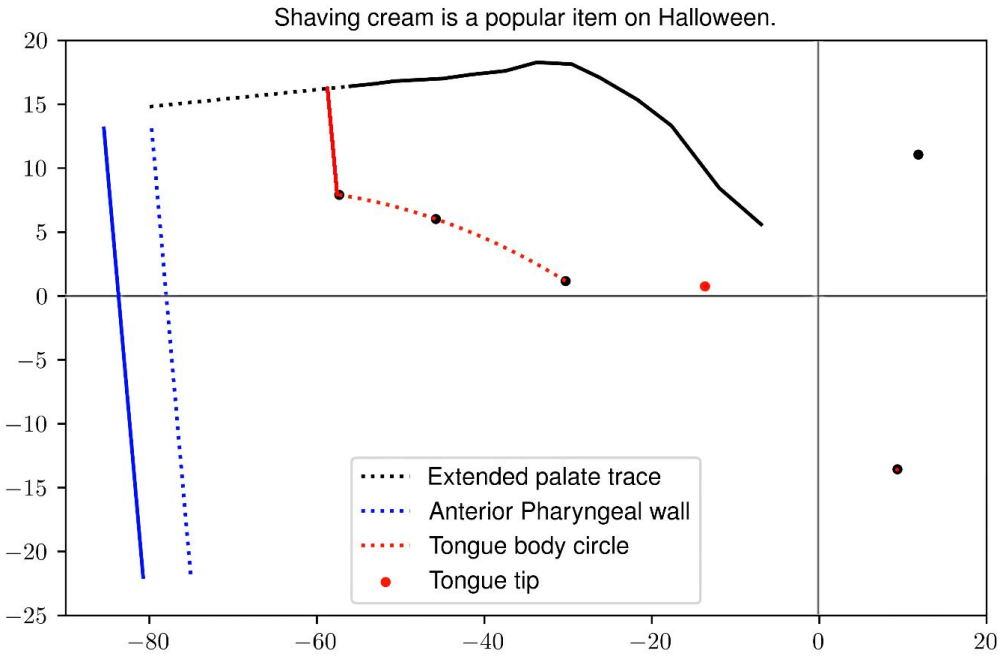
Tongue Tip Constriction Degree



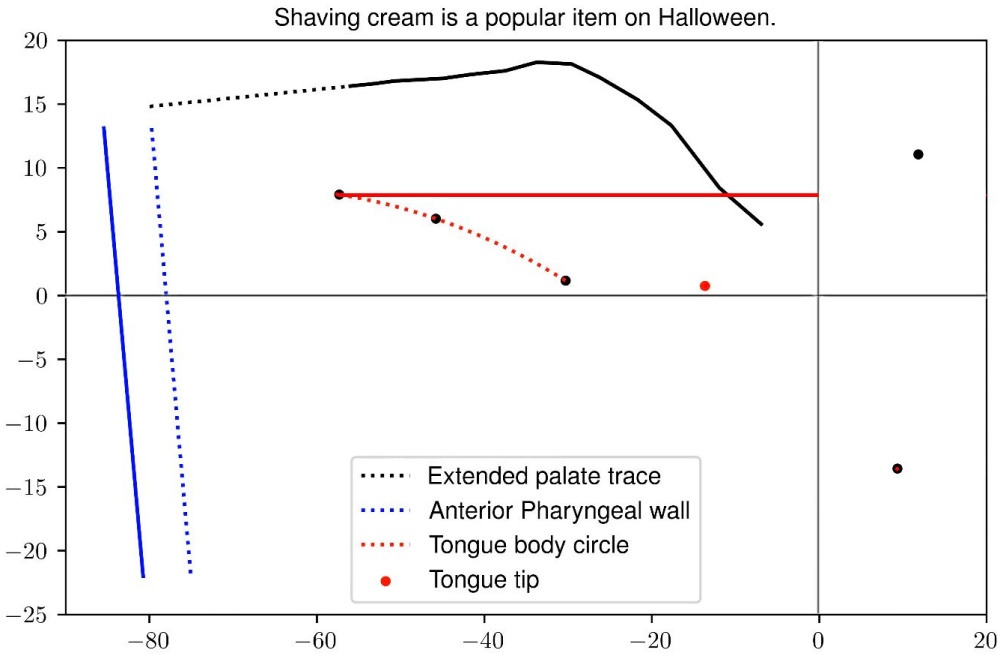
Tongue Tip Constriction Location



Tongue Body Constriction Degree



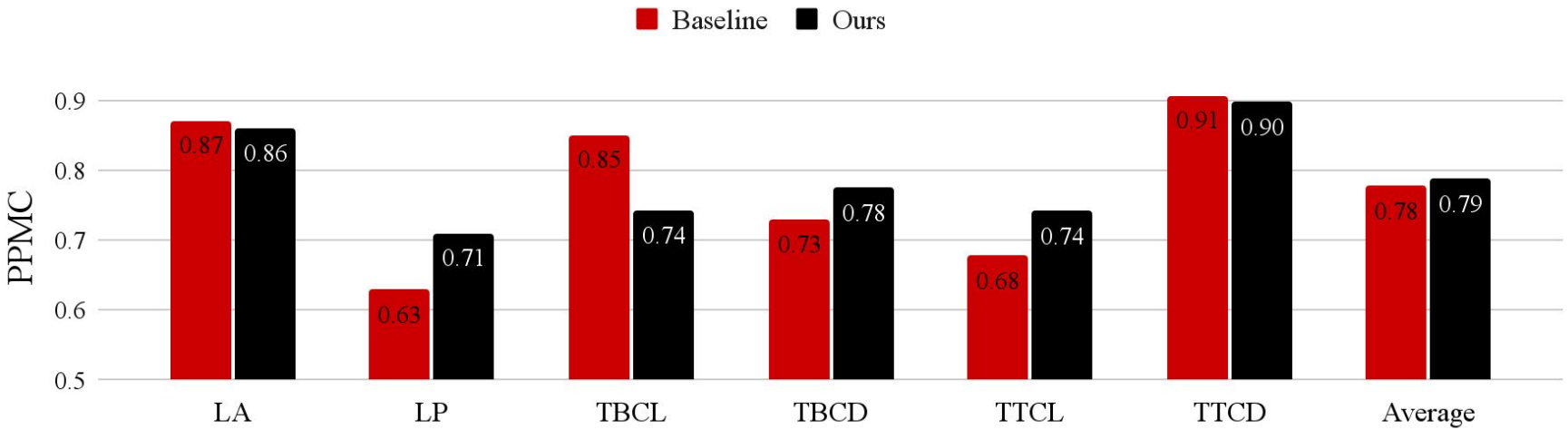
Tongue Body Constriction Location



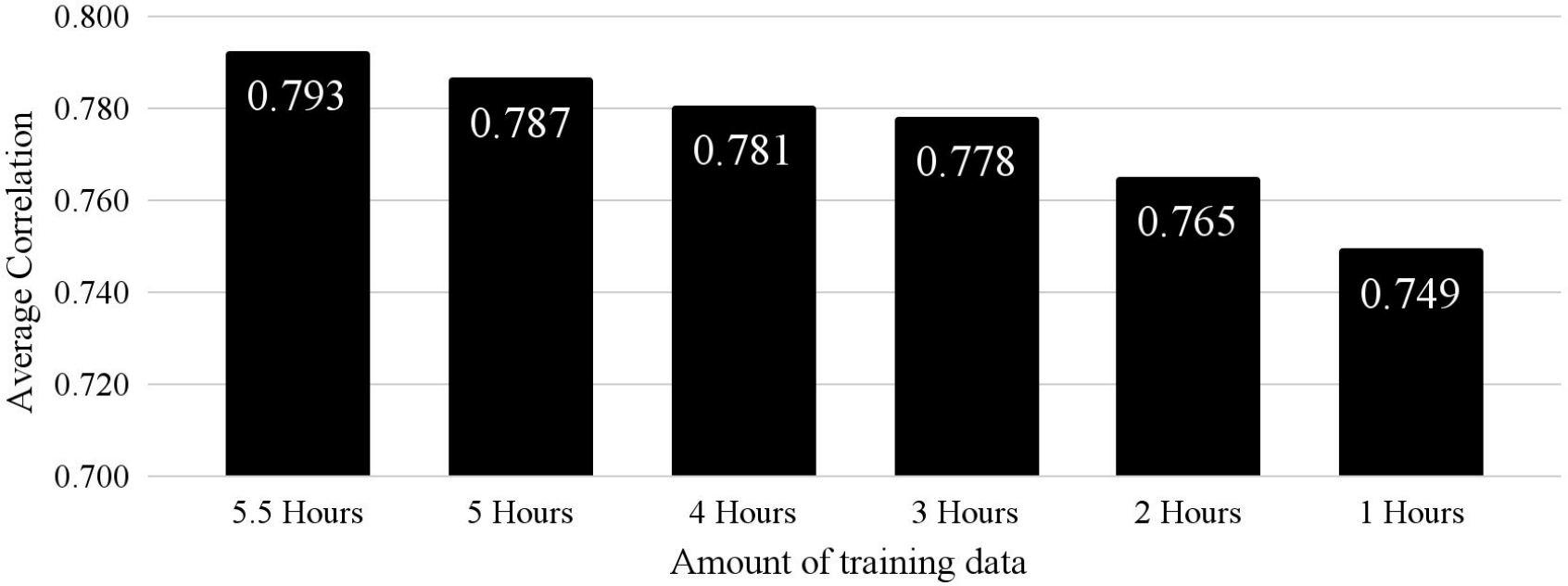
Experimental Results

- Showcasing how our transformations relate to the acoustics.
- We train a Speech Inversion (SI) model on the XRMB dataset.
- We transform the data using our proposed transformations and again using baseline transformations.

Experimental Results



Effect of the Amount of Training Data on Performance



Conclusion

- Introduced a novel geometric transformation, enhancing the derivation of TVs from XRMB data.
- Tangible enhancements in the performance of Speech Inversion (SI) models.
- Our exploration of the interplay between improved transformation and increased data volume underscores the significance of high-quality data.

References:

1. Sivaraman, Ganesh, et al. "Unsupervised speaker adaptation for speaker independent acoustic to articulatory speech inversion." The Journal of the Acoustical Society of America 146.1 (2019): 316-329.
2. Attia, Ahmed Adel, and Carol Y. Espy-Wilson. "Masked Autoencoders are Articulatory Learners." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.
3. Daniel, Mauro Miguel, et al. "Pharyngeal dimensions in healthy men and women." Clinics 62.1 (2007): 5-10.
4. Attia, Ahmed Adel, and Carol Y. Espy-Wilson. "Masked Autoencoders are Articulatory Learners." ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023.

Thank you

A. James Clark School of Engineering

Institute for Systems Research