Predicting Vocal Tract Shape Information from Tongue Contours and Audio Using Neural Networks

Sarah R. Li,[1] Alex Knapp,[1] Nicholas Schoenleb,[1] Jing Tang, PhD,[1] Suzanne Boyce, PhD CCC-SLP,[2] & T. Douglas Mast, PhD[1]
[1]Biomedical Engineering; [2]Communication Sciences and Disorders | University of Cincinnati

University of CINCINNATI

Introduction

- **Midsagittal ultrasound images** can show the tongue surface from much of the root to the tip in real time, providing useful articulatory information (e.g., for ultrasound biofeedback therapy (UBT) [1]).
  - However, missing information can cause difficulties in interpretation:
    - Structures toward which the tongue constricts in the vocal tract (e.g., **hard palate**) are not imaged; thus, the vocal tract constrictions that resulted in the acoustic production are uncertain.
    - The **tongue tip** is often obscured [2, 3].
- **Magnetic resonance images (MRI)** show the entire vocal tract.

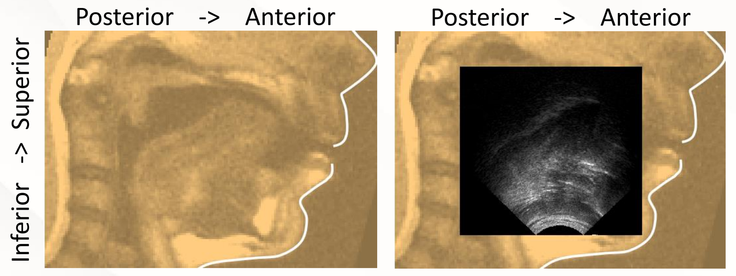


Fig. 1: (Left) Midsagittal MRI with darker pixels showing the vocal tract air space. (Right) Superimposed ultrasound image; a bright contour shows the tongue surface.

- Audio can be collected during ultrasound imaging.
  - Recent advancements in acoustic/articulatory prediction models [4, 5, 6] may be used to predict missing articulatory information.

Aim

Using MRI data, train a model that can combine acoustic and articulatory features accessible from ultrasound imaging to predict the midsagittal vocal tract shape. Trends for different model setups and production types were analyzed to identify preferred model choices and to understand the prediction accuracy of the model, with an eventual goal to aid interpretation during UBT.

Methods: Features

Data selection

- Vowels and central approximants used:
  - Vowels isolated from /bVt/ contexts (e.g., /aɪ/ from "bite")
  - /i/, /u/, /ɑ/ and /r/, /w/, /j/ in /VCV/ contexts (e.g., "eeree")
- 3678 productions (each ≈33 ± 29 frames; 120,926 total frames)
  - From 75 speakers (10, 10, and 55 in test, validation, and train sets)
- Semi-automatically determined timestamps (Fig. 3)

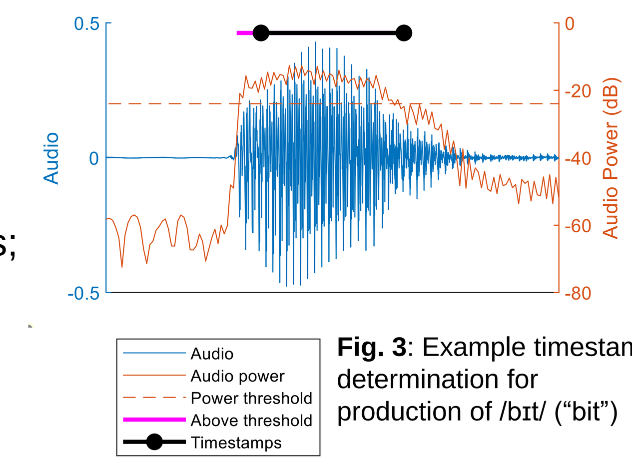


Fig. 3: Example timestamp determination for production of /bɪt/ ("bit")

Acoustic Features

- Audio with speech enhancement via Adobe Podcast Toolkit[i]
  A. Log power spectra (LPS)
  B. Mel-frequency Cepstral Coefficients (MFCC)
- C. LPS from unenhanced audio

[i]https://podcast.adobe.com/enhance

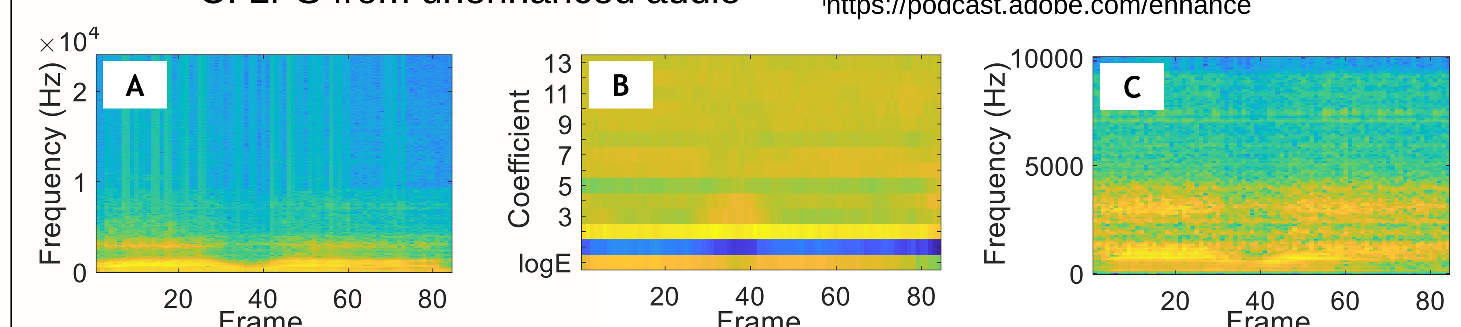
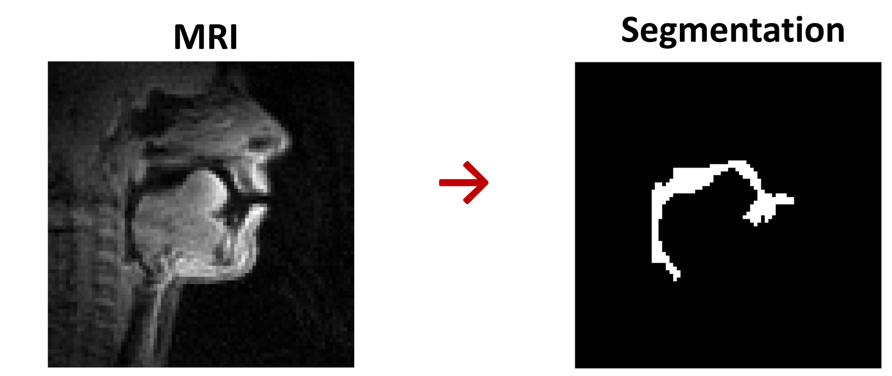
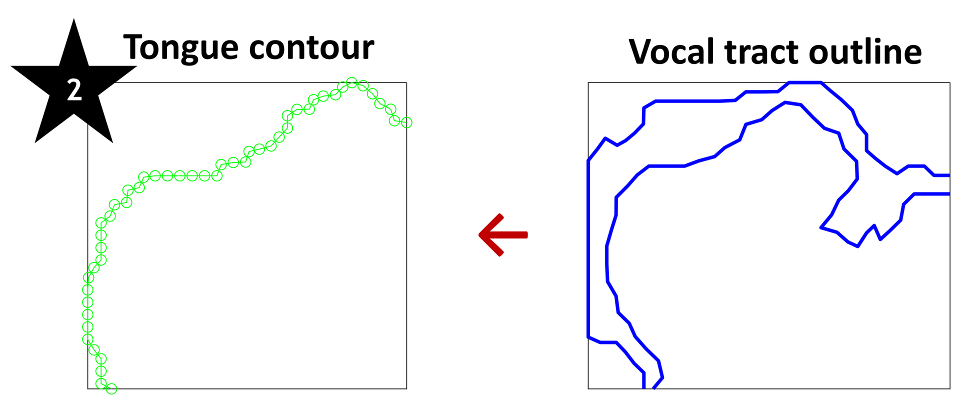


Fig. 4: Acoustic feature examples from a production of "aaraa"

Articulatory Features



Segmentation with U-Net [9] (newly trained on USC dataset using 622 manual segmentations; mean Dice coefficient of ≈0.92 ± 0.03 for the ≈5% test set)



Portion of vocal tract visible in ultrasound imaging (i.e., tongue contour) estimated geometrically

Rotation and translation augmentations were applied during model training.

Acknowledgements

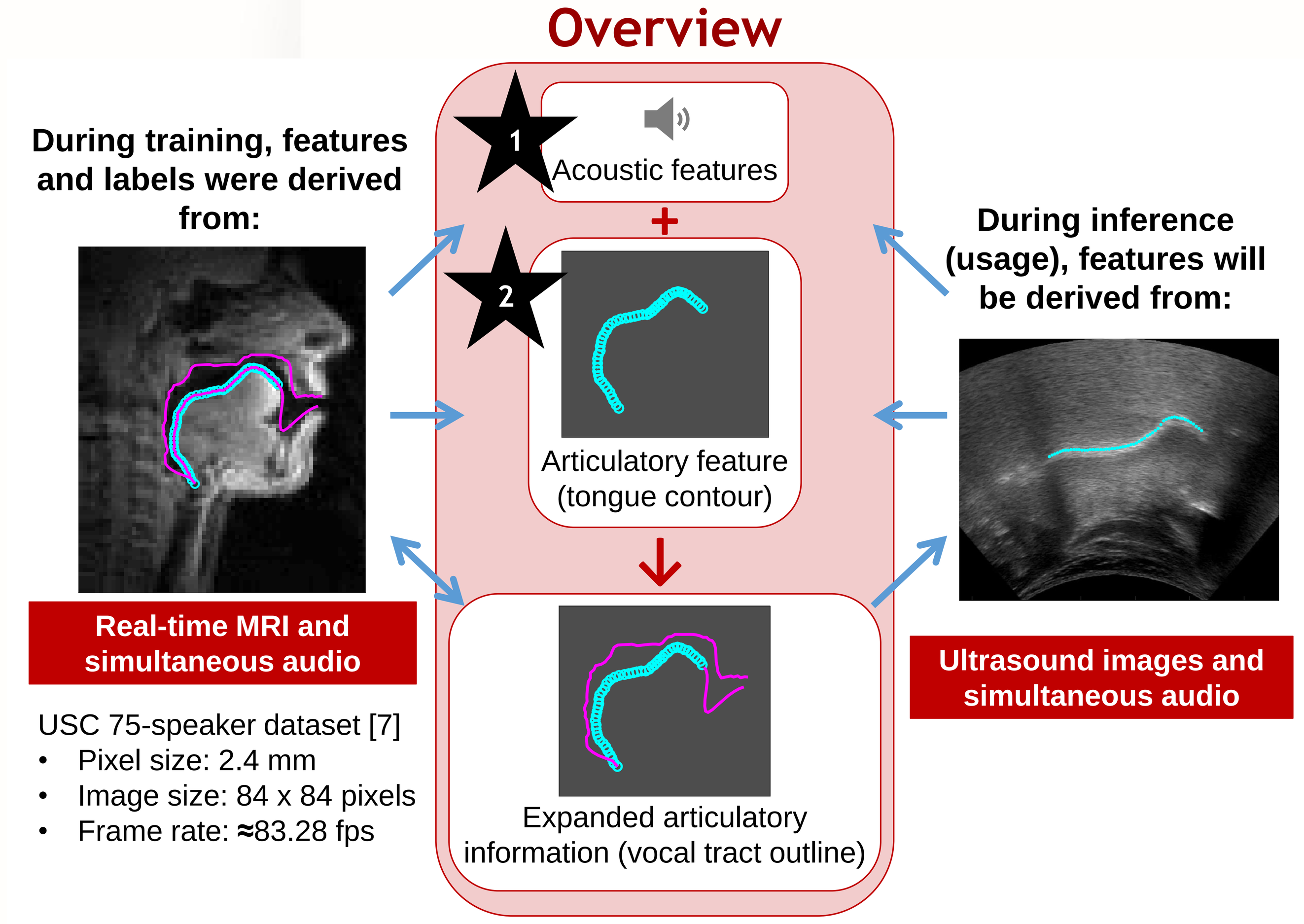
- NIH/NIDCD grants F31 DC020672 and R01 DC017301
- Mentors Dr. Steven Lulich and Dr. Shrikanth Narayanan
- Dr. Yongwan Lim for help with the USC 75-speaker dataset [7]
- M. Ruthven for publicly posted U-Net segmentation model of the vocal tract [9]
- Siemens Medical Solutions for lending the Acuson X300 ultrasound scanner

References

[1] E. Sugden, S. Lloyd, J. Lam, and J. Cleland, "Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders," *Int J Lang Commun*, vol. 54, no. 5, pp. 705–728, Sep. 2019.
[2] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clin Linguist Phon*, vol. 19, no. 6–7, pp. 455–501, Jan. 2005.
[3] S. R. Li, T. D. Mast, and S. Boyce, "Quantifying tongue tip visibility in ultrasound images of /r/ tongue shapes using numerical ultrasound simulations," *J Acoust Soc Am*, vol. 153, no. 3_supplement, p. A372, Mar. 2023.
[4] Y. M. Siriwardena, A. A. Attia, G. Sivaraman, and C. Espy-Wilson, "Audio data augmentation for acoustic-to-articulatory speech inversion," in *EUSIPCO*, Helsinki, Finland: IEEE, Sep. 2023, pp. 301–305.
[5] V. Ribeiro, K. Isaieva, J. Leclere, P.-A. Vuissoz, and Y. Laprie, "Automatic generation of the complete vocal tract shape from the sequence of phonemes to be articulated," *Speech Commun*, vol. 141, pp. 1–13, Jun. 2022.
[6] Z. Yue, E. Loweimi, Z. Cvetkovic, H. Christensen, and J. Barker, "Multi-modal acoustic-articulatory feature fusion for dysarthric speech recognition," in *ICASSP*, Singapore: IEEE, May 2022, pp. 7372–7376.
[7] Y. Lim et al., "A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3D volumetric images," *Sci Data*, vol. 8, no. 1, p. 187, Dec. 2021.
[8] P. Wu et al., "Speaker-independent acoustic-to-articulatory speech inversion," in *ICASSP*, Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5.
[9] M. Ruthven, A. M. Peplinski, D. M. Adams, A. P. King, and M. E. Miquel, "Real-time speech MRI datasets with corresponding articulator ground-truth segmentations," *Sci Data*, vol. 10, no. 1, p. 860, Dec. 2023.

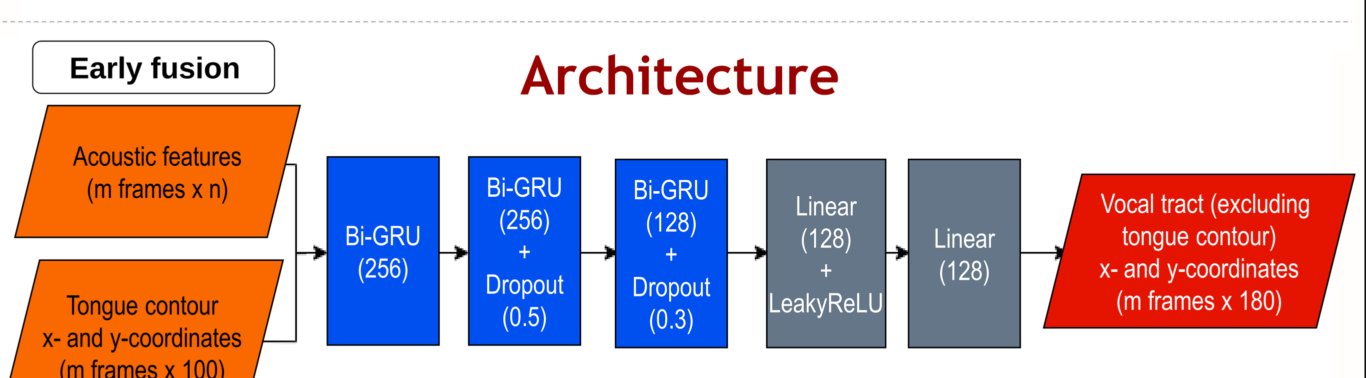
Methods: Model

Overview



USC 75-speaker dataset [7]
- Pixel size: 2.4 mm
- Image size: 84 x 84 pixels
- Frame rate: ≈83.28 fps

Architecture



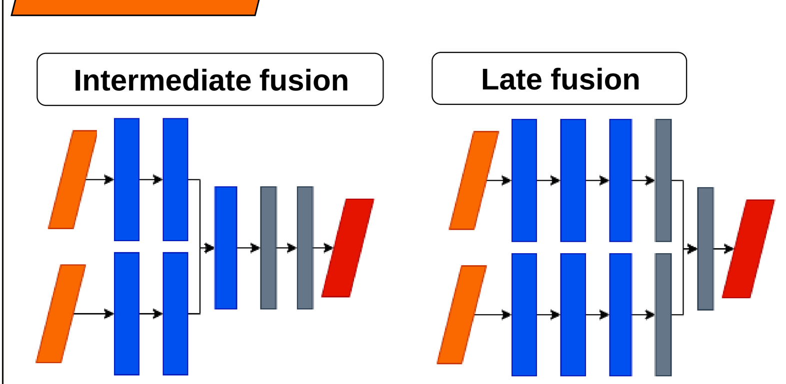


Fig. 2: Neural network architecture with bidirectional gated recurrent units (BiGRU) [4, 8], with different fusion location choices shown. The final prediction output (entire vocal tract outline) was formed by concatenating the model output (vocal tract excluding the tongue contour) with the input tongue contour.

Results, Discussion, and Conclusion

Results: Predictions

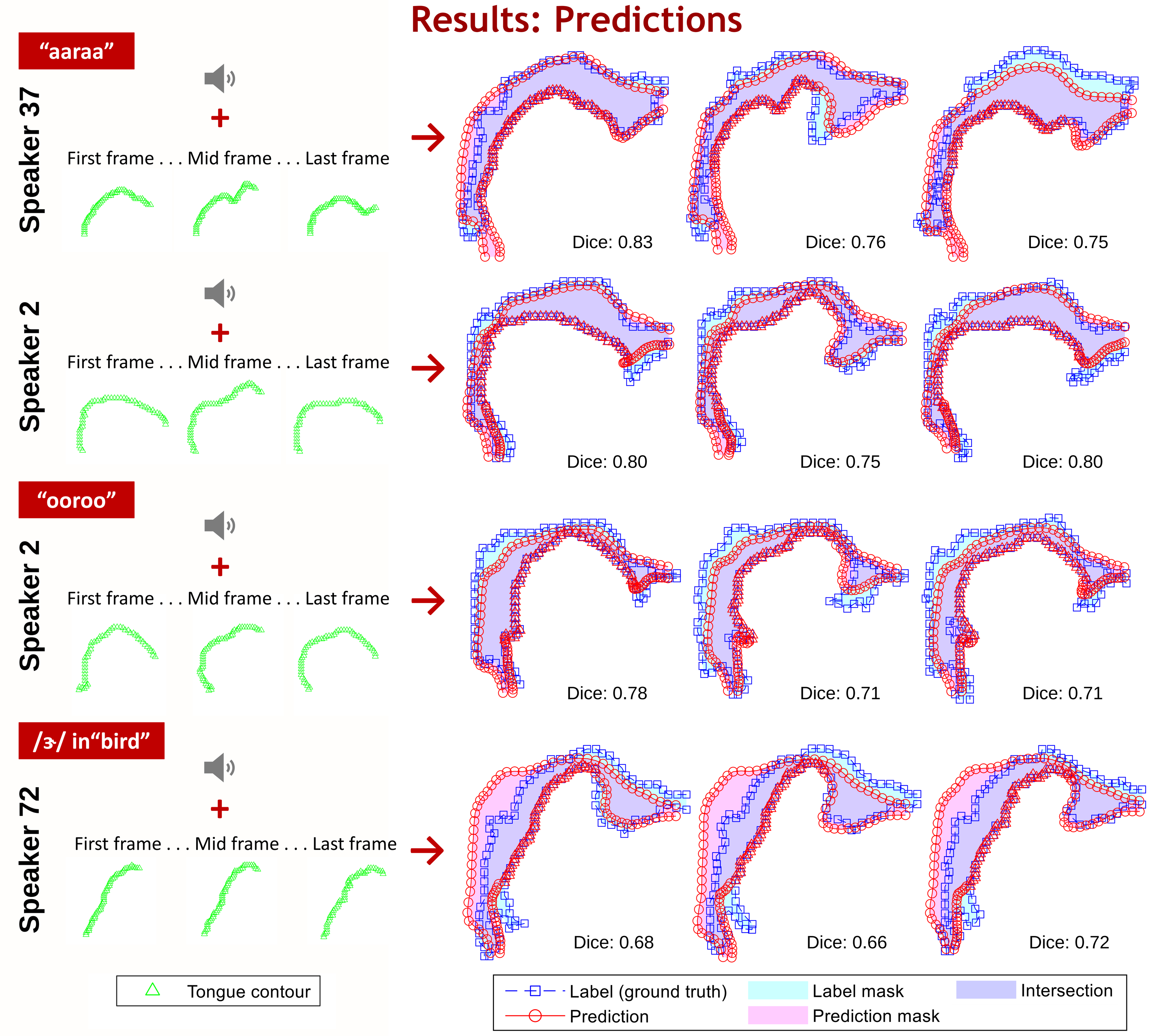
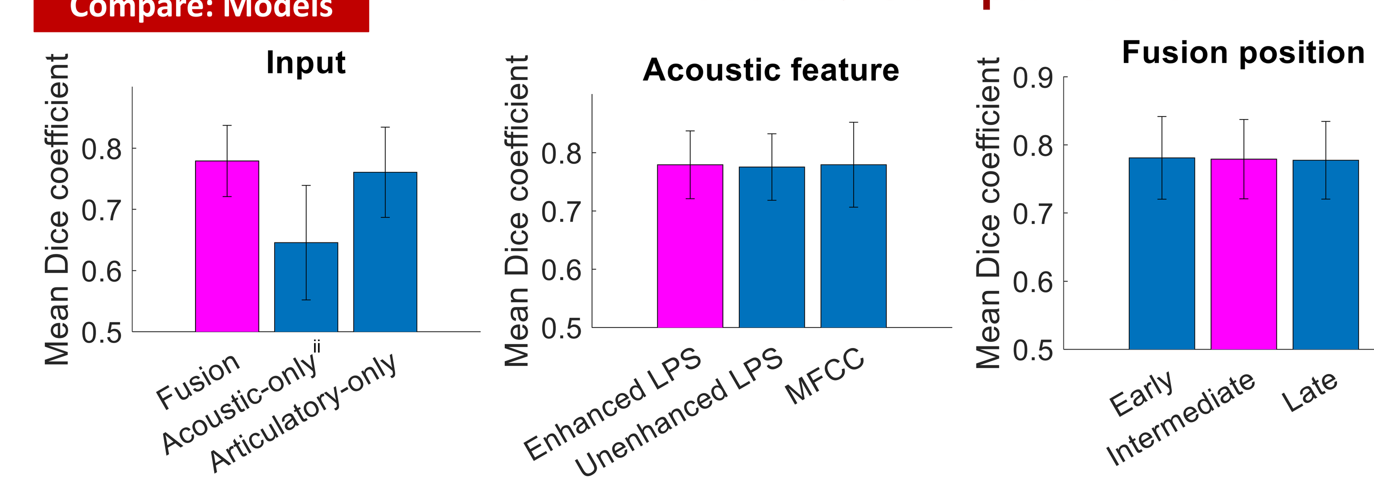


Fig. 5: Example vocal tract predictions (right, showing only three frames) resulting from the input (left), from the test set of the selected model (magenta in Fig.6). Dice coefficients for each frame in a production were calculated after transformations to masks.

Results: Comparisons

Compare: Models



All network models trained with Adam optimizer, with a learning rate of 1e-4 (100 epochs) then 1e-5 (100 epochs). Only one network was trained per feature / architecture type.

Fig. 6: The mean ± standard deviation over all frames for different models

[ii]Masks were created after Procrustes transformation (translation and rotation) for Acoustic-only model

Compare: Production types

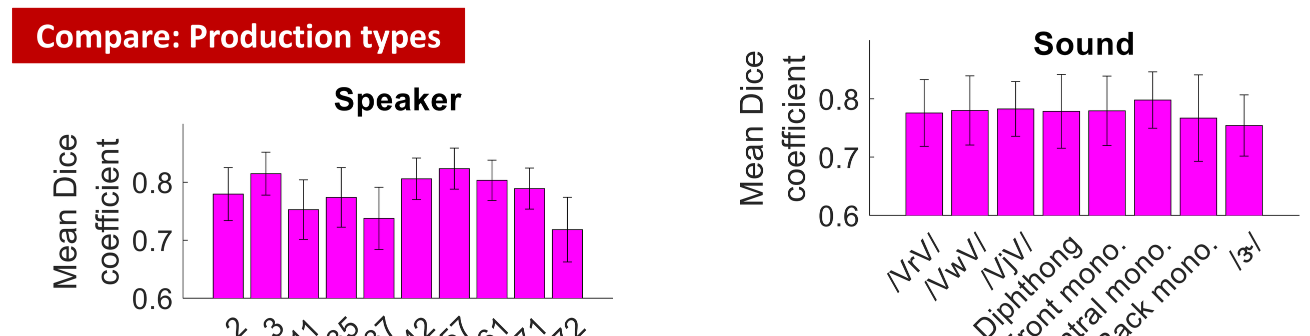


Fig. 7: The mean ± standard deviation over all frames from a speaker or of a sound

Discussion

- Fusion of articulatory and acoustic features improves predictions of vocal tract shape.
- Contribution from acoustic features may be limited due to degraded quality (MRI audio recording) or because acoustics result from a 3D vocal tract (vs. 2D imaged articulatory features).
- The acoustic feature and fusion choices investigated did not result in differing performance.
- Prediction performance varied by speaker more than by specific sound types.

Conclusion

Articulatory and acoustic features from MRI that are likely accessible from ultrasound imaging (i.e., audio and tongue contour) were combined to predict expanded articulatory information with moderately high accuracy (mean Dice coefficient of ≈0.78). Additional exploration of trends and model parameters may be necessary to improve prediction accuracy for future use in articulatory studies.