Data Selection and Adaptation for Naturalness in HMM-based Speech Synthesis

Erica Cooper, Alison Chang, Yocheved Levitan, Julia Hirschberg

Columbia University, New York, USA

Research Questions

- Can we identify metrics for selecting the **best** utterances in a found-data corpus for voice training, or for **excluding** utterances that will detract from the quality of the voice?
- Can we select a **subset** of training utterances from a corpus of found data to produce a **better** voice than one trained on all of the data?
- Can we **adapt** a voice towards the best utterances in a corpus, to improve the quality of the voice?

Data and Tools

- **Boston University Radio News Corpus (BURNC):** 7+ hours of professionally-read radio broadcast news from 3 female and 4 male speakers
  - Multiple speakers, non-TTS speaking style
  - 4 hours 22 minutes of speech from female speakers
- **Hidden Markov Model Based Speech Synthesis System (HTS):** Toolkit for training HMM-based statistical parametric voices
- **Amazon Mechanical Turk (AMT):** A popular crowdsourcing platform

Experimental Setup

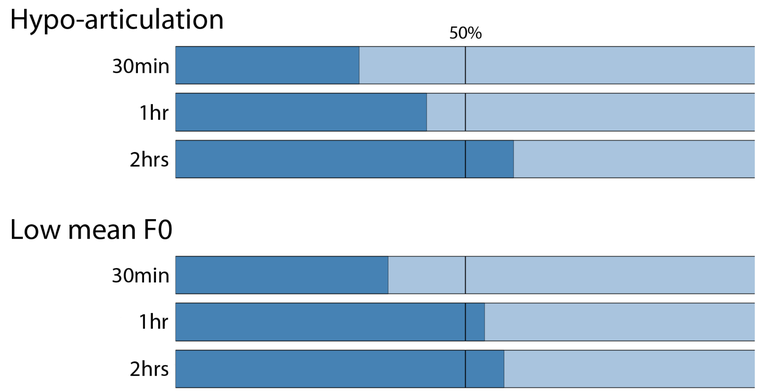
- **Baseline:** Speaker-independently trained voice using all of the female data
- **Subsets:** Train voices only on subsets of the data, selected based on certain criteria
- **Adaptation:** Adaptively train voices on all of the data, adapting to subsets of the data selected on certain criteria
- **Evaluation:** Amazon Mechanical Turk forced-choice pairwise naturalness preference test between test voice and baseline voice

Prior Results

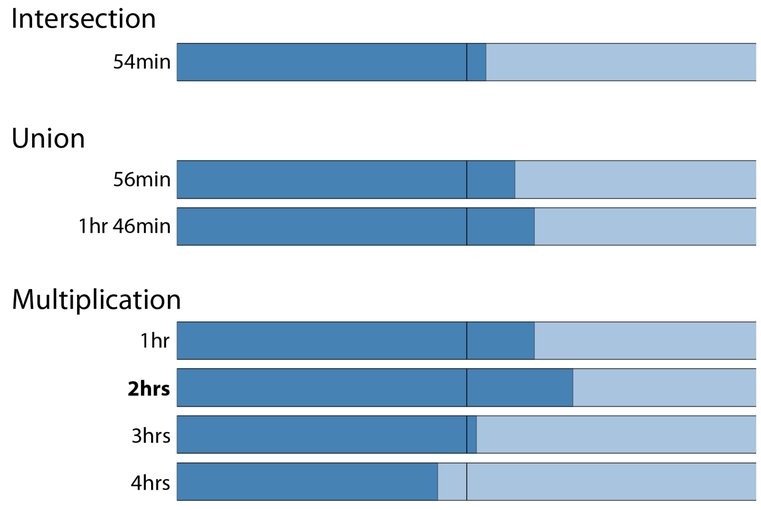
Best hour-long subset voices:

- **Hypo-articulated** utterances (articulation = mean energy / speaking rate)
- **Low mean f0** utterances

Pairwise Preferences: Varying Subset Sizes

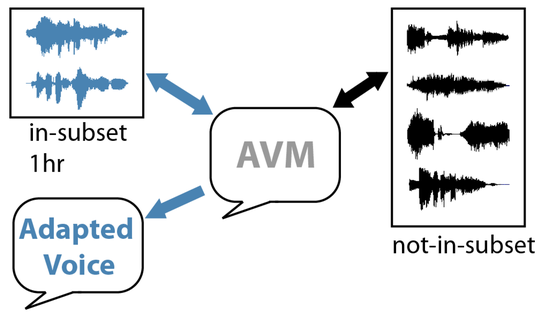


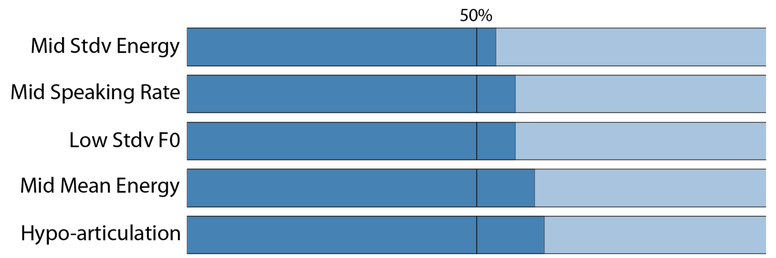
Combination of Best Approaches



Outlier Removal

Subset Adaptation





Conclusions and Future Work

- Level of articulation is a consistently informative feature
- Combination of best features gives best results
- What other features / combinations?
- Do findings generalize? Types of found data, more languages
- Lower-level features such as frame-level acoustic features
- Higher-level features such as speaker characteristics

Acknowledgments

This work was supported by NSF 1539087 "EAGER: Creating Speech Synthesizers for Low Resource Languages" and by Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.