Data Selection for Naturalness in HMM-based Speech Synthesis

Erica Cooper, Yocheved Levitan, Julia Hirschberg

Columbia University

Research Questions

- Can we identify which utterances in a corpus are the **best** to use for voice training, based on acoustic/prosodic features, and which utterances should be **excluded** because they will introduce noise, artifacts, or inconsistency into the voice?
- Can we use **found** data such as radio broadcast news to build HMM-based synthesized voices?
- 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?
- Which **voice training and modeling approaches** work best for this type of data?

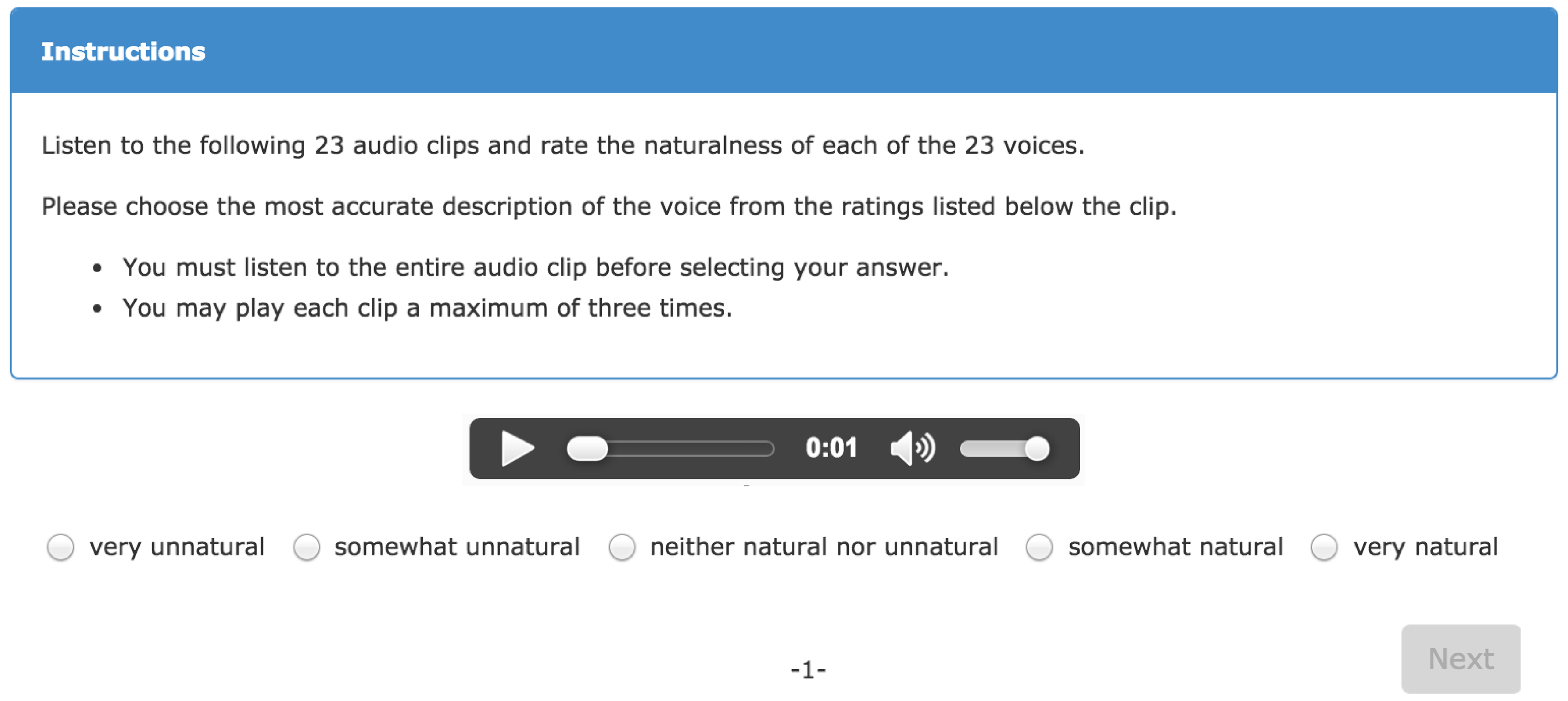
Data and Tools

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

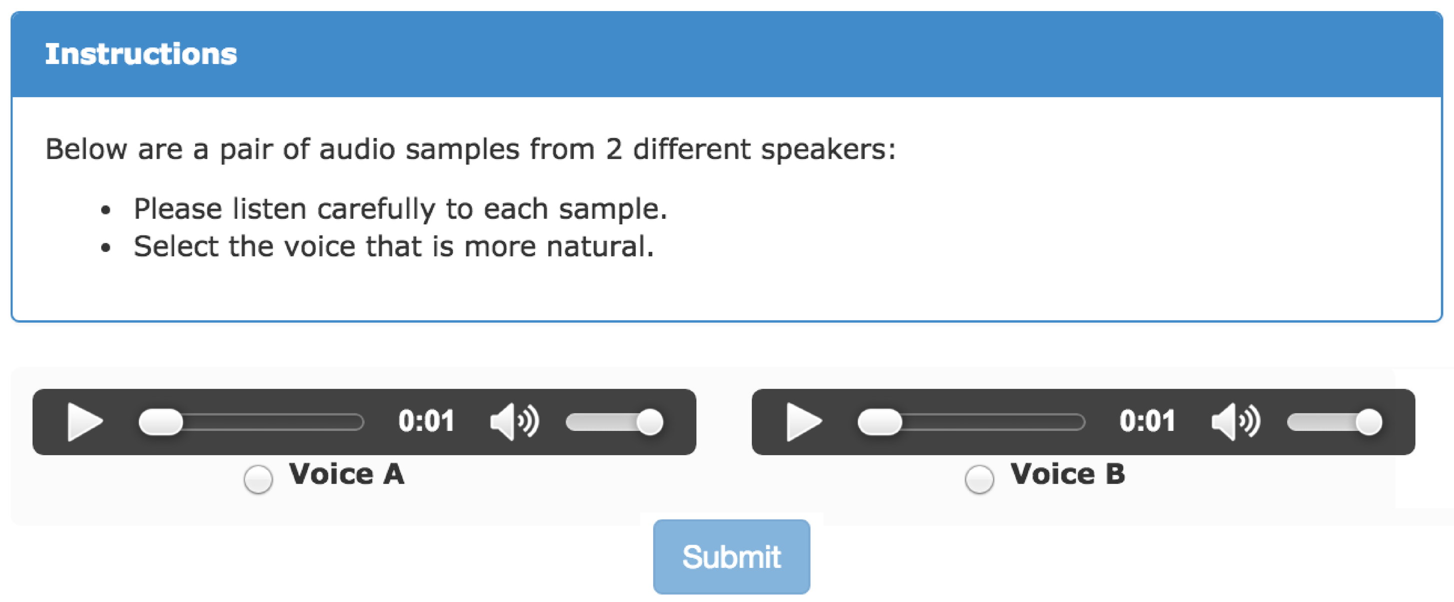
Experiments

- **Baselines:** Voices trained speaker-independently on all of the female data (4hrs 40min) or all of the male data (5hrs 15min)
- **1-hour Subsets** of female or male utterances based on features:
  - Mean/stdv of energy/f0 (high, middle, low)
  - Speaking rate (fast, middle, slow)
  - Hyperarticulation and hypoarticulation
  - Utterance length (long, medium, short)
- **Voice Modeling Approaches** compared to SI:
  - Speaker adaptively trained average voice model (SAT AVM)
  - Voices for individual speakers (speaker-dependent)
  - Monotone f0 contour and interpolated f0 contour

Naturalness Evaluation: Mean Opinion Score



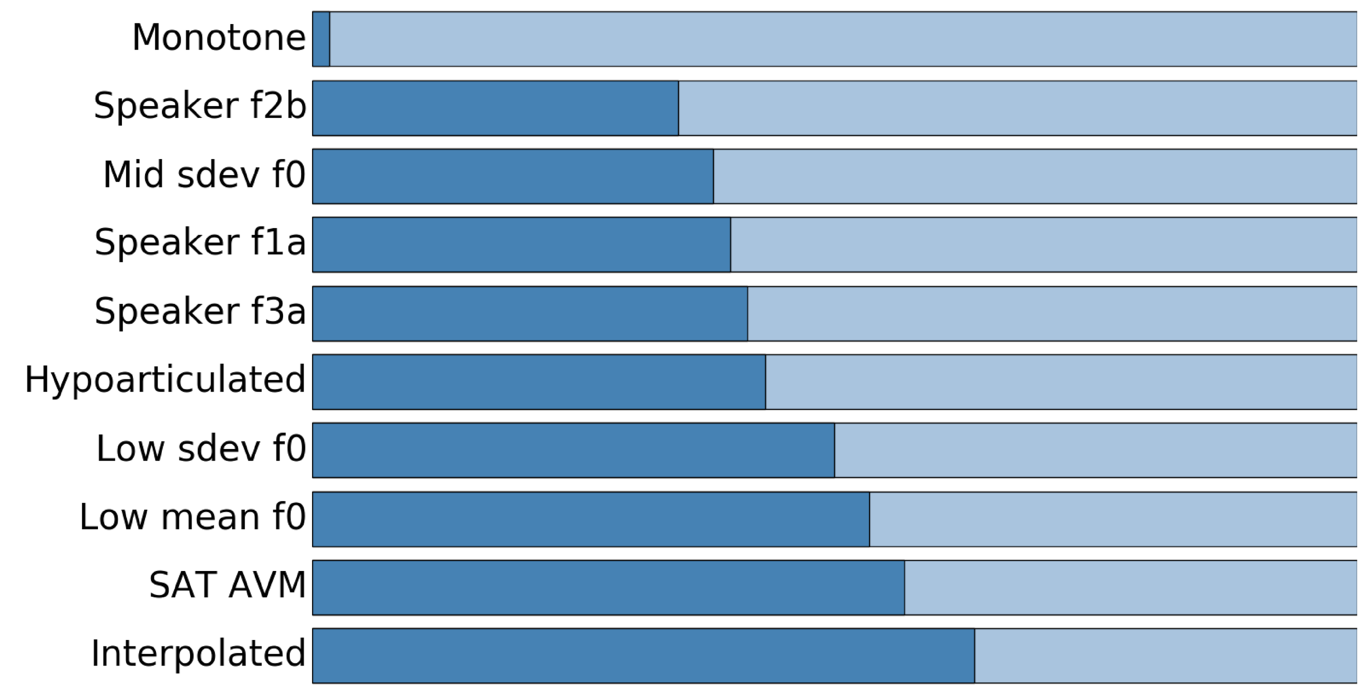
Naturalness Evaluation: Pairwise Comparison



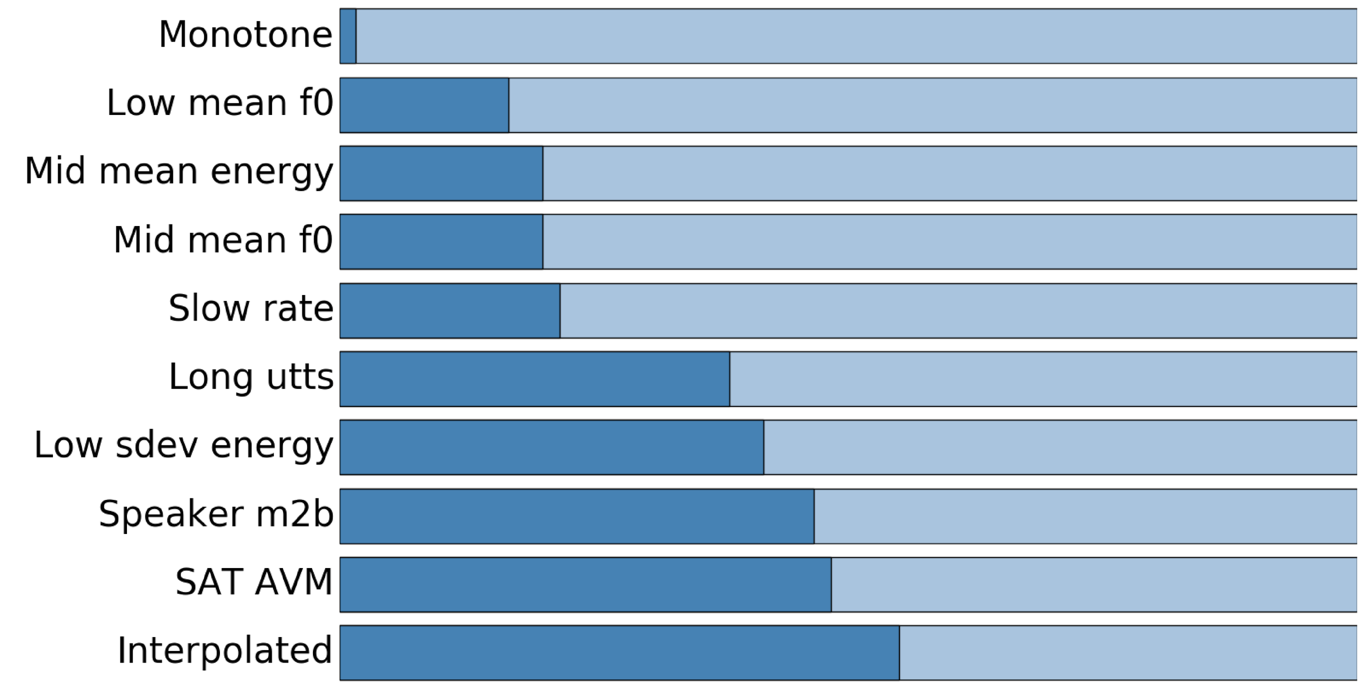
Female Voices: Mean Opinion Score

Voice	Rating	Voice	Rating
Robotic	1.03	Low mean energy	2.41
High mean f0	1.97	Mid mean energy	2.41
Hyperarticulated	2.08	Longest utts	2.5
High mean energy	2.08	Fast rate	2.55
Mid length utts	2.08	Mid mean f0	2.55
Slow rate	2.13	Mid sdev f0	2.6
High sdev energy	2.13	Low sdev f0	2.6
Mid sdev energy	2.28	Baseline	2.68
Shortest utts	2.33	Hypoarticulated	2.7
High sdev f0	2.37	Low mean f0	2.7
Low sdev energy	2.37	Natural speech	4.95
Mid rate	2.4		

Female Voices: Pairwise Preferences



Male Voices: Pairwise Preferences



Conclusions and Future Work

- Interpolation reduced "choppiness" – more direct modeling of prosody needed in the future
- Not enough single-speaker data to train on just one speaker
- SAT AVM did not produce a better voice with our data
- Voices that do badly (hyperarticulation, slow speaking rate)
- Future work: removal of outliers, combination of approaches
- Additional sources of found data: audiobooks, podcasts, course lecture videos, radio shows, speech recognition corpora
- Build voices for **low-resource languages** using found data