



A Comparison of Speaker-based and Utterance-based Data Selection for Text-to-Speech Synthesis

Kai-Zhan Lee, Erica Cooper, Julia Hirschberg

Columbia University, New York, USA

Research Questions

- ▶ Can we identify metrics for selecting the **best** parts of a found-data corpus for voice training?
- ▶ Is it better to select this data at the **speaker** level or the **utterance** level?
- ▶ Can selection metrics be **combined** to produce an even better training data set?

Data and Tools

- ▶ **MACROPHONE:** Short utterances read over the phone
 - ▷ 4005 female speakers with average 40.7 utterances each
 - ▷ 83 hours of female speech
 - ▷ Transcribed noise
- ▶ **Festival:** Modular frontend processing for text to speech
- ▶ **Praat:** Toolkit for phonetic and acoustic analysis of audio
- ▶ **Merlin:** Toolkit for training neural network based speech synthesis models
- ▶ **IBM Watson:** Online speech recognition API
- ▶ **Amazon Mechanical Turk (AMT):** A popular crowdsourcing platform

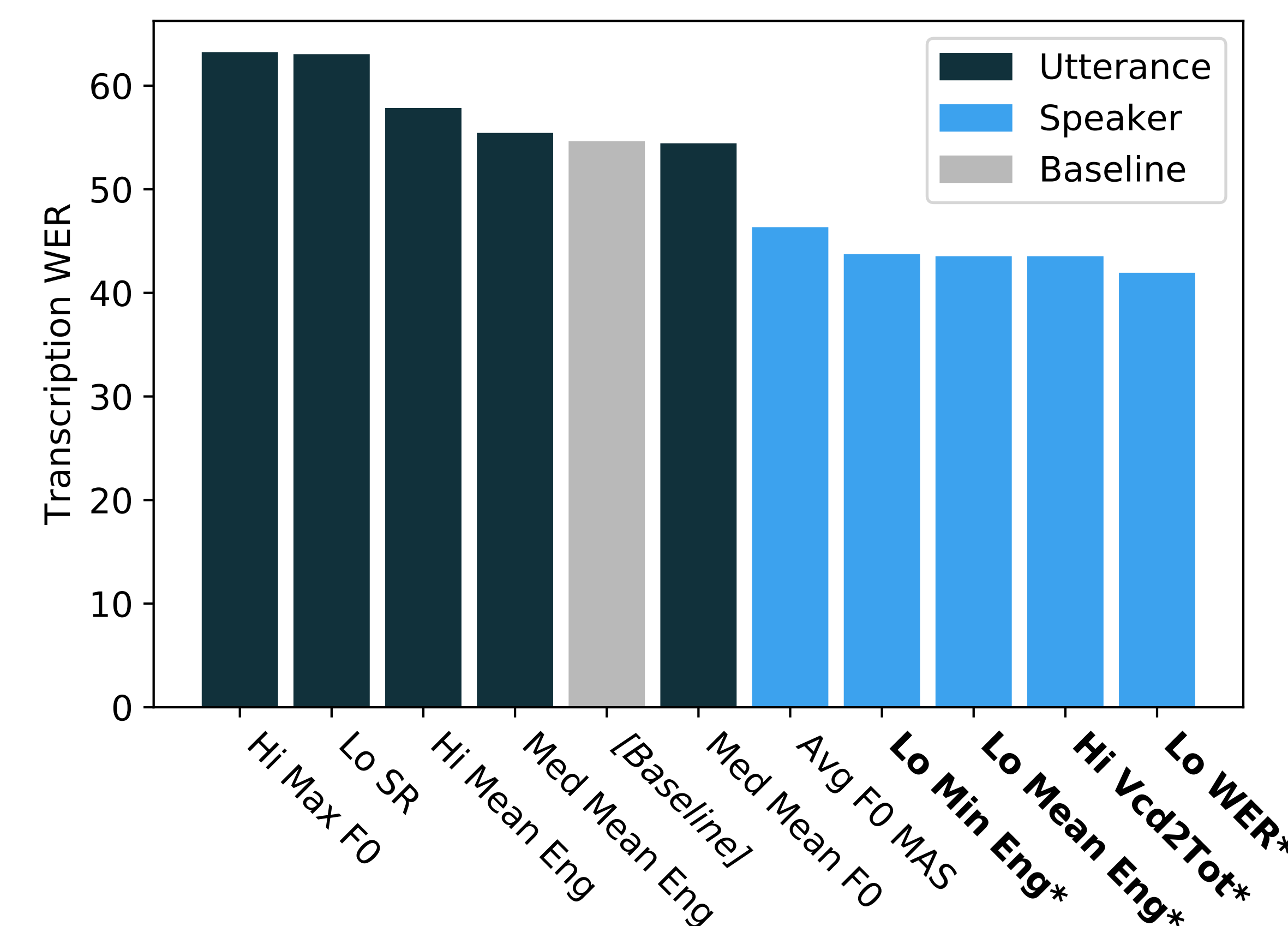
Acoustic and Prosodic Features

- ▶ F0 (min, max, mean, median, standard deviation)
- ▶ Mean absolute F0 slope (MAS)
- ▶ Energy (min, max, mean, standard deviation)
- ▶ Ratio of voiced to total frames
- ▶ Speaking rate (syllables per second)
- ▶ Level of articulation (mean energy / speaking rate)
- ▶ Watson WER

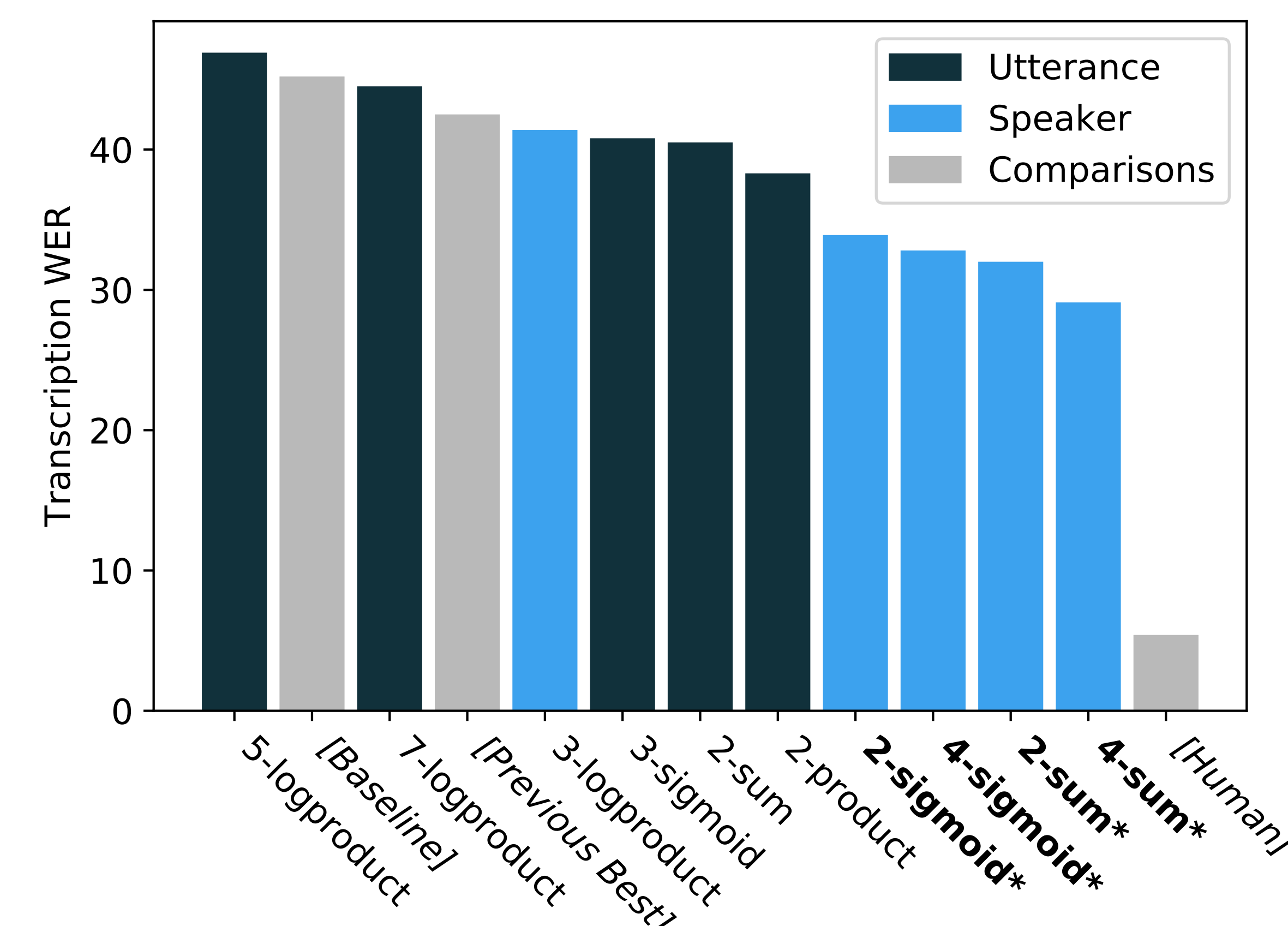
Experimental Setup

- ▶ **Baseline:** Voice trained on just the first 10 hours of female MACROPHONE data
- ▶ **Experimental voices:** Trained on 10 hours of data selected at the speaker or utterance level based on high, median, mean, and low values for each acoustic / prosodic feature
- ▶ **Evaluation:** IBM Watson (preliminary); Amazon Mechanical Turk transcription task; MCD (future)

Single Features: Speaker vs. Utterance Selection



Joint Features: Speaker vs. Utterance Selection



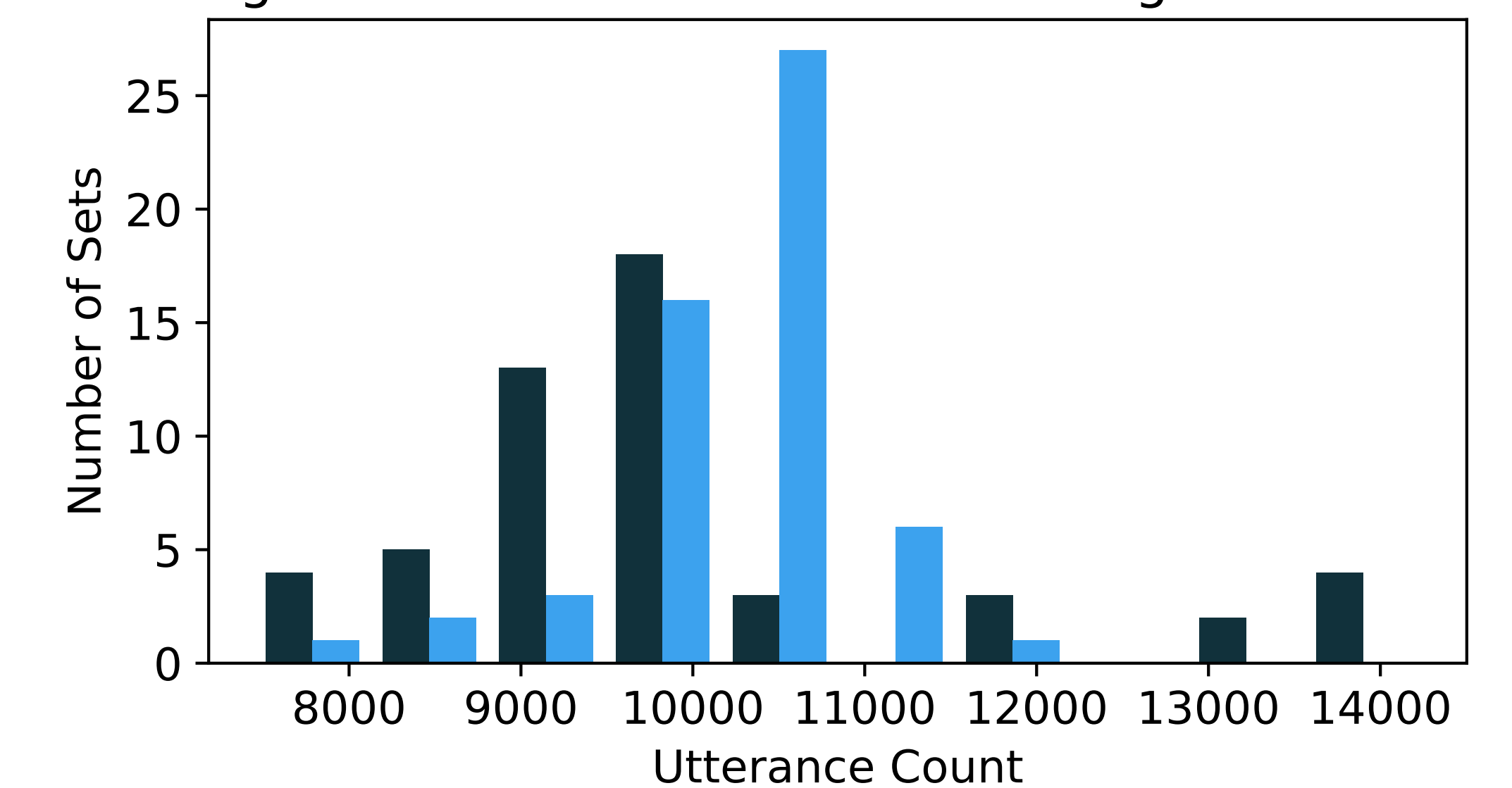
* Significantly more intelligible than baseline

Evaluation Method Correlations

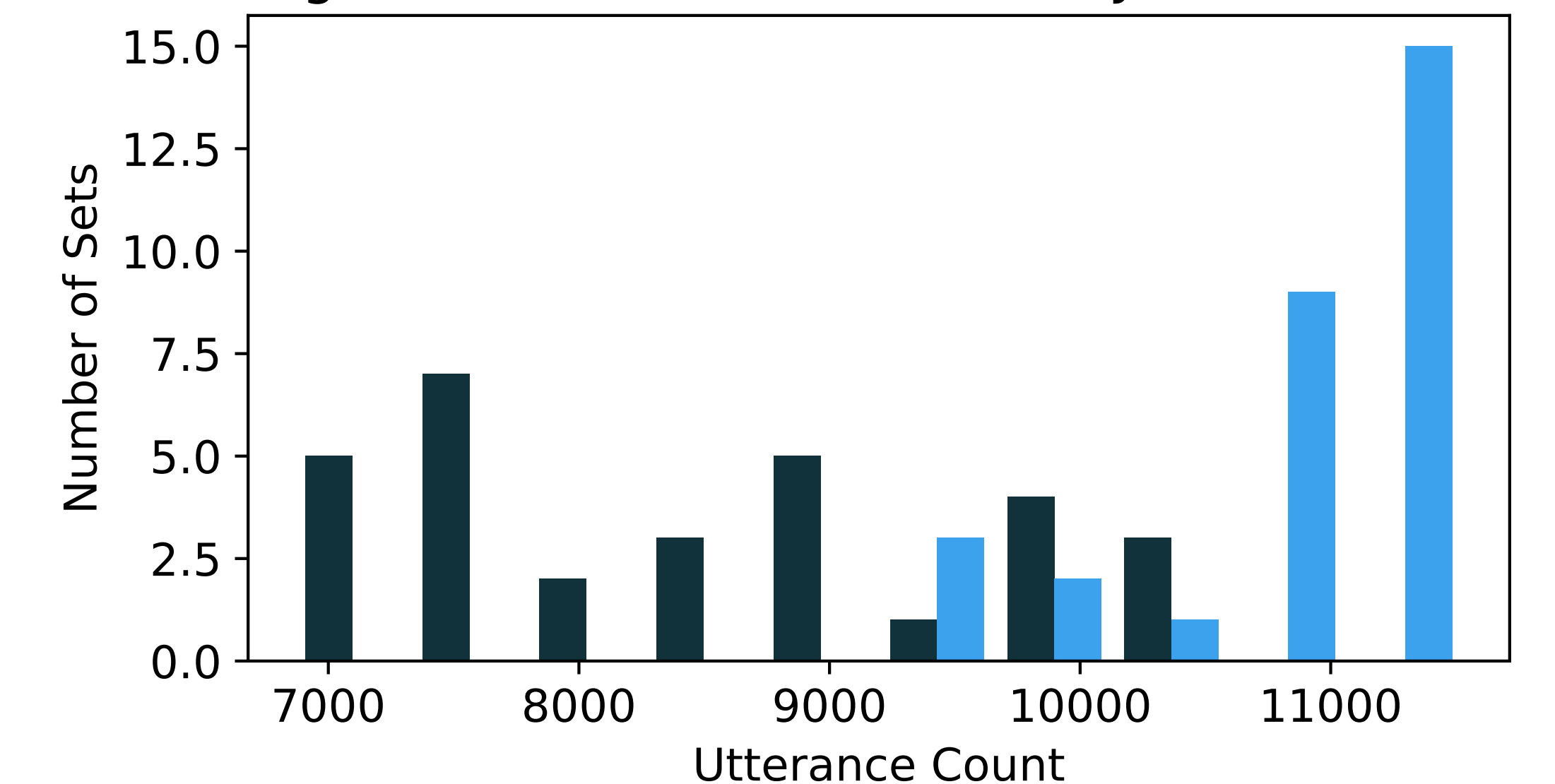
Comparison	Single-Feature	Joint-Feature
Watson, MCD	0.312	0.748
Watson, MTurk	0.814	0.899
MCD, MTurk	0.430	0.912

Subset Statistics: Utterance Counts

Histogram of Utterance Counts for Single Feature Sets



Histogram of Utterance Counts for Joint Feature Sets



Conclusions and Future Work

Conclusions:

- ▶ Selecting on speakers rather than utterances produces more intelligible voices
- ▶ Combining selection features results in further improvement
- ▶ Some speakers may be better suited for TTS than others

Future work:

- ▶ Low-resource languages
- ▶ MCD for objective evaluation
- ▶ Further explore what characteristics define “good” subsets
- ▶ Automatic selection of subsets

Acknowledgments

This work was supported by the National Science Foundation under Grants IIS 1548092 and 1717680.