# Utterance Selection for Optimizing Intelligibility of TTS Voices Trained on ASR Data

Erica Cooper [1], Xinyue Wang [1], Alison Chang [2], Yocheved Levitan [1], Julia Hirschberg [1]

[1]Columbia University, New York, USA
[2]Google, USA

## Research Questions

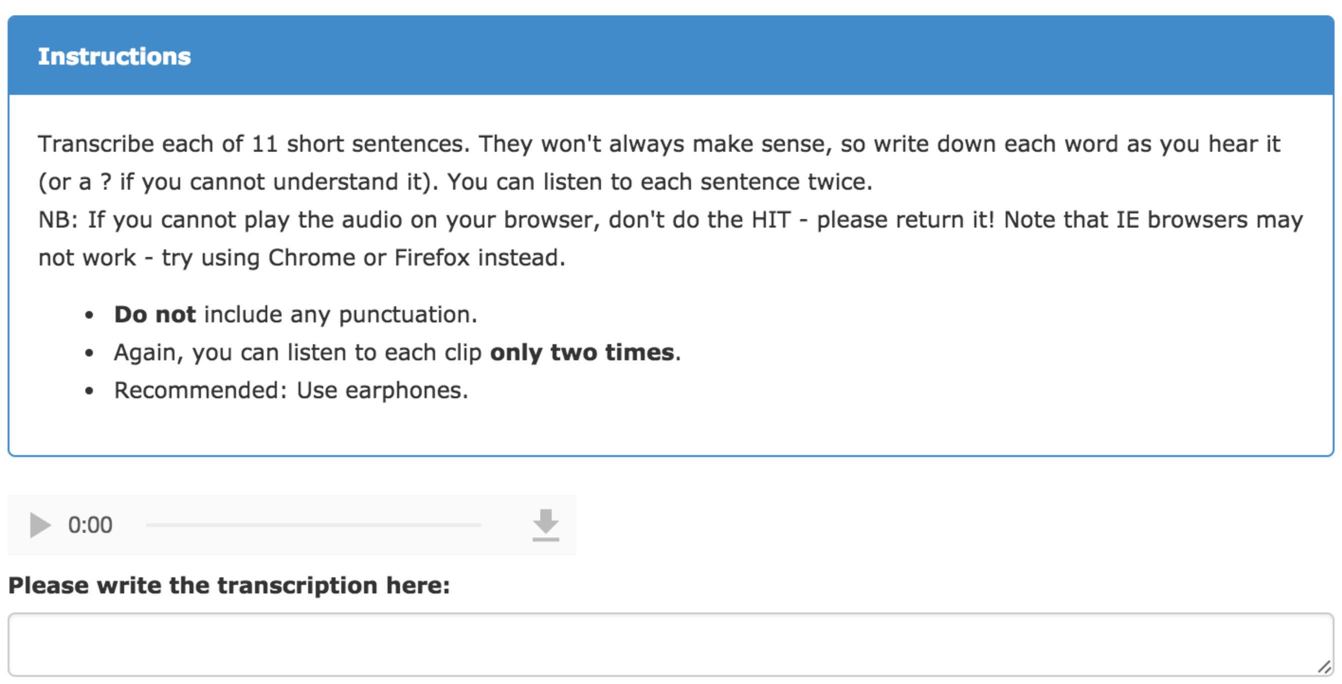
- Can we repurpose ASR data to create TTS voices?
- Can we select training utterances to optimize intelligibility?
- Does automatic evaluation using ASR correlate with human judgments?

## Data and Tools

- **MACROPHONE:** short utterances read over the phone
  - Multiple speakers
  - Using 83 hours of female speech
  - Transcribed noise
- **HTS:** Toolkit for training HMM-based TTS voices
- **Amazon Mechanical Turk (AMT):** Crowdsourcing platform

## Approach: Subset Selection

- **Baseline:** Train speaker-independent voice using first 10 hours of female data
- **Subsets:** Train voices only on subsets of the data, selected from high, medium, or low levels of various features:
  - Mean and standard deviation of f0 and energy
  - Speaking rate, level of articulation, and utterance length
  - Clipping, transcribed noise, spelled words
- **Evaluation:** Amazon Mechanical Turk




- **Constraint:** Transcribers are allowed to evaluate a given sentence only once.

## 2-hour Subsets Out of 10 Hours

Baseline: **67.7%** WER

| Feature | Low | Med | High |
|---|---|---|---|
| Mean f0 | 98.6 | 85.7 | 100.3 |
| Stdv f0 | 83.1 | 80.0 | 87.1 |
| Mean energy | 98.6 | 95.7 | 70.6 |
| Stdv energy | 100.9 | 85.4 | 79.7 |
| Speaking rate | - | 99.1 | **54.3** |
| Articulation | 76.0 | 87.7 | - |
| Utterance length | 96.6 | 85.4 | 96.9 |

## Removing Noise Out of 10 Hours

| Subset | Hours | WER |
|---|---|---|
| 3 or more words | 7:34 | 79.7 |
| No clipping | 9:57 | 77.7 |
| No transcribed noise | 5:53 | **58.9** |
| No spelled words | 9:24 | 94.3 |

## 2- and 4-hour Subsets Out of 83 Hours

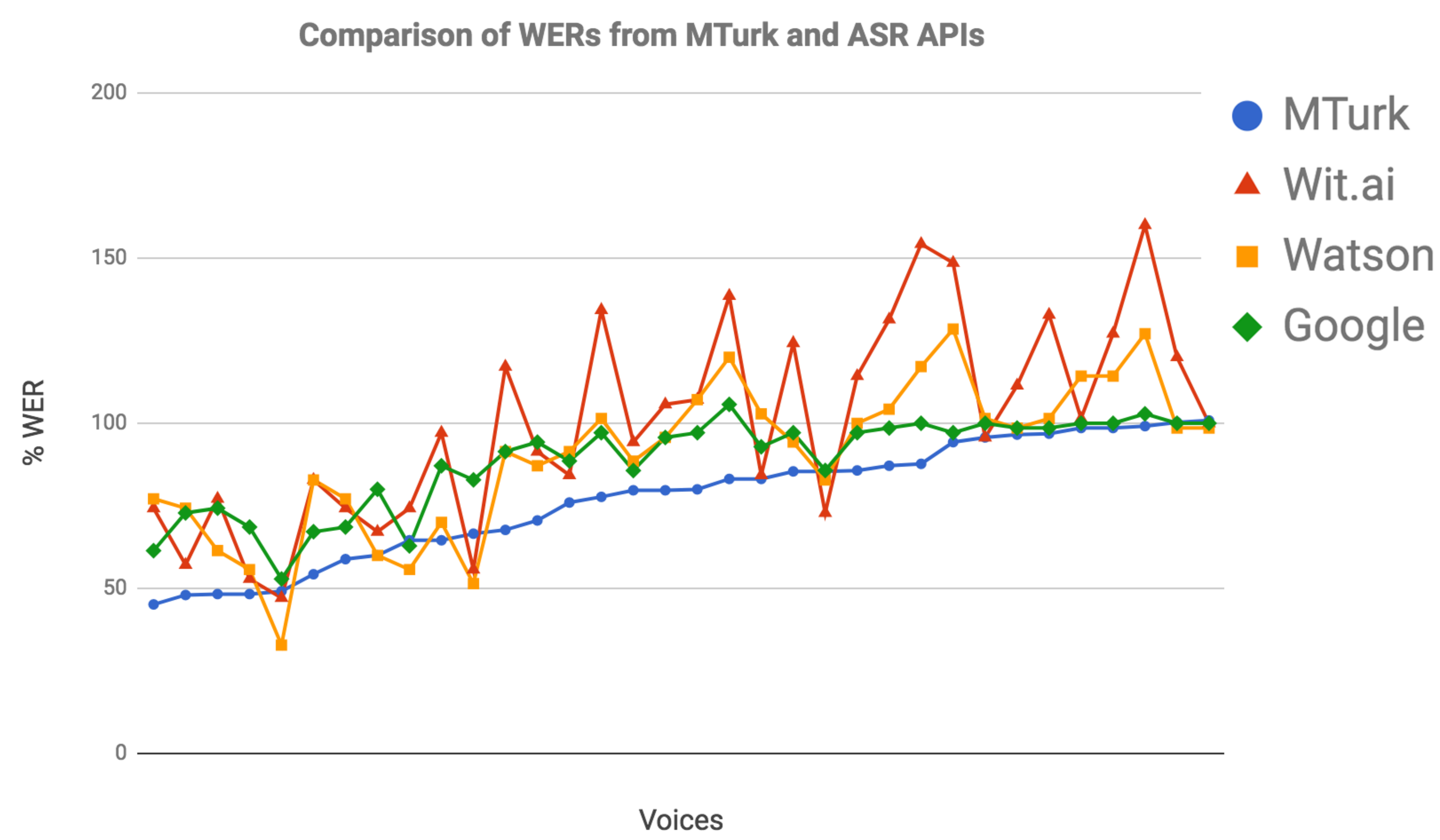
| Feature | 2hrs | 4hrs |
|---|---|---|
| High mean energy | **60.0** | **48.3** |
| High stdv energy | 83.1 | **64.6** |
| Fast speaking rate | **66.6** | **48.3** |
| Hypo-articulation | **64.6** | **49.1** |
| Middle stdv f0 | **48.0** | **45.1** |

Best voices:

- 4/83 and 2/83 hrs middle stdv f0
- 4/83 hrs fast speaking rate
- 4/83 hrs high mean energy
- 4/83 hrs hypo-articulated

## Automatic Evaluation Using ASR

IBM Watson Google Cloud Wit.ai




| Eval | Correl (r) | Std.Dev (%) |
|---|---|---|
| MTurk | — | 4.52 |
| wit.ai | 0.728 | 1.20 |
| Watson | 0.797 | 0.00 |
| **Google** | **0.876** | **0.00** |

**Limitations:**

- "Black box"
- Built for semantically sensible utterances
- Models may change

## 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. We also thank Meredith Brown for her helpful advice regarding evaluation with Mechanical Turk.