# A Comparison of Human and Machine Estimation of Speaker Age

Mark Huckvale & Aimee Webb
Speech, Hearing & Phonetic Sciences
University College London

# Speaker Profiling

- Determining characteristics of speaker from an audio recording
  - Gender, Region, Social Class
  - Age, Height, Weight
  - Health, Fatigue, Stress, Emotional state
- Applications in
  - Forensics
  - Marketing
  - Adaptation of Speech Technologies

# This study

- Objectives
  - Prediction of age of speaker from recording
  - How good are human listeners?
  - How good are machines?
- Method
  - Use same recordings for humans and machines
  - Use same objective measure of success:

Mean Absolute Error (MAE)

  - "how close is the average estimate to the actual age?"

# Previous studies of Human Listeners

**Table 1.** Previous studies on human listener age estimation

| Study | MAE (yr) | Notes |
|---|---|---|
| Braun et al, 1999 [4] | 10.5 | German speakers & listeners |
| Braun et al, 1999 [4] | 8.5 | Italian speakers & listeners |
| Krauss et al, 2002 [5] | 7.1 | Limited age-range |
| Amilon et al, 2009 [6] | 9.7 | |
| Moyse et al, 2014 [7] | 10.8 | |

# Previous studies of Machine Listeners

**Table 2.** Previous studies on machine age estimation

| Study | MAE (yr) | Notes |
|---|---|---|
| Bocklet et al, 2008 [10] | 0.8 | Children 7-10 yrs |
| Feld et al, 2009 [11] | 7.2-12.8 | Same & cross-language |
| Doby et al, 2011 [12] | 9.29-10.00 | Depending on gender |
| Bahari et al, 2011 [13] | 7.48 | Null model = 8.88 |
| Bahari et al, 2012 [14] | 7.9 | |
| Bahari et al, 2014 [8] | 6.08 | Null model = 10.3 |

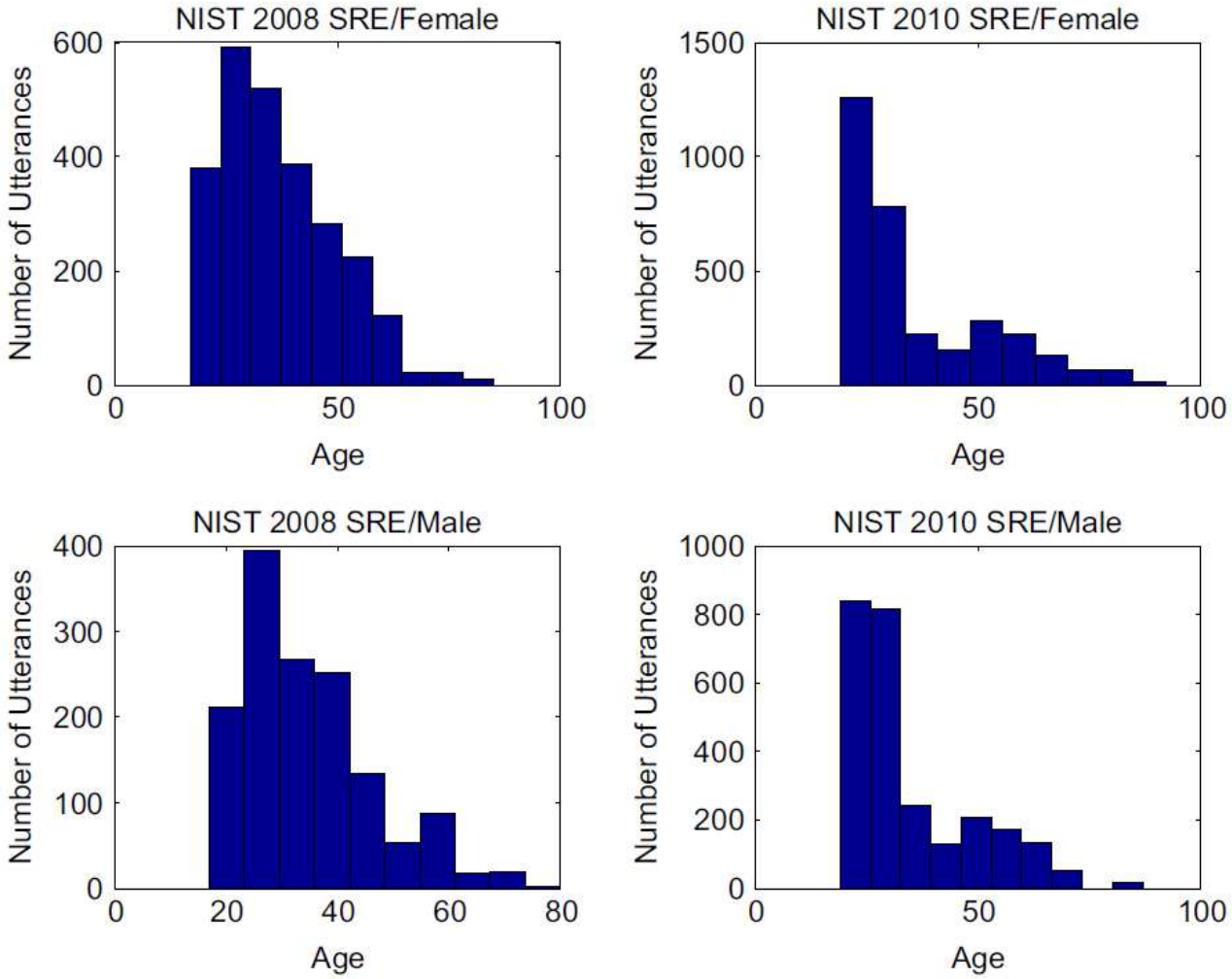



**Fig. 3.** Age histogram of telephone speech utterances for NIST 2010 and 2008 SRE Databases.

# Speech Corpus

- Accents of the British Isles 2 Corpus
  - 262 speakers
  - 13 accent areas of the British Isles
  - Read passage (median duration = 39.2s)
  - Wide bandwidth audio of good quality, recorded using a close-talking microphone at 22050 samples/sec.
- Test Set
  - 52 speakers
  - Equal representation of men and women for all 5-year age bands between 15 and 80
- Training Set
  - Other 210 speakers

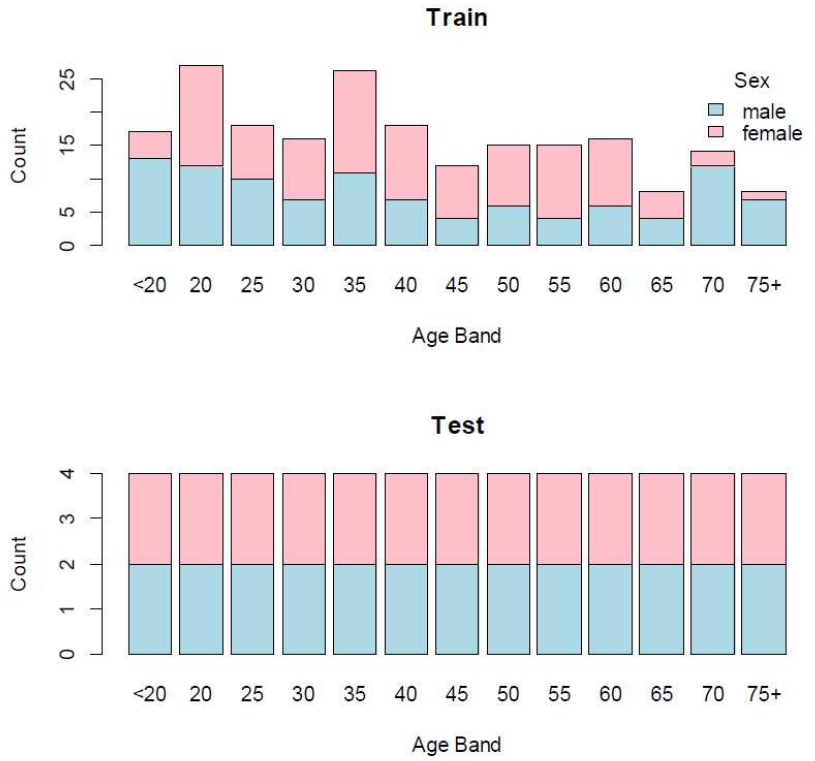
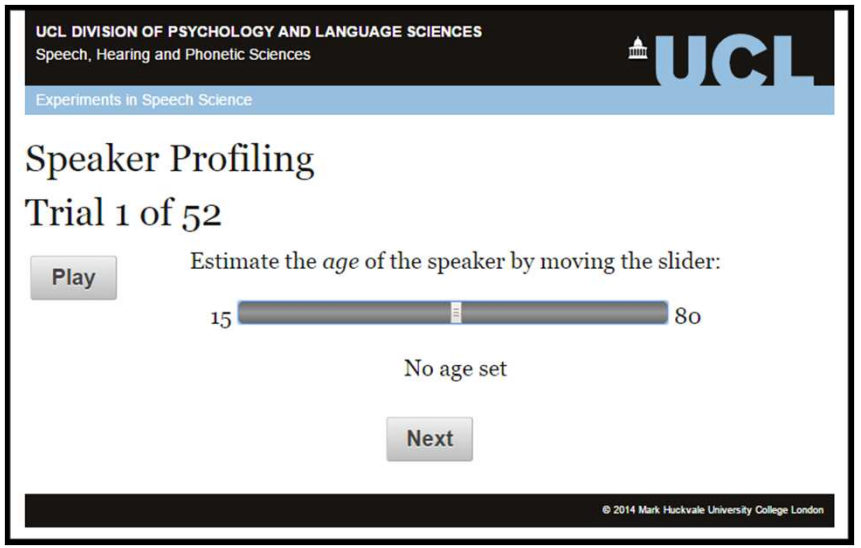



**Fig. 1.** Age and gender distribution for the train and test sets.

# Listening Experiment

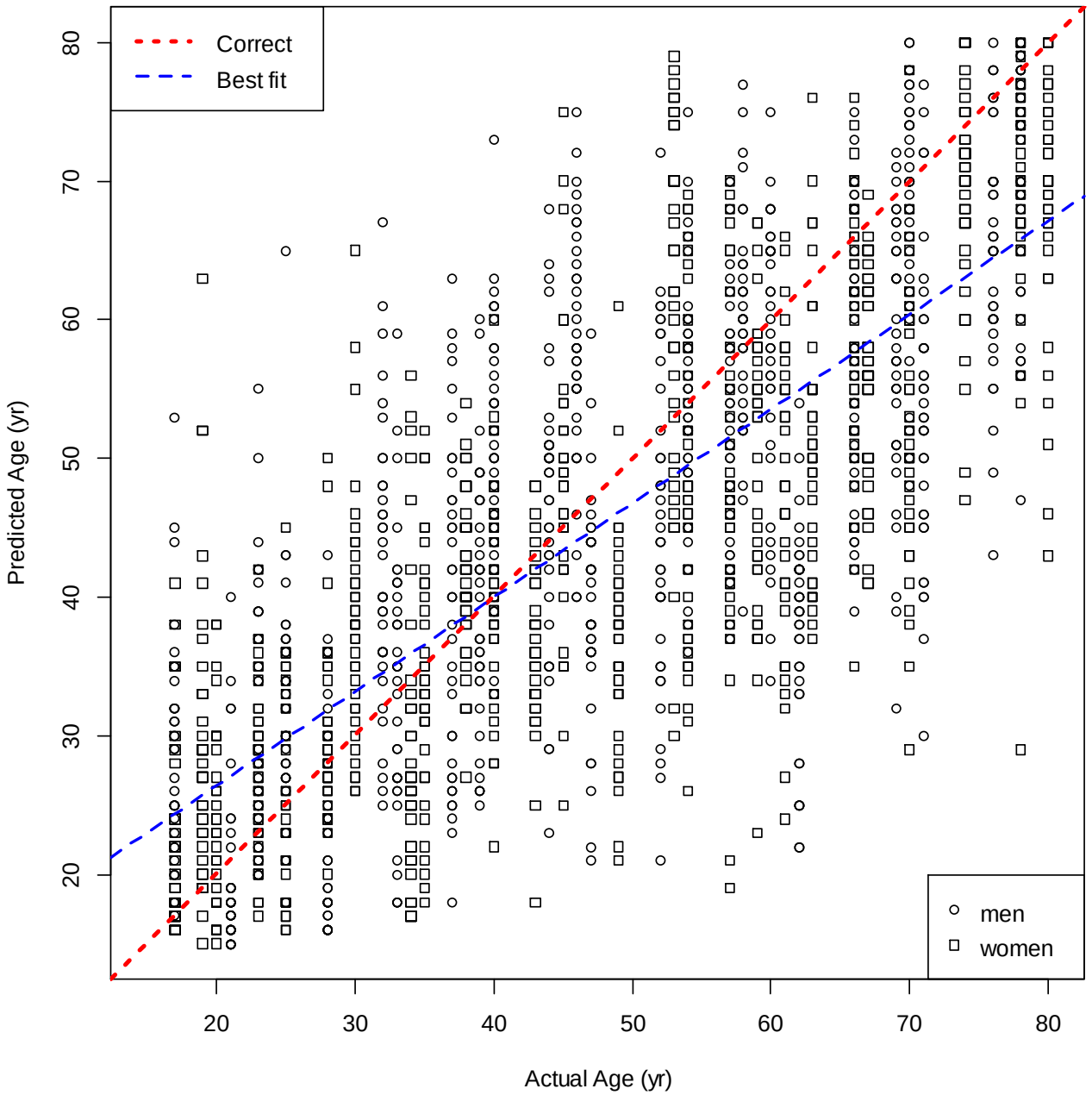
- Web experiment
  - Listen as much as wanted
  - Specify age using slider
  - Range: 15-80




- 36 British English listeners

**Table 3.** Distribution of listeners by age and gender

| Num-ber | 20-29 | 30-39 | 40-49 | 50-59 | 60-69 |
|---|---|---|---|---|---|
| Male | 4 | 4 | 3 | 3 | 2 |
| Female | 5 | 4 | 3 | 6 | 2 |


Correct
Best fit
men
women
Predicted Age (yr)
Actual Age (yr)
20
30
40
50
60
70
80

# Human Listener Performance

**Table 4.** Mean Absolute Error of prediction as a function of age and sex of the speaker

| MAE(yr) | **20-29** | **30-39** | **40-49** | **50-59** | **60-69** | **70-79** |
|---|---|---|---|---|---|---|
| Male | 7.42 | 9.32 | 10.52 | 7.99 | 13.85 | 11.14 |
| Female | 5.63 | 8.00 | 8.71 | 12.07 | 12.10 | 11.30 |

**Table 5.** Mean Absolute Error of prediction as a function of age and sex of the listener

| MAE (yr) | **20-29** | **30-39** | **40-49** | **50-59** | **60-69** |
|---|---|---|---|---|---|
| Male | 8.34 | 8.22 | 11.36 | 10.01 | 13.24 |
| Female | 9.95 | 9.39 | 10.57 | 9.38 | 10.10 |

- Overall MAE=9.79yr, Male=10.51yr, Female 9.51yr
- Speakers in age-band 20-29 significantly better recognised
- Older listeners worse than younger listeners at predicting age

# Listener Panels

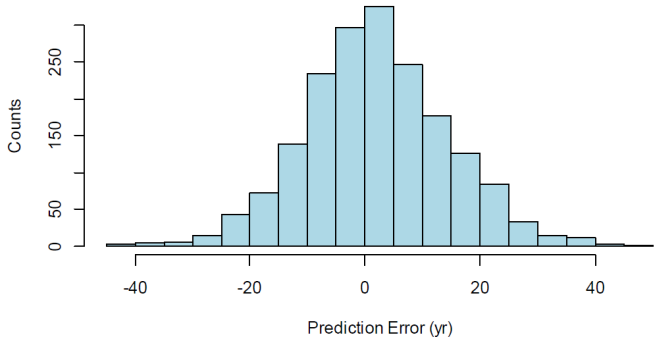
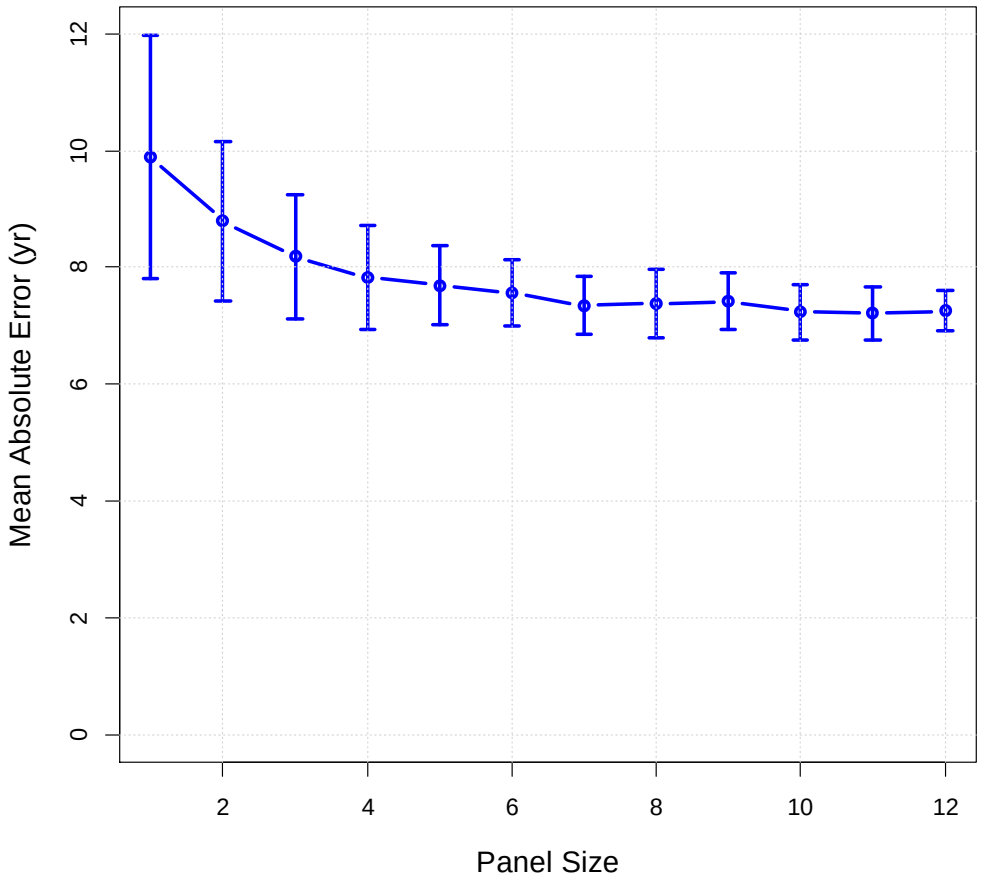



**Fig. 4.** Distribution of age prediction errors by human listeners.

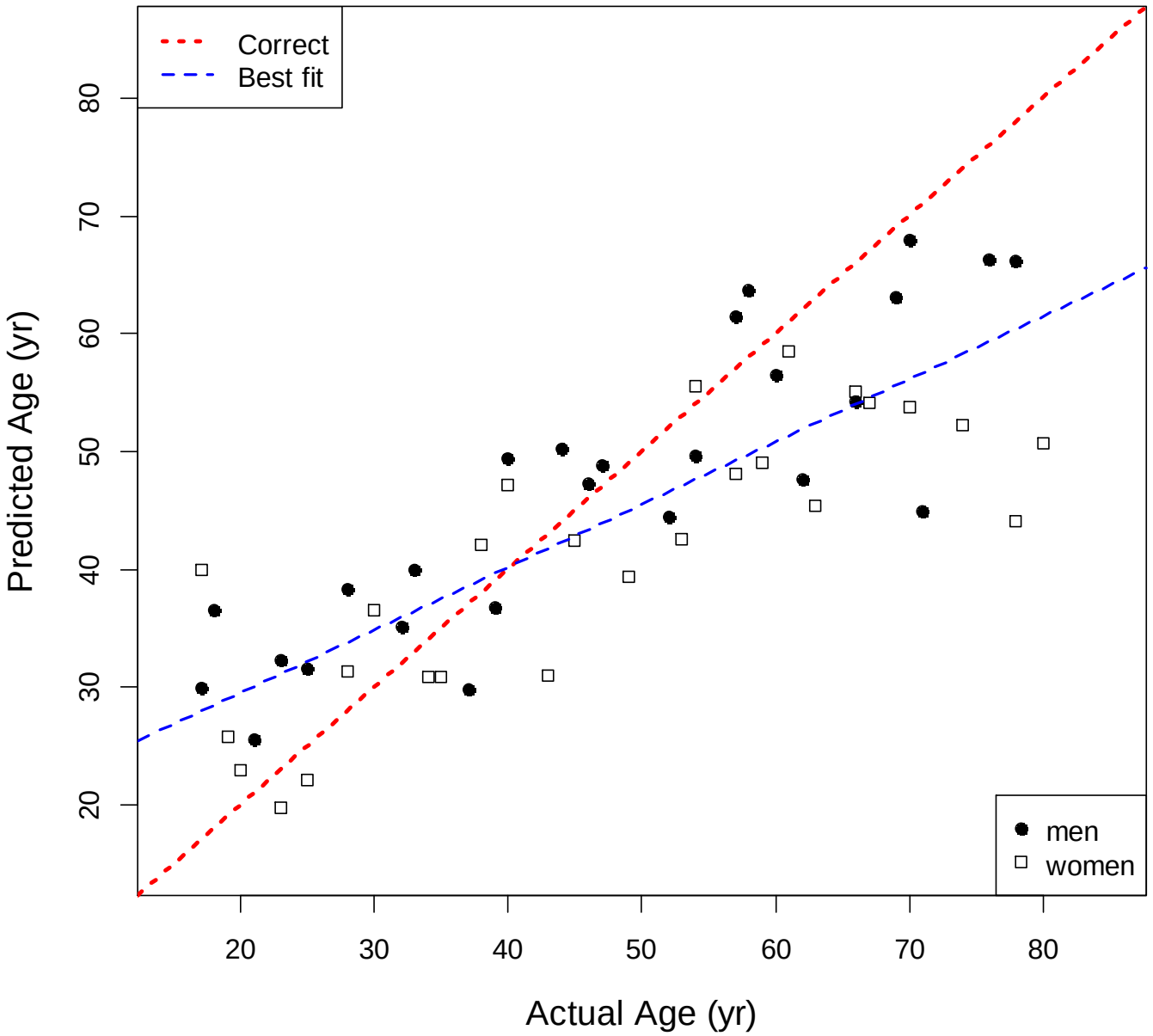


# Machine Estimation

- Feature Analysis
  - OpenSMILE Toolkit
  - InterSpeech 2014 Paralinguistics Challenge
  - 65 low-level descriptors
    - e.g energy, spectral envelope, pitch and voice quality
  - Summative statistics collected per audio file
    - e.g. means, medians, quantiles, differences
  - 6373 features per recording
- Feature Selection
  - Only features which have absolute correlation with age > 0.1
  - Leaves 2538 features

# Machine Estimation

- Machine Learning
  - Support Vector Regression ("e1071" in R)
  - Radial basis function kernel
  - Optimal control parameters found on training set
  - Separate systems for
    - Male speakers
    - Female speakers
    - All speakers


Correct
Best fit
men
women
Predicted Age (yr)
Actual Age (yr)
20
30
40
50
60
70
80

# Machine Performance

**Table 6.** Mean Absolute Error of prediction as a function of age and sex of the speaker

| MAE(yr) | **20-29** | **30-39** | **40-49** | **50-59** | **60-69** | **70-79** |
|---|---|---|---|---|---|---|
| Male | 7.64 | 4.89 | 4.68 | 5.54 | 8.87 | 12.40 |
| Female | 3.12 | 4.46 | 7.86 | 7.72 | 11.00 | 23.97 |

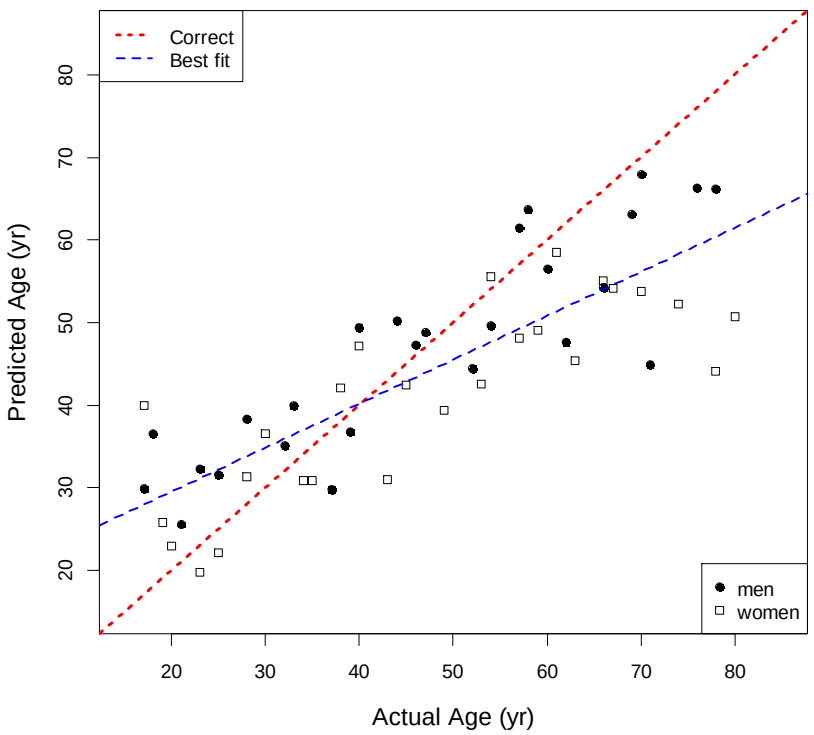
| Measurement | MAE (years) |
|---|---|
| Overall MAE | 9.13 |
| Male speakers only | 7.98 |
| Female speakers only | 10.29 |
| Gender independent model | 9.18 |
| Null model | 16.7 |

# Rebalance Training Set


SMOTE: Synthetic Minority Oversampling Technique

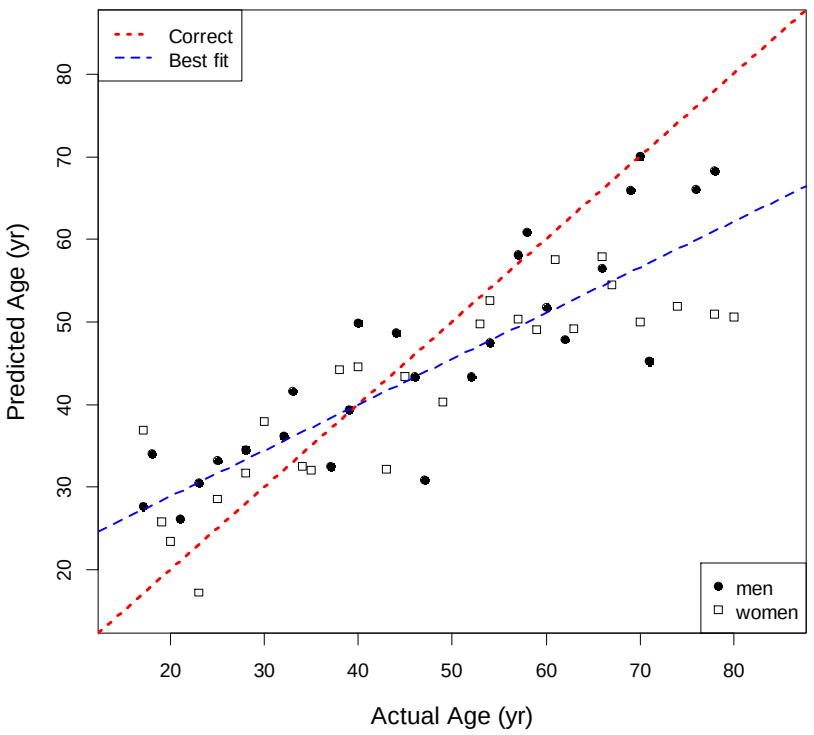
# Effect of balancing training set

Before




MAE=9.13years

After




MAE=8.64years

# Comparison of Human & Machine

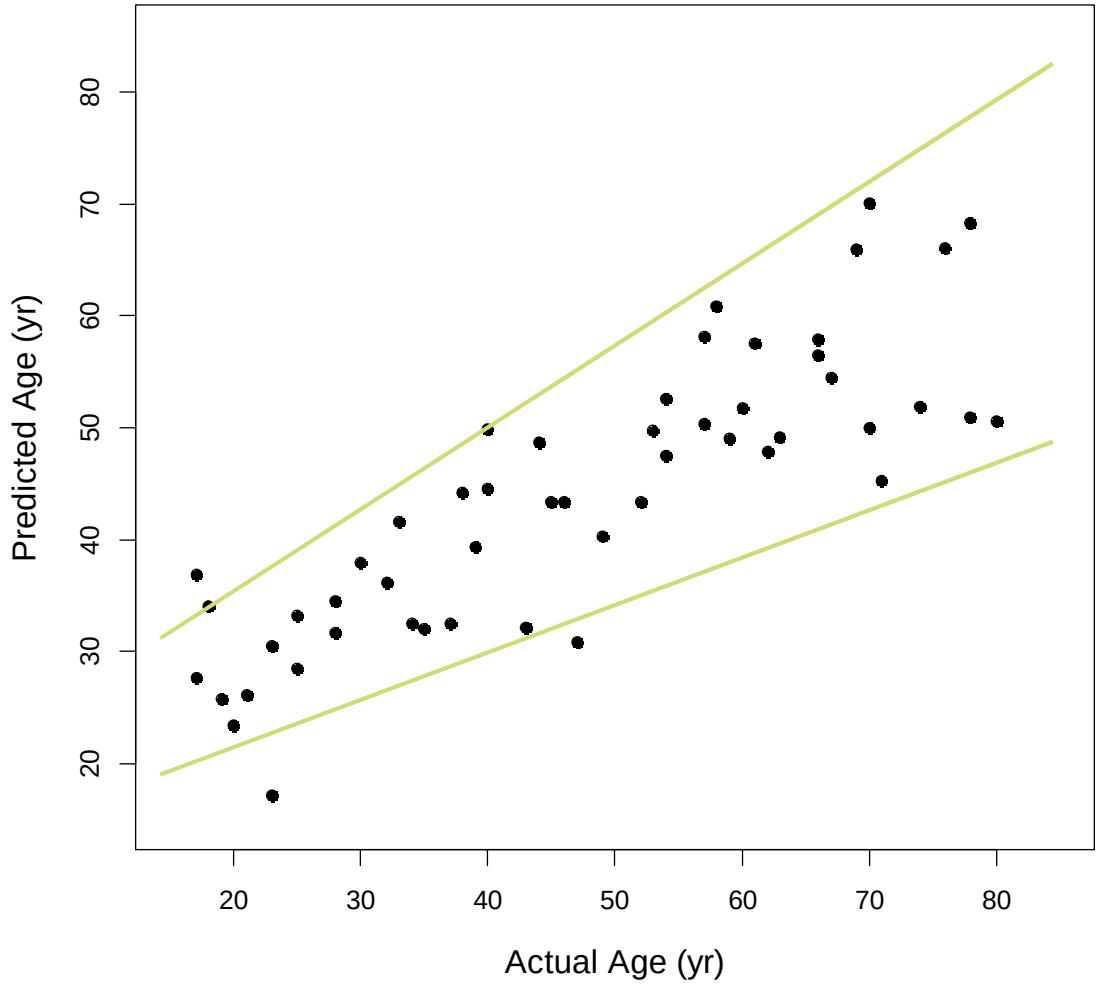
## Human Listeners

- Overall MAE=9.79years
- Listener Panels
  - Size 10: ~7.5yr
  - Size 2: ~8.5yr

## Machine

- Overall MAE=8.64years
- Better than 2/3 of individual listeners

# Increasing variation with age?




Age 58

Age 62

# Summary

- Human and machine performance on age estimation from speech more similar than previous studies suggested
  - Once we make a fair comparison
- Machine systems have an edge over the average listener
- Panels of listeners have an edge over the machine
- Some indication that increasing variability in voice with increasing age makes prediction of age more difficult for older speakers