GenderListener

A tool to automatically label speaker gender in audio

Cynthia E. Correa

March 13, 2018

CapitalOne Lightning Talks

Metadata of 2,500 TED talks

- Topics
- Comments
- Views
- Speakers
- Cool facts

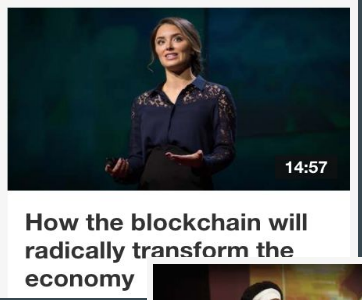
Audio files of 1,500 TED talks

→ Project idea

I wanted to know speaker gender for my armchair sociology research.

Picture source:: The Independent

GenderListener



Input: any audio file

Output: male / female label

Model: Classify as male or female based on analyzing the sound spectrum of TED talks

Broader uses:

- *actual* sociology research
    - continuous + anonymous labels
- improve Alexa and Siri responses
- customer service call routing
- labeling old data and discovering trends

Audio signal processing

Sampled 4 minutes from each talk with PyAudioAnalysis

Extracted these features ==>

Index	Name
1	Zero Crossing Rate
2	Energy
3	Entropy of Energy
4	Spectral Centroid
5	Spectral Spread
6	Spectral Entropy
7	Spectral Flux
8	Spectral Rolloff
9–21	MFCCs
22–33	Chroma Vector
34	Chroma Deviation

What information is in speaker’s sound?

Voice

- Pitch and variation, roughness
- Changes with age, weight
- Related to anatomy

Speech

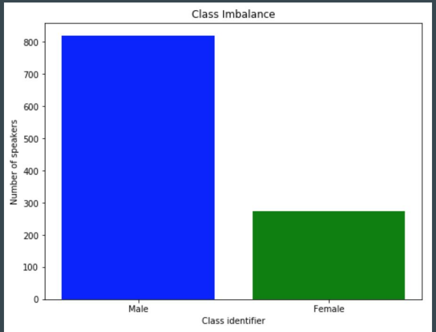
- Pace, intonation
- Changes with region and identity
- Related to gender

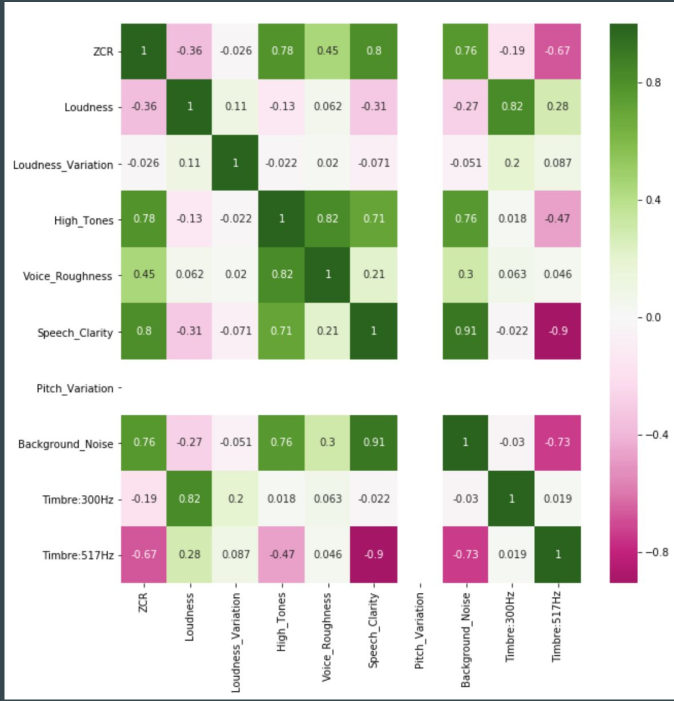
Model

- Supervised binary classifier
- Trained on 1,096 TED talks
- I got gender labels based on speaker's first name using gender_guesser and used those as my ground truth

Model

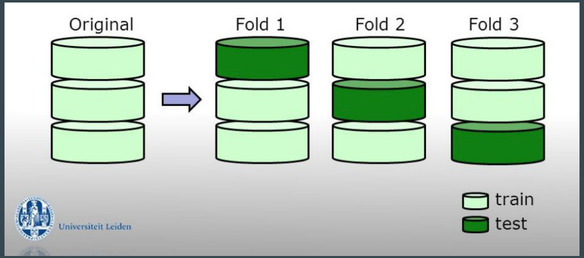
- Class balancing
  - Random oversampling
- Feature Engineering:
  - Drop uncorrelated features
  - Drop redundant features



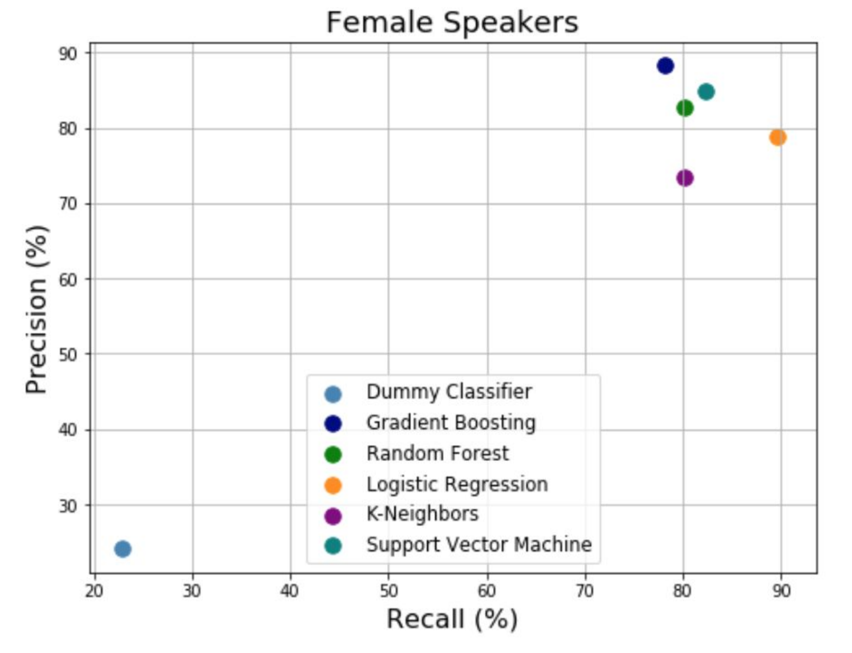


Model

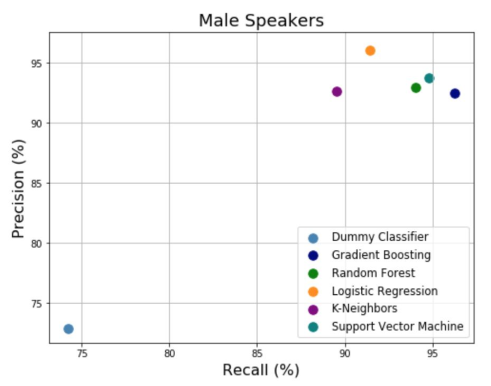
- Standardized features
- L2 regularization
- Stochastic average gradient descent
- Hyperparameter grid_search
  - Log loss (cross entropy)
  - Stratified 3-fold cross-validation



Model Comparison

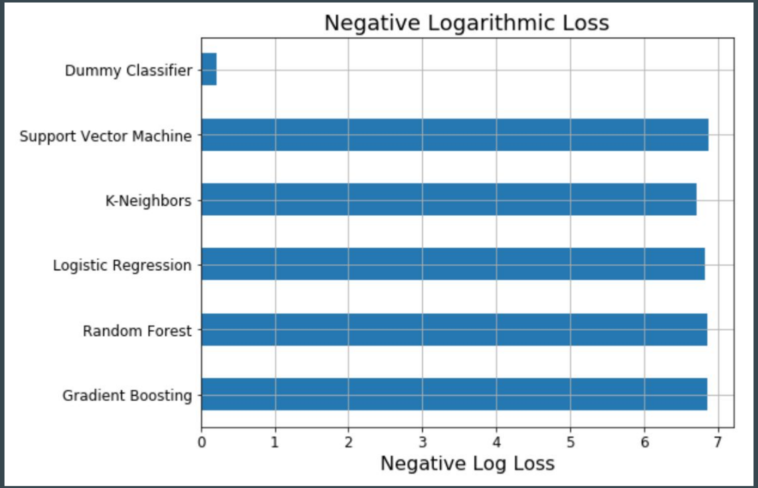


- Highest recall of female speakers:
  - Logistic regression
- Highest recall of male speakers:
  - Gradient Boosting



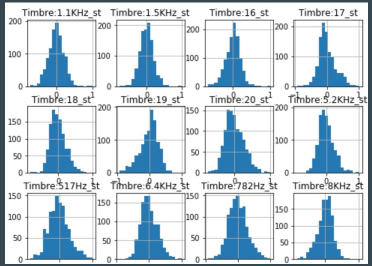
Model Comparison

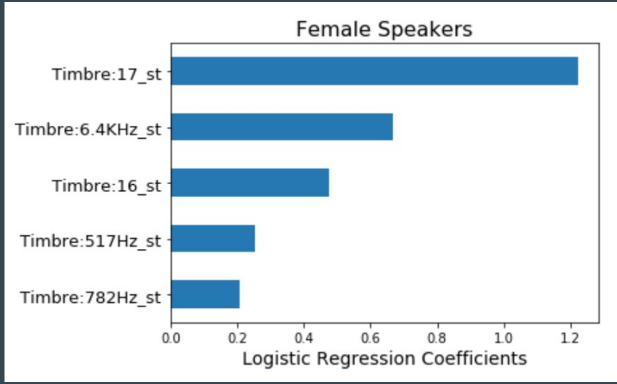
- Logistic regression and gradient boosting among the best



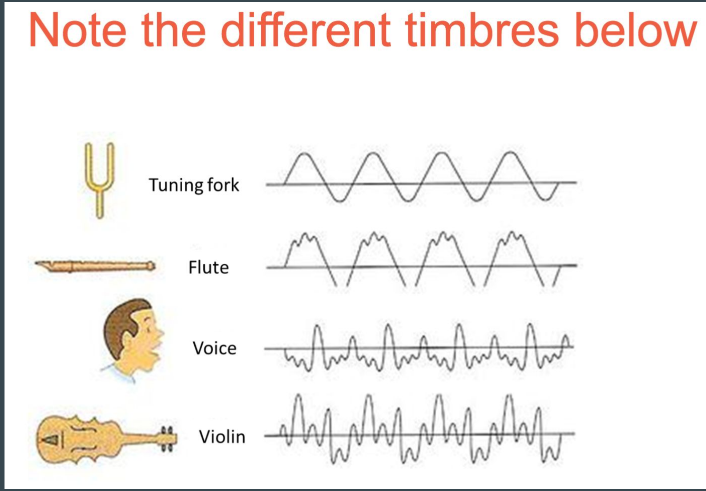
Feature Importance

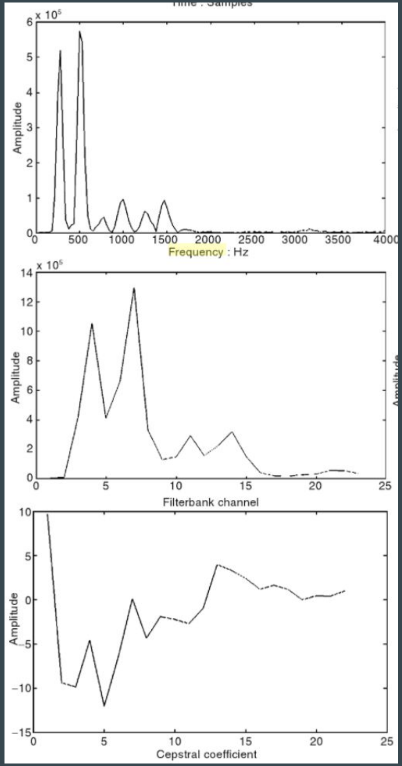
- Well-suited for Logistic Regression
- Top features:
  - Timbre at the 17th Frequency filterbank
  - Timbre at the 6.4KHz filterbank





Timbre

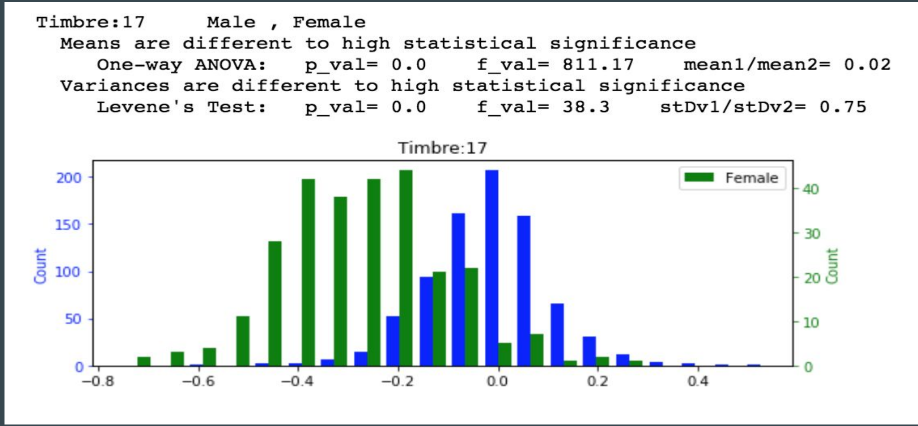




(Mini-Demo if time allows)

Feature Importance

- For the most predictive sound features, the distributions were not binary but *bimodal*
- Means for males and females are different
- There is much overlap over the entire spectrum



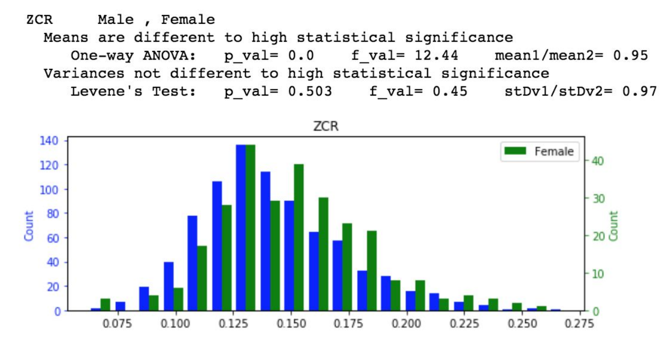
Demo

Can you identify the talks by
Doris, Samantha, and Jennifer?

Interesting Finding:

Pitch was not the biggest difference between “male” and “female” speakers!

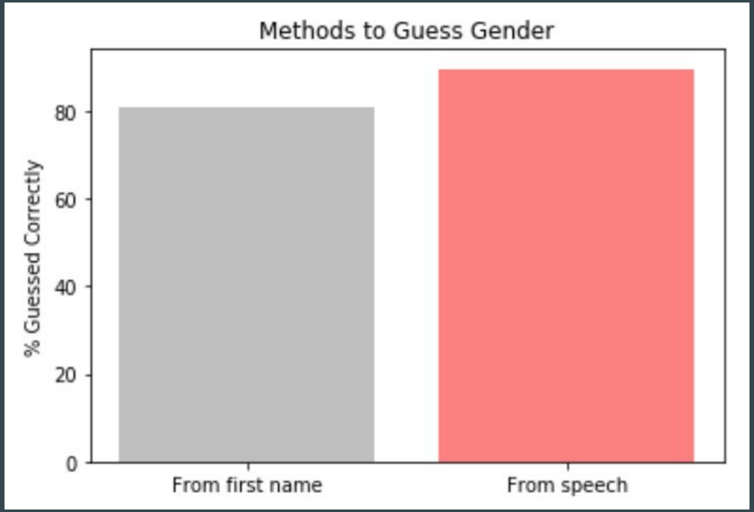
- 9 other features were more important.



Labeling gender by sound vs. labeling by first name

GenderListener pros:

- Higher accuracy (90% vs 80%)
- Can label throughout audio
- Don't need to know speaker's name



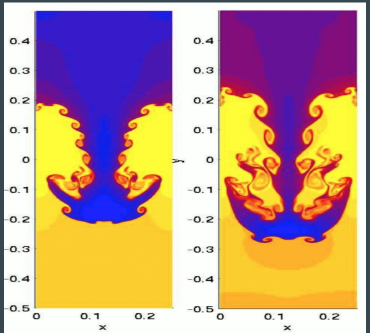
GenderListener cons:

- Labels are correlated with speaker's anatomy, not just with their gender

Summary

- ❏ Broader uses of GenderListener
- ❏ How model was built
- ❏ Which features were most predictive
- ❏ Comparison to alternative method

About Cynthia Correa



Signal → Discovery

Additional Slides

Index	Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum.
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9–21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22–33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

Complete list of implemented audio features. Each short-term window is represented by a feature vector of 34 features listed in the Table.

doi:10.1371/journal.pone.0144610.t002