Automatic Identification of Gender from Speech

Sarah Ita Levitan[1], Taniya Mishra[2], Srinivas Bangalore[2]
[1]Columbia University, [2]Interactions LLC

Research Questions

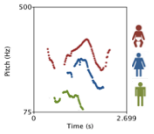
- Can we automatically identify speaker gender using short segments of speech (<3s)?
- Which feature combinations, representations and machine learning models are best for gender identification from speech?
  - Prosodic vs. cepstral features
  - Summary statistics vs. feature trajectories
  - Numeric vs. categorical classification
- Are our models robust across languages and corpora?
- What is the impact of including child speech with adult speech on gender identification?

Corpora

Name	HMIHY	aGender
Language	English	German
# utterances	5,002	46,157
Mean utt. len	6s	2.6s
Source	Telephone	Telephone
Gender	M,F	M,F, child

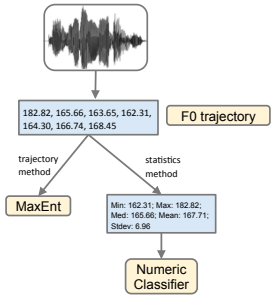
Features

- Fundamental frequency (f0)
- Cepstral coefficients (MFCCs)
- Energy
- Jitter, shimmer (voice quality)



Pitch contours extracted from male, female, and child speech

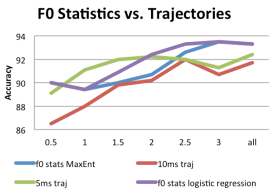
F0 Stats vs. Trajectories



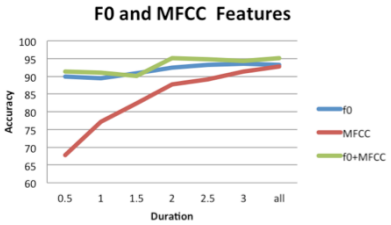
Method

- Extract f0 trajectories using Praat
- If statistics approach, compute statistics and use numeric classifier to make predictions
- If trajectory method, bin f0 tokens and use categorical classifier (MaxEnt)
- Trajectory approach avoids preprocessing step

- Experiment with 2 f0 trajectories, sampling speech at 10ms and 5ms intervals
- Model f0 trajectories as *text input* , binning each token (round to nearest 10)
- Use MaxEnt text classifier (LLAMA) which computes up to trigrams on f0 values as features
- Achieves >90% accuracy with one second of speech

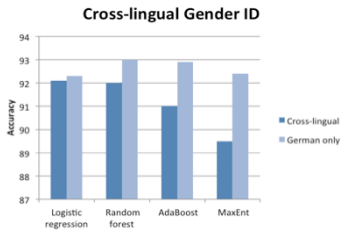


F0 vs. MFCC Features



- f0 and MFCC classification results of a logistic regression learner using varying length segments of speech
- f0 features are more predictive than cepstral features
- combined approach is most accurate: 95.2% with 2s of speech

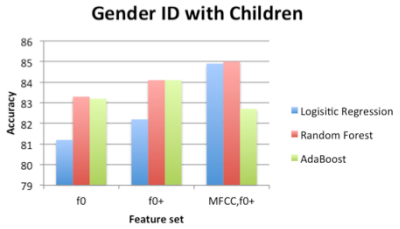
Cross-lingual Gender Identification



- Cross-lingual and same-language results of four classifiers, using f0 statistics
- Logistic regression and Random forest classifiers are most robust across corpora and languages, LLAMA (MaxEnt) is least
- Exclude utterances spoken by children in German data

➢ Can train a gender identification model on small amount of English speech and produce accurate predictions on German speech

Gender Identification with Children



- 3-way classification between male, female, and child speech
- Compare with benchmarked data from 2010 Interspeech Paralinguistics Challenge
  - Challenge baseline: 76.99%
  - Challenge winner: 84.3%
  - Our best model: **85.0%**
- f0+ represents f0 statistics features supplemented with energy statistics and jitter and shimmer

Conclusions

- Simple f0 statistics are highly predictive of speaker gender
- Novel trajectory approach using categorical classifier achieves >90% accuracy with *one second* of speech
- Combined f0 and MFCC features result in highest accuracy
- Cross-lingual evaluation (English models tested on German data) shows model robustness across corpora and languages
- State-of-the-art gender identification with children, using Random Forests with simple acoustic-prosodic features

sarahita@cs.columbia.edu, {tmishra,sbangalore}@interactions.com