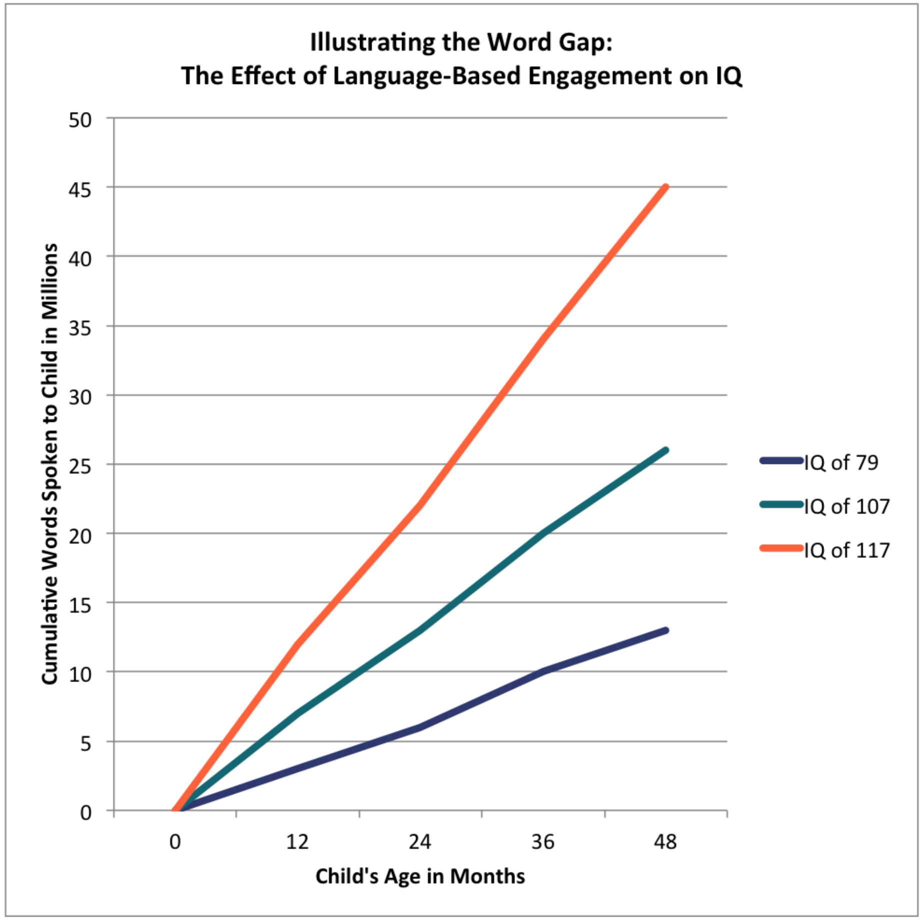
VersaMe: Television Audio Processing and Speech Detection

Richard Liu

August 21, 2015

VersaMe Overview

- VersaMe is a startup that is creating a wearable device that tracks the number of words a toddler is exposed to
- Research suggests that the biggest indicator of future trajectory in life is the number of words spoken to the child between 0 and 4 years old
  - Often called the "word gap"
  - Betty Hart and Todd Risley, 1995



Versame Projects

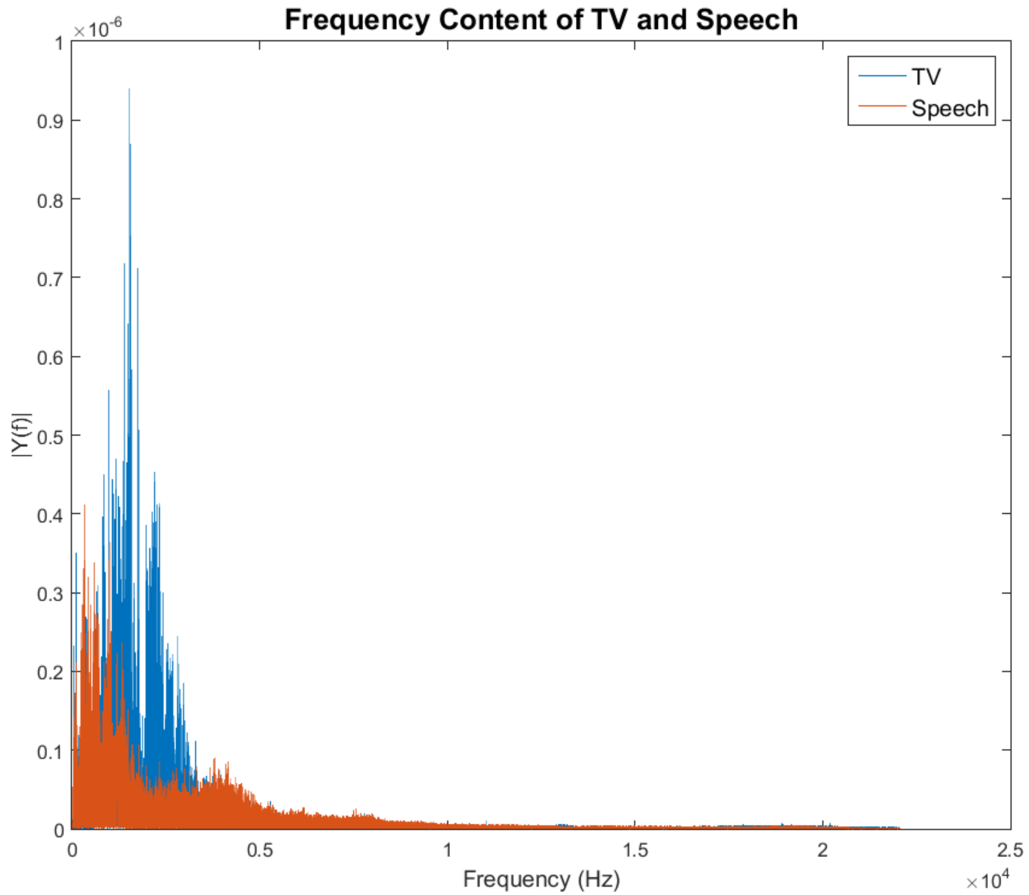
- Research project #1 (TV detection): Given an audio recording, is there a way to determine which portions are the TV?
- Research project #2 (Speech activity detection, SAD): Given an audio recording, is it possible to segment it into speech and nonspeech?
  - Or classify it into speech, silence, sound (audible nonspeech)

TV Detection

- TV audio consists of music and narrator/character speech
- Features – where can we see a difference?
  - Mel Frequency Cepstral Coefficients (MFCC’s)
  - Zero crossing rate – could possibly detect the presence of music
  - Prosodic features (longer term) – Jitter, Shimmer, Harmonic-to-noise ratio, Pitch, Formants, etc

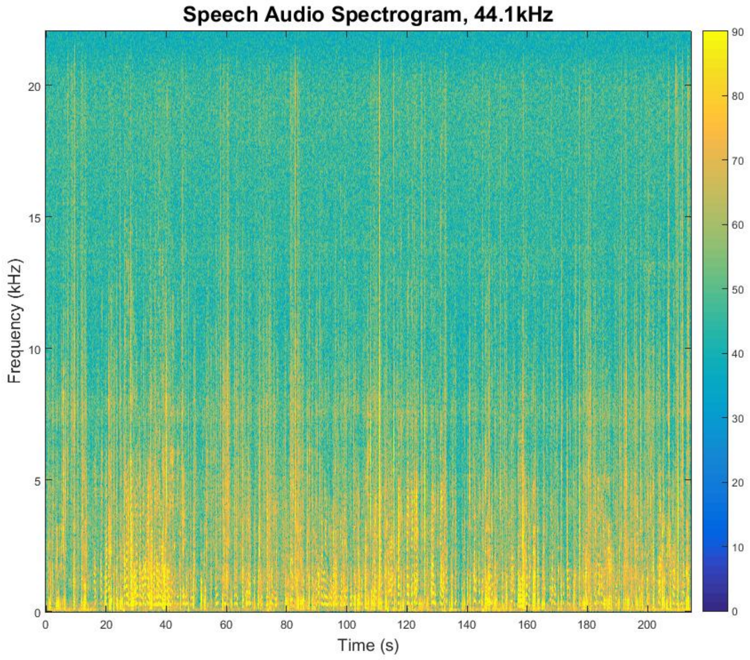
TV Detection – Frequency Content

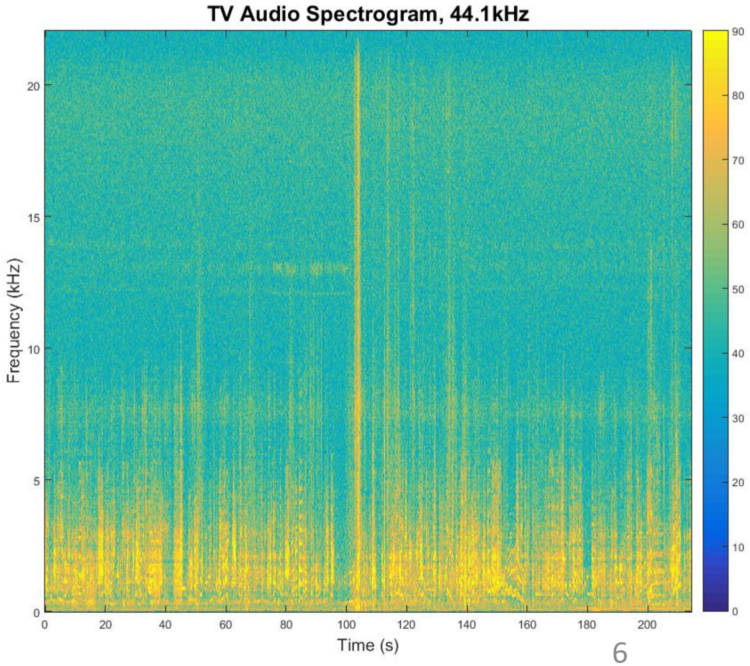
- TV Audio has a lot of energy between 1-4 kHz
- Speech is more present at lower frequencies (father's voice) and at slightly higher frequencies (children's voices)
- Data used: 44.1 kHz audio, ~43 seconds of TV (Curious George), ~57 seconds of household audio



TV Detection - Spectrograms

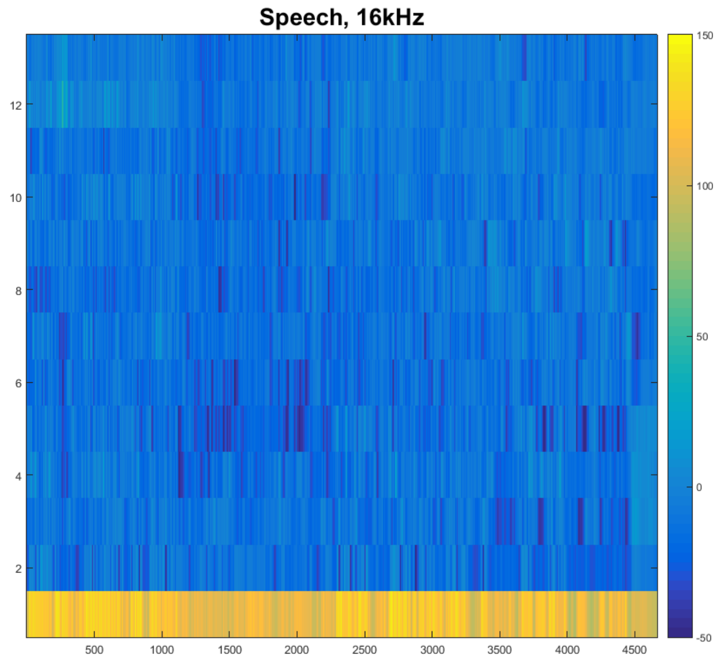
- Spectrograms of speech audio compared to TV audio are shown to the right
- Confirms previous observations that energy from speech is spread across all frequencies, while energy from TV is limited to a small band
- Data used: ~220 seconds of household speech, ~220 seconds of TV audio (Curious George)
- The spectrograms shown are for 44.1kHz audio, but a wearable device would not be able to constant record at 44.1 kHz

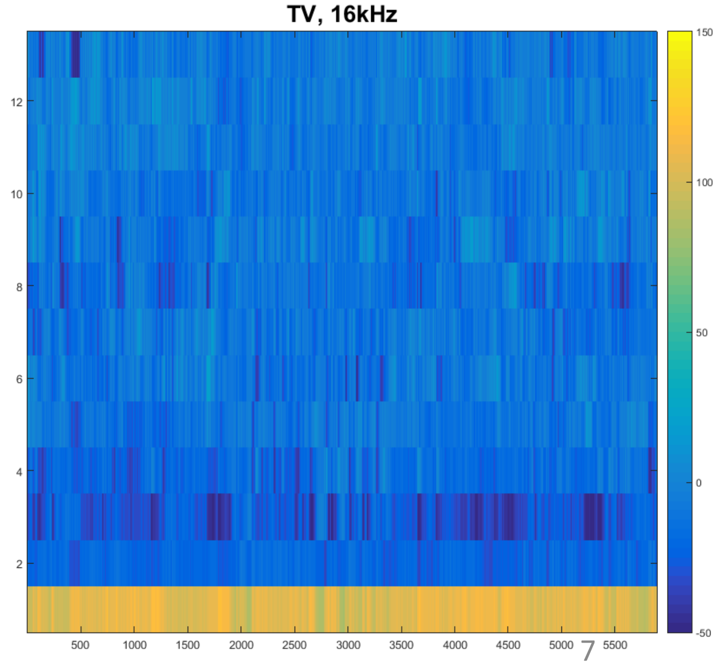




TV Detection – MFCC’s

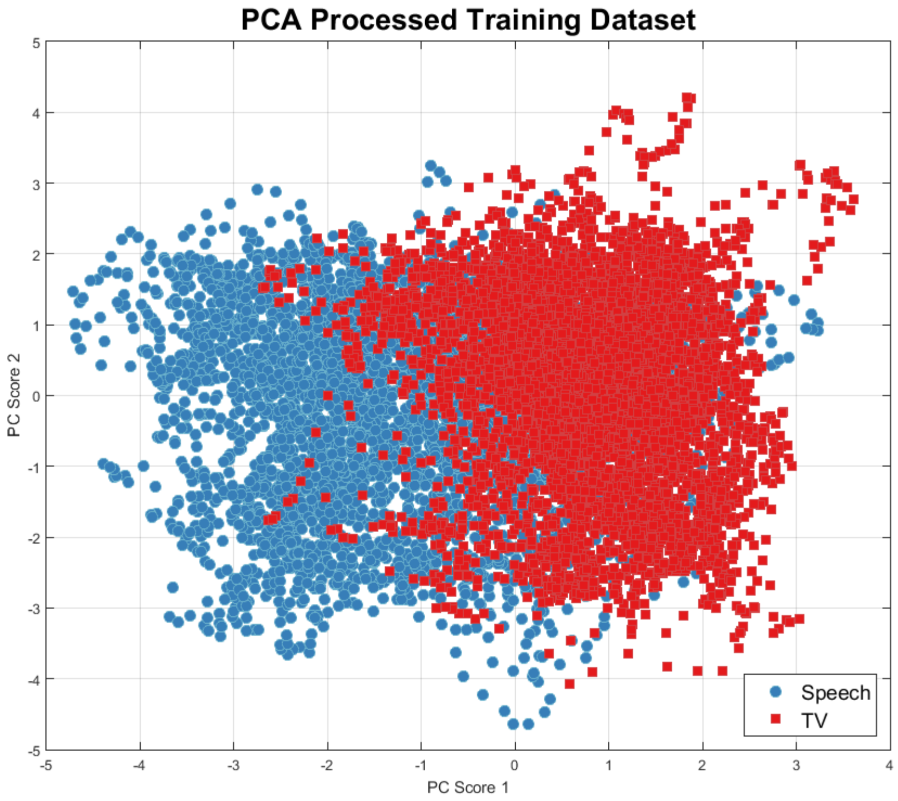
- Mel Frequency Cepstral Coefficients model the way the ear hears sounds
    - More distinction at higher frequencies
- Spectrograms of MFCC’s of household speech (top) and TV (bottom)
    - Data contains ~59 seconds of TV (Curious George) and ~47 seconds of household conversation (father, son, mother, daughter), 16kHz audio
- Visual differences in third and sixth band





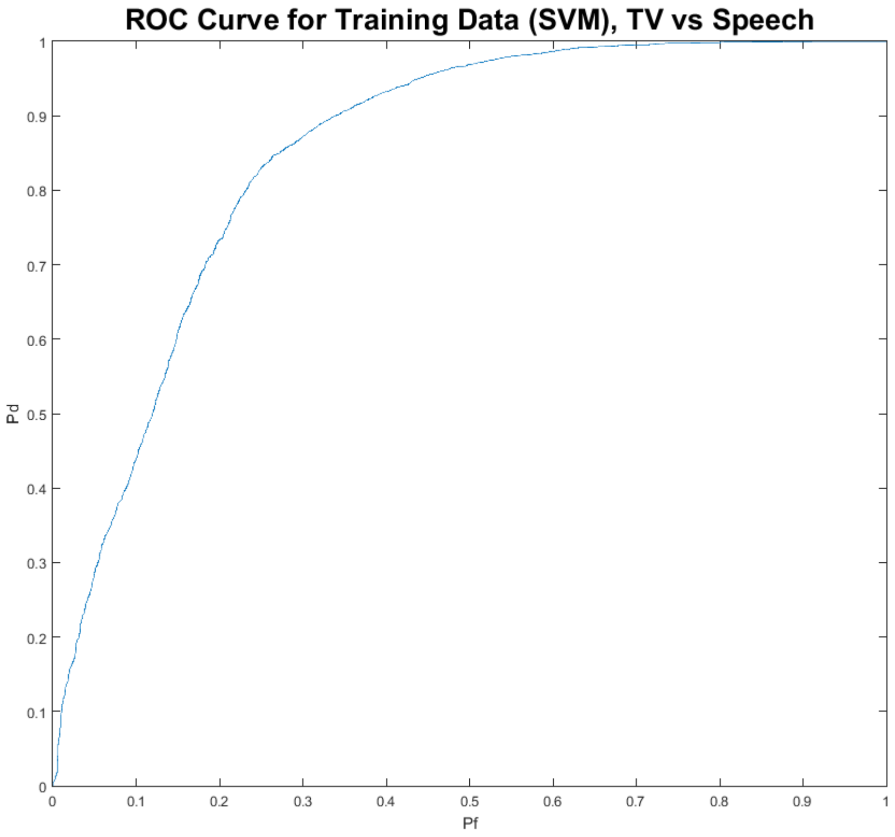
TV Detection - Clustering

- After running PCA on 16kHz audio, there was a large region of overlap between TV and speech
  - Perhaps silence, present in both TV audio and speech audio (ie. Pauses between words)
  - Possibly due to narrator/speech from TV characters



TV Detection – Classification

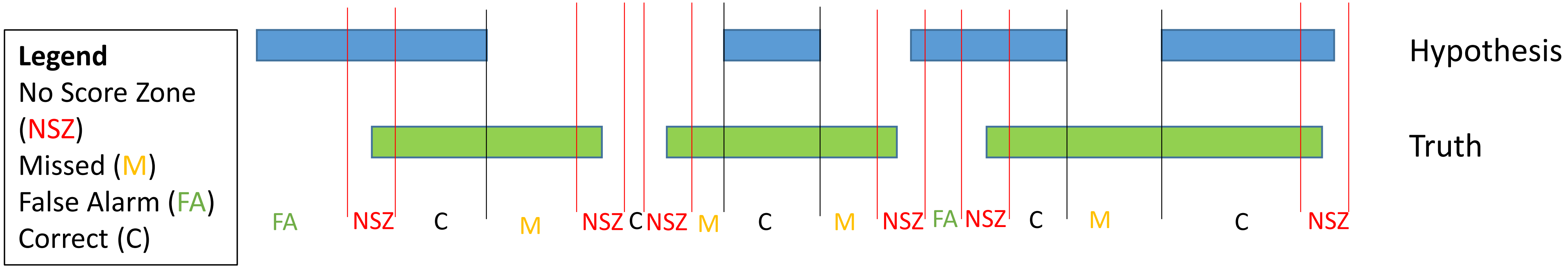
- Training and testing Linear SVM on this dataset yielded the following ROC curve (bottom right)
  - Features used: 12 MFCC's + Energy
  - Data used: ~59 seconds of TV and ~47 seconds of household conversation (same as previous)
  - AUC of 0.8506



Speech Activity Detection

- Motivation: Being able to properly segment between speech, silence and audible nonspeech is an important preprocessing step
  - Enables the isolation of speech utterances and allows us to toss out silence
    - Will allow for easier classification between TV and human speech
    - Lays the groundwork for emotion classification
- Difficulties with project: Not enough data, and no proper ground truth!
  - Best “ground truth” was human annotated audio

SAD – Scoring Method



Legend
No Score Zone (NSZ)
Missed (M)
False Alarm (FA)
Correct (C)

Hypothesis

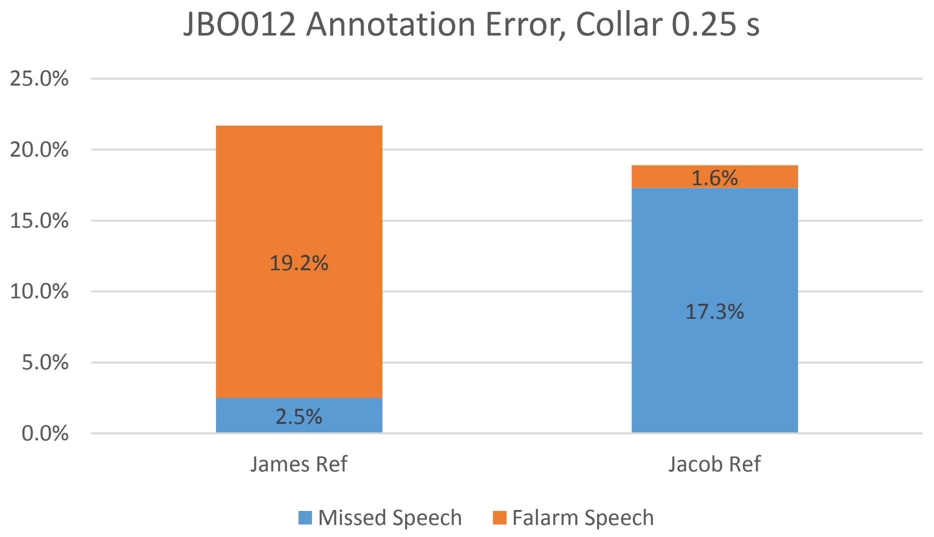
Truth

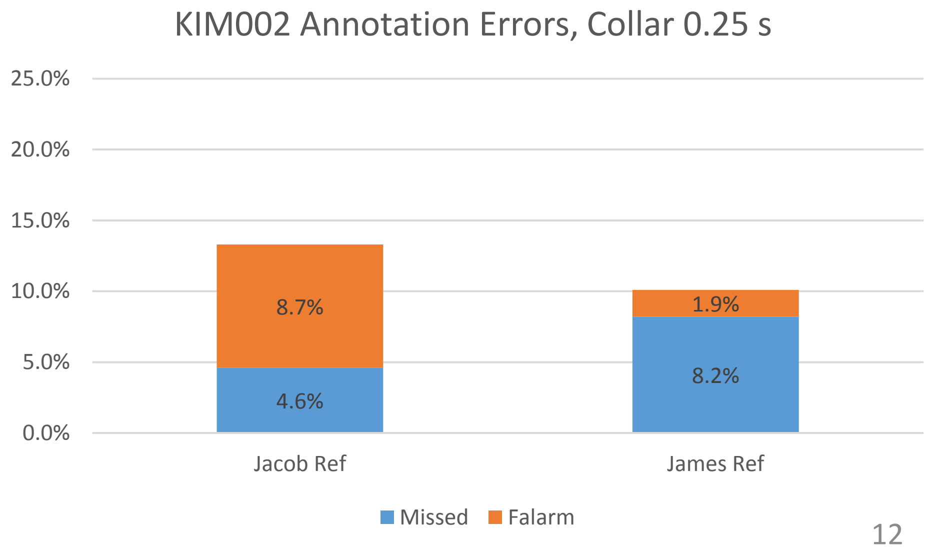
FA NSZ C M NSZ C NSZ M C M NSZ FA NSZ C M C NSZ

- Evaluated SAD performance by calculating Missed Speech time and False Alarm Speech time comparing output transcription and groundtruth (“reference”) transcription
  - Usually expressed as percentage, with error time/total scored time
- Collar – usually 0.25 seconds, refers to the amount of time before or after segment boundaries that are not scored (“no-score zone”)

SAD – Human Annotation Errors

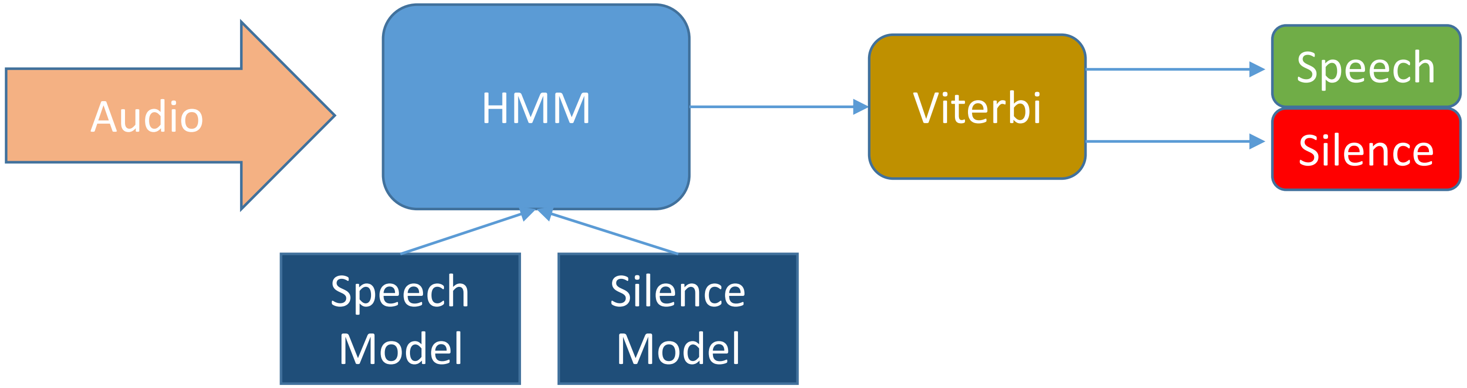
- No real groundtruth – even human annotators can have major discrepancy between them
- Experiment: asked two students, James and Jacob, to transcribe two audio files for speech
  - Bottom left – particularly difficult file with lots of background noise
  - Bottom right – easier file





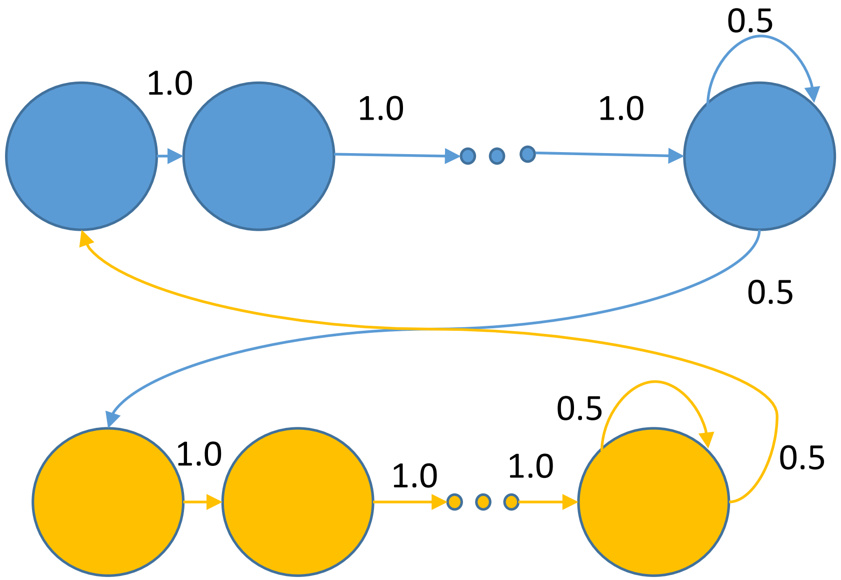
SAD – System Overview

- Hidden Markov Model (HMM) with two classifications – speech, silence, each modeled with Gaussian Mixture Models
  - Used Viterbi algorithm to classify into speech/silence
- Trained speech model on ~4 minutes of individual samples provided by father, toddler daughter, baby daughter and son, silence model trained on silence from these files as well (all audio 16kHz and Wiener-filtered)
  - Features used: 19 MFCC's + zero-crossing rate (ZCR), Deltas and Double Deltas (60 total)
  - 4 component GMM, Diagonal covariance matrix



SAD – System Overview (cont’d)

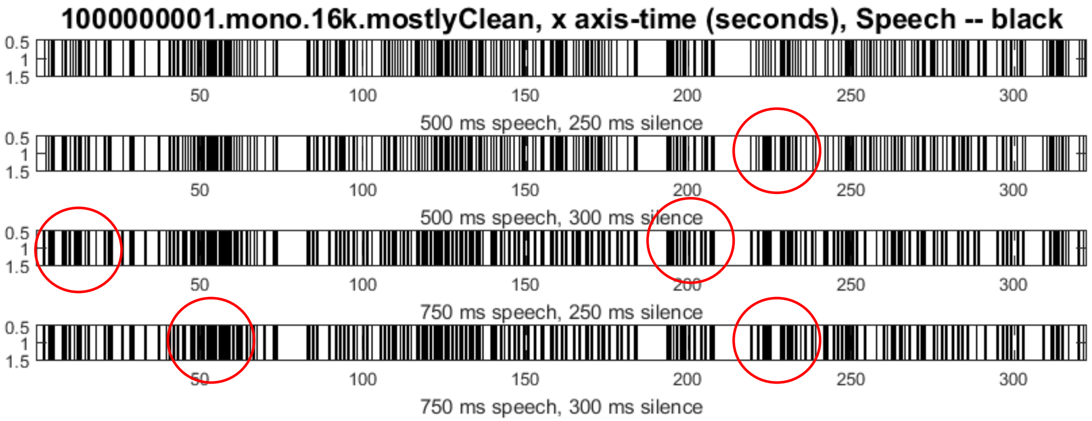
- “Minimum Duration Constraint” – if HMM goes into a state, it must remain in that state for a minimum duration



- Each state is modeled as a sequence of substates with the following transition probabilities
  - 1 from each substate to the next
  - 0.5 from the last substate to itself
  - 0.5 from the last substate to the first substate of the other state
- Each substate modeled with the same GMM

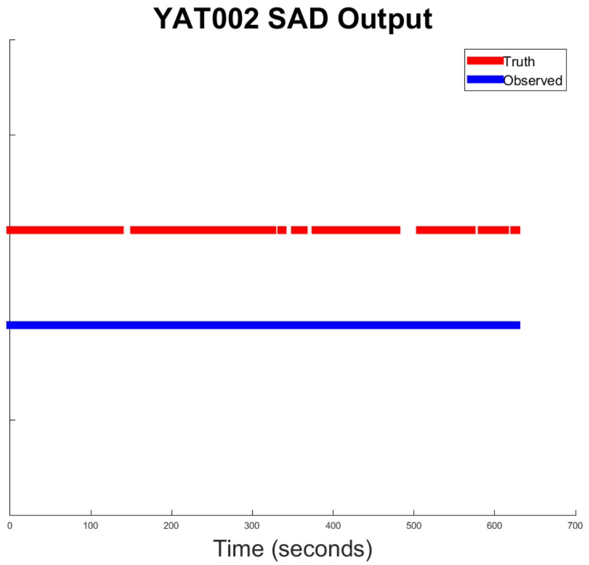
SAD – Minimum Duration Constraint Parameters

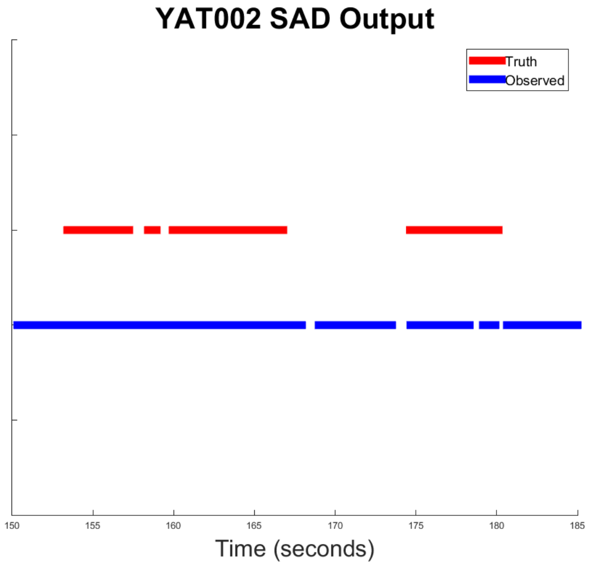
- Initially chose minimum duration of 500ms for speech, 250 ms for silence
  - Noticed output “speech” was too choppy – words often got broken up in the middle
  - Varied constraints between 500/750 ms for speech and 250/300 ms for silence to see the effects
    - Speech is shaded in black
- Data used: ~5 minutes of household speech, 16 kHz

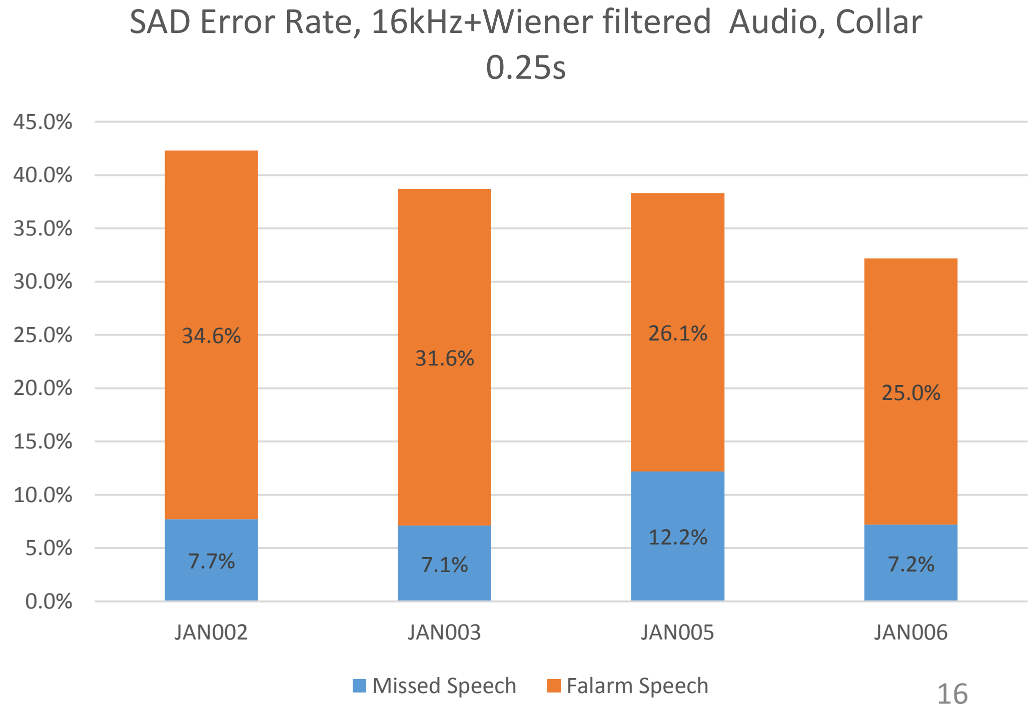


SAD - Performance

- Error rate of **43.4%** (11.5% Missed Speech, 31.9% Falarm speech) on a particular file (YAT002_16k_nr.wav)
  - However, did not sound bad empirically
  - Audio file had slight background noise
- To the right, visualizing the errors between observed and groundtruth in MATLAB
  - Total (~626 seconds) on top left, zoomed in on top right
- 2/3 of error rate comes from false alarm


YAT002 SAD Output


YAT002 SAD Output

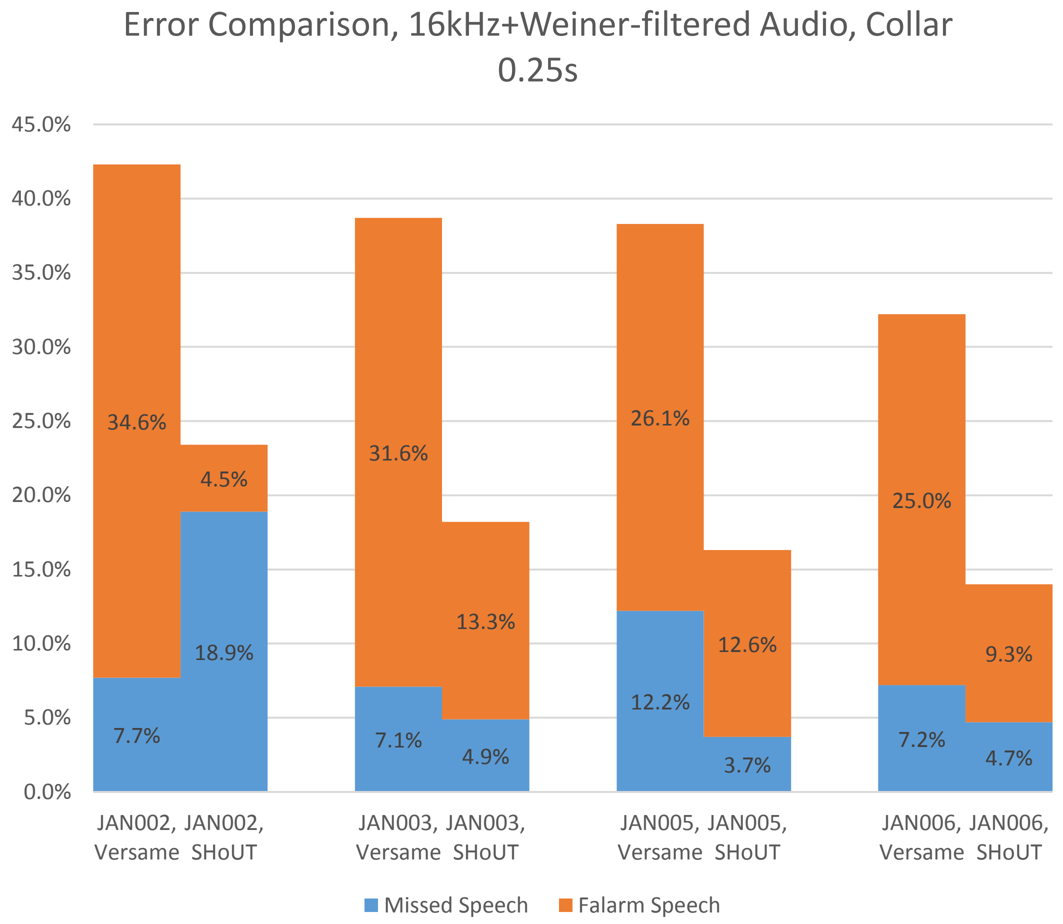

SAD Error Rate, 16kHz+Wiener filtered Audio, Collar 0.25s

SAD – SHoUT System Description

- SHoUT is a speech activity detection toolkit that can also perform diarization
  - Also a GMM-HMM setup
  - SAD classifies into Speech, Silence, Sound
- Features: 12 MFCCs + ZCR, deltas, double deltas (39 in total)
- Default speech models trained on ~3.5 hours of audio, 200 adult male/200 adult female from broadcast news corpus
- Implements adaptive models – The audio is initially segmented using default models, and then the system uses the initial segmentation to create new models and re-segment
  - Sound and Speech models are compared, and if the models are similar, Sound is discarded and audio is re-segmented using Speech and Silence models

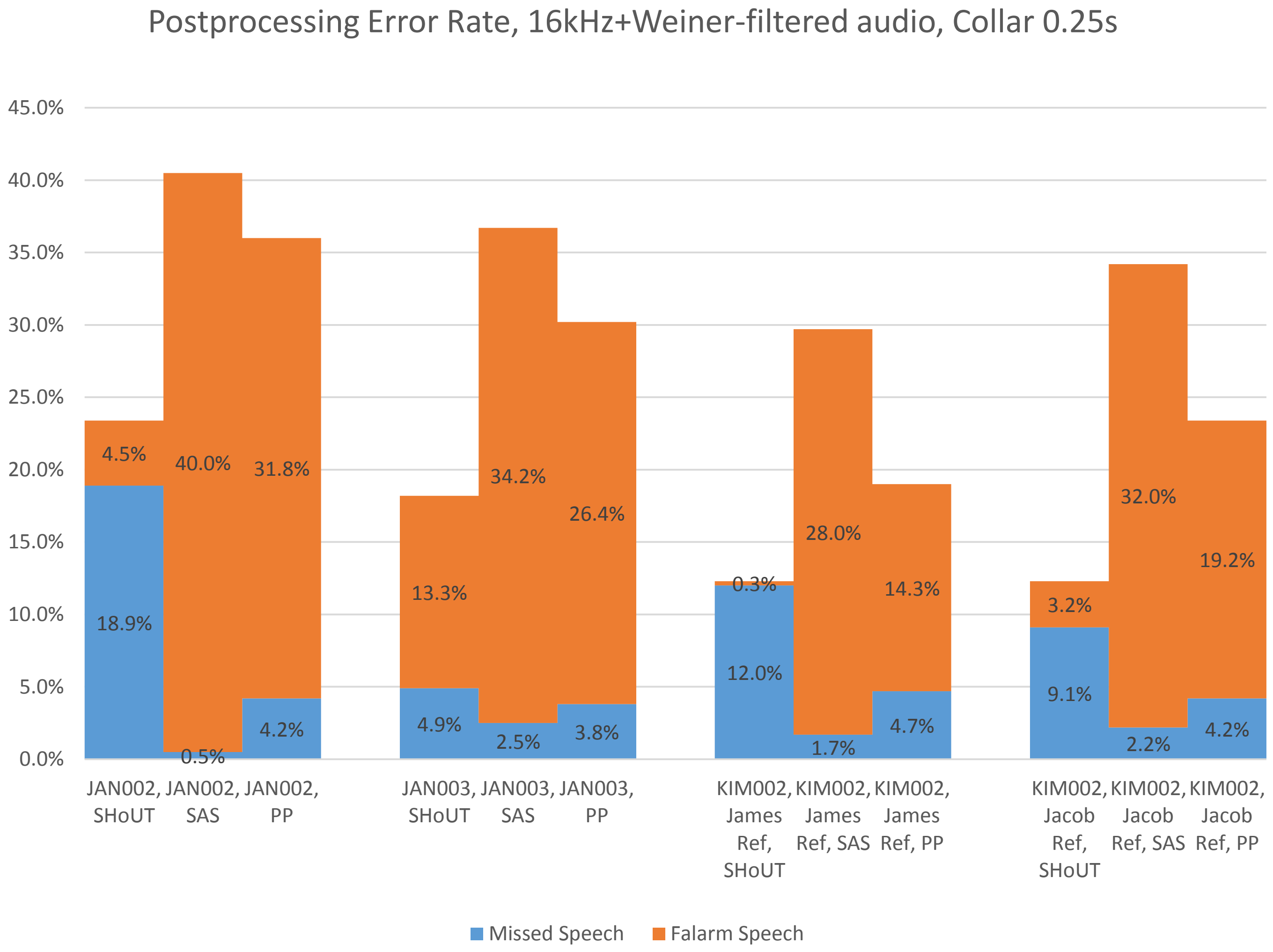
SAD – SHoUT Performance

- Vastly reduced error rate, although 15%-20% still not preferable
- In JAN002 and 003, it's likely that much of the "missed speech" was classified into the "Sound" category (SHoUT)
  - Audio indicates this is true
  - Oftentimes some speech is misclassified as sound, especially child speech



SAD – Postprocessing “Sound” (SHoUT)

- Sound as speech (SAS) – take everything classified as Sound and label it as Speech
- Postprocessed (PP) – take everything classified as Sound, run it through original GMM-HMM SAD, then label the “speech” output as Speech
- False alarm time – what is it exactly?



Future directions

- TV:
  - Extract longer-term prosodic features
  - If we run unsupervised clustering on TV audio, will we see music and character speech in separate clusters?

- SAD:
  - What’s behind the massive false-alarm rate?
    - Is it audible nonspeech or silence?
  - Train Sound models for Versame SAD
  - Add children to Shout default speech models