Emotional Speech Recognition

Kisang Pak
E6820: Speech & Audio Processing & Recognition
Professor Dan Ellis
What is emotional speech recognition?

- A technique which can recognize emotions in a speech
- Common emotions: anxiety, boredom, dissatisfaction, dominance, depression, disgust, frustrated, fear, happiness, indifference, irony, joy, neutral, panic, prohibition, surprise, sadness, stress, shyness, shock, tiredness, task load stress, worry
- A system usually recognizes 3-5 emotions
Common technique

Input Signal
- Speech samples

Feature Extraction
- Pitch
- Speech Energy
- Formant Frequencies

Classification
- HMM
- Binary decision tree
- ANN (Artificial Neural Networks)

Result
- Based on the results of the classification
18. - 22. Maximum, minimum, mean, median, interquartile range
23. Pitch existence in the utterance expressed in percentage (0-100%)
24. - 27. Maximum, mean, median, interquartile range of duration of plateaux at minima
28. - 30. Mean, median, interquartile range of values of plateaux at minima
31. - 35. Maximum, mean, median, interquartile range, upper limit (90%) of duration of plateaux at maxima
36. - 38. Mean, median, interquartile range of values of plateaux at maxima
39. - 42. Maximum, mean, median, interquartile range of durations of rising slopes
43. - 45. Mean, median, interquartile range of values of rising slopes
46. - 49. Maximum, mean, median, interquartile range of durations of falling slopes
50. - 52. Mean, median, interquartile range of values of falling slopes
53. Number of inflections in F0 contour
Feature Extractions: Energy

54. Maximum, minimum, mean, median, interquartile range
55. Maximum, mean, median, interquartile range of durations of plateaux at minima
56. Mean, median, interquartile range of values of plateaux at minima
57. Maximum, mean, median, interquartile range, upper limit (90%) of duration of plateaux at maxima
58. Mean, median, interquartile range of values of plateaux at maxima
59. Maximum, mean, median, interquartile range of durations of rising slopes
60. Mean, median, interquartile range of values of rising slopes
61. Maximum, mean, median, interquartile range of durations of falling slopes
62. Mean, median, interquartile range of values of falling slopes

\[ E_s(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} |f_s(n; m)|^2 \]

\( f_s(n; m) = s(n)w(m-n) \)

\( s(n) \): speech signal, \( w(m-n) \): window (i.e. hamming) of length \( N_w \)
Feature Extractions: Formants

Neutral
Formant 1 Frequency 355.6
Formant 2 Frequency 1400.4
Formant 3 Frequency 2588.6
Formant 4 Frequency 3505.9
Formant 5 Frequency 4653.3
Formant 6 Frequency 5338.3
Formant 7 Frequency 6279.6
Formant 8 Frequency 7000.2

Anger
Formant 1 Frequency 562.9
Formant 2 Frequency 743.9
Formant 3 Frequency 1458.5
Formant 4 Frequency 2882.6
Formant 5 Frequency 3731.8
Formant 6 Frequency 4196.8
Formant 7 Frequency 5381.2
Formant 8 Frequency 6419.5
Formant 9 Frequency 7215.3

Joy
Formant 1 Frequency 412.1
Formant 2 Frequency 674.6
Formant 3 Frequency 1567.9
Formant 4 Frequency 2653.4
Formant 5 Frequency 3661.1
Formant 6 Frequency 4372.9
Formant 7 Frequency 5489.9
Formant 8 Frequency 6422.8
Formant 9 Frequency 7038.4
Example)
Initial State: Anger
Observation
Fo = 250 Hz
Gender: Male

Emotion Classification: HMM
My technique: Overview

- Emotions: Sadness, Neutral, Anger, Happy, (Frustrated), (Surprised)
- Language: English
- Features to be used: Pitch, Energy, Formants,
- Classification: Modified Binary Decision
  *(why not HMM???)*
- Goal: 50% Correction Rate (independent, gender unknown)
My technique: Overview

- **Non-hyper (Sadness, Neutral)**
  - Pitch Track
  - Sadness
  - Neutral

- **Hyper (Anger, Frustrated Happy, Surprise)**
  - Energy (i.e. rising slopes)
  - f0
  - Negative (Anger, Frustrated)
    - Frequency
    - Local maximas
  - Positive (Happy, Surprise)
    - Happy
    - Surprise

- **Speech**
  - Women
  - Men
  - p.d.f. pitch contour
    - f0

- **Men**
  - Energy (i.e. rising slopes)
    - f0
My technique: Example (Gender Differentiation)

1. Fundamental Frequencies
   (Time-Domain Analysis using autocorrelation)

2. PDFs of mean value of pitch contour
My technique: Example
(Non-hyper vs. Hyper)

Neutral

Angry
My technique: Example (Neutral vs. Sadness)
Challenges (or Opportunities)

- Database (Main source: movies, TVs)
  
  *Enough angry speeches, insufficient happy speeches in Hollywood movies*
  
  *TV sitcoms might be good (i.e. Friends, Seinfeld)*

- No standard methodologies

- Characterize emotions according to pitch, energy, formants, etc

- Input is very subjective
Final Product

1. MatLAB

2. Stand Alone Application in LabVIEW
Bonus works “Dream big!”

- Emotional Speech Synthesize
  - neutral
  - angry
  - joy