Emotional Speech Recognition

Kisang Pak

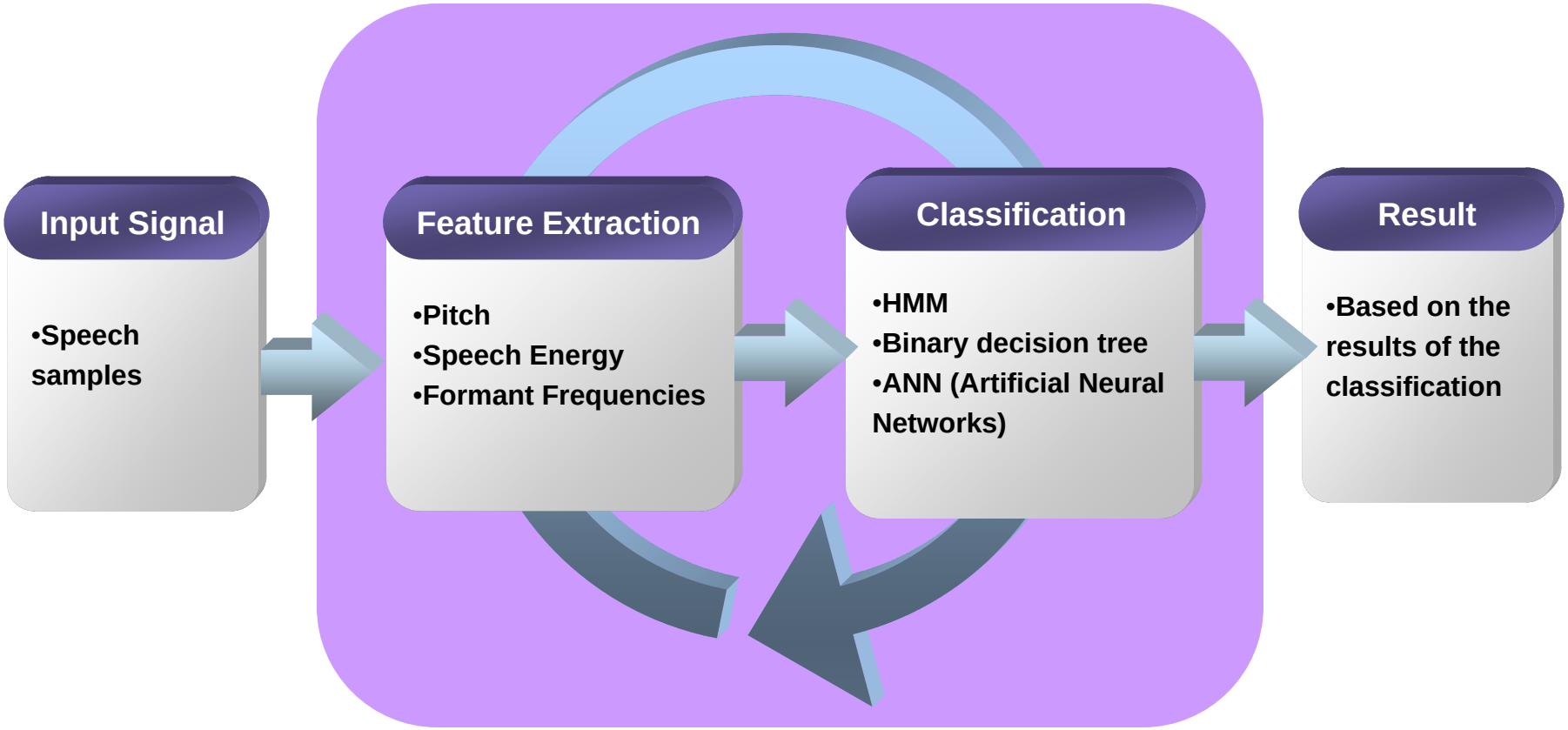
E6820: Speech & Audio Processing & Recognition

Professor Dan Ellis

What is emotional speech recognition?

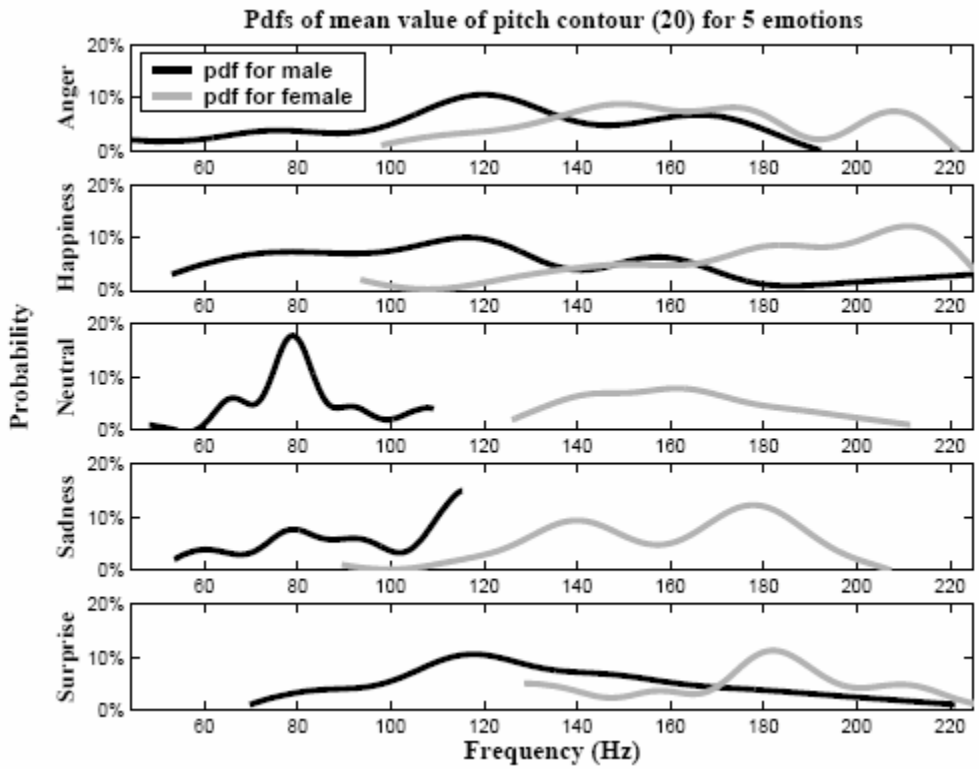
- A technique which can recognize emotions in a speech
- Common emotions: anxiety, boredom, dissatisfaction, dominance, depression, disgust, frustrated, fear, happiness, indifference, irony, joy, neutral, panic, prohibition, surprise, sadness, stress, shyness, shock, tiredness, task load stress, worry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

Common technique



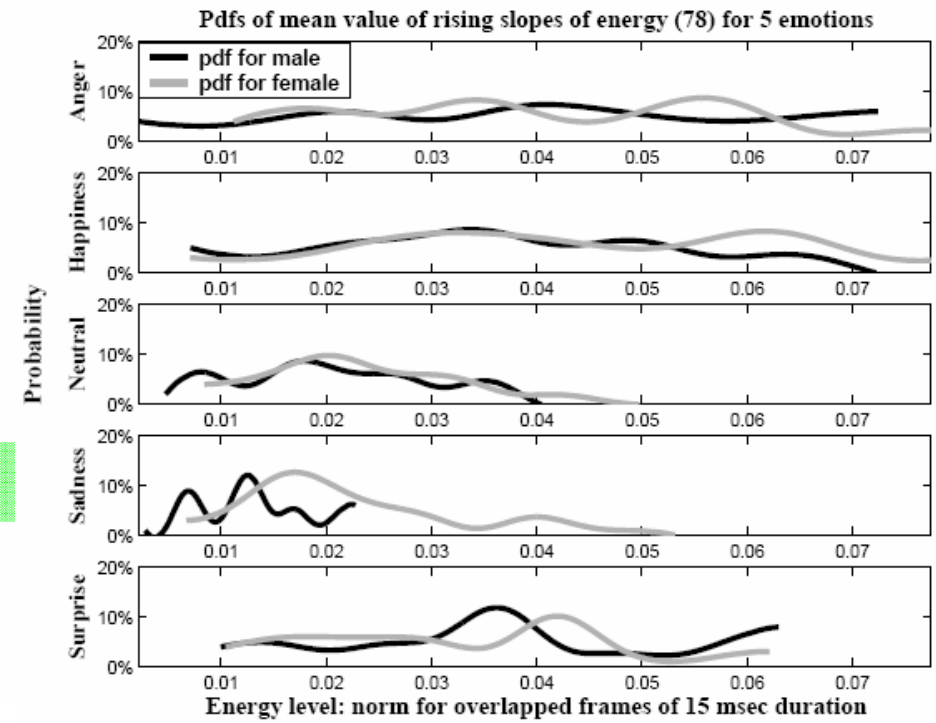
Feature Extractions: Pitch

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

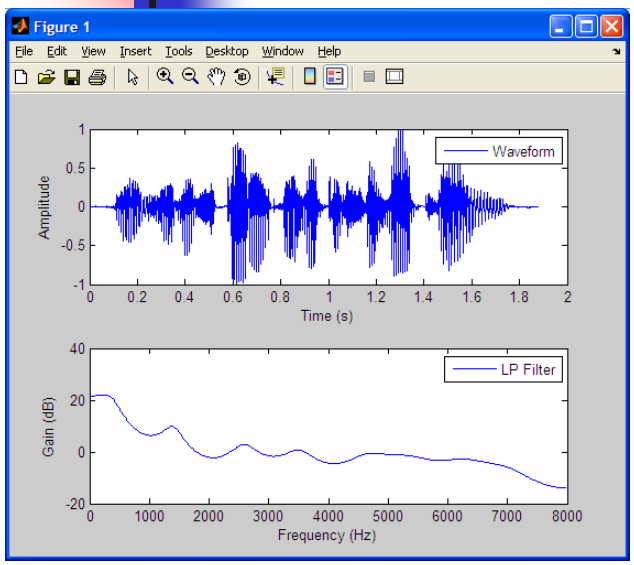
54. - 58. Maximum, minimum, mean, median, interquartile range
59. - 62. Maximum, mean, median, interquartile range of durations of plateaux at minima
63. - 65. Mean, median, interquartile range of values of plateaux at minima
66. - 70. Maximum, mean, median, interquartile range, upper limit (90%) of duration of plateaux at maxima
71. - 73. Mean, median, interquartile range of values of plateaux at maxima
74. - 77. Maximum, mean, median, interquartile range of durations of rising slopes
78. - 80. Mean, median, interquartile range of values of rising slopes
81. - 84. Maximum, mean, median, interquartile range of durations of falling slopes
85. - 87. Mean, median, interquartile range of values of falling slopes



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

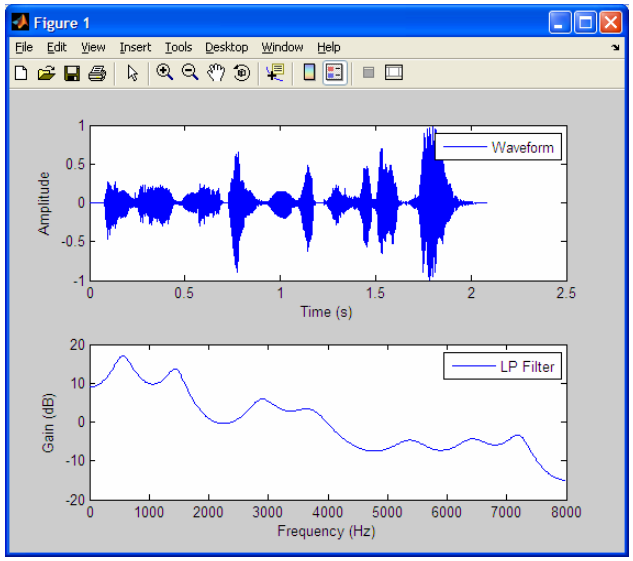
, fs(n:m)=s(n)w(m-n)
s(n): speech signal, w(m-n): window (i.e. hamming) of length Nw

Feature Extractions: Formants



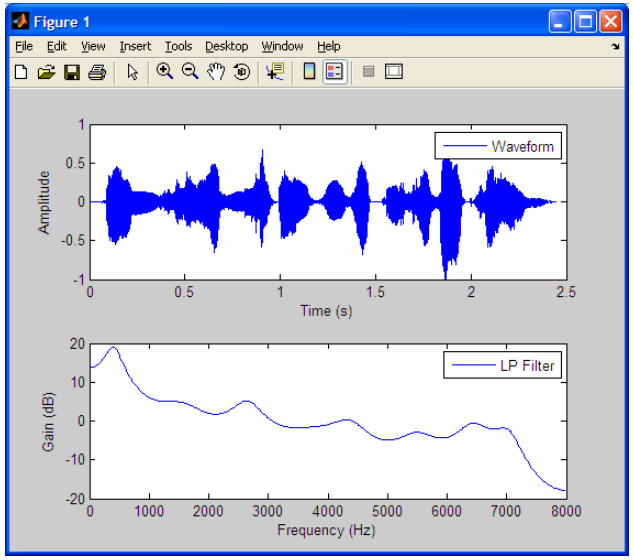
Neutral

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

Emotion Classification: HMM

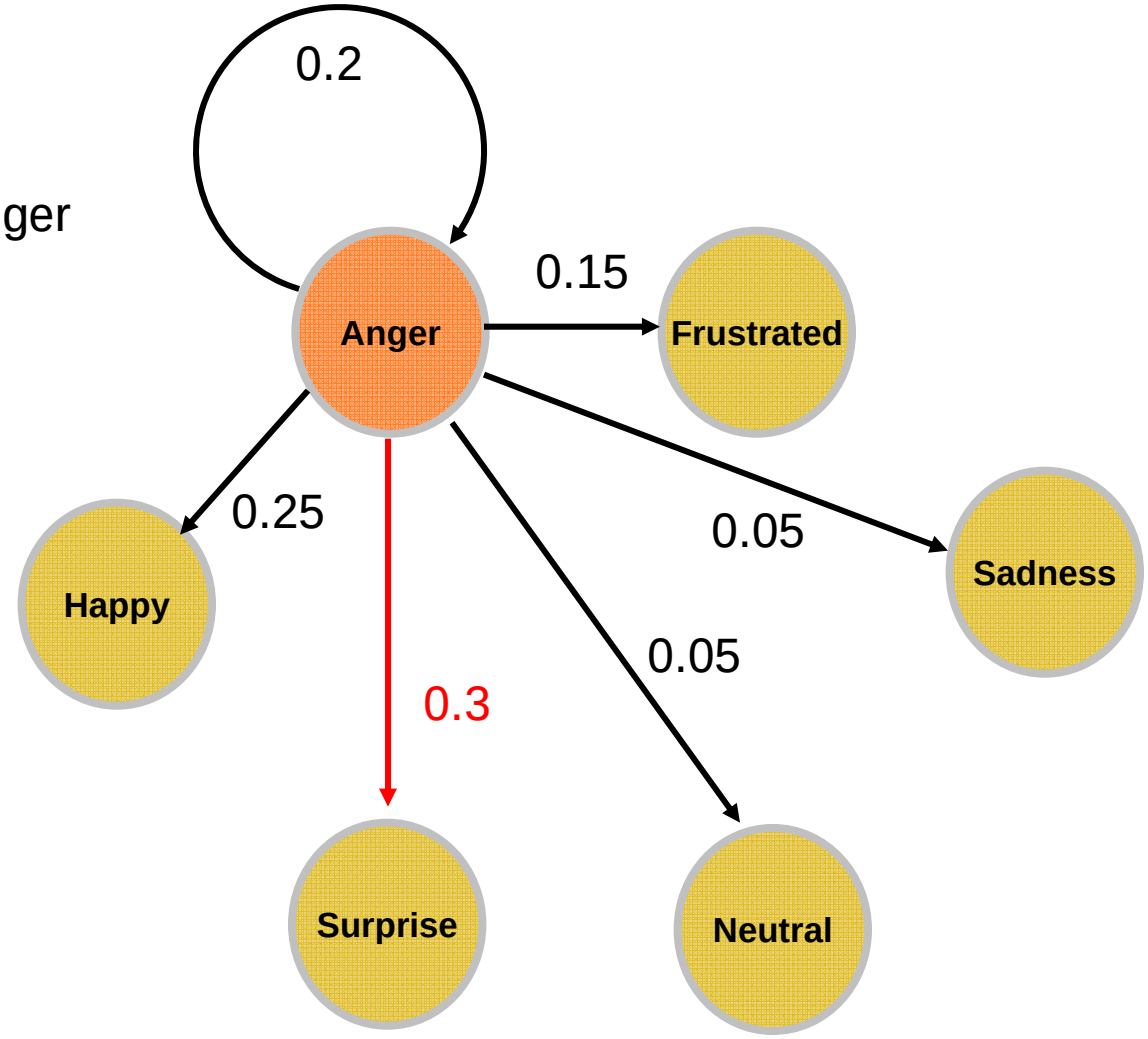
Example)

Initial State: Anger

<u>Observation</u>

Fo= 250 Hz

Gender: Male

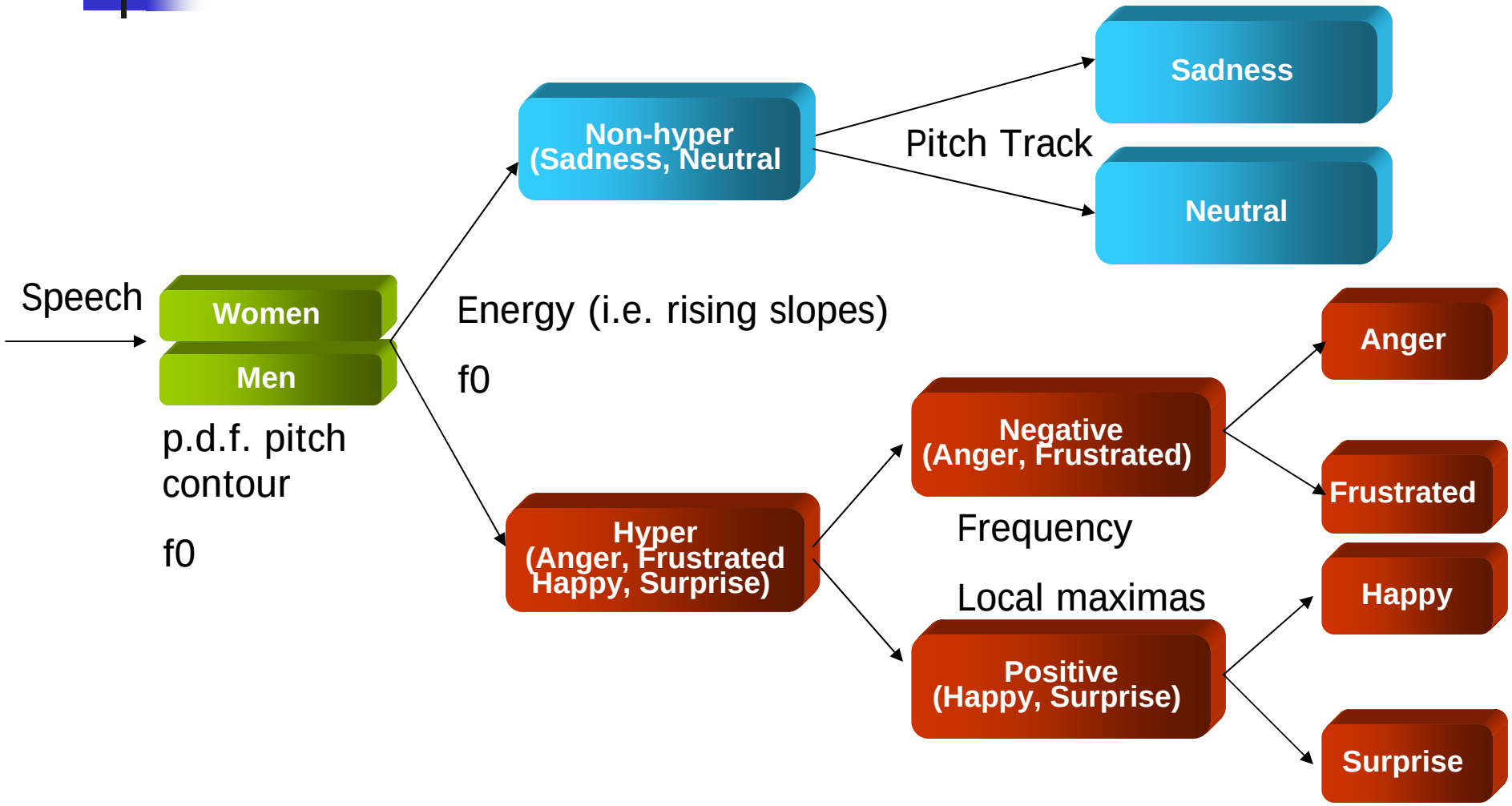


My technique: Overview

- 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



My technique: Example (Gender Differentiation)

1. Fundamental Frequencies

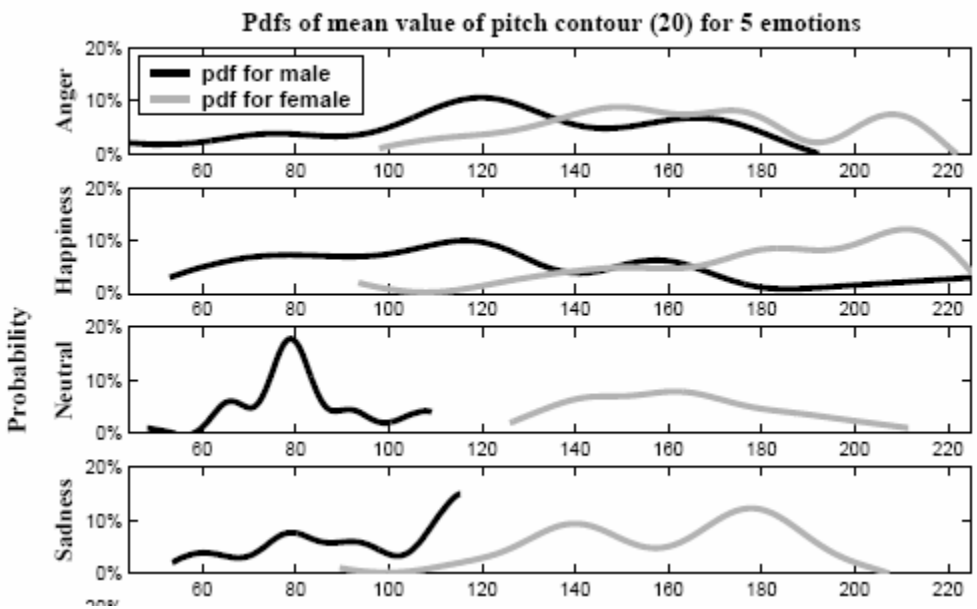
   (Time-Domain Analysis using autocorrelation)

	Male	Female
Sad	104	176
Neutral	110	202
Happy	281	410

	Male	Female
1	127	186
2	101	182
3	119	207

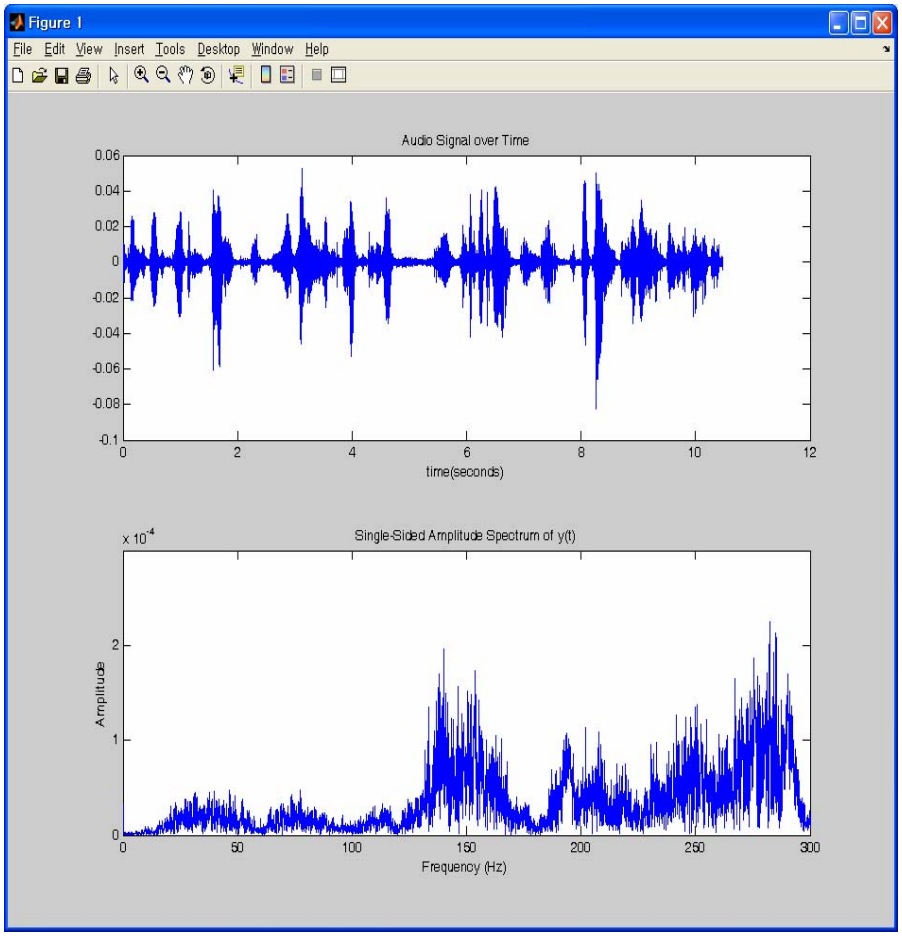
2. PDFs of mean value

of pitch contour

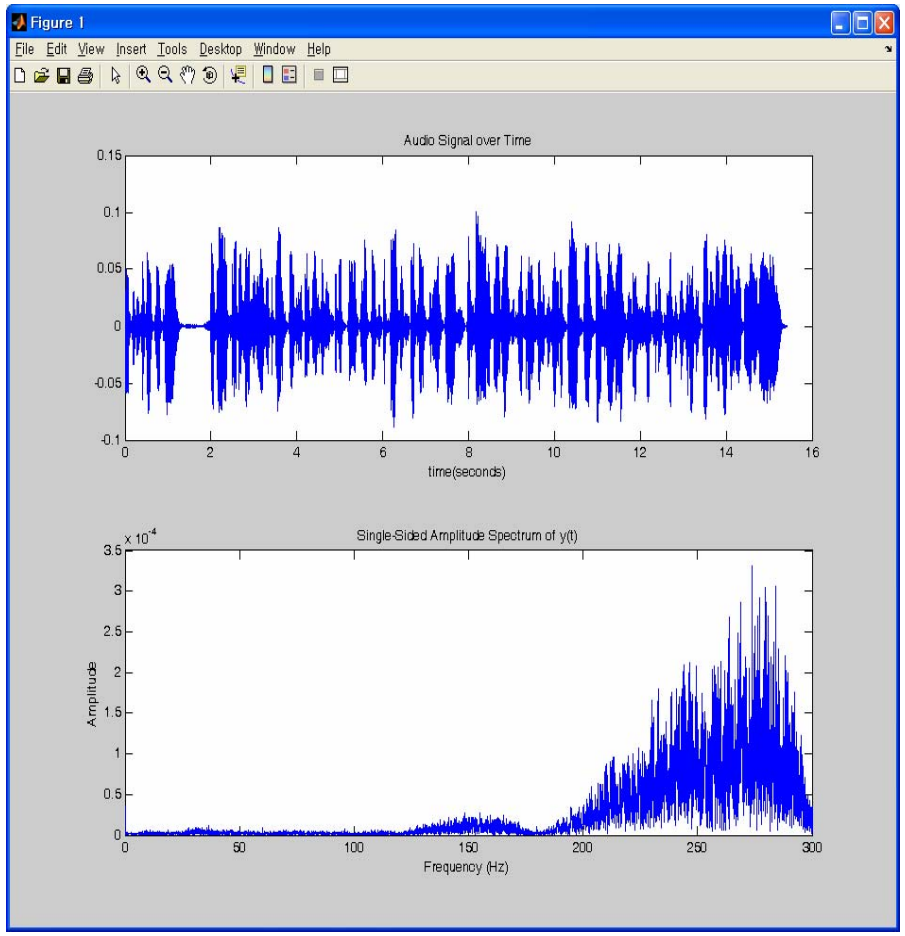


My technique: Example (Non-hyper vs. Hyper)

Neutral

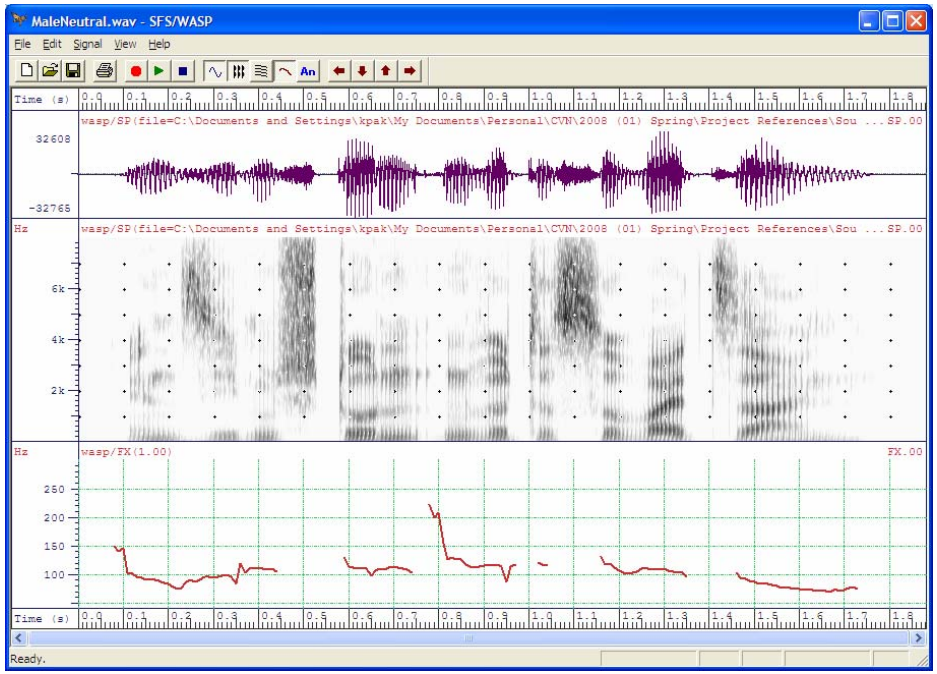


Angry

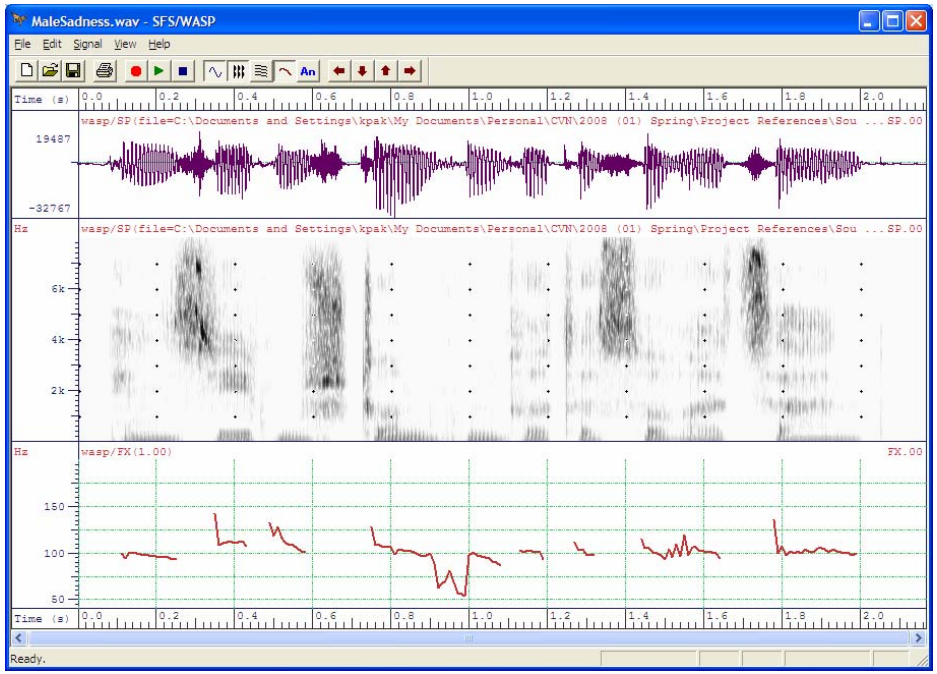


My technique: Example (Neutral vs. Sadness)

Neutral



Sadness



Challenges (or Opportunities)

- Database (Main source: movies, TVs)

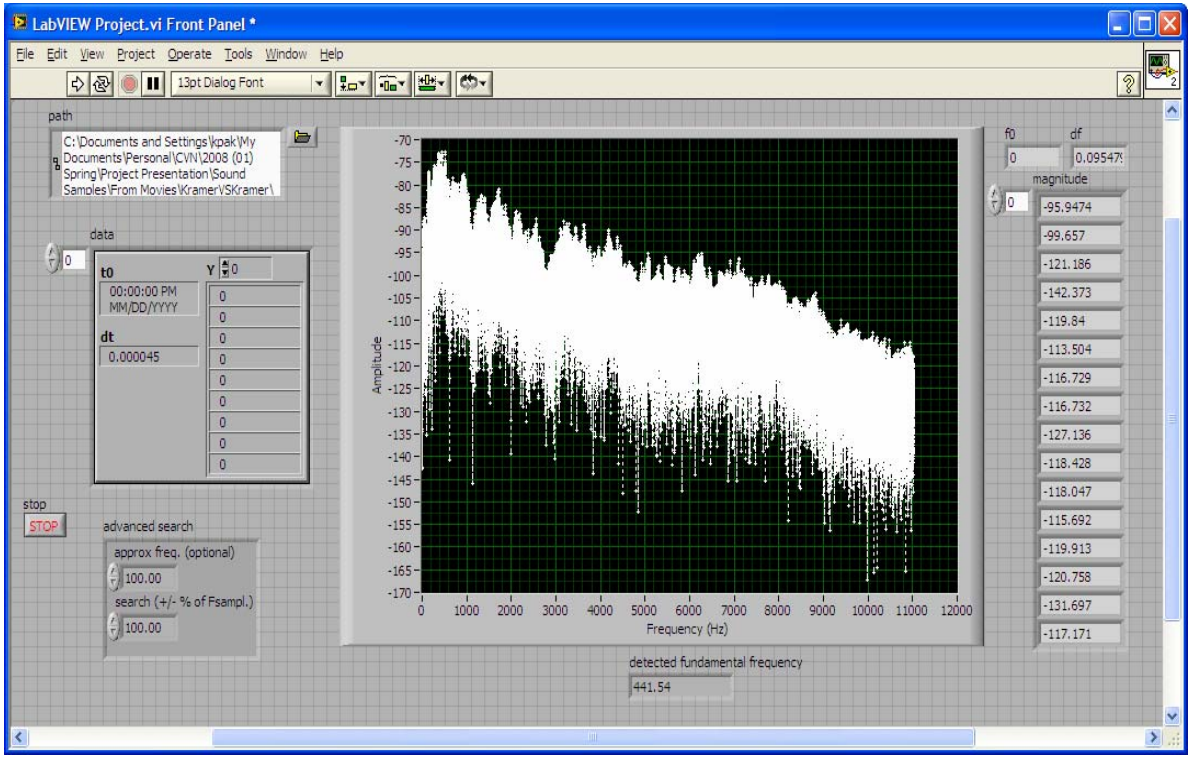
  Enough angry speeches, insufficient happy speeches in Hollywood movies

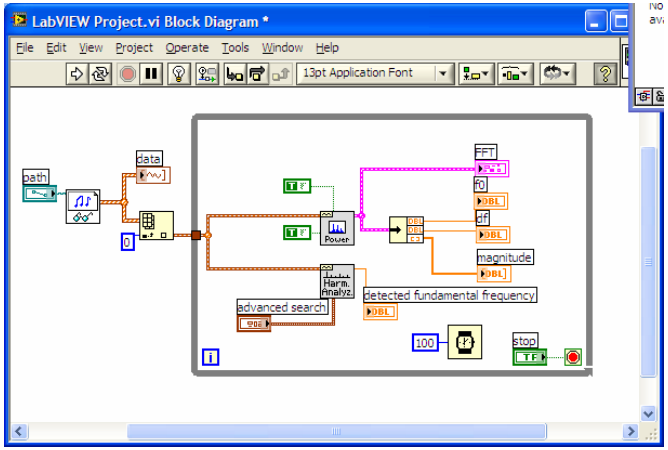
  TV sitcoms might be good (i.e. Friends, Seinfeld)
- No standard methodologies
- Characterize emotions according to pitch, energy, formants, etc
- Input is very subjective

Final Product

1. MatLAB

2. Stand Alone Application in LabVIEW





Bonus works “Dream big!”

- Emotional Speech Synthesize

  - neutral
  - angry
  - joy

Discussions