LING 83600/CSCI 76000
Prof. Levitan

HW4: Emotional Speech Analysis

In this homework you will practice extracting features from speech and using them to analyze emotional speech samples.

Deadline: November 21, 11:59pm

Submission instructions: Submit a document with your answers to the HW questions (.pdf or .doc) as well as your code (.ipynb or .py). Your code must run with python3.

Data

You are provided with samples from the MSP-Podcast corpus.
MSP-Podcast corpus description
MSP-Podcast samples

Task

Use Praat/Parselmouth to extract the following features:
Min, max, mean f0
Min, max, mean intensity
Jitter
Shimmer
Speaking rate

Fill out the table below with the results.

Resources

Praat
Parselmouth: a Python library for the Praat software

Feature extraction notes

- For pitch extraction, set pitch floor to 75Hz, and pitch ceiling to 600Hz.
- For jitter, extract local jitter only, and set period floor to 0.0001s, period ceiling to 0.02s, and maximum period factor to 1.3.
- For shimmer, extract local shimmer only, and set period floor to 0.0001s, period ceiling to 0.02s, maximum period factor to 1.3, and maximum amplitude factor to 1.6.
- To calculate HNR (harmonics-to-noise ratio), extract harmonicity (cc) first. Set time step to 0.01, minimum pitch to 75Hz, silence threshold to 0.1, and number of periods per window to 1.0.
- Speaking rate can be approximated with #words/duration.

Feature extraction results (5 pts)

Emotion	Min f0	Max f0	Mean f0	Min int.	Max int.	Mean int.	Jitter	Shimmer	HNR	Speaking rate
Happy	76.91	314.49	206.92	39.02	77.78	65.58	0.016	0.061	14.61	4.09
Angry	92.93	510.24	242	27.66	80.20	62.27	0.027	0.0982	7.78	3.23
Sad	87.13	268.53	212.87	26.92	81.45	62.87	0.014	0.059	15.56	2.31
Afraid	73.82	439.59	150.18	6.05	85.36	61.14	0.029	0.156	8.17	2.92
Surprised	83.80	434.82	240.95	40.92	82.07	65.014	0.03	0.091	12.49	3.56
Disgusted	84.36	254	162.45	36.8	78.96	69.51	0.019	0.078	10.87	4.46
Neutral	82.98	191.53	130.56	33.93	79.78	68.11	0.027	0.07	11.17	2.92

Questions

1. Choose 3 emotion categories and compare the characteristics of those speech samples using the results table above. Which features are most useful for characterizing each of the emotions? What are the limitations of this analysis? (3 pts)

Frequency: Disgusted, surprised, and neutral all have nearly the same min frequency but as would be expected, surprise has the highest max frequency, since people's voices can get very high pitched when they are excited and people get excited when surprised. Disgust can sometimes have a level of excitement but when someone is neutral, their pitch is not expected to vary much and they do not show a lot of fluctuation and excitement,

Intensity: It makes sense that surprise has the highest min intensity since I would expect people to raise their voice more when surprised then when disgusted and neutral but it makes sense that disgusted has a lower max intensity because people often talk quietly when disgusted. But I was surprised that surprise has the lowest mean intensity.

Jitter and Shimmer: while surprised and neutral have similar jitter, disgusted is a bit lower which is surprising. I would expect neutral to have the lowest jitter since jitter represents stability of vocal cord vibrations and measures frequency changes. Someone who is surprised is expected to have slightly higher jitter. Surprised has the highest shimmer, then disgust, then neutral. This

makes sense to me since someone who is surprised may have a lot of fluctuation in their pitch, someone who is neutral should not show fluctuation, and disgusted would be expected to lie somewhere in between.

HNR and Speaking Rate: I believe HNR is more a reflection of the speaker than of their expression but I am surprised to note that they seem to all have noticeable hoarseness. As is to be expected, neutral has the lowest speaking rate but I had assumed surprised would have a higher speaking rate although some people do not talk very quickly when surprised and some people talk quicker when disgusted so it could go either way

Intensity and frequency were useful in analyzing the audio files, specifically the surprised audio file since they depict how excited the speaker is by the pitch and amplitude of their voice. They were also useful in analyzing the disgusted audio file since it shows appropriate fluctuation. Jitter shimmer and HNR would be useful in detecting the speaker's health, and possibly in detecting how calm they are and if they are possibly being deceptive. Also with speaking rate, often people speak quicker when nervous, this was helpful in analyzing neutral since often when people are neutral and calm they speak at an average rate.

Some limitations of these methods is that often the way different people talk varies, some people have a higher pitch, some people speak louder, and some people speak quicker. Often emotions that have similar arousal and valence can be difficult to detect the difference between those emotions.

2. Do you think the MSP-podcast corpus is useful for real-world emotion classification? Why or why not? (1 pt)

It is useful because it is real people talking about real things so you will have a huge variety, it is not useful because the people speaking know they are being recorded so they may either be guarded or over dramatic depending on how they want to appear to their audience.

3. What other features might be useful for automatically recognizing emotions? How would you measure them? (1 pt)

It would be interesting to extract a change in a person's voice. To detect this you would need many samples of that specific person's voice. It would also be interesting to detect sarcasm, this would probability be detected using speaking rate, intensity and pitch.