CS 224S HW 4

Veni Johanna (veni), Tim Sakhuja (tsakhuja), Gina Pai (gpai)

The provided example of hot anger, as one might expect, is marked by an elevated intensity. Both the mean, maximum values, and range of intensity in the recording (73.3, 82.6, and 54.138 dB respectively) are the highest of all five recordings. Angry speech also appears to invoke an extreme in pitch: the pitch ceiling, pitch floor, and mean pitch of the angry speech recording are also the highest of all samples. Though the aggregate pitch metrics of the other three recordings are rather clustered, one would expect that features relevant to pitch (such as F0) would be useful for classifying angry speech. Angry speech is typically marked by yelling and other strained vocalization. Because such vocalizations alter the shape of the vocal tract and therefore the filter applied to speech, one would also expect that cepstral features such as MFCC would be relevant to angry speech classification. Additionally, the maximal pitch occurs much earlier in than in the angry speech clip than in the others, likely because hot angry speech is characterized by punctuated words and syllables, and suggests that the position of the maximal pitch may be a good feature.

Sad speech lies at the other end of the intensity spectrum. The sad speech recording has the lowest mean and maximum intensity values of all recordings (63.4 and 69.6 dB) in addition to the lowest range, suggesting that features related to intensity (max and mean RMSEnergy, for example) would be useful for classifying sad speech. In addition, the pitch variance of sad speech is visibly smaller than those of the other emotional speech clips, suggesting pitch standard deviation and features as good classifiers.

The examples of neutral and happy speech have very similar intensity profiles, but differ in pitch profile. The happy speech example has slightly elevated mean pitch and pitch ceiling (137.2 and 172.9 Hz vs. 104.6 and 142.4 Hz). Audibly, happy speech carries a more varied tone, suggesting that features that capture variance in tone (like variation in F0 and zero crossing rate, which approximates frequency) would be relevant to distinguish the two types of speech.

The pitch and intensity profiles of the despairing speech example overlap with most of the other examples, making the task of classifying them based on pitch and intensity features alone difficult. Audibly, the example of despairing speech is marked by pronounced breathiness and listlessness which might cause changes in shape of the vocal tract, suggesting that MFCC features and zero-crossing features (which can be used to differentiate between voiced and unvoiced speech) might be relevant.

Our best accuracy is **68.4848%**, which is reached by using LibSVM's C-SVC (multi-class classification), polynomial kernel with cost 5.0 and gamma 0.03125, using features:

- 1: max RMS
- 2: min RMS
- 3: range RMS
- 6: arithmetic mean of RMS
- 10: stddev of RMS
- 13:157: 144 features derived from the 12 MFCC
- 158: min zero-crossing rate of time

- 159: range zero-crossing rate of time
- 162: arithmetic mean of zero-crossing rate of time
- 171: range of voicing probability
- 172: absolute position of the maximum value of voicing probability
- 181: max F0
- 182: min F0
- 183: range F0
- 186: arithmetic mean of F0
- 190: stddev F0

We started experiments using LibSVM because we wanted to experiment on the kernel usage on the classifier, which isn't supported by LIBLINEAR. As this paper mentioned, using different kernels can significantly impact a classifier's result. LibSVM supports 4 types of kernels: linear, polynomial, radial basis function (RBF), and sigmoid - we tried all four, paired with its most optimum c and g values as earned through running easy.py, a convenience tool inside LibSVM's library to estimate these values. As part of train and predict, we also scale the data with svm-scale, which significantly increases our accuracy. Using these tools, we arrive at the classifier setup that best suits us.

On the other hand, when using LIBLINEAR classifier with the same feature set, our best accuracy is 58.7879%, which is reached by L1-regularized logistic regression classifier. We choose logistic regression classifier because this is the classifier used in the awkwardness / flirtation study mentioned in lecture. We experimented on the different values of cost before setting with 1.6, hence giving the classifier less regularization than the default 1.

We selected our feature set based on our intuitions about emotional speech. The first five features, for instance, capture the intensity profile of speech, which we would expect to vary with the emotional content of speech. In addition, min, max, mean, range and stdev of energy (RMS) and F0 is mentioned during lecture as important features in distinguishing emotions. Slide 34 of Lecture 7 underlined this importance; Liscombe et al. concluded that positive-activation emotions such as anger, frustration, happiness and confidence, have high F0 and RMS. Using F0-min in itself is useful for distinguishing happiness(81.25% acuracy), F0-max itself is useful in distinguishing sadness, and F0 and RMS, along with other features, are useful in distinguishing anger. This reasoning leads us to choose min, max, mean, range and stdev of RMS and F0.

We include all features of all 12 MFCCs because we expect MFCCs to play an important role in distinguishing the shape of the vocal tract (and the resulting filter it applies to speech) with changing emotional state. It seems reasonable that emotional state would affect phone production; both the energy put into phoneme production and mouth and throat shape are affected by emotion. Experimentally, using all MFCCs and all of their features yielded better performance than a using a subset.

We included max, min, and mean zero-crossing rate because zero-crossing rate is a good approximation for pitch, which we would expect to vary with the emotional content of the speech, as indicated above.

We lastly included the range and position of the maximal value of voicing probability since we found a paper in which it was used (along with log-energy) to classify depressed speech ("Characterising Depressed Speech for Classification", Sharifa, Alghowinem).