Expression of affect in spontaneous speech

Acoustic correlates and automatic detection of irritation and resignation

Dan Wilkey

Outline

- Questions and goals
- Background
- Experimental design
- Results
- Conclusions / Future Work

Questions and goals

- Does the affective content of spontaneous speech mirror that of acted speech?
- Can humans detect the difference between irritated (anger family), resigned (sadness family), and neutral speech?
- Can a recognizer be trained to make the same distinctions at least as well as humans?

Background

- Cowie, Scherer, and others attest that ‘big 6 emotions’ such as anger, fear, joy, and sadness can be distinguished reliably using only acoustic and prosodic cues from acted speech by humans
- Davitz and others have argued that affect in acted speech mirrors sponatneous
  - Problems with acted speech

Background

- Laukka and others have tried to elicit emotional speech in labs with limited success
- Studies attempting automatic detection of affect in speech have used few categories with rarely better than chance results
  - Classification gets more difficult as the speech becomes more closely natural

Experimental Design

- Large corpus of human-machine telephone conversations in Swedish
  - Information hotline (airlines, ferries, post)
- Only chose subjects with at least one neutral and one affective utterance to allow for speaker normalization
- Selected 200 utterances from 61K
  - 112 neutral, 31 emphatic, 21 resigned, 67 irritated

Experimental Design

- Automatically extracted 73 acoustic measures from each utterance using praat scripts
  - Used PCA to reduce the # of vars to 23
- Listeners rated utterance from 0 to 7 for irritation, resignation, neutrality, and intensity. Majority vote was ground truth
  - Used leave-one-out method to determine human classification accuracy

Human Results

Table 3
Summary of the results from the listening test.

	Classification results		
	Irritation	Resignation	Neutral
Mean rating (M/SD)			
Irritation	4.85 (0.79)	2.48 (0.96)	1.99 (0.78)
Resignation	1.78 (0.69)	4.53 (0.71)	1.68 (0.70)
Neutral	2.51 (0.68)	2.67 (0.52)	4.57 (0.67)
Emotion intensity	4.95 (0.79)	3.75 (0.66)	3.19 (0.55)
N speech samples	36	23	133
N speakers	31	15	61

- Irritation and intensity showed high correlation

Experimental Design

- 8 features showed statistical significance for irritated vs. neutral speech
- 6 features showed significance for resigned vs. neutral speech
- Conducted multiple regression analyses to train their classifier

Classifier Results

- Chance performance: 33%
- Human performance: 57.7%
- LDA classifier, no adaptation: 62.3%
- LDA classifier, adaptation: 54.3%
- Humans were better at detecting neutrality, though the classifier was better at detecting both emotions
  - What does this say about human analysis?

Classifier Results

- Poor performance with normalization likely due to small number of utterances per speaker
  - Different set of features were most important with the normalization vs. without

Classifier Results

Table 8
Confusion matrices for the detection of irritation, resignation, and neutral for (a) automatic detection using LDA (both with and without speaker adaptation) and (b) human perception.

		Recall (%)		
		Irritation	Resignation	Neutral
No adaptation	Irritation	**69.7**	6.1	24.2
	Resignation	0.0	**64.3**	35.7
	Neutral	28.7	21.2	**50.0**
Adaptation	Irritation	**57.6**	9.1	33.3
	Resignation	14.3	**42.9**	42.9
	Neutral	16.3	21.2	**62.5**
Human performance	Irritation	**56.7**	7.4	35.9
	Resignation	18.6	**45.3**	36.0
	Neutral	15.5	13.3	**71.2**

Conclusions / Future Work

- The effect sizes (intensity) of emotions in this study were much less than that of acted speech studies
  - However, similar features were useful for classification
- Acoustic correlates of resignation and irritation were very similar to sadness and anger
  - Supports emotion family theory

Conclusions / Future Work

- Statistical methods employed do not exactly match human methods due to differing strengths/weaknesses, though both are effective
  - Could be related to priors
- Greater variety in the context of the speech used may add to robustness
- It would be useful to simply know what percentage of speech contains affect

Conclusions / Future Work

- Combining this approach with facial expressions and bodily gestures may improve accuracy
- Better means of annotating speech data may prove useful
  - Mutually exclusive categories don't appear to be the best fit
- Could be used in reverse to determine affect inference process of humans

Questions?