Abstract 2pSC14

Improving the Accuracy of Speech Emotion Recognition Using Acoustic Landmarks and Teager Energy Operator Features

Reza Asadi, Harriet Fell
Northeastern University

Introduction

- Computers that can recognize human emotions could react appropriately to a user's needs and provide more human like interactions.
- Some of the applications of emotion recognition:
  - Diagnostic tool for medical purposes
  - Onboard car driving systems to keep the driver alert if stress is detected[1]
  - Similar system in aircraft cockpits
  - Online tutoring
  - Interaction with virtual agents or robots[2]
- Common approach for interpreting emotions from speech[3]:
  - Gather acoustic information in the form of sound signals
  - Extract related information from the signals
  - Find patterns which relate acoustic information to the emotional state of speaker
- Our contributions:
  - Use new combinations of acoustic feature sets to improve the performance of emotion recognition from speech
  - Provide a comparison of feature sets for detecting different emotions

Methodology

- Extract 3 different acoustic feature sets:
  - Mel-Frequency Cepstral Coefficients
  - Teager Energy Operator features
  - Acoustic Landmarks
- Classify an emotional speech database using these features sets
- Compare the results of using different features sets
- Compare the accuracy of the classification with a similar study

Acoustic Landmarks

- Acoustic landmarks are locations in the speech signal where important and easily perceptible speech properties are rapidly changing[4].
- The number of landmarks in each syllable might reflect underlying cognitive, mental, emotional, and developmental states of the speaker[5].

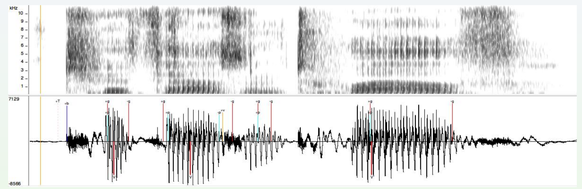

Figure 1. Spectrogram (top) and acoustic landmarks (bottom) detected in a neutral speech sample

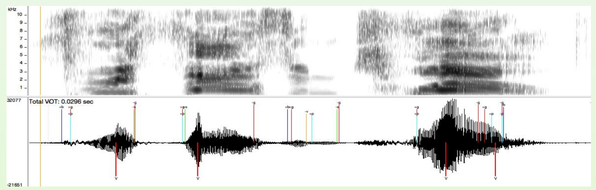

Figure 2. Spectrogram (top) and acoustic landmarks (bottom) detected in an anger speech sample

Mel-Frequency Cepstral Coefficients

- The Cepstrum is a signal analysis tool which is useful in separating source from filter in acoustic waves.
- Since the vocal tract acts as a filter on the glottal wave we can use the cepstrum to extract information only related to the vocal tract.
- The mel scale is a perceptual scale for pitches as judged by listeners to be equal in distance from one another.
- If we map frequency powers of energy in original speech wave spectrum to mel scale and then perform cepstral analysis we get Mel-Frequency Cepstral Coefficients (MFCC).

Teager Energy Operator features

- While speaking in emotional states of panic or anger, physiological changes like muscle tension alter the airflow pattern and can be used to detect stress in speech.
- Teager Energy Operator (TEO) computes the energy of vortex-flow interaction at each instance of time.
- Previous studies show that TEO related features can be used to recognize emotions in speech[6].

Classification

- The data used in this study came from the Linguistic Data Consortium's Emotional Prosody and Speech Transcripts database[7].
- Support Vector Machine(SVM) classifiers were used for automatic classification.
- Principal Component Analysis (PCA) was used to reduce the number of features.
- The target emotions included *anger, fear, disgust, sadness, joy,* and *neutral*.

Results

- The MisClassification Error rate (MCE) for the classifier using only the MFCC features is 27.68%.

	Actual Emotions					
	Anger	Anxiety	Disgust	Sadness	Joy	Neutral
Anger	124	1	5	1	23	0
Anxiety	0	135	24	24	18	10
Disgust	8	13	124	8	17	5
Sadness	2	22	10	109	11	7
Joy	4	7	10	8	105	3
Neutral	0	3	6	4	5	54
Total	138	181	179	154	179	79

Table 1. Confusion matrix of the classification using all feature sets

	Anger	Anxiety	Disgust	Sadness	Joy
Accuracy	-	TEO	-	TEO	TEO
Recall	LM	TEO	-	-	TEO
Precision	-	-	TEO	TEO	LM
F1	-	TEO	-	TEO	TEO

Table 2. Best feature sets to add to MFCCs features for each emotion and measurement (LM=Landmarks)

Comparison with the Baseline

- Smailagic et al.[8] performed similar experiments on the same data set using similar features and SVM classifiers.

	Anger	Anxiety	Disgust	Sadness	Joy	Neutral	Average
Baseline Accuracy(%)	90.02	79.57	87.16	76.18	76.06	83.45	82.07
Our Work Accuracy(%)	**95.51**	**86.92**	**89.17**	**89.56**	**88.57**	**96.39**	**90.77**
Increase in Accuracy(%)	5.49	7.35	2.01	13.38	12.51	12.94	8.70

Table 3. Comparison of the accuracy of the multiclass classification with the baseline study

The Future Works

- Test the classifier on non-acted and spontaneous speech samples.
- Test the robustness using noisy recordings.
- Improve the performance of the system for realtime applications.
- Multimodal emotion recognition by adding facial expression classification.

References

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review, " International Journal of Speech Technology, vol. 15, no. 2, pp.99–117, 2012.
[2] C. Chastagnol, C. Clavel, M. Courgeon, and L. Devillers, "Designing an emotion detection system for a socially intelligent human-robot interaction," in Natural Interaction with Robots, Knowbots and Smartphones. Springer, 2014, pp. 199–211.
[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," Speech communication, vol. 48, no. 9, pp. 1162–1181, 2006.
[4] J. Slifka, K. N. Stevens, S. Manuel, and S. Shattuck-Hufnagel, "A landmark-based model of speech perception: History and recent developments," From Sound to Sense, pp. 85–90, 2004.
[5] H. J. Fell and J. MacAuslan, "Automatic detection of stress in speech."in MAVEBA, 2003, pp. 9–12.
[6] A. Georgogiannis and V. Digalakis, "Speech emotion recognition using non-linear teager energy based features in noisy environments," in Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE, 2012, pp. 2045–2049.
[7] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional Prosody Speech and Transcript," 2002, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28.
[8] A. Smailagic, D. Siewiorek, A. Rudnicky, S. N. Chakravarthula, A. Kar, N. Jagdale, S. Gautam, R. Vijayaraghavan, and S. Jagtap, "Emotion recognition modulating the behavior of intelligent systems," in Multimedia (ISM), 2013 IEEE International Symposium on. IEEE, 2013,pp. 378–383.