Comparing Emotions Using Acoustics and Human Perceptual Dimensions

Keshi Dai
College of Computer
& Information Science
Northeastern University
Boston, MA 02115 USA
daikeshi@ccs.neu.edu

Harriet Fell
College of Computer
& Information Science
Northeastern University
Boston, MA 02115 USA
fell@ccs.neu.edu

Joel MacAuslan
Speech Technology
& Applied Research
54 Middlesex Tpk.
Bedford, MA 01730 USA
joel@S-T-A-R-corp.com

CHI 2009, April 4 – 9, 2009, Boston, MA, USA
ACM 978-1-60558-246-7/09/04.

Abstract

Understanding the difference between emotions based on acoustic features is important for computer recognition and classification of emotions. We conducted a study of human perception of six emotions based on three perceptual dimensions and compared the human classification with machine classification based on many acoustic parameters. Results show that the six emotions cluster differently according to acoustic features and to perceptual dimensions. Acoustic features fail to characterize the perceptual dimension of valence. More research is needed to find acoustic features that have a close relation to human perception.

Keywords

Emotional speech, emotion classification, acoustics, human perception of emotion

ACM Classification Keywords

H.1.2 [Models and Principles]: User/Machine Systems – H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Introduction

In a previous study [2], we developed a classifier that can automatically distinguish emotions in speech. We found that some emotion pairs are likely to be misclassified. Why does our classifier think these emotions are similar even though humans may think they are quite different?

The goal of this paper is to compare different types of emotions in terms of acoustic features and human perception. We first analyzed and visualized acoustic features extracted by our classifier. We then conducted a study to explore the human perception of the relationship between these emotions.

Background

As early as 1872, Darwin proposed that vocal expression is a primary carrier of affective signals in animal and human communication [3]. A large number of later studies have found specific acoustic profiles in voice that are associated with different emotions [7]. Fundamental frequency, formant frequency, speech rate, and voice intensity are considered as common vocal cues relating to emotions. Recent studies also showed that landmarks and Mel-frequency cepstral coefficients could be useful features [2, 4, 6].

Several machine learning algorithms such as Naïve Bayes Classifiers, Neural Networks, Support Vector Machines, Hidden Markov Models, and ensemble classification methods have been applied to classify emotions based on acoustic features [2, 4, 6, 10]. However, most studies focus on classification accuracy, making little effort to explain the sources of misclassification. These studies also ignore the difference between the human perception and computer recognition that leads to the inconsistency between human labeling and computer labeling. We will address these two points in this paper.

Data

Our data are from the Emotional Prosody Speech and Transcripts (EPST) corpus [8]. This corpus contains 15 emotions produced by 8 professional actors (5 female, 3 male) reading 4-syllable semantically neutral utterances. In this paper, we focus on 6 emotions: *happy*, *hot anger*, *neutral*, *interest*, *panic*, and *sadness*. Because actors were allowed to repeat the emotional phrase until they are satisfied, we only use the last utterance for each emotional phrase. We have 70 utterances for each emotion except *neutral* and 49 utterances for *neutral*.

Acoustic Features of Emotion

Our feature extraction generates a total of 47 acoustic features. These features can be categorized into 5 types: landmark features, syllable features, timing features, pitch features, and energy features.

Landmarks are the abrupt spectral changes in speech signal [9]. Based on landmarks, we extracted the following features: landmarks per word and landmarks per utterance, voice onset time, and landmark rate: the rate of each landmark type in an utterance.

A syllable is typically made up of a vowel with optional initial and final margins. It can also be seen as a sequence of landmarks. We recognized 38 possible syllable types and extracted 4 types of features: syllable rate, syllables per utterance, landmarks per syllable, and syllable duration statistical information.

Pitch is the perceptual correlate of the fundamental frequency of voice. We extracted pitch contour (10 percentile, 50 percentile, and 90 percentile values of the pitch), pitch statistical information, and pitch slope.

Energy is derived from the first derivative of the smoothed speech signal. We generated the energy contour, slope and other statistical information.

Timing features capture prosodic characteristics of the utterances. We extracted mean, minimum, maximum, and the standard deviation of voiced and unvoiced duration in each utterance. The ratio of the voiced and unvoiced duration is also measured.

Comparing Emotions Using Acoustics

A simple way to get a feel for an emotion is to do a parallel coordinates plot of the extracted data values of the utterances as in figure 1. Each feature corresponds to an axis, and 47 axes are organized as uniformly spaced vertical lines. Different colors of the line correspond to different emotions.

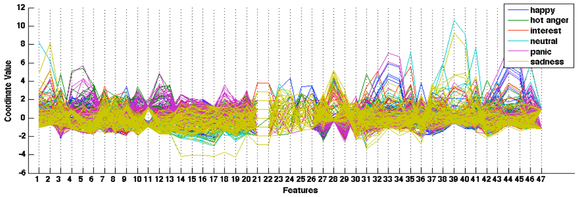

figure 1. Parallel coordinates plot of 6 emotions.

It is difficult to perceive the structure of the data due to a large number of observations for each emotion. Either dimension reduction or a new way to depict an emotion's attributes should be applied to reduce the complexity of data.

We summarized the attributes for an emotion by taking the average of the features over all utterances with that emotion, and used the average values as emotion representatives. Figure 2 shows the parallel coordinate plotting of the emotion representatives for 6 emotions. However, the relationship between emotions is still hard to observe due to the great number of features.

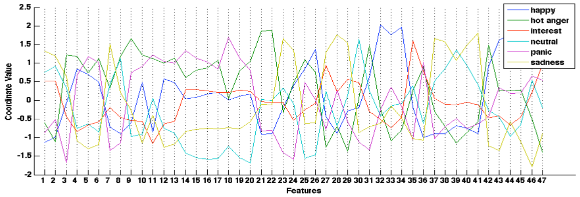

figure 2. Parallel coordinates plotting of emotion representatives for 6 emotions.

Classical multidimensional scaling (MDS) is a data analysis technique to help visualize the dissimilarities between each pair of data [2]. We form a dissimilarity matrix of Euclidean distances for all pairs of emotions based on the average representatives. After double centering this matrix, eigendecomposition is conducted to obtain the coordinate matrix whose configuration minimizes the loss function. Figure 3 shows the eigenvalues and the root mean square error of the reconstruction based on the coordinate matrix using different numbers of components.

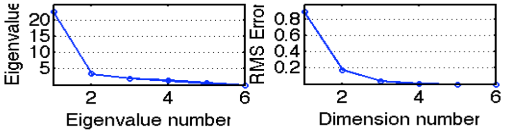

figure 3: plot of eigenvalues (left) and root mean square (RMS) error (right).

The plot of eigenvalues (figure 3 left) indicates that two components are enough to represent all features. After shifting all other emotions by choosing *neutral* as the origin, we can get a clearer view of the relationship of the 6 emotions (figure 4).

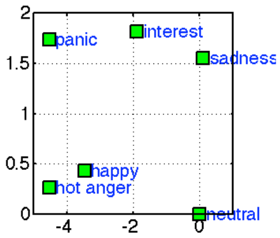

figure 4. The visualization of 6 emotions using classical MDS based on acoustic features.

Comparing Emotions Using Perceptual Dimensions

To explore human emotion recognition, we invited 20 subjects (8 males and 12 females) to describe their perception of 6 emotions from the EPST corpus. Subjects were all college students and English speakers, 13 of them native speakers.

This study was conducted in a standard study room using a computer-based survey. Participants were asked to listen to 42 emotional audio clips and describe the emotion they perceived. They were asked to choose which of the six EPST emotions best described the emotion they perceived or select “None” and describe the emotion using their own words. Participants also rated each audio clip in three dimensions: valence, potency, and activation.

Emotion classification by human subjects

Human subjects labeled a total of 840 audio clips. 61 clips (7.26%) were labeled as “None”, which implies subjects thought there was no appropriate emotion type in the list for these clips. Among those audio clips that subjects labeled, 502 clips were labeled as same as the corpus or 64.44% accuracy. In the previous study, a neural network classifier achieved 48.95% accuracy in classifying 6 emotions. A support vector machine classifier was also trained on the same dataset and obtained 52.04% accuracy. If misclassification includes those labeled as “None” by subjects, humans correctly classified 502 out of 840 clips (59.76%).

	E1	**E2**	**E3**	**E4**	**E5**	**E6**
E1	55	7	30	22	8	8
E2	3	104	10	4	6	0
E3	13	1	53	40	12	12
E4	0	0	3	107	1	24
E5	7	12	7	2	94	7
E6	1	0	0	31	6	89

table 1. The confusion matrix of the human classification (excluding “None” clips). E1: *happy*, E2: *hot anger*, E3: *interest*, E4: *neutral,* E5: *panic*, E6: *sadness*.

Table 1 is the confusion matrix of the human subject classification. We see that only 55 *happy* were correctly classified, 30 of them were considered as *interest*, 22 were classified as *neutral*. Most *hot anger* and *panic* clips are correctly classified. A large number of *interest* clips were seen as *neutral* and *sadness* was often misclassified as *neutral*. Based on this, we can conclude that *happy*/*interest*, *happy*/*neutral*, *interest*/*neutral*, and *neutral*/*sadness* are four confusing pairs, which is consistent with the results in the previous study [2].

The classification results (taking “None” as mislabeling) for native and nonnative English speakers are different. The classification accuracy for native speaker is 62.09% and for nonnative speaker is 55.44% but both groups confused the same pairs.

Perceptual dimensions of emotion

In our study, we used three perceptual dimensions: valence, potency, and activation [5]. Valence indicates how positive or negative an emotion is and ranges from unpleasant(1) to pleasant(5). Potency depicts the coping potential or power and ranges from weak(1) to strong(5). Activation relates to a subjective sense of mobilization and is from sleepy(1) to excited(5).

A simple factorial analysis (ANOVA) indicates that there is no significant difference on the perceptual dimension rating between native speaker and nonnative speakers. The *p*-values for 3 dimensions are 0.10, 0.42, and 0.88. Table 2 is a summary of statistics of perceptual dimension ratings for the six emotions by all subjects.

	Valence			**Potency**			**Activation**		
	mea	**std**	**med**	**mea**	**std**	**med**	**mea**	**std**	**med**
E1	4.5	0.5	5	3.2	0.8	3	4.0	0.8	4
E2	1.3	0.3	1	4.5	0.8	5	4.3	0.7	4
E3	3.8	0.7	4	3.6	0.8	4	3.9	0.7	4
E4	3.2	0.7	3	2.8	0.8	3	2.6	0.7	3
E5	1.8	0.8	2	2.9	1.3	3	4	0.9	4
E6	2.0	0.9	2	1.8	0.9	2	2.0	0.8	2

table 2. The statistics of perceptual dimensions for different emotions. E1: *happy*, E2: *hot anger*, E3: *interest*, E4: *neutral,* E5: *panic*, E6: *sadness*. mea: mean, std: standard deviation, med: median.

As shown in table 2, *Happy* and *interest* are quite similar. They both are above-medium potency and activation, but *happy* has higher valence value. Although *hot anger* and *panic* share high valence and activation, they are different in potency. *Neutral* sits in the center of every dimension, and all dimensions for *sadness* are below medium. Although every emotion is distinct from others in at least one dimension, emotions in the confused pairs have relatively close values.

Comparing figure 4 with table 2, we can see that emotions differing in potency or activation are also far from each other in figure 4, e.g. *hot anger* and *neutral*, *hot anger* and *panic*, and *sadness* and *hot anger*. This shows acoustic features can describe potency and activation very well, but they fail to represent valence because *hot anger* is close to *happy*, *interest* and *sadness* are both far from *neutral,* and *interest* is close *to sadness* in figure 4.

Classical MDS is also used to visualize and compare emotions in terms of human perception. We use the mean value of the emotion in each dimension over all ratings. Euclidean distance is used to measure the dissimilarity of emotions. Because the dimensionality of the data is 3, we can use classical MDS to easily reduce the data to 2 dimensions with small information loss (root mean square error=0.01).

Figure 5 is 2D visualization of 6 emotions with *neutral* as the origin. It shows the relative relations of the 6 emotions based on the human perception. Comparing figure 5 with figure 4, the 6 emotions cluster differently in the two figures because acoustic features do not capture distinctive the perceptual information in valence as we addressed before.

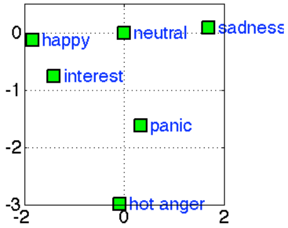

figure 5. The visualization of 6 emotions using classical MDS based on perceptual dimension ratings.

Conclusions and Future Work

In this paper, we analyzed acoustic features for 6 emotions and conducted an emotion perception study to explore the human perception of these emotions. We also compared classification results by humans and the computer. Furthermore, we found the six emotions cluster differently according to acoustic features and to perceptual dimensions, because the acoustic features we extracted do not completely and accurately model the human perception of emotion, especially in valence dimension. This may also result in the better classification performance by human subjects than our classifier. On the other hand, the difference between acoustic features and perceptual dimensions can also be caused by the information loss due to mapping data into lower dimensions or averaging to summarize the large number of observations for each emotion.

Future work will concentrate on finding features that have a close relation to human perception. These should be features that people believe they use to understand emotions in speech. Also, more sophisticated methods should be applied to measure the dissimilarity of emotions instead of using the Euclidean distance based on the average value. More important, real emotions should also be studied instead of acted emotions.

References

[1] Borg, I., and Groenen, P. *Modern Multidimensional Scaling: Theory and Application*. Springer, New York, USA, 2005.

[2] Dai, K., Fell, J.H., and MacAuslan, J. Recognizing Emotion in Speech Using Neural Networks. In *Proc*. IASTED 2008 on Assistive Technologies, ACTA Press (2008), 31-36.

[3] Darwin, C. *The Expression of Emotion in Man and Animals*. John Murray, London, UK, 1872.

[4] Kwon, O., Chan, K., Hao, J., Lee, T. Emotion Recognition by Speech Signals. In *Proc*. EUROSPEECH 2003, ISCA (2003), 125-128.

[5] Laukka, P. Vocal Expression of Emotion: Discrete-Emotions and Dimensional Accounts. PhD thesis, 2004, Uppsala University (2004).

[6] Lee, C. M., Yildirim. S., Bulut, M., Kazemzadeh, A. Busso, C., Deng, Z., Lee, S., and Narayanan, S. Emotion Recognition Based on Phoneme Classes. In *Proc*. ICSLP 2004, ISCA (2004), 889-892.

[7] Lewis M., and Hayiland J.M. *Handbook of Emotions*. Guilford Press, New York, USA, 1993.

[8] Linguistic Data Consortium (LDC2002S28), *Emotional Prosody Speech*, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28, University of Pennsylvania.

[9] Liu, S., Landmark Detection of Distinctive Feature-Based Speech Recognition. *Journal of the Acoustical Society of America*, 100(5), 1996, 3417-3430.

[10] Morrison, D., Wang, R., De Silva, L. C. Ensemble Methods for Spoken Emotion Recognition in Call-centres, *Speech Communication*, 100(5), 2006, 98-112.