Acoustic and physiological feature analysis of affective speech

Dandan Cui, Lianhong Cai

Key Laboratory of Pervasive Computing (Tsinghua University), Ministry of Education
Beijing 100084, P.R.China
{cuidd02}@mails.tsinghua.edu.cn
{clh-dcs}@tsinghua.edu.cn

Abstract. The paper presents our recent work on the acoustic and physiological feature analysis of affective speech. An affective speech corpus is first built up. It contains passages read in neutral state and ten typical affective states selected in Pleasure Arousal Dominance (PAD) space. Physiological data, including electrocardiogram, respiration, electro dermal data, and finger pulse, are also collected synchronized with speech. Then based on the corpus, the relationship between affective states and acoustic\physiological features is studied through correlation analysis and co-clustering analysis. The analysis results show that most acoustic features and physiological features are significantly correlated with the arousal dimension, whereas only respiration features are more correlated with the pleasure dimension.

1 Introduction

Affective computing refers to computing that relates to, arises from, or deliberately influences emotions [1], especially in human computer interaction. Giving a computer the ability to recognize and express emotion in speech is one of the most appealing demands in speech recognition and synthesis technologies. It makes the humancomputer interaction much more interesting and convenient. The modeling of affective speech needs the support of good data, both in quantity and in quality, that is, affective speech corpus. During corpus building, the conflict between naturalness and control could never be survivable. However, weighing those two aspects according to specific research target to decide what to collect and how, will be the right way out. The ESP corpus in Japan aims for 250 hours of spontaneous speech, and the total is 1000 hours [2]. And interesting spontaneous speech tone modes are found, which may benefit the expressive speech synthesis, especially for dialog systems, smartly.

In China, research on affective speech has been paid more and more attention in recent years, but still needs more. Our project of research on affective computing theories and methods started up in year 2004.

For affective speech modeling in the 3_dimension PAD emotion space, via which the synthesis of speech with any emotion, and the free conversion between emotions will be achieved, a corpus is designed and implemented first. The script is in passages, including sentences across emotions. The speech data is uttered by 20 persons.

Psychological and physiological activities are thought to be closely interconnected. Physiological data may reflect emotional states that can not be observed by human eyes or ears [1], so it is collected as well as speech.

Given the data, the first work for us to do is extracting and analysis of features. Statistical results of 16 acoustic and physiological features and the correlation between features and PAD dimensions show that features are likely to be relative to different aspects of emotional states. And furthermore, co-clustering analysis gets a result of three stable groups of features. Affective speech modeling in this 3D space is fairly promising.

The rest of this paper is organized as follows: Section 2 introduces the design of corpus, including design of emotion description and design of text script. Section 3 illustrates the corpus implement procedure and the data we get. Then in section 4, we extract the acoustic and physiological features, and take statistics. Before concluded with discussions in section 6, co-cluster analysis is also carried out in section 5.

2 Corpus design

Corpus design often refers to the design of text script for the speakers to utter. The design of description for emotional states is also very important. In this section we will focus on those two key stages, and show the details.

2.1 Description of emotion states

Emotion states can be described in either categories or dimensions [1]. In China, 5 basic emotional categories is mostly used: anger, fear, joy, sorrow, and surprise [3]. Other researchers refer to continuous dimensions to illustrate more emotions and relations. Two dimensions are widely agreed on: “arousal”, and “valence”. Yet, they cannot distinguish all the basic emotions, e.g. fear and anger.

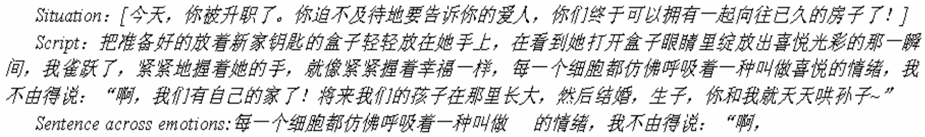
We choose PAD space [4] to be the base of our descript system which is getting increasing applications in many fields of humancomputer interaction [5]. The three dimensions are pleasure (P), arousal (A), and dominance (D). The dominance dimension can help to distinguish emotions such as fear and anger. Actually, it can be really helpful in research on affective speech, because speaking style vary greatly, depending upon who we are speaking to. Furthermore, it is a quantified model so that can distinguish all emotion states in theory. Feature analysis can be much more convenient too. Reliable and valid measure method is offered. And its Chinese Version is also developed [6]. We choose ten typical emotions that are commonly used in affective computing and everyday life. As shown in Table 1, they are distributed uniformly within the eight quadrants of the PAD space.

Table 1. Emotions in the corpus

PAD	Emotion	Description
+P +A +D	exuberant	extroverted, outgoing, happy, sociable
+P -A +D	Relaxed	comfortable, secure, confident, resilient to stress
+P -A -D	Docile	pleasant, unemotional, and submissive; likeable; conforming
-P -A +D	Disdainful	Contemptuous of others, sometimes anti-social
-P +A +D	Disgusted	dislike others, emotional in negative ways
-P -A +D	Angry	angry, emotional in negative ways, possibly violent
-P +A -D	Fearful	insecure, possibly extreme, physically active
-P +A –D	Anxious	worried, nervous, insecure, tense, illness prone
+P +A –D	Surprised	attached to people, interpersonally positive and sociable
-P -A -D	Sad	lonely, socially withdrawn, physically inactive

2.2 Text Script

Although emotion in broadcast/TV clips and everyday talks are thought to be more natural, our research on affective speech conversion and synthesis requires the speech data to be under strict design and control. So the form of speech is designed as reading/recite of text script. Since emotional states can not change as quickly and whose are inspired under a certain situation can be more natural, we designed the script as passages with certain situations. 10 passages is designed for each category, each passage contains 100 syllables or so. Situation selection is referred to a study on situation-activity combinations in everyday life represented by the author of PAD [7]. Different contents with different syllables or prosodic constituents may wreck some analyses. So in every passage, we designed a sentence that is emotionally unbiased. Here is an example:



Situation：[今天，你被升职了。你迫不及待地要告诉你的爱人，你们终于可以拥有一起向往已久的房子了！]

Script：把准备好的放着新家钥匙的盒子轻轻放在她手上，在看到她打开盒子眼睛里绽放出喜悦光彩的那一瞬间，我雀跃了，紧紧地握着她的手，就像紧紧握着幸福一样，每一个细胞都仿佛呼吸着一种叫做喜悦的情绪，我不由得说："啊，我们有自己的家了！将来我们的孩子在那里长大，然后结婚，生子，你和我就天天哄孙子~"

Sentence across emotions:每一个细胞都仿佛呼吸着一种叫做　的情绪，我不由得说："啊，

3 Corpus implement and the data

After recording of the script and corpus processing, we finally got the corpus. In this section, we will introduce the corpus implement procedure, and the current data we obtained in the following two subsections.

3.1 Corpus implement

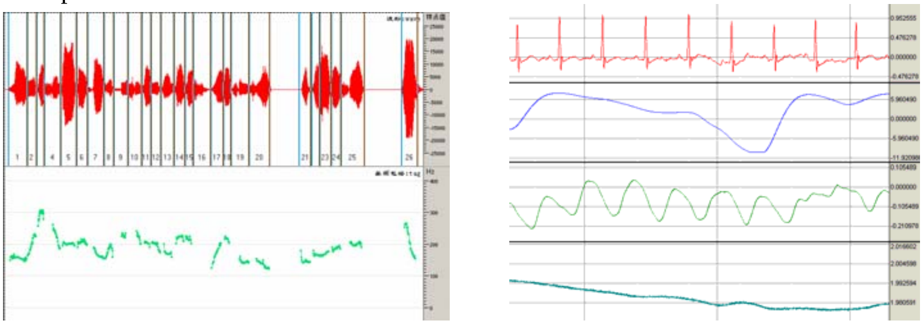
Speaker selection: The current 20 speakers, including 10 boys and 10 girls, are selected from more than 100 candidates via interviews, all of whom are chief subjects in a previous experiment. The age ranges from 18 to 25.

Recording: Besides the passage, the emotion, and situation are also given to the speaker. The speaker reads or recites the script, imaging himself in the situation. The emotion will can only be valid if all of the 3 listeners say OK and there is a long break before next emotion for reset. Besides speech, physiological data is also collected. It includes electrocardiogram, respiration, electro dermal data, and finger pulse.

Corpus Processing: The recorded data is segmented into passages, then annotated. Boundaries of prosodic constituents are marked. And the F0 values of speech data are also annotated. Besides the emotional categories, PAD values are also scored.

3.2 Data of the current corpus

After the corpus implement procedure introduced above, we got a data set of 2200 passages from 20 speakers and accordingly 2200 samples of sentence across emotions are got. The passage segmenting, prosodic boundary marking is finished. F0 value annotation for male speech is ready. And PAD scoring is in process. Fig 1 gives an example.



(a) (b)

Fig. 1. Data of the sentence "*每一个细胞都仿佛呼吸着一种叫做喜悦的情绪*". (a)For speech data, prosodic boundaries are marked (upper) and F0 is annotated (lower). (b) The physiological features are: electrocardiogram, respiration, finger pulse, and electro dermal data.

4 Feature extraction and statistics

Rich and well controlled data provides good base for feature extraction and analysis. First we extract 16 features that are thought to be valuable in affective analysis. The reliability is also considered. Then statistics are taken on the sentences across emotions in male data.

4.1 Extracting the features

We totally extract 16 features, 7 from speech data and 9 from physiological data. The number of physiological features is a bit more because we are not sure about which are significant for affective computing. In fact, the 9 features are extracted from either the electrocardiogram or respiration data, for the other two seem quite impenetrable.

Acoustic Feature: There have been many research results on affective speech. The prosodic features such as F0 and duration are mostly used. Inspired by music classification, Danning Jiang found several spectral features valuable [3]. We selected 7 features that are proved to be discriminative [3].

(1), (2): F0 and its first order difference (dF0). F0 can be read from our F0 annotation files with the extension of “.tag”.

(3): Duration (Dur) of syllable. This can also be read from ‘.tag’ files in which the boundary marks are saved.

(4): Short term energy (Ene).

(5): Spectral Centroid (SC). It reflects the high frequency content in spectrum and can be calculated by Eq.1.

$$SC = \frac{\sum_{n=1}^{N} nA(n)}{\sum_{n=1}^{N} A(n)} \tag{1}$$

(6): Spectral Flux (SF). It reflects the intensity of spectral flux and can be calculated by Eq.2.

$$SF = \sum_{n=1}^{N} (M_i(n) - M_{i-1}(n))^2 \tag{2}$$

(7): Band Periodicity (BP). It reflects the intensity of the periodicity in spectrum and can be calculated by Eq.3.

$$R_j(k) = \frac{\sum_{m=1}^{M} s_j(m-k)s_j(m)}{\sqrt{\sum_{m=1}^{M} s_j^2(m-k)}\sqrt{\sum_{m=1}^{M} s_j^2(m)}} \tag{3}$$

Physiological features: Physiological data changes follow an affective stimulus, and several valuable parameters are agreed on [8]. Considering both significance and reliability, we choose the following 9 out of electrocardiogram and respiration.

(1), (2): The peak value of R wave in electrocardiogram (PRH) and its first order difference (dPRH). The former may reflect the intensity of heart beat.

(3), (4): Interval of R wave peak (IRH) and its first order difference (dIRH). The former reflects heart rate and is represented in the unit of ms.

(5), (6): The min (minR) and max (maxR) value in respiration data which may reflect the range of respiration.

(7), (8): The mean (meaR) and median (medR) value in respiration data which may reflect the volume of respiration.

(9): The first order difference of value in respiration data (dR).

4.2 Statistical results

First, means are calculated within each utterance, except for the physiological features 5 to 8. Then means are calculated within each emotion.

Table 2. Satistical result of the features. The absolute value of dR is too small, so that we represent them in a thousand times larger form.

	Neutral	Relaxed	Docile	Surprised	Exuberant	Disdainful	Disgusted	Fearful	Sad	Anxious	Angry
F0	127.1	144.2	141.86	218	230.75	178.45	207.92	259.62	174.51	249.59	230.52
DF0	2.209	2.873	2.764	4.542	4.6340	2.923	3.831	5.08	3.170	4.883	4.762
Dur	0.138	0.123	0.121	0.094	0.103	0.113	0.102	0.088	0.129	0.096	0.078
Ene	39.315	34.113	33.917	43.836	43.925	40.725	41.891	42.924	40.162	45.616	45.357
SC	2274.7	2362.5	2346.4	2715.9	2774.8	2571.7	2877.1	3090.3	2484.9	3045.4	3155.4
SF	0.265	0.276	0.316	0.640	0.646	0.433	0.585	0.716	0.435	0.678	0.797
BP	0.684	0.777	0.701	0.630	0.644	0.706	0.631	0.644	0.714	0.643	0.662
PRH	0.963	0.974	0.945	0.959	0.959	0.957	0.959	0.955	0.937	0.937	0.918
dPRH	0.075	0.0823	0.0812	0.126	0.117	0.0914	0.110	0.128	0.092	0.115	0.114
IRH	687.49	674.29	682.91	640.08	621.57	676.29	625.78	596.08	646.71	609.73	606.34
dIRH	12.614	12.359	14.78	13.445	9.5273	12.813	11.108	9.3339	13.463	9.709	9.4175
minR	-0.5	2.8182	4.7273	-0.727	1.25	0.7	2.4167	4.2857	-2.636	5.75	5.3
maxR	6.7	7.2727	7.5455	7.5455	7.5833	7.7	7.8333	8	7.9091	8	8
meaR	5.386	6.3049	6.8825	5.9828	6.9499	6.5249	7.1031	7.7514	6.2398	7.8595	7.8301
medR	5.695	6.4243	6.8923	6.5794	7.4335	7.0652	7.4383	7.942	7.0098	7.9741	7.9599
dR*	2.087	1.451	9.98	2.896	1.884	1.792	1.614	1.222	2.27	0.994	1.042

As Table 2 shows, the features vary significantly with emotional categories. Being consistent with the results of previous research [3, 8], F0 rises when peoples are surprised, exuberant, disgusted, fearful, anxious, or angry and so does the speech rate and heart rate; the first order difference of F0 is highest when fearful; energy of anxiety and anger is highest and so on. Surprise and Exuberance are emotions that can not be well distinguished from each other, but their dIRH and respirational features are quite different.

As Danning Jian has reported in her dissertation [3], different features listed in the table seem to be correlated with different aspects of emotional states. In order to see more clearly, we calculated the correlation coefficient between each feature and the sign in each PAD dimension of the emotion it belonged to. The result is as below:

Table 3. The correlation coefficient between features and the signs of PAD

	P	A	D
F0	-0.38306	0.80821	-0.11271
DF0	-0.22349	0.83314	-0.13742
Dur	0.27241	-0.76929	-0.055276
Ene	-0.47356	0.82093	-0.011013
SC	-0.52391	0.80376	0.018542
SF	-0.38452	0.83607	-0.025527
BP	0.23345	-0.88642	0.19527
PRH	0.50317	-0.18226	0.21604
dPRH	-0.20754	0.84108	-0.14404
IRH	0.43712	-0.81251	0.086374
dIRH	0.39947	-0.74481	-0.28249
minR	-0.12943	0.3115	0.03945
maxR	-0.57247	0.32324	-0.15557
meaR	-0.45274	0.49517	-0.00038422
medR	-0.53809	0.52052	-0.010612
dR*	0.27194	-0.02983	-0.098619

As it shows in table 3, most of the features have high correlation coefficients with the dimension A, within which BP is the highest. For the dimension P, only SC, PRH, and maxR and medR have correlation coefficients above 0.5. It may implicate that respirational features are more relative with the intensity of pleasure in emotion. And the dimension D is almost unrelated with any of the features. Perhaps we need to consider new features in order to model the dominance in speech.

In addition, we have to explain why the F0 of neutral emotion in that table is fairly low. It is because that the “neutral” here is a bit different with its common meaning: the speakers were asked to speak like a robot in order to study the intonations and duration structures of emotions.

5 Co-cluster analysis

As we have found in the prior section, different features seem to be correlated with different aspects of emotional states which can be expressed as PAD dimensions, but not quite clear. The features’ correlations may help. Then, in this section, co-clustering is utilized to group both the speech data and the features, and the result is presented.

Co-clustering is a fairly new method. Not like the traditional clustering algorithms, it can group the features and samples simultaneously [9]. Thus, not only the sample clustering is optimized, but the features’ correlations are also mined out.

We co-cluster the utterances with the 16 features illustrated in section 4. Both the result of data and feature are satisfying, especially the later.

Speech clustering result: The number of clusters is appointed as 11. Co-clustering using the acoustic features gets a precision of 83.9%, 7.6% higher than a traditional model (PNN). Optimization of feature set in co-clustering shows its superiority. Co-clustering with all the 16 features raises the precision to 87.2%. The physiological features really help, but more features are still needed as we stated in section 4.

Feature grouping result: We tried to change the min group number from 2 to 3, and the max from 2 to 8, while the grouping result of features keeps unchanged. It is:

Table 4. Feature groups in the co-clustering result

Group ID	Features
1	F0, DF0, SF, dIRH, minR
2	Dur, BP, PRH, IRH, dPRH, dR
3	Ene, SC, maxR, meaR, medR

As it shows in table 4, the grouping result is similar with the statistical result in section 4, but not the same. All the features in group 1 have high correlation coefficients with the dimension A, except for minR. While BP is in group 2, which contains most heart activity features, and seems fairly inexplicable. The third group contains most respirational features, which are found to be more relative with the intensity of pleasure. In short, the grouping result shows correlation with PAD dimensions, but not quite assured. Maybe the groups correspond to their combination but not individuals.

6 Conclusion and discussion

As we have introduced above, 11 typical emotion categories are chosen according to their distribution in PAD space, and the script is in passages. The corpus contains both speech and physiological data. Yet, as the research work progresses, more passages and more emotions need to be designed and recorded.

Both the statistics and co-clustering results show that different features are correlated with different aspects of emotion. They could be PAD dimensions or their combanition. Affective speech modeling in this 3D space is promising, but still has a long way to go.

7 Acknowledgements

We thank Professor Mehrabian for authorizing us to use the PAD model and the PAD Emotion Scales in our research and offering invaluable comment on this study.

The work in this paper is part of a project supported by The National Science Foundation of China (60433030).

References

1. Rosalind W. Picard: Affective Computing. Cambridge, Mass.: MIT Press(1997)
2. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a New Generation of Databases. Speech Communication, Vol. 40(2003) 33-60
3. Danning Jiang: Acoustic Feature Analysis and Modeling of Emotion Speech. PhD dissertation. Tsinghua University, China (2005)
4. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology: Developmental, Learning, Personality, Social, 14 (1996) 261-292.
5. Patrick Gebhard: ALMA – A layered model of affect. In: Proceedings of the Fourth International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS). Utrecht, the Netherlands (2005) 29-36
6. Xiaoming Li, Haotian Zhou: The Reliability and Validity of the Chinese Version of Abbreviated PAD Emotion Scales. In: International Conference on Affective Computing and Intelligent Interaction. Beijing, China (ACII). Beijing, China (2005)
7. Mehrabian, A.: Emotional Correlates of Preferences for Situation-Activity Combinations in Everyday Life. In: Genetic, Social, and General Psychology Monographs. I23 (4) (1997).
8. Kenneth Hugdahl: Psychophsiology: the mind-body perspective. Cambridge, Mass.: Harvard University Press (1995)
9. I.S. Dhillon, S. Mallela, and D.S. Modha: Information Theoretic Co-clustering. In: Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington. DC, USA (2003) 89-98