Learning to Infer Public Emotions from Large-scale Networked Voice Data

Zhu Ren[1,2], Jia Jia[1,2], Lianhong Cai[1,2], Kuo Zhang[3], Jie Tang[1]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2] TNList and Key Laboratory of Pervasive Computing, Ministry of Education
bamboo.renzhu@gmail.com, {jjia,clh-dcs,jietang}@tsinghua.edu.cn
[3] Sogou Corporation, Beijing, China
zhangkuo@sogou-inc.com

Abstract. Emotions are increasingly and controversially central to our public life. Compared to text or image data, voice is the most natural and direct way to express ones' emotions in real-time. With the increasing adoption of smart phone voice dialogue applications (e.g., Siri and Sogou Voice Assistant), the large-scale networked voice data can help us better quantitatively understand the emotional world we live in. In this paper, we study the problem of inferring public emotions from large-scale networked voice data. In particular, we first investigate the primary emotions and the underlying emotion patterns in human-mobile voice communication. Then we propose a partially-labeled factor graph model (PFG) to incorporate both acoustic features (e.g., energy, f0, MFCC, LFPC) and correlation features (e.g., individual consistency, time associativity, environment similarity) to automatically infer emotions. We evaluate the proposed model on a real dataset from Sogou Voice Assistant application. The experimental results verify the effectiveness of the proposed model.

Keywords: public emotions, acoustic features, correlation features, factor graph model.

1 Introduction

It is an emotional world we live in. Emotions, which are associated with subjective feelings, cognitions, impulses to action and behavior [1], can be recognized from many different information sources, e.g., human voice [3], facial expression [15], physiological signal [16], or their multimodal combination [5]. Compared to individual emotions, **public emotions** pay attention to the major emotions of the public induced by social events. Previous studies have shown the success of using networked text [2] or image data [17] to infer public emotions. Nowadays, with the rapid development of smart phone voice dialogue applications (e.g., *Siri*[1] and *Sogou Voice Assistant*[2]), people can share voice messages to their friends or make requests to the voice

[1] http://www.apple.com/ios/siri/, an intelligent personal assistant and knowledge navigator which works as an application for Apple's iOS.

[2] http://yy.sogou.com, an smart phone voice dialogue application developed by Sogou (one of China's largest internet service providers).

adfa, p. 1, 2011.

assistant easily. Voice is the most direct way to express emotions. And emotions can be conveyed by not only linguistic information but also acoustic information. For example, a user who intends to share happiness with his friend on special days may send voice messages saying "Happy New Year" or "Happy birthday" with pleasant tone of higher pitch. Since people's voice data can be regarded as microscopic instantiations of emotions, the collection of all the public voice data uploaded over a given time period or around a special social event can unveil the trends of public emotions at a macroscopic scale.

Previous researches have been conducted for empirical analyses of emotion based on text or image data from social networks. Some of these analyses focus on public emotions around specific events [17], while others further analyze broader social and economic trends [7]. However, due to the lack of availability of large-scale networked voice data, few have been done in studying public emotions from voice signals. As voice is the fastest and the most natural method of communication [9], it can express people's emotional states in a much more vivid and efficient way. Therefore, using available large-scale networked voice data to perform public emotion analyses can significantly reduce the costs and efforts. Furthermore, it can benefit lots of fields, e.g., improve the user-friendly voice communication applications, or help companies formulate marketing strategies [2].

In this paper, employing a mobile voice assistant application as the basic of our experiments, we systematically study the problem of inferring public emotions from networked voice data. The problem is non-trivial and poses a set of unique challenges. First, the emotion patterns in human-mobile voice communication are quite different from that in human-human voice communication. It is unclear how to identify the underlying emotion patterns behind the human-mobile voice communication. Second, former studies have confirmed that acoustic features [3-5] can reflect the individual's emotions, while in different social environment, the acoustic features might be quite different. Third, technically, how to design a principled model to automatically infer public emotions by considering both the acoustic features and social environment?

To address the above challenges, we make our efforts and make contributions on three aspects:

- **Emotion Patterns.** Based on the observations of networked voice data, we investigate the primary emotions in human-mobile voice communication by combining the linguistic information with acoustic information. Furthermore, we identify two interesting emotion patterns behind the human-mobile voice communication.
- **Features.** Besides the selected acoustic features which can reflect emotions, we take into consideration three social correlation features (individual consistency, time associativity and environment similarity), which can be combined with acoustic features for performance improvement in different social environment.
- **Model.** We formulate our problem into a novel partially-labeled factor graph model (PFG) to infer public emotions by incorporating both acoustic features and correlation features.

Our experiments are based on a real networked voice dataset, which is from *Sogou Voice Assistant*. The experimental results demonstrate that the proposed model can

achieve better performance than alternative method using SVM. Discussion and analysis of the experimental results rationally verify the contribution of combining the acoustic features with the correlation features to improve the performance.

The rest of this paper is organized as follows: Section 2 gives the basic formulation of our problem. Section 3 introduces the data observation and dataset setup. Section 4 presents the PFG model for inferring public emotions. Section 5 carries out the experiments employed to analyze the feature contribution and evaluate the performance of the proposed model. We'd like to show some interesting case studies in this section too. Finally, Section 6 summarizes this paper.

2 Problem Formulation

Fig. 1 gives an illustration of inferring public emotions from networked voice data. Each utterance in the voice dataset not only has its own acoustic features, but also correlations with other utterances, such as individual consistency (blue line), time associativity (pink line), and environment similarity (green line). Part of the utterances in the dataset are labeled with emotions, and our task is predicting the emotions of unlabeled utterances. For further clarification, in this section, we give some essential definitions and subsequently present the problem formulation.

The network of input utterances can be represented as $G = (V, E)$, where $V = \{u_1, \dots, u_N\}$ is the set of $|V| = \mathrm{N}$ utterances, $E \subset V \times V$ is the set of $|E| = \mathrm{K}$ relationships between utterances. Each edge e_{ij} indicates u_i having a correlation with u_j (e.g. u_i and u_j recorded by the same speaker or recorded in the same city within a short time). We aim at learning a model that can effectively infer emotions from networked voice data. For this reason, we first define the speaker's emotions and the partially labeled network as follows.

***Definition** 1*. **Emotions**: The emotion category of an utterance u_i is denoted as $y_i \subset A$, where A is the emotion space that contains the primary emotions in human-mobile communication. The investigation on primary emotions will be described in details in Section 3..

***Definition** 2*. **Partially labeled network**: The partially labeled network is denoted as $G = (V^L, V^U, E, \Gamma)$, where V^L and V^U are respectively the set of labeled and unlabeled utterances with $V^L \cup V^U = V$; E is the correlations between utterances; Γ is an

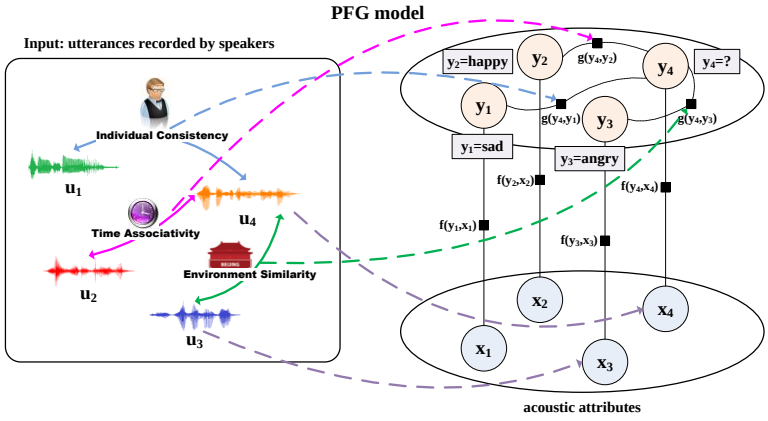

Fig. 1. The illustration of inferring public emotions from large-scale networked voice data using partially labeled factor graph model.

attribute matrix associated with utterances in V with each row corresponding to an utterance, each column representing an attribute and an element x_{ij} denoting the value of the j^{th} attribute of utterance u_i. The label of utterance u_i is denoted as $y_i \subset A$.

Problem. **Learning task**: Given a partially labeled network G, the objective is to infer the emotion categories of utterances by learning a predictive function

$$f: G = (V^L, V^U, E, \Gamma) \rightarrow \boldsymbol{\Lambda} \quad (1)$$

where $\boldsymbol{\Lambda} = \{\boldsymbol{S_1}, \dots, \boldsymbol{S_M}\}$ is the set of inferred results, with each $\boldsymbol{S_m}$ belonging to one emotion category in A.

3 Data Preparation and Emotion Pattern Analysis

3.1 Data Collection and Observation

We collected a corpus of large-scale networked voice data from *Sogou Voice Assistant*. The raw dataset contains 6,891,298 utterances recorded in Chinese by 405,510 users during year 2013. Each utterance has some basic information (e.g. user ID, record time, geographical position) and the corresponding speech-to-text information provided by Sogou Corporation.

For training and evaluating the proposed model for inferring emotions, we firstly need to know the primary emotions of networked voice data, and establish an experimental dataset with emotional labels as ground truth. Due to the massive scale of our dataset, manually annotating the emotion category for each utterance is not practical. Considering that linguistic information contained in voice data can help us understand speaker's emotion, we conduct the investigation on primary emotions as following steps: 1) we screen all the utterances' speech-to-text information and find the emotional words in them; 2) we compute the occurrence frequency of each emotional word, and drop the emotional words appearing few times; 3) in view of previous work on Chinese emotional words categorization [4,6], we classify all the selected emotional words into representative categories.

The classification results show the emotional words finally cluster to five main categories: *Happy*, *Sad*, *Angry*, *Disgusted*, and *Bored*. We can see that these five primary emotions in human-mobile communication are different from Ekman's six basic emotions proposed for human-human communication. Specifically, *Fear* and *Surprise* in Ekman's six emotions are replaced by *Bored*. Following the above steps, from the raw dataset, we finally pick out 48,211 utterances whose speech-to-text information con-

Table 1. Emotional word examples and selected utterance number of each category.

Category	Emotional word examples	Utterance number
Happy	happy ('高兴'), joyful ('快乐'), delighted ('开心'), sweet ('甜蜜'), etc.	12067
Sad	heart-broken ('伤心'), grieved ('痛苦'), sorrow ('悲哀'), miserable ('难受'), etc.	4754
Angry	angry ('生气'), rage ('愤怒'), idiot ('笨蛋'), bastard ('可恶'), etc.	13407
Disgusted	disagreeable ('讨厌'), disgusting ('恶心'), despise ('鄙视'), dissatisfied ('不满'), etc.	4320
Bored	bored ('无聊'), tired ('累'), toilsome ('辛苦'), etc.	13663

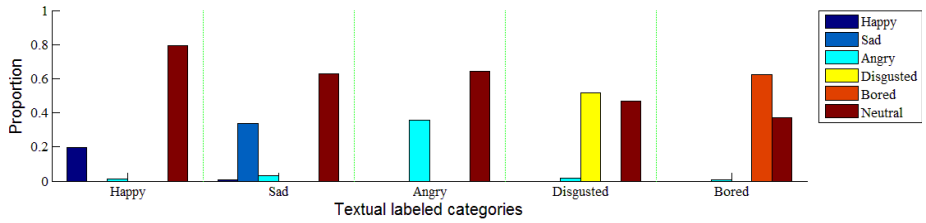

Fig. 2. The proportion of manually labeled emotions in each textual labeled category.

tains emotional words that only belong to one of the five primary emotion categories. We annotate these utterances with that specific category as their textual labels. Table 1 shows the emotional words and selected utterance number of each category.

In order to further explore how an utterance's textual label is consistent with its real emotion, we randomly selected 200 utterances from each textual labeled category (1,000 utterances in total). Then we invite three human labelers to annotate each utterance with an emotion category manually. Besides the above five emotions, we also allow the labelers to give *Neutral* annotation. When they have disagreement, they stop and discuss until they have final agreed views. The manually labeled results are regarded as the real emotions for these utterances.

We observe the distinctions between textual labeled categories and manually labeled emotions. The observation results are quite interesting. The proportion of manually labeled emotions in each textual labeled category is shown as Fig. 2. We can easily find two phenomena as follows:

- **Phenomenon I**: In some cases, the textual labels are not consistent with the real emotions. For example, some of the utterances with textual labels *Happy* are actually *Angry* emotion. It may happen when a user says "I'm really happy with what you said", but actually he is angry and what he says means an irony. This phenomenon indicates the **Emotion Pattern I**: There exist insincerities in human-mobile communication (this pattern exists in human-human communication too). It also means that we cannot use the textual labels as real emotion categories directly.
- **Phenomenon II**: A large part of the utterances, whose speech-to-text information contains emotional words, are actually neutral voice data. This phenomenon indicates the **Emotion Pattern II**: The expressions in human-mobile communication are more rational and implicit. People pay much attention to linguistic information rather than paralinguistic information to express their meanings. It leads us to find that each textual labeled category is approximately consisted by two parts, the real emotional voice data, and the neutral voice data.

Therefore, for each textual labeled category, we can further conduct a 2-cluster classification using acoustic features to separate it into two parts. Then we can establish an experimental dataset with emotional and neutral voices respectively. Furthermore, we can use the manually labeled utterances as the reference of which part is the real emotional voice data while which part is the neutral one.

3.2 Experimental Dataset Setup

According to the above observations, we setup our experimental dataset. Based on previous research about emotional speech analysis [3,4,11], we use 113 acoustic features to conduct the 2-cluster classification on 48,211 utterances:

- Energy features (13): the energy envelop applied with 13 functionals (mean, std, max, min, range, quartile1/2/3, iqr1-2/2-3/1-3, skewness, kurtosis).
- F0 features (13): the fundamental frequency contour, which are extracted using a modified STRAIGHT procedure[10], applied with 13 functionals the same as Energy.
- MFCC features (26): the mean and standard deviation of mel-frequency cepstral coefficients 1-13.
- LFPC features (24): the mean and standard deviation of log frequency power coefficients 1-12, which are extracted using the method in [11] with α=1.4.
- Spectral Centroid (SC) features (13): the spectral centroid contour applied with 13 functionals the same as Energy.
- Spectral Roll-off (SR) features (13): the spectral roll-off contour applied with 13 functionals the same as Energy.
- Syllable Duration (SD) features (11): the syllable duration sequence, which is extracted using the method in [12], applied with 11 functionals (mean, std, max, min, range, quartile1/2/3, iqr1-2/2-3/1-3).

Algorithm 1: The feature selection algorithm.

Input:
features[1..n][1..d]: acoustic features matrix with each row corresponding to an utterance, each column representing one kind of feature
labels[1..n]: the manual labels of n utterances from one text category
threshold: a const in {0.05, 0.01, 0.001, 0.0001, 0.00001, 0.000001}
Output:
The order numbers of the selected feature set

```
dnum ← 0
for j ← 1 to d do
    calculate the p-value for testing the hypothesis of no correlation between
    features[1..n][j] and labels[1..n]
    if p-value < 0.05 then
        dnum ← dnum + 1
        fset[dnum].p ← p-value
        fset[dnum].ind ← j
     end if
end for
sort structure array fset[1..dnum] along its element p in ascending order
initialize tag[1..dnum] = 0
calculate a matrix P of p-values for array features[1..n][1..d]
current ← 1
tag[current] ← current
repeat
    for j ← 1 to dnum do
        if tag[j] = 0 then
            if P[fset[current].ind][fset[j].ind] < threshold then
                tag[j] ← current
            end if
        else
            if P[fset[current].ind][fset[j].ind] < P[fset[tag[j]].ind][fset[j].ind] and tag[j] != j then
                tag[j] ← current
            end if
        end if
    end for
    find the minimum value of k satisfying the equation tag[k] = 0
    current ← k
    tag[current] ← current
until all the elements in tag[1..dnum] are not zero
find the same values as in tag[1..dnum] without repetitions into tset[1..gnum]
return fset[tset[1..gnum]].ind
```

Table 2. The results of feature selection and emotion classification.

Category	Manual labeled sample numbers	Feature selection		Emotion Classification	
		Threshold	Selected feature set	F1-Measure	Sample numbers in final dataset
Happy	H: 102 NH: 161	0.001	f0_quartile3, lfpc8_mean, energy_range, sc_skewness, f0_kurtosis, lfpc4_std, mfcc11_mean, sr_max	70.78%	H: 5721 NH: 6346
Sad	S: 167 NS: 133	0.01	lfpc9_std, syldur_quartile3, mfcc3_mean	75.84%	S: 2562 NS: 2192
Angry	A: 100 NA: 129	10E-4	mfcc2_std, lfpc11_mean, lfpc10_std, mfcc9_mean, energy_skewness, mfcc8_mean, mfcc6_std, mfcc12_mean, mfcc2_mean, mfcc7_std	70.97%	A: 7892 NA: 5515
Disgusted	D: 103 ND: 97	0.01	lfpc7_mean, mfcc5_std, mfcc9_mean, lfpc11_std, syldur_iqr1-2	74.64%	D: 2328 ND: 1992
Bored	B: 125 NB: 75	10E-4	sr_std, f0_iqr1-2, mfcc5_std, mfcc4_mean, mfcc6_mean, mfcc3_mean, lfpc7_std, lfpc9_mean, energy_iqr2-3, mfcc6_std, energy_iqr1-2	79.20%	B: 6968 NB: 6695

All the spectral features (MFCC, LFPC, SC, SR) are extracted from voiced segments of the utterances with the 20ms frame length and 10ms frame shift. Each kind of feature is normalized first, hence the mean is zero and the standard deviation is one.

We make use of the correlation coefficients and their significances between acoustic features and manual labels for the feature selection (Algorithm 1 gives the details). The manual labels of utterances are defined as X / NX, where X∈{H, S, A, D, B} and NX is *Neutral*. By changing the threshold in Algorithm 1, we can obtain diverse feature sets. As we pay more attention to clustering most of the positive samples into one class, we apply the F1-Measure of positive samples to evaluate the performance of clustering. For simplicity, we run the classic k-means clustering for each category, and compute the F1-Measure by using the Hungarian algorithm [8] to assign which class is corresponding to the positive samples. The best classification results and the selected feature sets they used are summarized in Table 2.

By the above method, we finally establish an experimental dataset with emotional labels as ground truth. The dataset contains 48,211 utterances consisted of five primary emotions as well as *Neutral*: *Happy* (5721), *Sad* (2562), *Angry* (7892), *Disgusted* (2328), *Bored* (6968), and *Neutral* (22740).

The emotional labels in our experimental dataset certainly still have some errors. But in the statistical level, it can be ignored in some ways such as using large-scale data. The most importance is the above method can help us avoid the impossible mission of manual annotation on large-scale networked voice data. Since our prime concern is the accuracy of public emotions inferring, which is quite different from the task of individual emotion recognition, we believe the above method provide the good balance between efficiency and performance.

4 Proposed Method

4.1 Prediction Model

As networked voice data are disposed in this paper, we take advantage of some social correlation information to improve the performance and identify three kinds of correlation features:

- Individual Consistency (IC): whether two utterances are recorded by the same speaker.
- Time Associativity (TA): whether two utterances are recorded within the same hour of one day.
- Environment Similarity (ES): whether two utterances are recorded in the same city.

Since the correlations between utterances are hard to be modeled by traditional classifiers such as SVM, we utilize a partially-labeled factor graph model (PFG) [13] to learn and infer public emotions. All the utterances recorded by ordinary speakers can be formalized as variables and observation factor functions in a factor graph, basing on the theory of FGM [14]. Each utterance u_i can be mainly described as one kind of primary emotion, which can be mapped as an emotional node n_i in the PFG model. The labels of emotional nodes are denoted as $Y = \{y_1, \dots, y_N\}$, where y_i is a hidden variable associated with n_i. As the emotions in G are partially labeled, they can be divided into two subset Y^L and Y^U corresponding to the labeled and unlabeled emotions. For each emotional node n_i, we define the emotional attributes into a vector $\boldsymbol{x_i}$, considering that the speaker-independent acoustic features may contain information about emotions. At the same time, we can also find the basic intuition that the relationships between the utterances (correlation features) can constitute the correlations between hidden variables in our model. Based on the above intuitions, we define the following two factors:

- **Attribute factor**: $f(y_i, \boldsymbol{x_i})$ represents the posterior probability of the emotion y_i given the attribute vector $\boldsymbol{x_i}$.
- **Correlation factor**: $g(y_i, \mathrm{R}(y_i))$ denotes the correlation among the emotions, where $\mathrm{R}(y_i)$ is the set of correlated emotions to y_i.

Given a partially-labeled network $G = (V^L, V^U, E, \Gamma)$, we can define the joint distribution over Y as

$$\mathrm{P}(Y|G) = \prod_i f(y_i, \boldsymbol{x_i}) g(y_i, \mathrm{R}(y_i)) \tag{2}$$

Since the two factors can be instantiated in different kinds of ways, we only give a general definition for them by using exponential-linear function. In particular, we define the attribute factor as

$$f(y_i, \boldsymbol{x_i}) = \frac{1}{\mathrm{Z}_\alpha} \exp\{\alpha^{\mathrm{T}} \cdot \boldsymbol{x_i}\} \tag{3}$$

where α is a weighting vector of Γ and Z_α is a normalization factor.

The correlation factor can be naturally modeled in a Markov random field. Thus, by the fundamental theorem of random fields, it can be defined as

$$g(y_i, \mathrm{R}(y_i)) = \frac{1}{\mathrm{Z}_\beta} \exp\left\{\sum_{y_j \in \mathrm{R}(y_i)} \beta_{\mathrm{ij}} \cdot h_{ij}(y_i, y_j)\right\} \quad (4)$$

where $h_{ij}(y_i, y_j)$ is a feature function that captures the correlation between emotional nodes n_i and n_j; β is the weighting of this function; Z_β is also a normalization factor.

Finally, the joint probability defined in (2) can be written as

$$\mathrm{P}(Y|G) = \frac{1}{\mathrm{Z}} \exp\left\{\sum_{y_i \in Y} \left[\alpha^{\mathrm{T}} \cdot \boldsymbol{x_i} + \sum_{y_j \in \mathrm{R}(y_i)} \beta_{\mathrm{ij}} \cdot h_{ij}(y_i, y_j)\right]\right\} \quad (5)$$

where $\mathrm{Z} = \mathrm{Z}_\alpha \mathrm{Z}_\beta$ is a normalization factor.

Learning the predictive model is to estimate a parameters configuration $\varphi = (\{\alpha\}, \{\beta\})$ from the partially-labeled dataset, so that it can maximize the log-likelihood objective function $\Theta = \log \mathrm{P}(Y|G)$, i.e. $\varphi^* = \arg\max \Theta(\varphi)$.

4.2 Model Learning

After learning the parameter values, we turn to address the problem of estimating the remaining free φ and inferring speakers' emotions. Specifically, we first calculate the gradient of each parameter with regard to the objective function:

$$\frac{\partial \Theta(\varphi)}{\partial \alpha_{\mathrm{j}}} = \mathrm{E}\left[f_j(y_{ij}, x_{ij})\right] - \mathrm{E}_{\mathrm{P}_{\alpha_{\mathrm{j}}}(y_i|x_{ij},G)}\left[f_j(y_{ij}, x_{ij})\right] \quad (6)$$

where $\mathrm{E}\left[f_j(y_{ij}, x_{ij})\right]$ is the expectation of feature function $f_j(y_{ij}, x_{ij})$ given by the data distribution and $\mathrm{E}_{\mathrm{P}_{\alpha_{\mathrm{j}}}(y_i|x_{ij},G)}\left[f_j(y_{ij}, x_{ij})\right]$ is the expectation of feature function $f_j(y_{ij}, x_{ij})$ under the distribution $\mathrm{P}_{\alpha_{\mathrm{j}}}\left(y_i|x_{ij}, G\right)$ given by the estimated model. Similar gradients can be derived for parameter β. Then we update the parameters by $\varphi_{\mathrm{j}}^{\mathrm{new}} = \varphi_{\mathrm{j}}^{\mathrm{old}} + \gamma \cdot \frac{\partial \Theta(\varphi)}{\partial \varphi}$, where γ is the learning rate.

Given the observed value $\boldsymbol{x}$ and the learned parameters φ, the inference task is to find the most likely y,as follows

$$y^* = \arg\max_y \mathrm{p}(y|\boldsymbol{x}, \varphi) \quad (7)$$

Finally, we utilize the loopy belief propagation to compute the marginal probability of each emotional node and then predict the type of an unlabeled emotion as the label with largest marginal probability.

5 Experiments and Discussions

5.1 Experimental Setup

We use the dataset described in Subsection 3.2 in our experiments, which contains 48,211 utterances that carefully chosen from networked voice data recorded by 25,370 speakers and labeled with one of the six primary emotions. We perform five-fold cross validation and quantitatively evaluate the performance of inferring public emotions in term of *Accuracy* and *F1-Measure* computed as

$$\text{Accuracy} = \frac{\sum_i \text{correctly predicted sample number of category } i}{\sum_i \text{sample number of category } i} \quad (8)$$

$$\text{F1-Measure} = \frac{\sum_i \text{F1-Measure of category } i}{\text{category number}} \tag{9}$$

For the purpose that justifying whether the correlation information can help infer public emotions from utterances, we define a baseline method using the classic machine learning technique Support Vector Machine (SVM). We compare the performance achieved by PFG model utilizing both acoustic features and correlation features (*Individual Consistency (IC)*, *Time Associativity (TA)*, and *Environment Similarity (ES)*) with the baseline method.

5.2 Results and Discussions

Performance Comparison. Table 3 shows the *Accuracy* and *F1-Measure* for proposed PFG and SVM. The *Accuracy* of the proposed PFG model achieves 86.11%, while the SVM model achieves only 49.15%. For *F1-Measure* shown in Table 3, PFG also shows clearly the best performance and yields an 27.04-83.51% improvement compared with SVM. These results demonstrate the effectiveness of our proposed method on inferring emotions from networked voice data. Furthermore, the proposed PFG model utilizes both acoustic features and correlation features, while SVM model cannot express the relationships between utterances. So the experimental results also verify that the correlation information among utterances can compensate the deficiency of acoustic features and help infer public emotions in our problem.

Table 3. Performance of emotion prediction with different method.

Category \ Model	F1-Measure (%)		
	SVM	PFG	
		Acoustic	All
Happy	7.83	56.27	**84.92**
Sad	6.50	49.65	**69.46**
Angry	8.56	69.34	**92.07**
Disgusted	8.47	26.07	**67.64**
Bored	5.04	51.24	**68.02**
Neutral	64.98	89.74	**92.02**
Average	16.90	57.05	**79.02**
Accuracy (%)	49.15	73.86	**86.11**

Feature Contribution Analysis. Fig. 3 shows the *F1-Measure* of each kind of acoustic or correlation feature when inferring emotions. Comparing the *F1-Measures* of different correlation features, we find that *TA* makes the improvement in the majority of the emotion categories, while *ES* benefits the prediction of several categories as well. Since the mean number of utterances per speaker is 1.9 and 62.03% of the utterances are recorded by the speakers who only have one utterance in the dataset, there

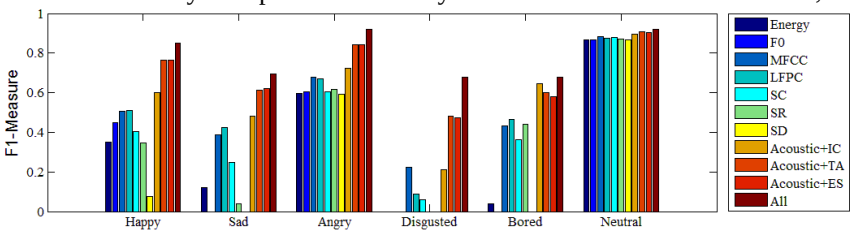

Fig. 3. Feature contribution analysis.

are lesser *IC* correlations than *TA* and *ES*, which leads to its lower performance. The aforementioned results confirm that public emotions are closely related to time and environment. For acoustic features, the spectral features (*MFCC*, *LFPC*) make a great contribution to inferring most emotions, which is consist with the feature selection results described in Table 2.

Case Study. To further demonstrate the effectiveness of the proposed model, we would like to show two interesting case studies. Since Beijing, Shanghai and Shenzhen are the top 3 active cities with the most users in *Sogou Voice Assistant*, we use the utterances that respectively uploaded in these three cities as our case study data. The proposed PFG model is used to infer emotions from them. By analyzing the results, we find the trends of public emotions related to specific time period or social event. We take positive emotion *Happy* and negative emotion *Angry* as examples:

- We all know that Beijing suffered the severest fog and haze in history during the week of Jan 21 to 27, 2013. What was the major public emotion inferred from the voice data in that period in Beijing? The answer is people felt less happy and more agonizing day by day, shown in Fig.4a(1). Comparing with Shanghai and Shenzhen's results shown as Fig. 4a(2) and Fig. 4a(3), Beijing obviously had more people in bad moods due to the environmental problem. These results are rational and common in our daily life.
- A strong earthquake struck the southwestern Chinese province of Sichuan on April 20, 2013. Were the public emotions affected by the emergency? The answer is yes. Taking Beijing as an example, we can find an increasing of negative emotion from April 20 (Fig. 4b). After the earthquake, people were generally worried about the disaster and suffering from the tragedy of losing compatriots, so their anxieties caused the negative public emotions. These results also indicate the networked voice data can reflect the changes of public emotions on emergency in real-time.

6 Conclusions

With the increasing adoption of smart phone voice dialogue applications, we can now use the large-scale networked voice data to achieve the goal of inferring public emotions. Compared to text or image data, the most advantage of voice is that it is the most natural and direct way to express ones' emotions in real-time. Our main contributions are: 1) we reveal the five underlying primary emotions in human-mobile communication, which are quite different from the widely-used Ekman's six emotions in human-human communication; 2) we experimentally analyze the fundamental acoustic features, and combine them with social correlation features that can better reflect emotions in different social environment; 3) we formulate the problem into a PFG model for inferring public emotions from large-scale networked voice data, turning out good results.

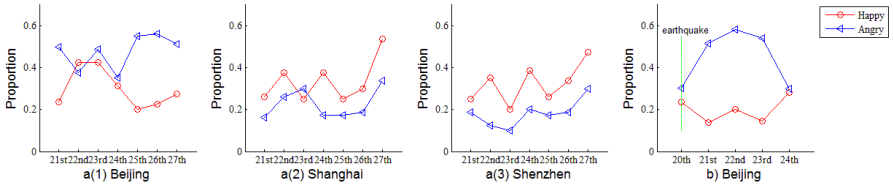

Fig. 4. Case study: a) The major emotional trends in three cities from Jan 21 to 27, 2013. b) The major emotional trends in Beijing from April 20 to 24, 2013.

For future works, we are planning to investigate and model more social phenomenon such as conformity in human-mobile communication for further improving the inferring accuracy.

Acknowledgements. This work is supported by the National Basic Research Program of China (2012CB316401), National Natural, and Science Foundation of China (61370023, 61003094). And this work is partially supported by 973 Program of China (2013CB329304). We also thank Microsoft Research Asia-Tsinghua Univertity Joint Laboratory for its support.

References

1. Plutchik, R: Emotion: A Psychoevolutionary Synthesis. Harper & Row, New York (1980)
2. Tang, J., Zhang, Y., Sun, J., Rao, J., Yu, W., Chen, Y., and Fong, A.: Quantitative study of individual emotional states in social networks. IEEE Transactions on Affective Computing 3(2), 132-144 (2012)
3. Wu, D., Parsons T. D., and Narayanan, S.: Acoustic feature analysis in speech emotion primitives estimation. In: Proc. of INTERSPEECH 2010, Makuhari, Japan, pp. 785-788 (2010)
4. Cui, D.: Analysis and Conversion for Affective Speech (in Chinese): [doctoral dissertation], Beijing: Tsinghua University (2007)
5. Nicolaou, M. A., Gunes, H., Pantic, M.: Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing 2(2), 92–105 (2010)
6. Mei, J.: Tongyici Cilin (version 2, in Chinese). Shanghai Dictionary Press, Shanghai (1996)
7. Bollen, J., Mao, H., Pepe, A.: Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In: Proc. AAAI 2011, San Francisco, California, USA, pp. 450-453 (2011)
8. Kuhn, H. W.: The Hungarian Method for the assignment problem. Naval Research Logistics Quarterly 2, 83-97 (1955)
9. Ayadi, M. E., Kamel, M. S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Patter Recognition 44, 572-587 (2011)
10. Kawahara, H., Cheveigne, A. De, Banno, H., Takahashi, T., and Irino, T.: Nearly Defect-free F0 Trajectory Extraction for Expressive Speech Modifications based on STRAIGHT. In: Proc. of INTERSPEECH 2005, Lisboa, pp. 537-540 (2005)
11. Nwe, T., Foo, S., Silva, L. De: Speech emotion recognition using hidden Markov models. Speech Commun 41, 603–623 (2003)
12. Wang, D., and Narayanan, S.: An acoustic measure for word prominence in spontaneous speech. IEEE Transactions on Speech, Audio, and Language Processing 15(2), 690-701 (2007)
13. Tang, W., Zhuang, H., and Tang, J.: Learning to infer social ties in large networks. In ECML/PKDD'11, pp. 381–397 (2011)
14. Frey, B., and Dueck, D.: Mixture modeling by affinity propagation. In Y. Weiss, B. Schölkopf, and J. Platt, editors, NIPS, pp. 379–386 (2006)
15. Fasel, B., and Luettin, J.: Automatic facial expression analysis: a survey. Pattern Recognition 36(1), 259–275 (2003)
16. Fairclough, S. H.: Fundamentals of physiological computing. Interacting with Computers 21, 133–145 (2009)
17. Jia, J., Wu, S., Wang, X., Hu, P., Cai, L., Tang, J.: Can We Understand van Gogh's Mood? Learning to Infer Affects from Images in Social Networks. In: Proc. of ACM Multimedia, Nara, Japan (2012)