INFERRING USERS' EMOTIONS FOR HUMAN-MOBILE VOICE DIALOGUE APPLICATIONS

Boya Wu[1,2,3], *Jia Jia*[1,2,3], *Tao He*[4], *Juan Du*[1], *Xiaoyuan Yi*[1], *Yishuang Ning*[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Key Laboratory of Pervasive Computing, Ministry of Education
[3]Tsinghua National Laboratory for Information Science and Technology (TNList)
[4]College of Computer Science, Sichuan University, Chengdu 610065, China
stella.1991@163.com,jjia@mail.tsinghua.edu.cn,ithet1007@gmail.com
duj09225@gmail.com,yxyz2012yxy@163.com,ningys13@mails.tsinghua.edu.cn

ABSTRACT

In this paper, we tackle the problem of inferring users' emotions in real-world Voice Dialogue Applications (*VDAs*, Siri[1], Cortana[2], etc.). We first conduct an investigation, indicating that besides the text information of users' queries, the acoustic information and query attributes are very important in inferring emotions in VDAs. To integrate the information above, we propose a Hybrid Emotion Inference Model (**HEIM**), which involves a Latent Dirichlet Allocation (LDA) to extract text features and a Long Short-Term Memory (LSTM) to model the acoustic features. To further improve accuracy, a Recurrent Autoencoder Guided by Query Attributes (RAGQA) which incorporates other emotion-related query attributes is proposed in HEIM to pre-train LSTM. The accuracy of HEIM on a data set collected from Sogou Voice Assistant[3] (Chinese Siri) containing 93,000 utterances achieves 75.2%, which outperforms state-of-the-art methods for 33.5-38.5%. Specifically, we discover that on average, the acoustic information enhances the performance for 46.6%, while query attributes further enhance the performance for 6.5%.

Index Terms— Emotion, voice dialogue applications, Long Short-Term Memory

1. INTRODUCTION

Voice dialogue applications (VDAs) are gaining popularity worldwide (Siri[1], Cortana[2], Google Now[4], etc.). Statistics from Nuance[5] indicate that about 57% of people worldwide having VDAs use it at least once a day. In VDAs, users' queries are conveyed through voices, which contain not only query contents, but also users' emotions that can significantly help VDAs to humanize responses.

At present, VDAs mainly generate responses by text-based natural language processing (NLP) techniques [1]. However, is the text information of users' queries sufficient to reflect users' emotions? We employ a data set consisting of 1,000 utterances collected from Sogou Voice Assistant[3] (Chinese Siri) and conduct an emotion labeling experiment (three human labelers, six emotion categories defined in [2]). Comparing results of labelers reading the texts only, we discover that when labelers read and listen to the utterances simultaneously, 47.8% of the emotion labeling results happen to change: 21.8% from unclear to certain and 26.0% from one to another. These utterances are very short with only 7.27 characters on average and lack of emotional keywords. For example, when a user say 'Oh why do you come here', she may be happy because her friends give her a surprise, or she may be angry the other way because the unforeseen incidents intrude on her day. The results indicate that in VDAs, the *acoustic information* of users' queries should not be ignored.

Meanwhile, there are tremendous amounts of users in VDAs, bringing in a great diversity of users' dialects and expression preferences. This diversity increases the difficulty of inferring users' emotions. However, in the specific situation of VDAs, is there any other emotion-related attributes that can assist us to reduce the difficulty? Recent researches have verified there is a topical or geographical dependency in users' behaviors. [3] analyzes user behaviors in social media systems from topics related to users' interests. [4] confirms that the physical distance between locations is a strong constraint on the adoption of hash tags. But the application of these dependencies in terms of emotions is still largely undeveloped. We conduct an investigation and discover that these dependencies also exist in VDAs. Shown in Figure 1, the distributions of emotions vary significantly for different query topics and users' locations. We take them into consideration and summarize them as *query attributes*.

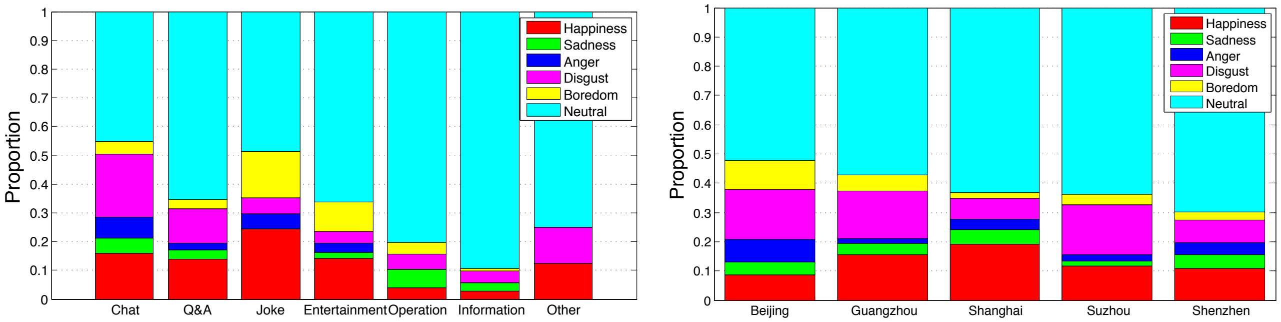
[1] http://www.apple.com/ios/siri/

[2] http://www.microsoft.com/en-us/mobile/campaign-cortana/

[3] http://yy.sogou.com

[4] http://www.google.com/landing/now/

[5] an American multinational computer software technology corporation that provides speech and imaging applications



(a) The emotion distribution of utterances of different query topics.

(b) The emotion distribution of utterances of different cities.

Fig. 1. Query attributes (query topical dependency, geographical dependency) are closely related to the emotion expressions.

In this paper, unlike most of previous works which tackle the emotion inference problem on acted corpora, we address the problem in real-world VDAs. We collect a real-world voice data set from Sogou Voice Assistant, which contains 6,891,298 Mandarin utterances recorded by 405,510 users. We propose a Hybrid Emotion Inference Model (**HEIM**). HEIM incorporates the text information, the acoustic information of users' queries and query attributes in a joint framework. We adopt Latent Dirichlet Allocation (LDA) [5] which is widely used in the text-based sentiment analysis and achieves good performance to model the text information. For the acoustic information, we propose a Long Short-Term Memory (LSTM) to model it. LSTM is capable of learning contextual dependencies, so it is well-suited to deal with time sequences like voices. Moreover, to further improve the accuracy, we propose an unsupervised Recurrent Autoencoder Guided by Query Attributes (RAGQA) to pre-train LSTM, which leverages query attributes into modeling. Since it is an unsupervised method, massive-scale unlabeled data in VDAs can be used to find better parameters. The accuracy of HEIM on a data set collected from Sogou Voice Assistant containing 93,000 utterances achieves 75.2%, which outperforms state-of-the-art methods for 33.5-38.5%. Specifically, we discover that on average, the acoustic information enhances the performance for 46.6%, while query attributes further enhance the performance for 6.5%. Besides, to demonstrate the adaptability of our method, we conduct experiments on the public Berlin Emotional Database (Emo-DB) [6]. Our method easily adapts to utterances of other languages and performs well. At last, an interesting case study based on 93,000 utterances is given to show the application of our work.

2. RELATED WORK

Previous researches concerning speech emotion recognition have unveiled that the acoustic information is closely related to speakers' emotions. [7] presents a wide range of features employed for speech emotion recognition and the acoustic characteristics of those features. [8] proposes an speech emotion recognition algorithm based on binary decision tree and Support Vector Machine (SVM). Similar works can be found in [9, 10]. But notably: 1) These researches mainly focus on inferring emotions from acted corpora. Few have been done to address the problem in real-world VDAs in mobile environment. 2) Instead of acoustic feature sequences, they mainly leverage the statistical values of acoustic features. Thus these methods are not able to capture the contextual information of utterances, which is very important because utterances evolve as time goes by.

Recently, Long Short-Term Memory (LSTM) is gaining its popularity. It learns to bridge minimal time lags in excess of 1,000 discrete time steps and is well-suited to learn from experience to process time sequences [11]. As voice is a typical kind of time sequences, LSTM can capture the correlations among frames and works efficiently on voices.

3. FORMULATION

Given a set of utterances $\mathbf{U}$, for each utterance $\mathbf{u} \in \mathbf{U}$, we denote $\mathbf{u} = \{\mathbf{x}, \mathbf{d}, l_c\}$. $\mathbf{x}$ is the set of acoustic features extracted from every frame. $\mathbf{d}$ represents the text of utterance. l_c represents the query attributes (query topic and user's location) provided by Sogou Corporation.

Definition. **Emotions**. Previous research [2] discovers that in human-mobile interaction, the emotion categories are different from theories about emotions related to facial expressions. According to their findings, we adopt {*happiness*, *sadness*, *anger*, *disgust*, *boredom* and *neutral*} as the emotional space and denote it as $\mathbf{E_S}$, $S = 6$.

Problem. **Learning task.** Given utterances set $\mathbf{U}$, we aim to infer the emotion for every utterance $\mathbf{u} \in \mathbf{U}$:

$$f : \mathbf{u} = \{\mathbf{x}, \mathbf{d}, l_c\} \rightarrow s \tag{1}$$

where $s \in \mathbf{E_S}$.

4. METHOD

In this paper, we incorporate the text information, the acoustic information of users' queries and query attributes in a joint framework named **HEIM** to infer users' emotions in VDAs. HEIM combines the text features generated by LDA with the acoustic features generated by LSTM, and uses a softmax classifier to make inferences. Besides, an unsupervised model named Recurrent Autoencoder Guided by Query Attributes (RAGQA) which leverages the emotion-related query attributes is proposed to pre-train LSTM.

To model the text information of users' queries, Latent Dirichlet Allocation (LDA) [5] is widely used in the text-based sentiment analysis and achieves good performance [12, 13]. We adopt the LDA method used in [13] to generate the text features. Given utterance $\mathbf{u}$'s text $\mathbf{d}$, it outputs a vector $g = \{g_1, g_2, ...g_K\}$, where K is the length of the vector. K is an adjustable parameter, and in our work we set $K = 20$ empirically. In the following subsections, we focus on introducing the modeling of the acoustic information of users' queries and query attributes.

4.1. LSTM pre-trained by RAGQA

Traditional machine learning methods used for speech recognition (e.g., SVM, KNN, CNN) mostly regard the statistical values of acoustic features as the input. For example, a kind of time-domain acoustic features called *Energy* is applied with functions like *max*, *min* and *mean* from the *Energy* sequence extracted from every frame. However, since voice is a kind time sequence signals which evolves as time goes by, the statistical values of acoustic features are insufficient to reflect the subtle changes of voice signals.

To employ the whole time sequences, Long Short-Term Memory (LSTM) has been proved to be an effective method [11]. It captures the correlations among frames, thus well-suited to deal with voices. We leverage a Long Short-Term Memory (LSTM) to generate the high-level representations of acoustic feature sequences. For every utterance, we extract 7 kinds of acoustic features which involve frequency-domain features (*Mel Frequency Cepstrum Coefficient*, *Log Frequency Power Coefficients*, etc.) and time-domain features (*Energy*, *F0*, etc.). In our work, the model contains an input layer and an recurrent layer. Given the input $\{x_t, h_{t-1}, c_{t-1}\}$ at time t, the current high-level representations of acoustic feature sequences refer to the activation h_t of the recurrent layer. It is calculated by the following equations [14]:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \tag{2}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \tag{3}$$

$$c_t = f_t c_{t-1} + i_t \mu(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{4}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{5}$$

$$h_t = o_t \mu(c_t) \tag{6}$$

In the functions above, $W_{\alpha\beta}$ represents the weight matrix connecting β layer to α layer and b_α represents the bias vector. i, o, f and c are the input gate, forget gate, output gate and memory cells. σ is the sigmoid function. For μ, we use a hyperbolic tangent function $f(x) = 1.7159\tanh(\frac{2}{3}x)$, which has been proved to be capable of improving convergence [15].

As t evolves, LSTM computes h_t iteratively. Finally we obtain the output h_T as high-level representations of acoustic feature sequences.

To further improve the accuracy, we propose an unsupervised method named Recurrent Autoencoders Guided by Query Attributes (RAGQA) to pre-train LSTM, which is illustrated in Figure 2. This pre-training process uses query attributes to help LSTM find a better solution in large parameters space.

A large collection of unlabeled utterances are given as the input of RAGQA. Like traditional recurrent autoencoder frameworks, the proposed RAGQA has the encoder part and the decoder part. The encoder part maps the input into the hidden layer, and the decoder part maps the activation of the hidden layer into the output layer. The structure of RAGQA's encoder is the same as that of LSTM, while the decoder is a nonlinear mapping. Worthy of attention, the traditional autoencoder only reconstructs the input x_t into $\hat{x}_t$. In RAGQA, we also reconstruct l_c and x_{t+1} into $\hat{l_c}$ and $\hat{x}_{t+1}$, where l_c represents the query attributes.

RAGQA aims to learn a function to make reconstruction $y_t = [\hat{x}_t,\ \hat{l_c},\ \hat{x}_{t+1}]$ as similar as $z_t = [x_t,\ l_c,\ x_{t+1}]$. For convenience, we define the set of the encoder parameters as $\theta = \{W_{\alpha\beta}, b_\alpha\}$ and the set of the decoder parameters as $\theta' = \{W', b'\}$, where W' is the weight matrix and b' is the bias. The training target of RAGQA can be formalized as minimizing the cost function (7):

$$\arg\min_{\theta,\theta'} \|y_t - z_t\|^2 + \frac{\lambda}{2}\|\xi\|^2 \tag{7}$$

$$\|\xi\|^2 = (W')^2 + \sum_{\alpha}\sum_{\beta}(W_{\alpha\beta})^2 \tag{8}$$

where λ is the parameters of weight decay. The reconstruction y_t is calculated in the decoder of the RAGQA by

$$y_t = \mu(W'h_t + b') \tag{9}$$

where h_t is the activation of the encoder and the calculation is the same as that of LSTM.

To train RAGQA, we apply the stochastic gradient descent method. After the training process, θ is used as the initial values of LSTM's parameters.

4.2. Hybrid emotion inference learning

HEIM combines text features g generated by LDA and high-level representations of acoustic features h_T generated by LSTM together, and then feeds them to a softmax classifier to calculate the probability of every emotion category. For emotion category s, the learning target of HEIM can be summarized as minimizing the cost function (10):

$$J_{(\theta,\mathbf{v})} = -\log p(s|h_T; g) + \frac{\lambda}{2}\|\epsilon\|^2 \tag{10}$$

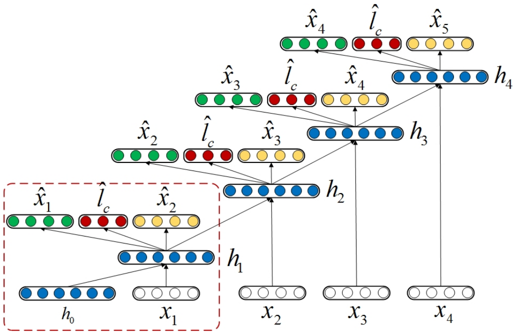


Fig. 2. A sequence feature learning example whose length is 5: $\{x_1, x_2, x_3, x_4, x_5\}$. RAGQA maps the activations of the recurrent layer to $\hat{x}_t$, $\hat{l}_c$, $\hat{x}_{t+1}$.

$$p(s|h_T; g) = \frac{v_s^h \cdot h_T + v_s^g \cdot g}{\sum_{q=1}^{S} \exp^{v_q^h \cdot h_T + v_q^g \cdot g}} \tag{11}$$

$$\|\epsilon\|^2 = \sum_{q=1}^{S}((v_q^h)^2 + (v_q^g)^2) + \sum_{\alpha}\sum_{\beta}(W_{\alpha\beta})^2 \tag{12}$$

where $\mathbf{v}$ represents the weight matrix that connects the recurrent hidden layer to the softmax layer.

After training the model, we can get the emotion for the utterance by finding the maximum probability $p(s|h_T; g)$.

5. EXPERIMENTS

5.1. Experimental setup

We establish a corpus of voice data from Sogou Voice Assistant[3] (Chinese Siri) containing 6,891,298 Mandarin utterances recorded by 405,510 users in 2013. Every utterance is assigned with its corresponding speech-to-text information, query topic and user's location provided by Sogou Corporation. The query topics are divided into seven categories: *Chat*, *Q&A*, *Information*, *Entertainment*, *Operation*, *Joke* and *Others*. The users come from more than 50 cities in China.

Due to the massive scale of our data set, manually labeling the emotion for every utterance is not practical. We randomly sample 3,000 utterances from the data set and invite three human labelers to label the emotions. The labelers are well trained and asked to label the emotions by reading and listening to the utterances simultaneously. When labelers have different opinions about the same utterance, they stop and discuss. If they cannot reach an agreement, the utterance is labeled *unclear* and discarded. Finally, 2,942 utterances are labeled. The emotion distributions of these utterances are: *Neutral:* 61.3%, *happiness:* 13.2%, *disgust:* 13.0%, *boredom:* 4.8%, *Anger*: 3.9% and *sadness:* 3.8%. Besides the labeled data, 90,000 unlabeled data are employed as the supplement for the labeled data in the pre-training process.

5.2. Experimental results

5.2.1. Performance compared to baseline methods.

Four methods are chosen as the comparison methods: Naive Bayesian (NB)[6], K-Nearest Neighbors algorithm (KNN)[6], Support Vector Machine (SVM)[7] and Deep Sparse Neural Network (DSNN). Because NB, KNN, SVM and DSNN can not handle time sequences, we compute the statistical values of the acoustic feature sequences. Table 1 shows the comparison results. The proposed HEIM outperforms all the baseline methods: +38.5% compared with NB, +34.9% compared with KNN, +36.2% compared with SVM and +33.5% compared with DSNN. Possible reasons can be summarized as follows: 1) For baseline methods, they can only accept the statistical values of the acoustic feature sequences. However, HEIM is capable of modeling acoustic feature sequences directly, so it captures the correlations between frames and better depicts the acoustic information. 2) For NB, KNN and SVM, these methods can only model the query attributes of labeled data. In HEIM, however, query attributes of massive-scale unlabeled data can be used to pre-train LSTM, helping LSTM find a better solution in the large parameters space.

However, when inferring utterances labeled *Neutral*, the performance of HEIM is lower than baseline methods. We assume that this is because *Neutral* utterances hold a rather large proportion of training samples. So it is natural that traditional machine learning methods like SVM turn out better. However, the performance of baseline methods for inferring utterances labeled other emotions is far from satisfaction. Besides, we analyze utterances that are classified into wrong emotion categories. Interestingly, we find out that utterances labeled *Disgust* and *Boredom* are mixed together. 33.6% of *Disgust* utterances are classified into *Boredom*, while *27.3%* of *Boredom* utterances are classified into *Disgust*. Utterances of these two types of emotions are both slow with low *F0* and *Energy*. Similar phenomena has also been reported by [7].

5.2.2. Modality contribution analysis.

We then conduct a series of comparison experiments to verify whether the acoustic information and query attributes benefit inferring users' emotions in VDAs. The text features are generated by LDA. The acoustic features are processed by LSTM. The query attributes are leveraged by RAGQA.

Experimental results are illustrated in Figure 3. The average performance of LDA which infers emotions from the text information of users' queries is unsatisfactory: 0.2206 in terms of F1-Measure. However, LSTM which only uses the acoustic information outperforms LDA (+41.5%), indicating that when inferring emotions in VDAs, the acoustic information of users' queries may be more crucial than the

[6] Tools provided by MATLAB, a widely used software developed by MathWorks, Inc.

[7] LIBSVM, a library for support vector machines.

Table 1. The F1-Measure of inferring emotions in VDAs.

Method	Happiness	Sadness	Anger	Disgust	Boredom	Neutral	Average
NB	0.4000	0.3617	0.2619	0.2514	0.2740	0.6548	0.3673
KNN	0.4693	0.3220	0.3947	0.2967	0.2132	0.7191	0.4025
SVM	0.4410	0.2625	0.4199	0.2787	0.1968	**0.7419**	0.3901
DSNN	0.4219	0.3238	0.4594	0.3412	0.2446	0.7111	0.4170
HEIM	**0.9715**	**0.9635**	**0.8905**	**0.5463**	**0.5170**	0.6226	**0.7519**

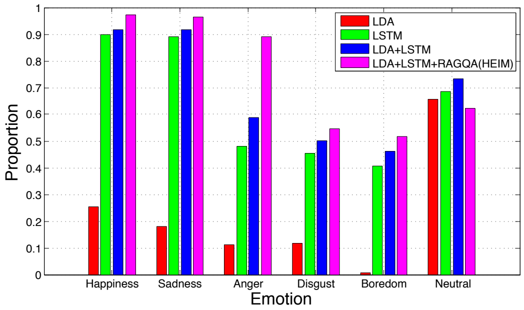
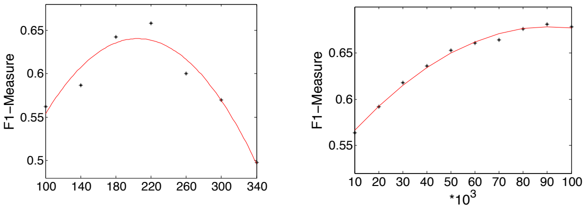


Fig. 3. Modality contribution analysis.



(a) The number of cells in LSTM. (b) The size of unlabeled data in RAGQA.

Fig. 4. Parameter sensitivity analysis.

text information. Moreover, HEIM which incorporates the text information, the acoustic information and query attributes jointly achieves the best performance. The acoustic information can enhance the performance for +46.6% on average, especially for the inference of *Happiness* (+66.4%) and *Sadness* (+73.7%), while query attributes can further enhance the performance for +6.5% on average, especially for the inference of *Anger* (+30.3%). These results validate the necessity and the effectiveness of taking the acoustic information of users' queries and the query attributes into consideration.

5.2.3. Parameter sensitivity analysis.

We show how changes of parameters in HEIM affect the performance of inferring users' emotions in VDAs.

- **The number of cells in LSTM.** Visualized in Figure 4(a), as the number of cells in LSTM increases, the performance turns better at first and then declines. The performance reaches the highest 0.658 in terms of F1-Measure when the number of cells is 220. Thus we set the number of cells as 220 in the experiments above.
- **The size of unlabeled data in RAGQA.** Visualized in Figure 4(b), as the scale of unlabeled data used for pre-training becomes larger, the performance gets better gradually. When the size of unlabeled data is over 90,000, the performance reaches convergence. The experiment containing 10,000 utterances lasts for 10-12 hours on 24 core 2.10GHZ CPU, 64.0GB memory. Considering time efficiency, we conduct experiments on a data set containing 90,000 unlabeled utterances.

5.3. Experimental results on public data set

To demonstrate the comparability and the adaptability of our method, we also report experimental results on the public Berlin Emotional Database (Emo-DB) [6]. To compare our method with speech emotion recognition works and due to the lack of query attributes in the database, we only extract the acoustic features. That means only LSTM is involved without pre-training process. As shown in Table 2, the accuracy reaches 0.882, showing +5.7% improvement compared with [9], +8.5% improvement compared with [7], and +7.1% improvement compared with [10], indicating that our method still shows advantages on the acted database and utterances of other language.

5.4. Case study

At last, we'd like to give an interesting case study of our method. 93,000 utterances described above are sampled and further classified into different groups according to users' locations. We use proposed HEIM to infer the emotions of utterances, and then we calculate the similarity of emotion distributions among cities, which is evaluated by Euclidean distance. In Figure 5, the x-axis and y-axis represent cities, which are sorted in a descending order according to their Gross Domestic Product (GDP) values. The value of coordinate *(a, b)* represent the similarity of emotion distributions between city *a* and city *b*, and is visualized by color, where cool color implies greater similarity. We can see that colors of the top-right submatrix are comparatively cool and much alike, indicating that the emotion distributions among these cities are very similar. These cities are all developing cities in China (GDPs ranked from 12 to 20), so we assume that devel-

Table 2. The accuracy for Emo-DB.

Ang	Bor	Dis	Fea	Hap	Sad	Neu
0.7142	0.8859	0.9234	0.9626	0.8086	0.9233	0.9621

[1] *Ang:* Anger; *Bor:* Boredom; *Dis:* Disgust; *Fea:* Fear; *Hap:* Happiness; *Sad:* Sadness; *Neu:* Neutral.

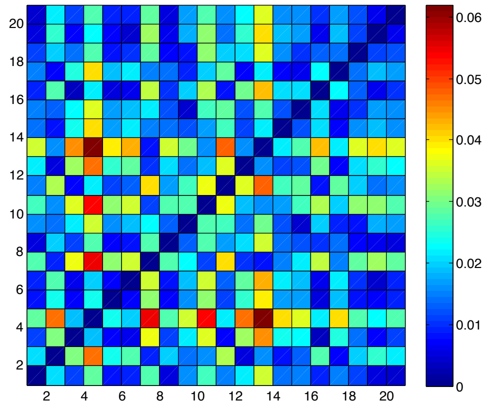


Fig. 5. The emotion similarity among cities.

oping cities may share similar emotion distributions, as they have many things in common like problems they encounter when searching for better development. On the contrary, colors of the bottom-left submatrix are comparatively random, showing that the emotion distributions of developed cities in China (GDPs ranked from 1 to 11) vary. We assume that this may be explained by the reason that these cities are 'grown-up' and well-developed, thus having their own characteristics and cultures.

6. CONCLUSION

In this paper, we take a significant step towards the problem of inferring users' emotions for real-world VDAs. The proposed Hybrid Emotion Inference Model (HEIM) shows great improvement on a massive-scale real-world data set. Based on our work, VDAs can take users' emotions into consideration when generating responses and be more humanized to understand users' intentions, thus optimizing the interaction.

7. ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Program of China (2012CB316401), National Natural, and Science Foundation of China (61370023). This work is partially supported by the National Basic Research Program of China (2011CB302201), and the National High Technology Research and Development Program("863"Program) of China (2012AA011602). We would also like to thank Microsoft Research Asia-Tsinghua Univertity Joint Laboratory: FY14-RES-SPONSOR-111 for its support.

8. REFERENCES

[1] J. Bellegarda, "Spoken language understanding for natural interaction: The siri experience," *Natural Interaction with Robots, Knowbots and Smartphones*, pp. 3–14, 2013.

[2] Z. Ren, J. Jia, Q. Guo, K. Zhang, and L. Cai, "Acoustics, content and geo-information based sentiment prediction from large-scale networked voice data," in *ICME*, 2014, pp. 1–4.

[3] H. Yin, B. Cui, L. Chen, Z. Hu, and Huang Z., "A temporal context-aware model for user behavior modeling in social media systems," in *ACM SIGMOD*, 2014, pp. 1543–1554.

[4] K. Kamath, J. Caverlee, K. Lee, and Z. Cheng, "Spatio-temporal dynamics of online memes: a study of geo-tagged tweets," in *WWW*, 2013, pp. 667–678.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[6] F. Burkhardt, A. Paeschke, M. Rolfes, and W. Sendlmeier, "A database of german emotional speech," *INTERSPEECH*, vol. 11, pp. 1517–1520, 2005.

[7] S. Ramakrishnan and I. Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, pp. 1467–1478, 2013.

[8] E. Yuncu, H. Hacihabiboglu, and C. Bozsahin, "Automatic speech emotion recognition using auditory models with binary decision tree and svm," in *ICPR*, 2014, pp. 773–778.

[9] P. Shen, C. Zhou, and X. Chen, "Automatic speech emotion recognition using support vector machine," *EMEIT*, vol. 2, pp. 621–625, 2011.

[10] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in *ICAECC*, 2014, pp. 1–4.

[11] Sepp Hochreiter and Jurgen Schmidhuber, "Long short-term memory," *Neural computation*, pp. 1735–1780, 1997.

[12] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE TKDE*, vol. 24, pp. 1134–1145, 2012.

[13] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, "How do your friends on social media disclose your emotions?," in *AAAI*, 2014, pp. 306–312.

[14] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *ASRU*, 2013, pp. 273–278.

[15] Y. A. LeCun, L. Bottou, G. B. Orr, and K. Muller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 9–48. 2012.