Inferring Emotions From Large-Scale Internet Voice Data

Jia Jia, *Member, IEEE*, Suping Zhou, Yufeng Yin, Boya Wu, Wei Chen, Fanbo Meng, and Yanfeng Wang

***Abstract*—As voice dialog applications (VDAs, e.g., Siri,[1] Cortana,[2] Google Now[3]) are increasing in popularity, inferring emotions from the large-scale internet voice data generated from VDAs can help give a more reasonable and humane response. However, the tremendous amounts of users in large-scale internet voice data lead to a great diversity of users accents and expression patterns. Therefore, the traditional speech emotion recognition methods, which mainly target acted corpora, cannot effectively handle the massive and diverse amount of internet voice data. To address this issue, we carry out a series of observations, find suitable emotion categories for large-scale internet voice data, and verify the indicators of the social attributes (query time, query topic, and users location) and emotion inferring. Based on our observations, two different strategies are employed to solve the problem. First, a deep sparse neural network model that uses acoustic information, textual information, and three indicators (a temporal indicator, descriptive indicator, and geo-social indicator) as the input is proposed. Then, to capture the contextual information, we propose a hybrid emotion inference model that includes long short-term memory to capture the acoustic features and a latent dirichlet allocation to extract text features. Experiments on 93 000 utterances collected from the Sogou Voice Assistant[4] (Chinese Siri) validate the effectiveness of the proposed methodologies. Furthermore, we compare the two methodologies and give their advantages and disadvantages.**

***Index Terms*—Emotion, Internet voice data, deep sparse neural network, long short-term memory.**

I. Introduction

EMOTIONS occupy an important role in our daily life. Comparatively speaking, voice is the most effective real-time method of expressing emotions, in contrast to text and image data. As voice dialogue applications (VDAs, e.g., Siri, Cortana, Google Now) are currently used worldwide, Nuance[5] finds that VDAs are used at least once a day by nearly 57% of users in the world. Thus, it is easier to obtain large-scale internet voice data today. Further, the emotions inferred from the voice data help VDAs give a more reasonable and humane response.

Considerable research efforts have been devoted to speech emotion recognition, and these studies focus on extracting effective features and utilizing diverse types of learning methods. Spectrograms, Mel Frequency Cepstral Coefficients (MFCCs), Energy, Formant, Linear Predictive Coding (LPC) and pitch of voice are extracted as features by Meddeb *et al.* [1]. The authors in [2] investigate a heuristic algorithm harmony search (HS) for feature selection and extract 3 feature sets: MFCC, Fourier Parameters (FP), and features extracted with the OpenSmile toolkit. [3] uses continuous hidden Markov models to classify speech emotions. A Support Vector Machine (SVM)-based speech emotion recognition framework is presented in [4]. As deep learning is developing rapidly, it performs well in generating features in similar areas (e.g., Facial Expression Recognition [5], Multimedia Computing [6]). [7] uses Convolutional Neural Networks (CNNs) to extract high-level features to identify different emotional states. In [8], Deep Neural Networks (DNNs) are utilized to analyze speech recognition.

However, the data sets of these studies come primarily from acted corpora instead of large-scale internet voice data. Therefore, inferring emotions from the large-scale internet data presents us with several challenges. First, the tremendous numbers of users in large-scale internet voice data lead to a variety of accents and expression patterns. Thus, there are insufficient training data available for specific users. Second, the previous studies on voice emotion recognition mainly used corpora data (e.g., Berlin Emotional Database [9], [10], RML Emotional Database [11]), which have limited amounts of data. For internet data, a more suitable methodology to deal with the massive amount of voice data is needed.

To address the above challenges, we collect a corpus of large-scale internet voice data from the Sogou Voice Assistant, which includes 6,891,298 utterances. The data set is recorded by 405,510 users in the year 2013 in the Chinese language. We first

Manuscript received March 1, 2017; revised July 20, 2017, December 10, 2017, April 17, 2018, and June 22, 2018; accepted November 20, 2018. Date of publication December 17, 2018; date of current version June 21, 2019. This work was supported in part by the National Key Research and Development Plan (2016YFB1001200), in part by the Innovation Method Fund of China (2016IM010200), in part by the National Natural and Science Foundation of China (61521002), and in part by the Tiangong Institute for Intelligent Computing, Tsinghua University, Beijing, China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hatice Gunes. *(Corresponding author: Jia Jia.)*

J. Jia, S. Zhou, Y. Yin, and B. Wu are with the Department of Computer Science and Technology, the Key Laboratory of Pervasive Computing, Ministry of Education, the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: jjia@tsinghua.edu.cn; 1874504489@qq.com; yinyf15@mails.tsinghua.edu.cn; wuboya10@gmail.com).

W. Chen, F. Meng, and Y. Wang are with the from Sogou Corporation, Beijing 100084, China (e-mail: chenweibj8871@sogou-inc.com; mengfanbosi0935@sogou-inc.com; wangyanfeng@sogou-inc.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2887016

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

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

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

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

[5] An image and speech applications corporation in America.

consider how to model the emotions of large-scale internet voice data and unveil six major emotion categories: disgust, happiness, anger, sadness, boredom and neutral. Then, to solve the problem of data diversity, we consider finding other emotion-related attributes to help infer emotion. The existence of topical and geographical dependencies in users behaviors has been proven in previous work [12]–[14]. Thus, we investigate whether query attributes, such as query time, query topic (e.g., Chat, Search, Joke), and user location (e.g., Beijing, Shanghai), are related to the emotions in large-scale internet voice data.

Based on our observations, we propose two methodologies to infer emotions from large-scale internet voice data. In the first methodology, we propose a Deep Sparse Neural Network (**DSNN**) to predict the users emotions. DSNN combines three different indicators (temporal, descriptive and geo-social indicators) with textual and acoustic features. In the second methodology, we propose a Hybrid Emotion Inference Model (**HEIM**) to solve the problem. HEIM uses Long Short-Term Memory (LSTM) to capture the acoustic features and Latent Dirichlet Allocation (LDA) to extract the text features. Additionally, to maximally utilize unlabeled data and further improve the accuracy, we apply an auto-encoder to pre-train the network in an unsupervised way. Additionally, we apply back-propagation optimization to fine-tune the DSNN. In HEIM, a Recurrent Auto-encoder Guided by Query Attributes (RAGQA), which combines other emotion-related query attributes, is employed to pre-train the LSTM. The experimental results confirm the accuracy of the two proposed methodologies. In terms of the F1-measure, the DSNN (0.4355) achieves a +0.0682 improvement compared with Naive Bayesian (0.3673), a +0.0330 improvement compared with KNN (K-Nearest Neighbors, 0.4025) and a +0.0454 improvement compared with SVM (Support Vector Machine, 0.3901). HEIM (0.7523) achieves a +0.3850 improvement compared with Naive Bayesian, a +0.3498 improvement compared with KNN and a +0.3622 improvement compared with SVM. The factor contribution analysis also proves the effectiveness of the three indicators in the DSNN and HEIM. Comparing these two feasible methodologies, we find that they both have advantages and disadvantages. For example, the performance of HEIM (0.7523) is much higher than that of the DSNN (0.4355) in terms of the F1-measure. However, the processing time of the DSNN is shorter than that of HEIM, so the DSNN is a better choice when considering timeliness.

The remainder of this paper is organized as follows. In Section II, we survey the existing research in the area of voice emotion. In Section III, we introduce the large-scale internet voice data that we establish. In Section IV, we formally define the problem. In Section V, we give a series of analyses and present our observations. In Section VI, we provide an overview of the proposed DSNN and HEIM. In Section VII, we conduct experiments and report the experimental results. In Section VIII, we conclude this work and discuss ideas for future work.

II. Related Works

A. Speech Emotion Inference

Previous research on speech emotion recognition has focused primarily on extracting effective features and utilizing diverse types of learning methods. In terms of inferring emotions from speech, it is believed that a proper selection of features significantly affects the classification performance [15]. A large number of acoustic features have been explored to infer speech emotions. Peipei *et al.* [16] extract Mel Frequency cepstrum coefficients (MFCCs), energy, pitch, Linear Prediction coefficients and Mel cepstrum coefficients (LPCMCCs), and linear prediction cepstrum coefficients (LPCCs) as features. To achieve accurate classification of speech emotion, [2] investigates a heuristic algorithm harmony search (HS) for feature selection and extracts 3 feature sets: MFCCs, Fourier Parameters (FPs), and features extracted with the OpenSmile toolkit. They also combine MFCCs with FPs as the fourth feature set. [17] also uses the OpenSmile toolkit to extract acoustic features to improve audio emotion recognition accuracy. [15] concludes that the speech features that are explored in emotional states recognition can be grouped into four categories: continuous features (pitch, energy, etc.), qualitative features (voice quality, tense, etc.), spectral features (LFPC, MFCC, etc.), and TEO (Teager energy operator)-based features. However, the best speech features are still unclear for speech emotion recognition, despite the above explored various features. Furthermore, some studies suggest that in addition to acoustic features, other types of features must be considered in emotion modelling. [18] combines three kinds of information - acoustic, lexical and discourse - to identify emotion states. The results show that the combination of all three features improves emotion classification markedly both for male and female samples compared with using only acoustic features. As deep learning is developing rapidly, it performs well in generating features in similar areas (e.g., facial expression recognition [5], [19] and [20]). [21] trains a deep convolutional neural network (DCNN) to learn the relevant, complex feature representation from short segments of speech data for speech emotion classification without hand-tuned features.

For learning methods, various types of classifiers have been used to perform emotion recognition from speech. Naive Bayesian is employed in [22]. [3] uses continuous hidden Markov models to classify speech emotions. Since the Gaussian Mixtures Model (GMM) performs well in capturing the distribution of the input features, it is shown to have the capability to develop an emotion recognition model with a large feature vector. Thus, in [23], [24], GMM is utilized to identify different emotional states. Additionally, K-Nearest Neighbors (KNN) is adopted in [18] and a Support Vector Machine (SVM) based speech emotion recognition framework is presented in [16]. In [25], a latent Dirichlet allocation (LDA) model is introduced for speech emotion recognition. Recently, increasing attention has been paid to the use of deep learning for speech emotion recognition which results in a better performance than that achieved by the traditional framework. In [26], [27], Deep Neural Networks (DNNs) are utilized to analyze the speech emotion. [28] and [29] adopt Convolutional Neural Networks (CNNs) to identify different emotional states. Furthermore, [30] introduces a model consisting of a 1-state HMM exclusively and a 1-layer Artificial Neural Network (ANN) for estimating the emission probabilities instead of a 5-layer DBN.

However, 1) the data sets of these studies come mainly from acted corpora instead of from large-scale internet voice data.

Thus, the traditional speech emotion recognition methods cannot effectively handle the massive and diverse internet voice data. 2) While the contextual information of utterances is ignored, these methods mainly focus on the statistical values of acoustic features. However, since utterances evolve as time passes, considering contextual information may be beneficial. 3) Although some studies have revealed that acoustic attributes are insufficient in speech emotion recognition [18], [15], whether social attributes can assist in inferring emotion from large-scale internet voice data is still unclear.

B. Large-Scale Internet Data Emotions

For large-scale internet data emotion analysis, previous works have been based mainly on text or image data. [31] and [32] use text data collected from Twitter[6] for the task of emotion analysis. [33] and [34] employ large-scale image data from Flickr[7] to study the emotion influence in large image social networks. [35] uses a Gaussian mixture model to analyze a large-scale Image-Emotion-Social-Net data set. [36] proposes a semi-supervised hierarchical classification (SSHC) algorithm for the emotional classification of color images from the internet cloud.

In addition, large-scale internet data emotion analysis has considered specific events, such as how Flickr users affect the distribution around Thanksgiving [37] and the response of microbloggers to the death of Michael Jackson [38]. Additionally, there has been further analysis on social and economic trends, such as consumer confidence and political opinions [39], [40] as well as the relationship between Twitter moods and stock market fluctuations [41].

Furthermore, for large-scale internet data emotions analysis, the challenge is how to leverage the large-scale unlabeled data. An auto-encoder is commonly used to make better use of unlabeled data. Many early works in semi-supervised learning for neural networks are built on auto-encoders. [42] imposes sparse and orthogonal constraints on the auto-encoder and makes it a highly discriminative descriptor. [43] proposes a deep feature learning framework based on stacked auto-encoders (SAEs) by integrating pairwise constraints to serve as a discriminative term. [44] constructs a model that includes a corresponding auto-encoder (Corr-AE) by correlating hidden representations of two uni-modal auto-encoders.

In our paper, we also use an auto-encoder to utilize large-scale unlabeled data to enhance the performance of speech emotion inferring. However, since it is difficult to acquire large-scale internet voice data, works on inferring emotions from large-scale internet voice data are still scarce.

C. Contributions

In this paper, we systematically study the problem of inferring users emotions from a large-scale internet voice data base. Here, we extend our previous work in [45], where we perform limited single model validation. We propose two different methodologies (DSNN and HEIM) to solve the problem and give a detailed analysis of the two methodologies compared with other traditional models (SVM, KNN, NB). Further, we discuss their advantages and disadvantages for different requirements and analyze their own contributions. Specifically, the performance of HEIM (0.7523) is much better than that of the DSNN (0.4355) in terms of the F1-measure. However, the processing time of HEIM is much longer than that of the DSNN. In addition, we investigate suitable emotional categories for large-scale internet voice data and the correlation between the query time, query topic and user location and the emotion inferring.

III. The Large-Scale Internet Voice Data Set

A. Data Collection

The Sogou Voice Assistant provides us with a corpus of large-scale internet voice data that includes 6,891,298 utterances. Each utterance is approximately 3 to 4 seconds long. The data set is recorded by 405,510 users in the year 2013 in Chinese. Basic information (e.g., the users ID, query time, query topic and users location) is attached to each utterance. Additionally, the corresponding speech-to-text information is available from Sogou Corporation.

B. Acoustic Feature Extraction

Seven main acoustic features are selected according to previous research on speech emotion recognition [46], [47]: Syllable Duration (SD), Energy, F0, Mel Frequency cepstrum coefficients (MFCCs), Log Frequency Power Coefficients (LFPC), Spectral Centroid (SC), and Spectral Roll-off (SR). In particular, the voice segments of each utterance to extract these features are a 10-ms frame shift and a 20-ms frame length.

At the frame level, we extract 29 acoustic features for each frame: Energy (1), F0 (1), MFCC (13), LFPC (12), SC (1), and SR (1). At the utterance level, we adopt the feature selection algorithm used in [48] to extract 113 acoustic features for each utterance:

- *Syllable Duration (SD) (11):* the syllable duration sequence, which is extracted using the method in [49], is applied with 11 functionals (mean, std, max, min, range, quartile1/2/3, iqr1–2/2–3/1–3.).
- *Energy (13):* the energy envelop is applied with 13 functionals (mean, std, max, min, range, quartile1/2/3, iqr1–2/2–3/1–3, skewness, kurtosis.).
- *F0 (13):* the fundamental frequency contour, which is extracted using a modified STRAIGHT procedure [50], is applied with 13 functionals (as for Energy).
- *MFCC (26):* the mean and standard deviation of the Mel frequency cepstral coefficients 1–13
- *LFPC (24):* the mean and standard deviation of log frequency power coefficients 1–12, which are extracted using the method in [14] with $\alpha = 1.4$.
- *Spectral Centroid (SC) (13):* the spectral centroid contour applied with 13 functionals (as for Energy).
- *Spectral Roll-off (SR) (13):* the spectral roll-off contour applied with 13 functionals (as for Energy).

The number in parentheses is the dimension of each acoustic feature.

[6] http://twitter.com/

[7] http://www.flickr.com/

We define three types of indicator features and formulate N-dimensional (N = 24,70,21) vectors to describe them.

- *Temporal Indicator (TI):* We set a vector $t = (t_1, \ldots, t_{24})$, where $t_i \in \{0,1\}$ shows whether the utterance is recorded within the time interval [i-1, i].
- *Descriptive Indicator (DI):* We use a vector $d = (d_1, \ldots, d_{70})$ for each utterance where $d_i \in \{0,1\}$ indicates whether the utterance belongs to the rank ith inquiry type.
- *Geo-social Indicator (GI):* We formulate a vector $g = (g_1, \ldots, g_{21})$, where $g_i \in \{0,1\}(i \in [1,2,\ldots,20])$ implies whether the utterance is recorded in the rank ith city and g_{21} shows if the utterance is recorded in other cities besides the top 20 cities.

C. Labeling

Because our data set is extremely large, it is not feasible to label each utterances emotion manually. Therefore, we ask three well-trained human labelers to label the emotions of 3,000 utterances, which are randomly chosen from the whole data set. When differences arise regarding a certain utterance, the labelers label the utterance following a discussion. If they cannot reach a satisfactory conclusion, this utterance is labeled unclear. In total, there are 58 utterances that are labelled unclear, and these utterances are not adopted as the ground truth in our later experiments. Therefore, in total, 2,942 utterances are labeled with clear emotions, and the emotion distributions are as follows: neutral: 61.3%, happiness: 13.2%, disgust: 13.0%, boredom: 4.8%, anger: 3.9% and sadness: 3.8%. In addition to the above labeled data, we use 90,000 unlabeled data samples in the pre-training process.

IV. Problem Definition

In this section, we will introduce the problem formulation of emotion inference on large-scale internet voice data. For each utterance $\mathbf{u}$ in a given set of utterances $\mathbf{U}$, we define $\mathbf{u} = \{\mathbf{x}, \mathbf{g}, l_c\}$. $\mathbf{x}$ represents the set of acoustic features. $\mathbf{g}$ is the textual information of an utterance and l_c stands for the social attribute (query time, query topic and user's location) which are provided by Sogou Corporation.

Definition: **Emotions**. Considering both textual and acoustic information, we investigate the main emotions in human-mobile voice interactions according to the observations of internet voice data from VDAs. It is worth nothing that we find that the emotion categories are different from theories about emotions regarding facial expressions in human-mobile voice communication. Based on our observations, we identify {*happiness*, *sadness*, *anger*, *disgust*, *boredom* and *neutral*} as the emotional space and define it as $\mathbf{E_S}$, where $S = 6$. Further, underlying human-mobile voice interaction, we denote two interesting emotion patterns, which are illustrated in detail in Section V.

Problem: **Learning task.** For utterances set $\mathbf{U}$, we focus on inferring the emotion for every utterance $\mathbf{u} \in \mathbf{U}$:

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

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

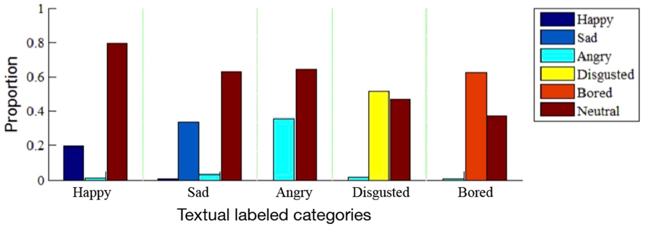


Fig. 1. The proportion of manually labeled emotions in each textual labeled category.

V. Preliminary Findings

The natural way for VDAs to give answers is to utilize text information with NLP techniques. To investigate the existence of other emotion-related information that can assist in inferring emotion, we conduct a series of observations and uncover several phenomena.

A. Emotion Categories

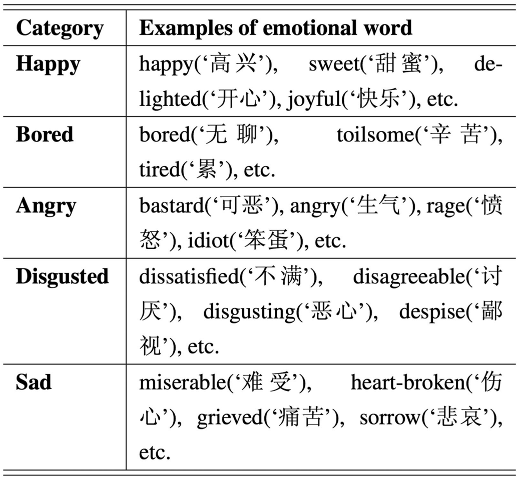
Before providing the proper methodology of our task, we first need to define the primary emotions found in internet voice data.

Considering that the text attribute in voice data is rather helpful for understanding users emotions, we investigate the primary emotions as follows:

- We pick out all the emotional words in a list of common Chinese emotional words from [46] after searching the given text information of 6,891,298 utterances;
- We adopt the emotional words with a high frequency of occurrence and remove those with a low frequency.
- we cluster the emotional words into corresponding categories based on previous work on Chinese emotional words categorization [46], [51].

As shown in Fig. 2, we finally divide the emotional words into five main categories, as follows: Happy, Bored, Sad, Angry, and Disgusted. Compared with Ekmans six basic emotions proposed for human-human communication, we can see that Fear and Surprise from Ekmans six emotions are replaced by Bored. In total, we collect 48,211 utterances. The principle is that text information of each utterance should include emotional words that belong to only one of the above five emotions. Additionally, we regard this emotion category as the textual label of this utterance.

To address how an utterances textual label matches its real emotion, we randomly choose 1,000 utterances (200 for each category). Then, we ask three well-trained human labelers to label each utterance with the above five emotions and a Neutral annotation. They label the utterances by listening to the utterances, and they do not look at the text. Additionally, when the labelers have different opinions, they engage in a discussion to reach an agreement. We consider the manually labeled results as the real emotion of each utterance. We compare the results of the textual and manually labeled emotions, and the proportion of manually labeled emotions in each textual labeled category is shown in Fig. 1. Two interesting phenomena are found as follows:



Category	Examples of emotional word
Happy	happy('高兴'), sweet('甜蜜'), delighted('开心'), joyful('快乐'), etc.
Bored	bored('无 聊'), toilsome('辛 苦'), tired('累'), etc.
Angry	bastard('可恶'), angry('生气'), rage('愤怒'), idiot('笨蛋'), etc.
Disgusted	dissatisfied('不满'), disagreeable('讨厌'), disgusting('恶心'), despise('鄙视'), etc.
Sad	miserable('难受'), heart-broken('伤心'), grieved('痛苦'), sorrow('悲哀'), etc.

Fig. 2. Examples of emotional word.

- *Phenomenon I:* In some cases, the textual labels do not match the real emotions. For instance, some of the utterances with happy textual labels have the emotion labeled anger. This happens when a user says, Im really happy with what you said, but he does not actually mean what the word means. This phenomenon indicates **Emotion Pattern I**: Irony exists in human-mobile communication. Additionally, it is not proper to directly use the textual labels as real emotion categories.
- *Phenomenon II:* A large part of the utterances are actually neutral voice data, even those whose speech-to-text information has an emotional word. This phenomenon shows **Emotion Pattern II**: Compared with talking to a human, people speak in a more implicit and rational manner when speaking to a mobile. Thus, users would like to apply linguistic information which conveyed by speech rather than para-linguistic information conveyed by text to express their intentions. In other words, acoustic information has more of an impact on emotion expression than text. Thus, each textual labeled category consists of both the real emotional and neutral voice data.

Therefore, according to these findings, we add 'Neutral' to the above five main categories to form our complete emotional space $\{$*happiness*, *sadness*, *anger*, *disgust*, *boredom* and *neutral*$\}$.

B. Observation on Query Time

We classify the manually labelled utterances as mentioned in III.C into different groups according to their published time, whose granularity is hourly. Then, we calculate the proportion of emotion distributions among different hours. In Fig. 5, the x-axis represents different times of a day, and the y-axis represents the proportion of the six types of emotions. From the figures, we summarize some interesting findings about the time-emotion correlation as follows.

- *Joy at night:* In Fig. 6(a), the proportion of happiness from 17:00 to 20:00 is relatively high, indicating that people may feel more relaxed and comfortable when they finish their work during the day and start to enjoy the night.

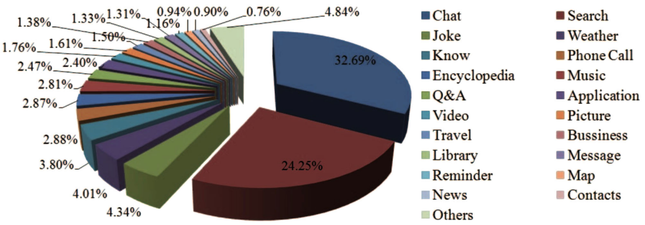


Fig. 3. The types with the top 20 query amount.

TABLE I
EXAMPLES OF 7 BASIC TYPES

Topic	Examples
Chat	Chat, Help
Search	Search, Q&A, Encyclopedia, Know, Library
Joke	Joke
Entertainment	Music, Video, picture, movie, TV series, fiction, plot, lyric
Operation	Call, SMS, reminder, Contacts, calculator, cellphone operation, translation
Consultation	weather, travel, business, news, map, lottery, date, people, shopping, food, programme, stock, car, holiday, holiday, exchange rate, oil prices, college
Other	customer, service, Post Bar, Precious metals, Constellation, Microblog, express

- *Dull before dawn:* In Fig. 6(b), the proportion of boredom is obviously higher and the proportion of anger is lower from 2:00 to 5:00, which is the typical time for sleeping. Such a phenomenon can be explained by people who suffer from insomnia. When people find it difficult to sleep in the early morning, they may feel bored and chat with the VDA.

C. Observations Regarding the Query Topic

Fig. 3 shows the types with the top 20 queries of 70 raw topics. To simplify our model, we divide the 70 original topic types into seven categories. As Table I shows, we define $\{$*Chat*, *Q&A*, *Joke*, *Entertainment*, *Operation*, *Information* and *Other*$\}$ as the 7 basic types.

We classify the labeled utterances in III.C into different groups according to the topics of the utterances and calculate the proportion of emotion distributions among different topics. In Fig. 7, the x-axis represents ten different types of topics, and the y-axis represents the proportion of the six types of emotion. From the figures, we summarize some interesting findings about the topic-emotion correlation as follows.

- *Fun seeker:* In Fig. 8(a), the proportion of boredom in the topics "Chat" and "Joke" is relatively high, indicating that people may treat the VDA as a funny friend when they are bored.
- *Healing music:* In Fig. 8(b), the proportion of sadness in the topic "Music" is obviously higher than the others, which indicates that music is a common way for people to comfort themselves when they are sad.

D. Observations Regarding the Users Locations

Fig. 4 shows the proportion of inquiries in the top 10 cities. We find that as the political and cultural center of China, Beijing has

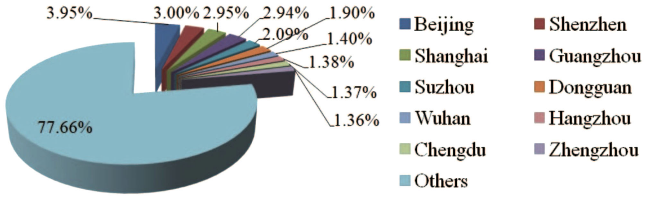


Fig. 4. The cities of the top 10 query amount.

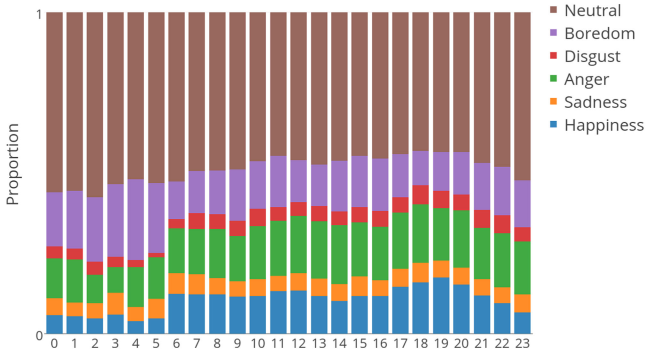


Fig. 5. Emotion proportion of different time in a day.

the highest number of queries. In addition, first-tier cities such as Shenzhen, Shanghai and Guangzhou have many voice assistant users due to their high level of technology development.

Additionally, utilizing the utterances labelled in III.C, Fig. 9 demonstrates the emotion distribution of utterances of different query locations. People living in Guangzhou have more abundant expressions of emotions: only 34.40% of the utterances in Guangzhou are neutral, which is the lowest compared with other locations. We also find that the negative emotions sadness, anger, disgust and boredom are more frequent than the positive emotion happy, which may occur because the stressful living environment of first-tier cities causes people to be negative.

Based on the above findings, we find that in addition to text information, acoustic information and social attributes are related to the users emotions; thus, we consider all of these factors in the modelling.

VI. Methodologies

Traditionally, text features and acoustic features are applied to infer the users emotions. However, our data set is large-scale and comes from non-specified users. Further, we find that the users dialects and expression preferences vary in terms of correlations such as locations, time, and topics. Thus, the problem cannot be solved effectively by considering only textual and acoustic information.

A. Our Proposed Methodologies

To take the correlations into account, two different strategies are employed to solve the problem.

In the first methodology (Deep Sparse Neural Network, Fig. 10), we utilize input features to discover latent features at the low level of the network. Then, we use a deep sparse network to learn the high-level features.

In the second methodology (Hybrid Emotion Inference Model), since voice is a type of feature that changes over time, we apply Long Short-Term Memory (LSTM) to extract the acoustic features.

B. Deep Sparse Neural Network

Recently, deep neural networks have been applied to model the features of large-scale speech and image data, and they perform well in classification tasks [52], [53]. Thus, in the first methodology (Fig. 10), to discover latent features, we propose a Deep Sparse Neural Network (DSNN) to extract features and infer users emotions. In DSNN, we utilize utterance level acoustic features as the input. Additionally, to better utilize voice data, we combine 3 types of indicator features together with 7 types of utterance level acoustic features.

Furthermore, we employ the LDA method used in [54] to extract text features. LDA is a generative probabilistic model for collections of discrete data such as text corpora [55]. It is widely applied in sentiment analyses based on text and performs well [54], [56]. Given utterance u's text d, it outputs an emotion distribution vector $g = \{g_1, g_2, \ldots, g_K\}$, where K is the length of the vector. K is an adjustable parameter, and in our work, we set $K = 20$. In addition, in the voice dialogue applications, the users pronunciation is very short, and every utterance contains 7 Chinese characters, on average. Therefore, in our modeling, we no longer perform segmenting for the text of each utterance but consider each Chinese character to be a word.

In total, there are 248 dimensional features as input: 20 dimensions for text features, 113 dimensions for utterance level acoustic features, 70 for descriptive features, 24 for temporal features and 21 for geo-social features in the model. To jointly model the input features, the network has two hidden layers, size 400 and 200 neurons respectively.

To utilize the unlabeled data maximally, we apply an auto-encoder to pre-train the network in an unsupervised manner. Additionally, we apply back propagation optimization to fine-tune the DSNN.

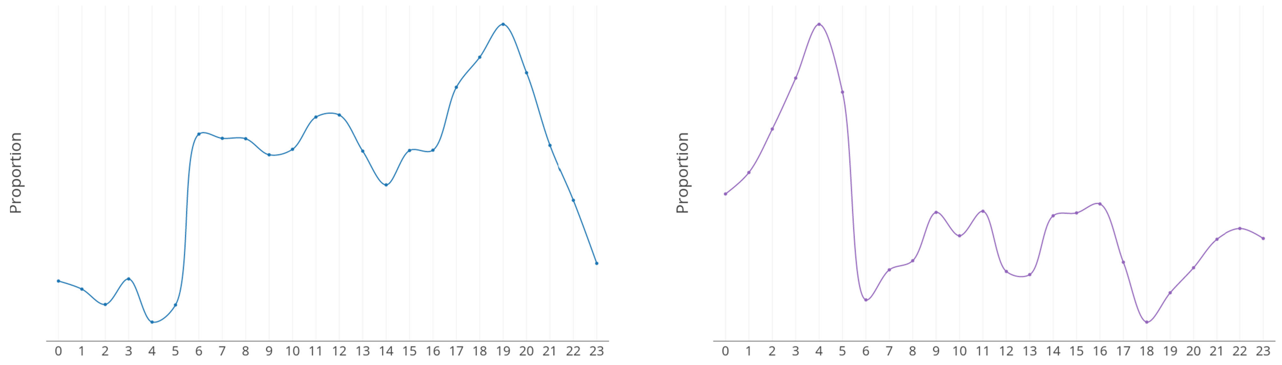
Pre-training can initialize the input features in the lower layers of the network in an unsupervised manner. A standard auto-encoder [57] first encodes the input with higher layer neurons and then decodes the input with the following representation:

$$\tilde{x} = f\left(w^{(2)} f\left(w^{(1)} x + b^{(1)}\right) + b^{(2)}\right) \tag{2}$$

The goal of an auto-encoder with data set x and reconstruction $\tilde{x}$ is given by:

$$\min \frac{1}{m} \sum_{i=1}^{m} \|\tilde{x}^{(i)} - x^{(i)}\|^2 + \frac{\lambda}{2} \sum_{l=1}^{2} \sum_{i=2}^{s_1} \sum_{j=1}^{s_2} (w_{ij}^{l})^2 + \beta \sum_{j=1}^{s_2} KL(\rho || \rho_j) \tag{3}$$

where $w^{(1)}$ and $w^{(2)}$ are weight matrices of the encoder and decoder and $b^{(1)}$ and $b^{(2)}$ are the biases of the encoder and decoder. $f(\cdot)$ is the Softplus activation function. λ and β are the weight decay and sparse penalty, and ρ is the sparse parameter.



(a) Happiness proportion of different time in a day.

(b) Boredom proportion of different time in a day.

Fig. 6. Findings of time observation.

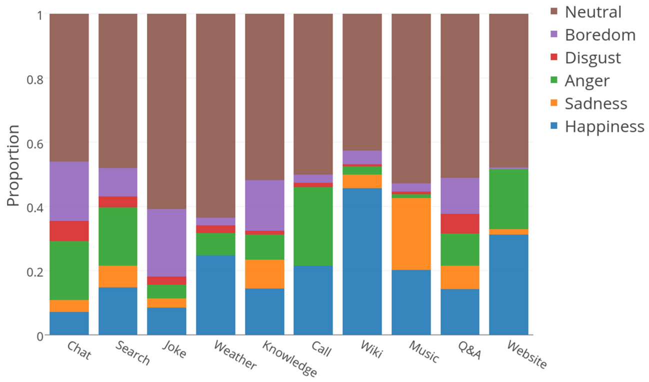


Fig. 7. Emotion proportion of different topics.

Inspired by the characteristics of the V2 region of the human visual cortex, many researchers have introduced depth Sparsity into their model [58]. Therefore, in our paper, we also adopt Kullback-Leibler (KL) divergence to represent Sparsity. $KL(\rho||\rho_j)$ is the Kullback-Leibler (KL) divergence given by

$$KL(\rho||\rho_j) = \rho \log \frac{\rho}{\rho_j} + (1-\rho) \log \frac{1-\rho}{1-\rho_j} \tag{4}$$

Fine-tuning is performed in a supervised manner with backpropagation optimization. The hypothesis of the network is defined by

$$a^{(4)} = \frac{e^{w_j^{(3)} a^{(3)}}}{\Sigma_{k=1}^{s_4} e^{w_k^{(3)} a^{(3)}}} \tag{5}$$

where $a^{(3)}$ is the activation of the highest level feature neurons with the feed-forward network. The overall objective function of the network is then given by

$$\min -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{s_4} y_j^i \log a^{(4)} + \frac{\lambda_s}{2} \sum_{i=1}^{s_4} \sum_{j=1}^{s_3} \left(w_{ij}^{(3)}\right)^2 \tag{6}$$

where $y_j^{(i)}$ is the ground truth indicating whether example (i) belongs to class j, with a zero for false and one for true.

We empirically set the parameters to $\lambda = 0.15$, $\beta = 4$, $\rho = 0.25$ and $\lambda_s = 0.1$.

C. Hybrid Emotion Inference Model

Traditional machine learning methods (e.g., NB, SVM, KNN) that are applied to recognize speech consider only the statistical values of acoustic features as the input. In fact, voice is a type time-varying feature, and it is ineffective for showing the variations of voice signals when we consider only the statistical values of acoustic features.

Therefore, in our second proposed method, we apply LSTM to model the problem. Long Short-Term Memory (LSTM) has proven to be effective when considering time sequences [59]. In [60] [61] [62] [63], to model conversation emotion, the authors use bidirectional Long Short Term Memory (BLSTM) networks to exploit long-range contextual information and achieve better performance than traditional classification methods. As LSTM can capture the correlations among frames, it is well suited to deal with voices. In our paper, we utilize 29 dimensional frame level acoustic features as the input features of LSTM. Given the input $\{x_t, h_{t-1}, c_{t-1}\}$ at time t, the current high-level representations of the acoustic feature sequences refer to the activation h_t of the recurrent layer. It is calculated by the following equations [64], which are standard LSTM equations [59]:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1} + b_i) \tag{7}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1} + b_f) \tag{8}$$

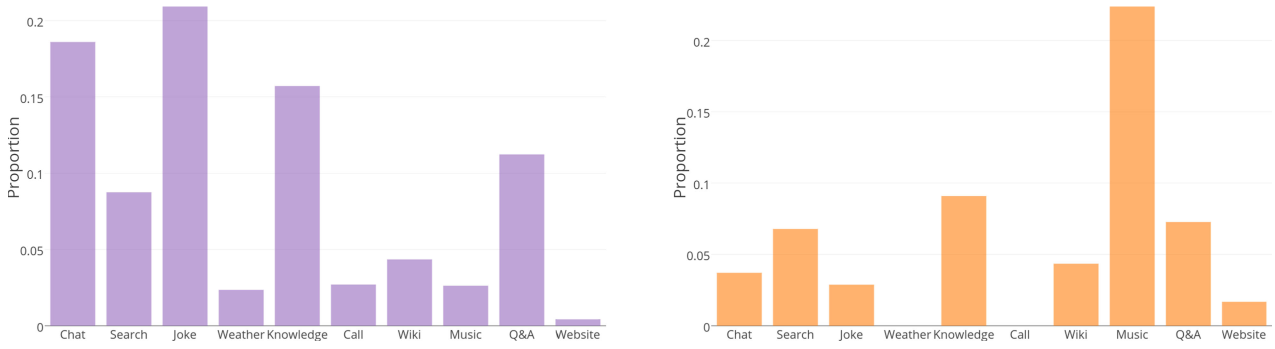
$$c_t = f_t c_{t-1} + i_t \mu(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{9}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{10}$$

$$h_t = o_t \mu(c_t) \tag{11}$$

In the functions above, $W_{\alpha\beta}$ indicates the weight matrix connecting the β layer to the α layer and b_α is the bias vector. i, o, f and c are the input gate, forget gate, output gate and memory cells. σ represents the sigmoid function. For μ, we use a hyperbolic tangent function $f(x) = 1.7159 \tanh(\frac{2}{3}x)$, which has been proven to be capable of improving convergence [65].

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



(a) Boredom proportion of different topics.

(b) Sadness proportion of different topics.

Fig. 8. Findings of topic observation.

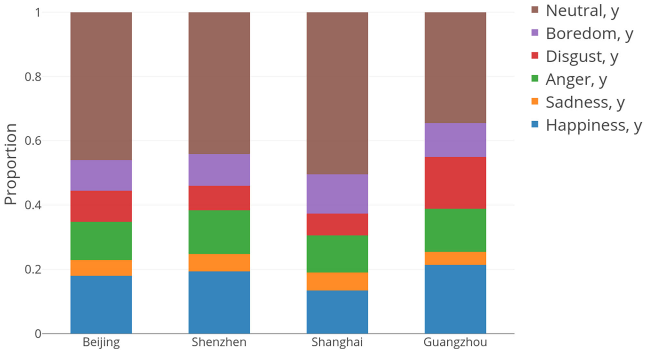


Fig. 9. The emotion distribution of utterances of different cities.

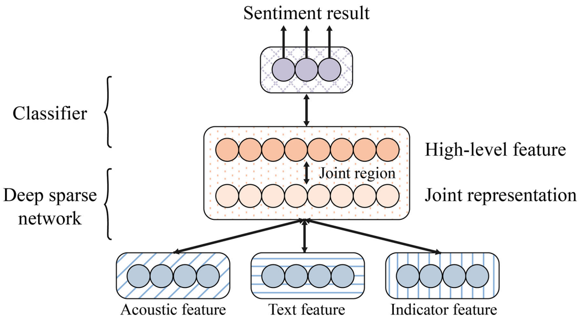


Fig. 10. The architecture of Deep Sparse Neural Network.

Similar to the first methodology (DSNN), we propose Recurrent Auto-encoders [66] Guided by Query Attributes (RAGQA), as illustrated in Fig. 11, to pre-train the LSTM in an unsupervised way. RAGQA utilizes query attributes to find better choices of parameters for LSTM. The query attributes include 70 dimensions of Descriptive Indicator features (DI) and 21 dimensions of Geo-social Indicator features (GI). Specifically, we have tried to include the temporal indicator features (TI) as query attributes, but they do not improve the performance further; therefore, they were left out for simplicity.

The structure of the RAGQAs encoder (Fig. 11) is the same as that of LSTM, and the decoder is a nonlinear mapping. It is worth noting that the traditional auto-encoder only reconstructs the input x_t into $\hat{x}_t$. However, in RAGQA, l_c and x_{t+1} are also reconstructed into $\hat{l}_c$ and $\hat{x}_{t+1}$, where l_c means the query attributes.

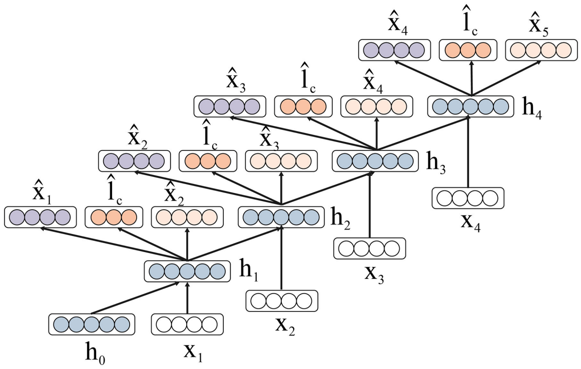


Fig. 11. The architecture of Recurrent Auto-encoders Guided by Query Attributes (RAGQA).

The goal of RAGQA is to learn a function to make a reconstruction $y_t = [\hat{x}_t,\ \hat{l}_c,\ \hat{x}_{t+1}]$ similar to $z_t = [x_t,\ l_c,\ x_{t+1}]$. For convenience, we identify 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 summarized as minimizing the cost function (12):

$$\arg\min_{\theta,\theta'} \|y_t - z_t\|^2 + \frac{\lambda}{2}\|\xi\|^2 \tag{12}$$

$$\|\xi\|^2 = (W')^2 + \sum_{\alpha}\sum_{\beta}(W_{\alpha\beta})^2 \tag{13}$$

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

$$y_t = \mu(W' h_t + b') \tag{14}$$

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

Similar to DSNN, we apply the LDA method to extract the text features. Given utterance u's text d, it outputs an emotion distribution vector $g = \{g_1, g_2, \ldots, g_K\}$, where K is the length of the vector; we set $K = 20$.

TABLE II
THE PRECISION, RECALL AND F1-MEASURE OF THE 5 METHODS FOR INFERRING EMOTIONS IN VDAs

	Method	Happiness	Sadness	Anger	Disgust	Boredom	Neutral	Average
Precision	NB	0.4400	0.2464	0.4400	0.1704	0.2041	**0.8871**	0.3980
	KNN	0.4336	0.3800	0.4321	0.3649	0.3043	0.6727	0.4313
	SVM	0.5152	0.8750	0.6039	**0.5124**	0.5254	0.6133	0.6075
	DSNN	0.5026	0.5821	0.4237	0.2238	0.2979	0.6788	0.4516
	HEIM	**0.9661**	**0.9651**	**0.8856**	0.4721	**0.5781**	0.7018	**0.7614**
Recall	NB	0.3667	0.6800	0.1864	0.4792	0.4167	0.5189	0.4413
	KNN	0.5113	0.2794	0.3633	0.2500	0.1641	0.7724	0.3901
	SVM	0.3856	0.1544	0.3218	0.1914	0.1211	**0.9386**	0.3522
	DSNN	0.5457	0.3900	0.4405	0.2338	0.2654	0.6834	0.4265
	HEIM	**0.9827**	**0.9620**	**0.8955**	**0.6481**	**0.4676**	0.5594	**0.7525**
F1-Measure	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	0.5232	0.4671	0.4319	0.2287	0.2807	0.6811	0.4355
	HEIM	**0.9743**	**0.9635**	**0.8905**	**0.5463**	**0.5170**	0.6226	**0.7523**

TABLE III
THE INPUT FEATURES OF 5 METHODS FOR INFERRING EMOTIONS IN VDAs

NB, KNN, SVM	text features, utterance level acoustic features and indicator features
DSNN	text features, utterance level acoustic features and indicator features
HEIM	text features, frame level acoustic features and indicator features

In HEIM, we first combine the textual features g generated by LDA with high-level representations of acoustic features h_T generated by LSTM. Then, we put them into a SoftMax classifier to compute the probability of each emotion category. For emotion category s, the learning target of HEIM can be formalized as minimizing the cost function (15):

$$J_{(\theta,\mathbf{v})} = -\log p(s|h_T; g) + \frac{\lambda}{2}\|\epsilon\|^2 \tag{15}$$

$$p(s|h_T; g) = \frac{v_s^h \cdot h_T + v_s^g \cdot g}{\sum_{q=1}^{S} \exp^{v_q^h \cdot h_T + v_q^g \cdot g}} \tag{16}$$

$$\|\epsilon\|^2 = \sum_{q=1}^{S}\left(\left(v_q^h\right)^2 + (v_q^g)^2\right) + \sum_{\alpha}\sum_{\beta}(W_{\alpha\beta})^2 \tag{17}$$

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

After we train HEIM, we can obtain the emotion of every utterance by finding the maximum probability $p(s|h_T; g)$.

VII. EXPERIMENTS

A. Experiment Setup

1) Comparison Methods: To demonstrate the effectiveness of our method, three learning methods, namely, Naive Bayesian (NB), the K-Nearest Neighbors algorithm (KNN) and Support Vector Machine (SVM) are chosen as the baseline methods. We conduct comparison experiments on the same data set.

NB: Naive Bayesian is frequently used in many classification problems and performs well. It is also used as the baseline method in [67], [68].

KNN: The K-Nearest Neighbors algorithm is a non-parametric method that is used for classification and regression [69]. The method is also used as the baseline method in [70].

SVM: The Support Vector Machine is a widely used classifier and has good performance. It is also used as the baseline method in [68], [71].

2) Evaluation Metrics: We compare the performance of our two proposed methods with that of the three baseline methods mentioned above in terms of precision, recall and the F1-measure. These evaluation metrics are frequently applied in retrieval problems. Worthy of attention, we utilize five-fold cross validation on 2942 labeled utterances in all experiments to evaluate the performance of emotion inferring. In detail, we use 2,354 utterances for training, and 588 utterances are used for testing each cross validation.

Note that for multi-class classification, the model classifies the utterance into one of the six categories. Thus, for an utterance whose true emotional tag is happiness, only when the inference result is happiness will it be calculated as a true positive.

B. Experimental Results

1) Multi-Class Classification: For our baselines, as shown in Table III, in NB, KNN, and SVM, the input features are acoustic features; text features, which are also calculated by LDA; and indicator features of 2942 labeled utterances. Since they cannot handle time sequences, we calculate the utterance-level acoustic features. Therefore, for NB, KNN, SVM and DSNN, we use utterance-level acoustic features, while in HEIM, we use the frame-level acoustic features. Table II summarizes the Precision, Recall and F1-measure.

We can see that the proposed DSNN outperforms all the baseline methods in terms of the F1-measure: +6.82% better than NB, +3.30% better than KNN, and +4.54% better than SVM. The proposed HEIM performs much better than the other four methods: +38.5% improvement over NB, +34.9% improvement over KNN, +36.2% improvement over SVM and +31.6% improvement over DSNN.

2) Analysis: For NB, KNN and SVM, only labeled data can be utilized. In DSNN and HEIM, however, they maximally utilize the unlabeled data in an unsupervised manner. In the DSNN, we apply an auto-encoder to initialize the lower layers in the neural network. While in HEIM, query attributes are used on the large amount of unlabeled data to pre-train the LSTM, which helps find better parameters for the LSTM.

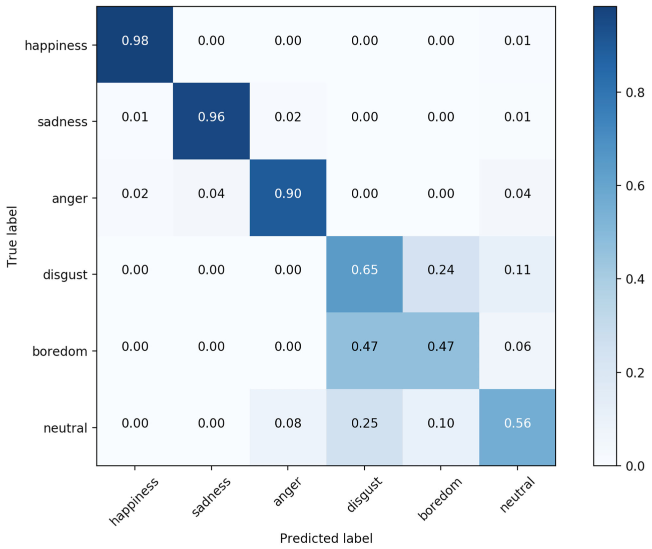


Fig. 12. The confusion matrix of HEIM.

NB, KNN, SVM and DSNN use time features as the input, which discover the latent features of the temporal correlations. However, they cannot handle time sequences and use only the statistical values of the acoustic feature sequences as the input. For HEIM, it utilizes LSTM to generate the acoustic features, which can better describe the contextual correlation between frames of voice information. Therefore, HEIM outperforms the other four methods.

Furthermore, by comparing the prediction results of the different categories, we find that boredom and disgust have lower performances than do the other categories. On the one hand, these two categories have a small proportion in the labeled training data, as mentioned in III.C. On the other hand, as shown in Fig. 12, the confusion matrix of HEIM method, the utterances labeled boredom and disgust are often mixed together and are difficult to distinguish. This is also reported in [72]. We assume that this occurs because both emotion categories have low *F0* and *Energy* values.

C. Factor Contribution Analysis

In the DSNN, we combine three indicator features (temporal indicator, descriptive indicator and geo-social indicator) with acoustic features and text features as the input for the whole network. In HEIM, we utilize Latent Dirichlet Allocation (LDA) to extract text features and LSTM pre-trained by RAGQA to process acoustic features. To examine the effect of each of the different input features on the final performance, we conduct experiments to evaluate the contribution of each type of feature in our proposed methods.

For the DSNN, Fig. 13(a) shows the F1-measure of 5 models with different input combinations. From the figure, it shows that text information helps achieve a better performance than statistic-level acoustic information for all 6 of the emotion categories, except boredom. In addition, comparing the F1-measures of these 5 methods, we can find that indicator features are beneficial to improving the performance of the F1-measure for all of the emotions except boredom.

For HEIM, the F1-measures of the 4 methods for the 6 emotion categories are shown in Fig. 13(b). We can see that only the acoustic features perform better than only the textual features (+26.1%). In contrast, for the DSNN model, as shown in Fig. 13(a), using text features alone leads to better performance than does using acoustic features alone, which indicates that when inferring emotions in VDAs, the detailed frame level acoustic features (as in the HEIM) may be more informative than the text features. In addition, when these features are computed at the utterance level (as in the DSNN), then they become less detailed and perhaps less informative than text. Further, we compare the Textual feature model to the Textual+Acoustic feature model and find that acoustic features can make a +46.6% improvement on average. In particular, it achieves a +66.4% improvement for *Happiness* and a +73.7% improvement for *Sadness*. These experiment results prove that it is necessary and effective to account for the acoustic information of utterances.

D. Use of Unlabeled Data

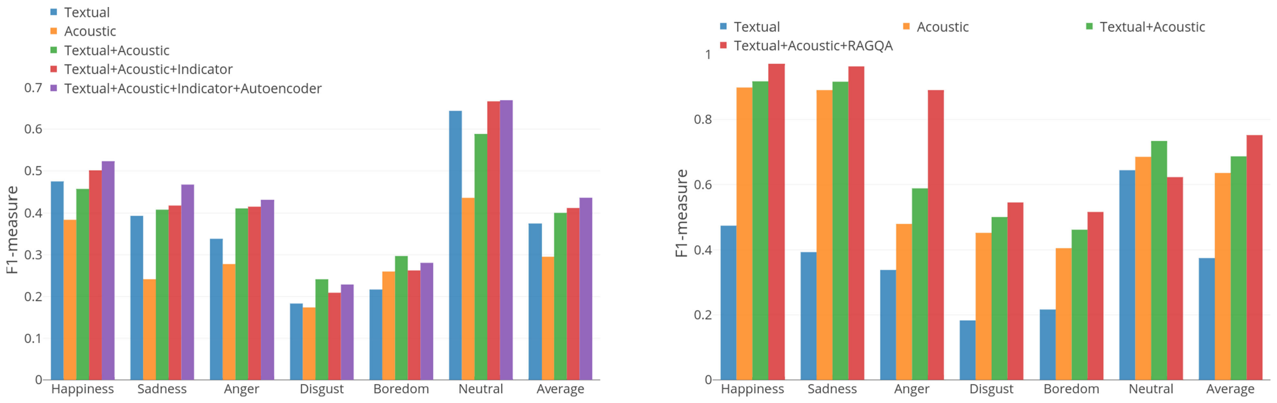
In terms of the DSNN, based on Fig. 13(a), we can see that pretraining with an auto-encoder helps improve the performance for all 6 emotion categories. More importantly, we have the best average result when we combine the auto-encoder with all three indicator features.

In Fig. 13(c), we can find that the application of RAGQA, which utilizes unlabeled data to pre-train the LSTM, also helps improve the F1-measure of all 6 emotion classes. Compared to the Textual+Acoustic feature model, the query attributes in HEIM can improve +6.5% for the performance, on average. Additionally, they are enhanced by +30.3% on the inference of *Anger*. As a result, we find that HEIM, which combines textual information processed by LDA, acoustic features of utterances generated by LSTM and query attributes, has the best performance.

E. Parameter Analysis

Compared with other parameters, such as the number of layers, batch size and number of epochs, we find that the number of cells in the LSTM influences the result notably. In addition, the scale of the unlabeled data set is a key factor that influences the performance. Fig. 14 illustrates the influence of the parameter changes in HEIM on the performance of emotion inference in VDAs.

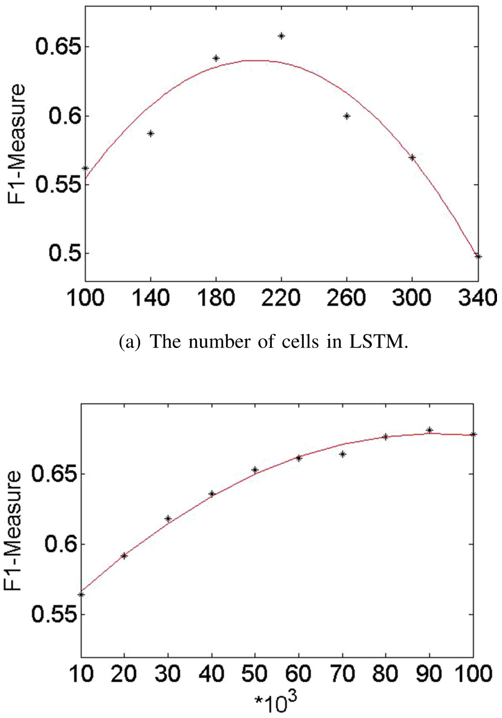
- *The number of cells in LSTM:* As shown in Fig. 14(a), the performance first becomes better and then declines with an increase in the number of cells in the LSTM. In terms of the F1-measure, the performance is the highest (0.658) when the number of cells in the LSTM is 220. Therefore, in the experiments above, we set the number of cells in the LSTM to 220.
- *The scale of the unlabeled data in RAGQA:* As shown in Fig. 14(b), performance gradually improves as the size of the unlabeled data applied for LSTM pretraining increases. Further, when the scale of the unlabeled data is larger than



(a) DSNN feature contribution analysis.

(b) HEIM feature contribution analysis.

Fig. 13. Feature contribution.



(a) The number of cells in LSTM.

(b) The scale of unlabeled data in RAGQA.

Fig. 14. Parameter sensitivity analysis.

90,000, the performance reaches convergence. Using a 24 core 2.10 GHz CPU with 64.0 GB memory, the training lasts for 10–12 hours on a data set containing 10,000 utterances. In fact, we have more than 6 million unlabeled utterances. However, for time efficiency, we perform experiments on the data set containing 90,000 unlabeled utterances.

F. Summary of DSNN and HEIM

Although the performance of HEIM (0.7523) is much better than that of the DSNN (0.4355) in terms of the F1-measure, the processing time of HEIM is much longer than that of the DSNN. In our experiment, during five-fold cross validation on 2942 labeled utterances, on a 24-core 2.10 GHz CPU with 64.0 GB memory, the processing time of the DSNN is 150 seconds, while that of the HEIM is 6000 seconds. It shows that the HEIM is 40 times more time-consuming than the DSNN is because HEIM considers text features additionally and LSTM keeps time sequence information. Thus, the DSNN is a better choice when processing time is limited, and we need to update the model frequently. HEIM is considered better when a higher classification performance is required.

G. Error Analysis

Finally, we analyze the possible sources of errors based on the emotion inference results of the proposed DSNN and HEIM.

- *Limited labeled data:* Due to the massive scale of our dataset, we cannot label every utterance manually. Thus, we only utilize 2,942 labeled utterances to train the DSNN, which may be not enough to obtain a well-trained model.
- *Unbalanced data:* As 61.3% of the labeled utterances belong to the Neural category, the data are extremely unbalanced, which has a negative impact on the classification results.
- *Limited emotion categories:* Inferring emotion categories is a very difficult task because emotion is highly subjective and complicated. Occasionally, different types of feelings can mix together. Thus, new types of emotional labels emerge. Currently, there is still no consensus on how to model emotion. Thus, the 6 categories that we adopt may not cover all the human feelings in the VDAs.

H. Findings

To study the problem of emotion inferring from Large-scale Internet Voice Data for VDAs, the main challenge is that the tremendous numbers of users in VDAs lead to a variety of users accents and expression patterns. Thus, there are insufficient training data available for specific users. Therefore, in our proposed DSNN and HEIM two methodologies, besides considering acoustic features and text features, we also take the Query Attributes of users (temporal indicator, descriptive indicator and geo-social indicator) into modeling. The experimental results on a real-world VDA dataset prove that these can be helpful for the unspecific speaker problem in VDAs.

Additionally, based on the dataset from the Sogou Voice Assistant, we also have some interesting findings which can be helpful for the study of emotion inferring from Internet Voice Data. For example, using MFCC feature alone achieves a $+22.35$ improvement over using Energy feature alone by HEIM method in terms of the F1-measure. Therefore, frequency domain acoustic features has more of an impact on emotion expression on Internet voice data than time domain acoustic features. Meanwhile, Energy feature is really important to distinguish Sadness from others. Normalizing volume of voice data before feature extraction which mainly affects the energy feature causes a 21% reduction by DSNN method in terms of the F1-measure of Sadness.

VIII. Conclusion

In this paper, we study the problem of classifying emotions in utterances from large-scale internet voice data coming from VDAs. At first, we consider how to measure the emotions of large-scale internet voice data and exploit five major emotion categories. Then, we investigate whether social attributes are related to inferring the emotion. Taking these observations into consideration, we propose two methodologies to solve the problem. In the first methodology (Deep Sparse Neural Network), we identify three indicators (temporal indicator, descriptive indicator and geo-social indicator) and combine them with acoustic and text features as the input of our whole network. In the second methodology (Hybrid Emotion Inference Model), we apply Long-Short Term Memory (LSTM) to generate the acoustic features. Additionally, to maximally utilize unlabeled data and further improve accuracy, we apply an auto-encoder to pre-train the network in an unsupervised way. Additionally, we apply back-propagation optimization to fine-tune the DSNN. In HEIM, a Recurrent Auto-encoder Guided by Query Attributes (RAGQA), which combines other emotion-related query attributes, is employed to pre-train the LSTM. Evaluating these two methodologies, we find that HEIM outperforms the traditional methods, such as BN, SVM and KNN, on a large-scale voice data set in the real world. However, in settings in which processing speed is a concern, the proposed DSNN method trains much faster than the HEIM and outperforms traditional methods such as BN, SVM and KNN that have comparable processing speeds. Thus, we suggest utilizing HEIM for accuracy and employing DSNN when focusing on timeliness.

For our future work, more information can be considered. For example, users of different genders may have different emotional expression characteristics in terms of their word choices or different acoustic characteristics. In addition, users emotions may be influenced by weather. Additionally, in this paper, we use 93,000 utterances, but 93,000 is not an enormous number. We want to expand the scale of our data set to one million; thus, we can employ more voice data to improve the performance of our methodologies.

As VDAs are becoming increasingly popular, users expect to communicate with them not only by instructions and queries but also through chats and conversations. According to our work, VDAs can take users emotions into account and better understand their intentions when giving responses, thereby optimizing the interactions.

References

[1] M. Meddeb, H. Karray, and A. M. Alimi, “Speech emotion recognition based on arabic features,” in *Proc. 15th Int. Conf. Intell. Syst. Des. Appl.*, 2015, pp. 46–51.

[2] Y. Tao, K. Wang, J. Yang, N. An, and L. Li, “Harmony search for feature selection in speech emotion recognition,” in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 362–367.

[3] B. Schuller, G. Rigoll, and M. Lang, “Hidden Markov model-based speech emotion recognition,” in *Proc. Int. Conf Multimedia Expo.*, 2003, vol. 1, pp. 401–404.

[4] Y. Pan, P. Shen, and L. Shen, “Speech emotion recognition using support vector machine,” *Int. J. Smart Home*, vol. 1, no. 20, pp. 6–9, 2013.

[5] T. Zhang *et al.*, “A deep neural network-driven feature learning method for multi-view facial expression recognition,” *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[6] G.-J. Qi, H. Larochelle, B. Huet, J. Luo, and K. Yu, “Guest editorial: Deep learning for multimedia computing,” *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1873–1874, Nov. 2015.

[7] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,” *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[8] G. Hinton *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[9] X. Zhou, J. Guo, and R. Bie, “Deep learning based affective model for speech emotion recognition,” in *Proc. Int. IEEE Conf. Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People, Smart World Congress*, 2016, pp. 841–846.

[10] I. Trabelsi, D. B. Ayed, and N. Ellouze, “Improved frame level features and SVM supervectors approach for the recogniton of emotional states from speech: Application to categorical and dimensional states,” *Int. J. Image Graph. Signal Process.*, vol. 5, no. 9, pp. 8–13, 2013.

[11] L. Gao, L. Qi, E. Chen, and L. Guan, “A fisher discriminant framework based on kernel entropy component analysis for feature extraction and emotion recognition,” in *Proc. IEEE Int. Conf Multimedia Expo Workshops*, 2014, pp. 1–6.

[12] D. Phung, S. K. Gupta, T. Nguyen, and S. Venkatesh, “Connectivity, online social capital, and mood: A Bayesian nonparametric analysis,” *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1316–1325, Oct. 2013.

[13] H. Yin, B. Cui, L. Chen, Z. Hu, and Z. Huang, “A temporal context-aware model for user behavior modeling in social media systems,” in *Proc. Assoc. Comput. Mach. Special Interest Group Manage. Data Int. Conf. Proc.*, 2014, pp. 1543–1554.

[14] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng, “Spatio-temporal dynamics of online memes:A study of geo-tagged tweets,” in *Proc. Int. Conf. World Wide Web*, 2013, pp. 667–678.

[15] M. E. Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[16] P. Shen, Z. Changjun, and X. Chen, “Automatic speech emotion recognition using support vector machine,” in *Proc. Inte. Conf Electron. Mech. Eng. Inf. Technol.*, 2011, pp. 621–625.

[17] Y. Fan, X. Lu, D. Li, and Y. Liu, “Video-based emotion recognition using CNN-RNN and c3d hybrid networks,” in *Proc. 18th ACM Int. Conf. Multimodal Interact*. ACM, 2016, pp. 445–450.

[18] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[19] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, “Disentangling factors of variation for facial expression recognition,” in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 808–822.

[20] D. Yu, M. L. Seltzer, J. Li, J. T. Huang, and F. Seide, “Feature learning in deep neural networks - studies on speech recognition tasks,” presented at the Int. Conf. Learn. Represent., Scottsdale, AZ, USA, May 2013.

[21] W. Zheng, J. Yu, and Y. Zou, “An experimental study of speech emotion recognition based on deep convolutional neural networks,” in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 827–831.

[22] M. H. Sedaaghi, C. Kotropoulos, and D. Ververidis, "Using adaptive genetic algorithms to improve speech emotion recognition," in *Proc. IEEE Workshop Multimedia Signal Process.*, 2007, pp. 461–464.

[23] N. Thapliyal, "Speech based emotion recognition with gaussian mixture model," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 5, pp. 65–69, 2012.

[24] X. Cheng and Q. Duan, "Speech emotion recognition using Gaussian mixture model," in *Proc. 2nd Int. Conf Comput. Appl. Syst. Model.*, 2012, pp. 1222–1225.

[25] M. Shah, L. Miao, C. Chakrabarti, and A. Spanias, "A speech emotion recognition framework based on latent dirichlet allocation: Algorithm and FPGA implementation," in *Proc. IEEE Int. Conf Acoust., Speech Signal Process.*, 2013, pp. 2553–2557.

[26] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 223–227.

[27] A. Stuhlsatz, C. Meyer, F. Eyben, and T. Zieike, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 5688–5691.

[28] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 801–804.

[29] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. Int. Conf Affect. Comput. Intell. Interact.*, 2015, pp. 827–831.

[30] D. Le and E. M. Provost, "Data selection for acoustic emotion recognition: Analyzing and comparing utterance and sub-utterance selection strategies," in *Proc. Int. Conf Affect. Comput. Intell. Interact.*, 2015, pp. 146–152.

[31] R. Cbalabantaray, M. Mohammad, and N. Sharma, "Multi-class twitter emotion classification: A new approach," *Int. J. Appl. Inf. Syst.*, vol. 4, no. 1, pp. 48–53, 2012.

[32] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting emotions in twitter messages," in *Proc. ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conf.*, 2014, pp. 27–31.

[33] X. Wang, J. Jia, J. Tang, B. Wu, L. Cai, and L. Xie, "Modeling emotion influence in image social networks," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 286–297, Jul.–Sep. 2015.

[34] B. Wu, J. Jia, Y. Yang, P. Zhao, and J. Tang, "Understanding the emotions behind social images: Inferring with user demographics," in *Proc. IEEE Int. Conf Multimedia Expo*, 2015, pp. 1–6.

[35] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multi-task shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, Mar. 2017.

[36] N. Li, Y. Xia, and Y. Xia, "Semi-supervised emotional classification of color images by learning from cloud," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2015, pp. 84–90.

[37] J. Jia *et al.*, "Can we understand van gogh's mood?: Learning to infer affects from images in social networks," in *Proc. ACM Int. Conf Multimedia*, 2012, pp. 857–860.

[38] E. Kim, S. Gilbert, M. J. Edwards, and E. Graeff, "Detecting sadness in 140 characters: Sentiment analysis and mourning michael jackson on twitter," *Web Ecol.*, vol. 3, pp. 1–15, 2009.

[39] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. 4th Int. AAAI Conf Weblogs Social Media*, Washington, DC, USA, May, 2010, pp. 122–129.

[40] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *Comput. Sci.*, vol. 44, no. 12, pp. 2365–2370, 2009.

[41] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2010.

[42] Y. Zhao, Y. Li, Z. Shao, and H. Lu, "Lsod: Local sparse orthogonal descriptor for image matching," in *Proc. ACM Multimedia Conf.*, 2016, pp. 232–236.

[43] X. Yao, J. Han, G. Cheng, and L. Guo, "Semantic segmentation based on stacked discriminative autoencoders and context-constrained weakly supervised learning," in *Proc. 23rd ACM Int. Conf Multimedia.*, 2015, pp. 1211–1214.

[44] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf Multimedia.*, 2014, pp. 7–16.

[45] B. Wu, J. Jia, T. He, J. Du, X. Yi, and Y. Ning, "Inferring users' emotions for human-mobile voice dialogue applications," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2016, pp. 1–6.

[46] D. Cui, "Analysis and conversion for affective speech," Ph.D. dissertation, Dept. Comput. Sci. Technol., Tsinghua University, 2007.

[47] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.

[48] Z. Ren, J. Jia, L. Cai, K. Zhang, and J. Tang, "Learning to infer public emotions from large-scale networked voice data," in *Proc. Int. Conf Multimedia Model*. Springer, 2014, pp. 327–339.

[49] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 690–701, Feb. 2007.

[50] H. Kawahara, A. D. Cheveign, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight," in *Proc. Eurospeech, Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sept. 2005, pp. 537–540.

[51] J. Mei, *Tongyici Cilin (Version 2)*. Shanghai, China: Shanghai Dictionary Press 1996.

[52] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. Int. Conf Mach. Learn*, Bellevue, WA, USA, 28 Jun.–Jul. 2012, pp. 689–696.

[53] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, no. 8, pp. 1967–2006, 2012.

[54] Y. Yang *et al.*, "How do your friends on social media disclose your emotions?" in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 306–312.

[55] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3 pp. 993–1022, 2003.

[56] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[57] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transfer Learn.*, 2012, pp. 37–49.

[58] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 873–880.

[59] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, 1997.

[60] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2362–2365.

[61] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[62] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream ASR framework for BLSTM modeling of conversational speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 4860–4863.

[63] M. Wöllmer, B. Schuller, and G. Rigoll, "A novel bottleneck-BLSTM front-end for feature-level context modeling in conversational speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011, pp. 36–41.

[64] A. Graves, N. Jaitly, and A. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2013, pp. 273–278.

[65] Y. A. LeCun, L. Bottou, G. B. Orr, and K. Muller, "Efficient backprop," in *Proc. Neural Netw., Tricks Trade*, 2012, pp. 9–48.

[66] F. Weninger, S. Watanabe, Y. Tachioka, and B. Schuller, "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4623–4627.

[67] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 83–92.

[68] J. Tang *et al.*, "Quantitative study of individual emotional states in social networks," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 132–144, Apr.–Jun. 2012.

[69] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Stat.*, vol. 46, no. 3, pp. 175–185, 1992.

[70] S. Emerich, E. Lupu, and A. Apatean, "Emotions recognition by speechand facial expressions analysis," in *Proc. Eur. Signal Process. Conf.*, 2009, pp. 1617–1621.

[71] J. Jia *et al.*, "Can we understand van gogh's mood?learning to infer affects from images in social networks," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 857–860.

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

Jia Jia is currently an associate professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Her research interests include affective computing and human-computer speech interaction.

Suping Zhou is currently working toward the doctoral degree with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Her research interests include affective computing and human-computer speech interaction.

Yufeng Yin currently an undergraduate with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include human-computer interaction and affective computing.

Boya Wu received the Master degree with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Her research interests include affective computing and social network analysis.

Wei Chen received the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications, Beijing, China, in 2011. He has been Chief Scientist of the voice interaction technology center with Sogou Corporation, Beijing, China, an innovator in search and a Leader in China's internet industry, since 2016. Over the past six years, he led his team to establish the company's voice-based multimodal human-computer interaction operating system, the Zhiyin OS. His research interests include multi-modal interaction technology, speech recognition and synthesis, and machine translation.

Fanbo Meng received the doctorate degree from the Department of Computer Science and Technology from Tsinghua University, Beijing, China. He has been working in Sogou Company for more than 4 years. His research interests include TTS acoustic feature modeling, virtual anchor, and waveform modeling such as WaveNet.

Yanfeng Wang is the General Manager of Sogou Voice Interaction Technology Center, Beijing, China. His research interests include speech recognition and synthesis, natural language processing, machine learning, and image processing.