감성과학
제23권 2호, 2020
<연구논문>

pISSN 1226-8593
eISSN 2383-613X
Sci. Emot. Sensib.,
Vol.23, No.2, pp.61-74, 2020
https://doi.org/10.14695/KJSOS.2020.23.2.61

Greeting, Function, and Music: How Users Chat with Voice Assistants

Ji Wang[1†] · Han Zhang[2] · Cen Zhang[3] · Junjun Xiao[4] · Seung Hee Lee[5]

Abstract

Voice user interface has become a commercially viable and extensive interaction mechanism with the development of voice assistants. Despite the popularity of voice assistants, the academic community does not utterly understand about what, when, and how users chat with them. Chatting with a voice assistant is crucial as it defines how a user will seek the help of the assistant in the future. This study aims to cover the essence and construct of conversational AI, to develop a classification method to deal with user utterances, and, most importantly, to understand about what, when, and how Chinese users chat with voice assistants. We collected user utterances from the real conventional database of a commercial voice assistant, NetEase Sing in China. We also identified different utterance categories on the basis of previous studies and real usage conditions and annotated the utterances with 17 labels. Furthermore, we found that the three top reasons for the usage of voice assistants in China are the following: (1) greeting, (2) function, and (3) music. Chinese users like to interact with voice assistants at night from 7 PM to 10 PM, and they are polite toward the assistants. The whole percentage of negative feedback utterances is less than 6%, which is considerably low. These findings appear to be useful in voice interaction designs for intelligent hardware.

Key words: Voice Interaction Design, Chat, Chinese User, Intelligent Hardware

1. Background

In the development of voice assistants, major companies have developed a conversational system that matches their product styles and technical characteristics. Although the functions and styles of the voice assistants from different companies vary greatly, the processing of user utterances is basically the same, and it divides into three categories: function, chat, and overall responses.

The function of the voice assistant, also known as the domain or skill, it is the same as the APP on the mobile phone. However, unlike the graphical interface, the app is visible to the user's eyes, that is, the “what you see is what you get” design specification. Like mobile phone manufacturers developing mobile phones, the development of voice assistants is also divided into system-owned and third-party support. In the mobile phone, such as calling, texting, calendar, alarm clock, and other functions, the conversion to the voice assis-

[1†] (Corresponding Author) Ji Wang: Graduate School of Comprehensive Human Sciences, University of Tsukuba / E-mail : wangjige@foxmail.com / TEL : 81-29-853-2858
[2] Han Zhang: Graduate School of Comprehensive Human Sciences, University of Tsukuba
[3] Cen Zhang: Graduate School of Comprehensive Human Sciences, University of Tsukuba
[4] Junjun Xiao: NetEase Hangzhou Network Co. Ltd
[5] Seung Hee Lee: Faculty of Art and Design, University of Tsukuba

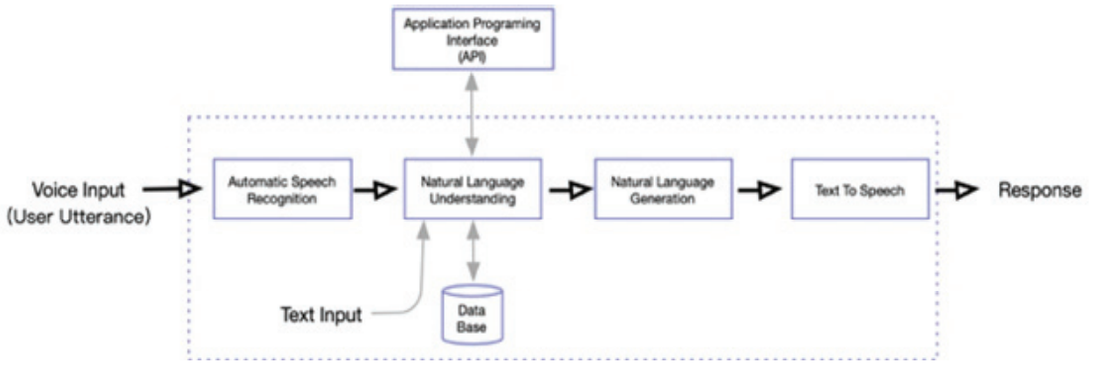

Fig. 1. Components and processes of a typical conversational system

tant is the weather, time, alarm clock, music are the essential mainstream function of the current voice assistant (Hoy, 2018). The development of each function is independent and requires much work. Let us take the example of letting the voice assistant play music to illustrate the implementation process of this function. If the voice assistant is compared to a human being, then to realize the function of letting it play music, first of all, it must be able to hear what the user asked, or to hear that the user is talking to it, here the voice assistant need to use automatic speech recognition (ASR). ASR algorithm is very sophisticated, involving signal processing, audio conversion, etc. we will not discuss its technical details here. The quality of the ASR algorithm is related to whether the voice assistant can accurately hear what the user asked. If the performance is not excellent, there will be a “hearing” problem, which will be catastrophic for the accurate execution of the user's intention in the whole system. After the ASR is successfully identified, the recognized speech is converted into text, and the text is processed by the next module: natural language understanding (NLU). NLU actually solves the problem of “understanding.” The basis for its understanding is “to hear these questions.” Here are some methods of artificial intelligence (deep learning, machine learning) and data (database of songs, singers, composers) to train the accuracy of the NLU algorithm, such as collecting a large number of voice commands when the user asks for play music, extracting keywords in the music command. After NLU understands the user's question, it moves to the next module: the cloud. The cloud is where data is stored. The path to the cloud is called application programming interface (API). Developing API requires a lot of work and investment, which limits the function of voice assistant at an early stage. After the request of API is working, another module natural language generation (NLG) will start. Its function is to reply to the user's question, generate the corresponding reply text. It finally converts the text into voice through the text to speech (TTS) module, and reply to the user (see Fig. 1).

The above is the implementation of the voice assistant's functions. In addition to calling specific functions, users have other types of interactions with voice assistants, such as greetings, complaints, and so on. All the user's utterance can develop into a function, but as we know about the function module above, it needs much investment, and the investment cannot meet the real use condition. Generally, these companies will choose to develop another “function” with a much more extensive coverage: chat.

The research of chit-chat in the world of artificial intelligence is much earlier than the voice assistant, ELIZA is an early attempt to create artificial intelligence through human interaction (Weizenbaum, 1966). Now the chat of the voice assistant can be understood as the traditional chat robot, from the text chat to the voice chat process, that is, based on its ability to pronounce (TTS).

The chat algorithm, similar to other functions (such as music, weather) algorithms, is based on a computational model, coupled with the training of a large amount of data, to get a model that can handle user questions, how to feedback these user questions. Cho et al. (2009) found that emotion voice feedback can increase participant's perceptional fun and hedonic quality. Duaquett et al. (2008) found the social robots can help autistic children focus on their attention and express their emotions. Nass and Brave (2005) suggested that computers embedded with human voices should be received as a natural and powerful modality in human-computer interaction. These may be the reason why user likes to use voice interaction with conversation artificial intelligent. The mainstream voice assistant has two ways to handle it: cloud response and custom response, the mainly difference between these two algorithms are the results of cloud response is unexpected; the custom response is controllable. Obtaining from the cloud, as the name suggests, is to find the most matching answer to this utterance based on the algorithm model. For example, a user asks, “How are you today,” which can find the same from the dialogue from a movie. Ask the question, and then transplant the similar question and answer into the reply. The advantage of this method is that the number of the answer is enormous, and the way of answering can change according to the nuances of the user's question. The problem of the cloud is also apparent: the answers of the voice assistant are unpredictable, which bring a huge risk for the company. Tay, a chatbot Microsoft introduced in the US market in the early days. After the launch for a few days, it learned swearing according to the user's tweets data and become a "racist" (Price, 2016). Under the pressure of public opinion in the United States, Microsoft had to delete it. Microsoft XiaoIce, which was online in China, also under public pressure. Therefore, commercial companies mostly adopt the second way: custom response. The custom response is straightforward. It is the staff of the company, select, edit, and expand the response to the relevant questions. The advantage of this is that the answers to all questions can be controlled, there will be no “wrong” answers, and customized answers to specific questions, such as asking about the company's related business, product, or founder. In the future, when voice interaction becomes the mainstream advertising portal, this method is also more suitable for customer advertising customization. The shortcomings of this method are also apparent. It requires a lot of workforce input. The quality of the answer depends on the level of the editors. In order to maintain the consistency of the voice assistant's personality, it requires a high level of management. In the initial stage of investment, there will also be a large number of user questions that cannot answer accurately. If the user's chat data appears in a large number of categories, for example, users often ask questions such as “XXX plus XXX equals?” can develop a new function “calculator” based on the question data in these chats data. Moreover, the user's utterances can be used as the primary database to train the function model to identify the “calculation” intent. In the early stage of developing voice assistant, there will be a large number of unanswerable questions (new questions, no corresponding answers) can only be through the “universal answer,” which enters the third and final layer: overall.

Overall, it is the last layer of the conversational system. The question of entering here is filtered by the function and chat algorithm model. There is no need for a new algorithm model to deal with these questions. Only a few answers need to edit to reply to these questions randomly, for example: “Sorry, I don't understand what you are saying” “Sorry, I may not know what you mean,” and so on. These user utterances are precious and can be used to test the functionality of the chat algorithm model, while the optimization of the ASR model base on them.

The human language is very complicated, and the development of artificial intelligence has not yet reached the point of omnipotence. Even human beings themselves will have problems such as inaudible, incomprehensible, and slippery, not to mention the fact that in the early stages of development, the algorithm

model and training data are not matured enough. Only by knowing what users are talking with the voice assistant, researchers can better enable voice assistant to help users in their lives.

According to the data from Canalys (2019), China overtakes US to become the largest smart speaker market, reached 10.6 million units by growing nearly 500% year on year in 2019 Q1, taking up 51% of current global share. However, few studies about the user behaviors of Chinese voice assistant can be found.

What's more, four main voice assistants in global voice market, Alexa, Siri, Google assistant, and Cortana, all take English as the first language, as they built by companies headquartered in bay North America. They applied the use of voice assistants in different countries just by translating English to local languages, which do not consider the peculiarity of different language. Yoo et al. (2020) made the compartive study by interviewing UI/UX design experts in Korea, they can only use the localization intelligent persomnal assistants from English to compare with Samsung's Bixby. So the experience of voice interaction research on Chinese can help to explore the future use and design guideline of voice assistant in other language.

Most commercial companies focus on the development of different voice skills according to the characteristics of their products. For example, Amazon, the world's largest e-commerce company, has used online shopping as its principal function even in the early days of developing Alexa. Most users will not use this feature, and for the data of the conversation, these companies will also be regarded as core business secrets, not announced. In academia, there are also studies on voice assistants and chat conversations. The main problem is that their research focuses on particular people, such as the elderly, children, office crowds, or using log research methods to track a small number of users' usage habits, the amount of data is relatively small. User groups are very limited. In order to resolve the contradiction between business and academics, we obtained the original user data from a commercial voice assistant to analyze what users in China have talked with the voice assistant. In the process of solving this problem, our classification and data processing method based on the current technology of voice assistant is also proposed, which will bring positive influence to developers in the development of commercial voice assistants in the future.

Table 1. Five categories of performative verbs

Category	Definition
Expositive-verbs	Asserting or expounding views, classifying usages and references;
Exercitives-verbs	Issuing a decision that something is to be so, as distinct from a judgement that it is so;
Verdictives-verbs	Delivering a finding, official or unofficial, upon evidence or reason as to value or fact;
Commissives-verbs	Committing the speaker to some course of action;
Behabitives-verbs	Involving the attitudinal reaction of the speaker to someone's conduct or fortunes

2. Introduction of Netease Sing Cloud Speaker

Netease Sing Cloud Speaker (In Chinese: 网易三音云音箱) is a brand of smart speakers developed by Netease Hangzhou Network Co. Ltd, based on the 400 million users of Netease Cloud Music Sing Cloud Speaker connect to the voice-controlled assistant Sing. The device can provide the service include: play music, setting alarms, setting reminders, provide weather information, time check, and free chat (service like an open domain). It can also connect with several IoT devices, such as light, fan, and air condition, acting as the control center of smart devices. It introduced to the market in June 2018, the voice interaction with the device is currently available in Mandarin, and Mandarin mixed with English. We randomly selected the data for a particular day in August 2018, as the product has been on the market for two months, and the user utterance data tends to be stable. We excluded the user utterance data of the function domain. The selected data are all from chat and overall, a total of 20,234 user utterances, including 13428 chat and 6806 overall data.

3. User Utterance Category

To understand what users and voice assistants are talking is a long-term job. In addition to getting the raw data of the user's utterance, we need to classify and define the types of user utterance. Many studies classify user conversations, their researches are based on different purposes, and the classification is varied.

As shown in Table 1, the classic categorization is from Austin (Austin, 1962), postulates five major speech act classes based on five categories of performative verbs (Moldovan et al., 2011).

Moldovan et al. (2011) also reported the speech act taxonomy in the online chat corpus. There are statements, system, greet, emotion, wh-question, continuer, accept, reject, bye, yes answer, no answer, emphasis, clarify, and other, total 15 classifications. This study focus on the online chat to indicate speakers' intentions.

Sato-Shimokawara et al. developed and researched chat robot for older people in Japan, and they estimated a category of the user utterances (Sato-Shimokawara et al., 2016), the categories included health, environment, society, music, event, family, washing, fashion, go-out, meal, game, work, media, machine, sleeping, cleaning, exercise, cooking, and other, totally 19 categories. Nevertheless, this study mainly focused on older people, and included all the functions, such as music, is treated as an independent domain in all the mainstream voice assistants.

Akasaki & Kaji (2017) classified the non-task-oriented utterances of Yhaoo! Voice Assist according to their dialogue acts into 11 categories: greeting, self-disclosure, order, question, invitation, information, thanks, curse, apology, interjection, and misc.

The above is a review of our existing classification of chat; their classification is based on the purpose of their own research, and does not follow a fixed norm. Combined with these studies, we also categorized existing user chat and overall questions. As we introduced above, chats often appear in the future development plan, and become an independent domain, because training a mature domain requires a large number and various types of user utterances. Those are currently difficult to become specific functions, and we will classify them according to the users' intentions. The purpose is to form a unified and standardized answer when editing such questions, and some classifications base on practical working necessary, for example, to optimize the ASR algorithm, or develop new NLU calculations. However, there will still be a part that cannot be classified, just like human gibberish about others, which we can only say, “Sorry, I do not understand what you say,” those are “meaningless” for both humans and artificial intelligence.

We divide some user utterances into major categories by direct taking (greeting, emotion, music), combine (media, game, exercise et al., to function), and rename (information to encyclopedia; question to ask back; interjection to modal particle; curse to negative feedback; other to meaningless; self-disclosure to personality) from previous studies. Considering the current technology used on voice assistant, we make two categories: recognition (ASR) and multi-intent (NLP). Moreover, NetEase Sing is a commercial voice assistant, so we need to take the commercial part into the classification: competitor and company. Least we create the last categories from intent analysis: joke, story, ridicule. Here are these categories and their definitions and examples.

Greeting: This category contains the user's positive, positive dialogue with the voice assistant, mainly as the beginning of a conversation, including morning and evening greetings, holiday blessings, greetings, gratitude, compliments, and some greetings with Chinese style, such as: “Have you eaten”.

Function: This classification contains the user's question about the voice assistant function, including three sub-categories. The first category is the voice operation of the existing function. Because the algorithm model is not perfect, mainly because of the NLU, it is not able to perform what the user wants to operate.

Music: This category consists of tasks related to music information. As we know, many song names are like daily chatting or items, so the natural language understanding model has the difficulty of deciding this query to be music or chat.

Meaningless: It contains utterances that have logical problems or identify errors, single phrases, single Chinese characters, and challenging to define categories that cannot understand in the dialogue.

Ask back: This category contains categories that require the user's intent to be retrieved again in order to perform the next step. For example:

"Play" (Play what?)

"Can you" (Can me what?)

"I want to listen" (Music, joke, or story?)

Emotion: This category includes conversations in which the user expresses emotions, seeks comfort and companionship, or treats voice assistants as human beings, seeking to establish intimate relationships and share conversations about their intimate relationships.

Personality: This category contains the context of the voice assistant's birth, family, personality, hobbies, abilities, behavioral habits, and other related conversations.

Negative feedback: This category includes insults, complaints, sexual provocations, or conversation with sigma.

Ridicule: This classification includes having the voice assistant perform tasks that significantly exceed its capabilities, using web buzzwords for conversations, requiring it to perform other voice assistant features, and other conversations that are fun but not insulting.

Encyclopedia: This category consists of tasks related to common sense related questions, including geography, characters, and commom knowledge.

Joke: This category includes letting voice assistants tell jokes. This category includes letting voice assistants tell jokes, including unspecified types of jokes and specific types of jokes, including telling a joke again or speaking many jokes at once.

Story: This category includes letting the voice assistant tell stories, including unspecified types of stories and specific types of stories, including retelling a story or telling multiple stories at once.

Modal particle: This category refers to users using modal particles as a dialogue, which is hard to know users' real intent. For example:

"La la la"

"Ha ha ha"

"Oh Oh Oh Oh Oh"

Recognition: It includes conversations that are meaningful in speech because some words are incorrectly identified and cannot understand by the voice assistant.

Competitor: It includes conversations related to other voice assistants, including their company, features, and their comparisons.

Company: This category includes companies that develop voice assistants, the company's history, founders, products, reviews, and more.

Multi-intent: This classification includes having the voice assistant perform two or more tasks, or answer two or more questions.

4. Methods

We collected 20,234 utterances (The original utterances are all in Chinese. Example utterances given in this thesis are Chinese and its English translations). The utterances are the results of automatic speech recognition from the real conventional data of a commercial voice assistant, NetEase Sing. Every single data included user utterances, resources (chat or overall), and time. To protect privacy during the data collection, we removed the information of personal privacy, including the name and IP of the user.

Follow the method of Akasaki & Kaji (2017), we recruited volunteers to annotate the 20,234 utterances with seventeen labels. The definition and example of every category were explained to the volunteers by the experimenter. Three volunteers are all native

Chinese speakers with master degree in Chinese language and literature, all females aged between 25 years and 27 years. The volunteers annotated the label to every utterance. For example, the volunteers annotated the label "Joke" when users were going to ask the voice assistant to tell a joke. Note that this voice assistant works primarily on a smart speaker, and thus, the utterances include many operational instructions such as alarm setting. Three volunteers assigned to each utterance, and the final labels were obtained by majority vote to ensure the quality issue inherent in crowdsourcing. As there are so many categories, when the utterance three volunteers annotated in three different labels, the experimenter and volunteers started a group discussion to decide the category it belongs to. The utterance will be annotated the label "Meaningless" when the ground discussion cannot reach an agreement.

Knowing what users and voice assistants are talking about is the beginning of the research, in order to further understand, we also analyze the time of user questioning to understand when users chat with voice assistants.

5. Results

We start by describing the top 20 questions asked by the users. The reason for analyzing such results is that in the dialogue design of the existing voice assistant, in order to reduce the mechanical sense and unnaturalness of the voice assistant, multiple answers are often set for frequently occurring questions, or multiple high-frequency functional queries are added. Sentences, including core information, are more similar to conversations between people. Alternatively, add multiple sentences to the high-frequency function query. For example, when the user asks the voice assistant to play a specific song of Taylor Swift, we can say, "OK, play …by Taylor Swift", or "Please enjoy this song by Taylor Swift."

Rank first, eleventh, and fifteenth user utterance is the name of the voice assistant. In the design of the voice assistant, whether it is woken-up or wake-up, it will enter the state of "listening." The time of this state lasts about 7 to 8 seconds without voice input, fine-tuned according to the strategy of each company.

The second most asked question by the user is "Good night". In the top 20, there are four other words that are used in daily greetings. They are "Hello," "Sleep," "Bye," and "Good morning".

In the third place is "Turn on," accounting for 1.1% of the total. In the design of the voice assistant, the voice assistant only needs to be powered on when it is used for the first time, connected to the network, and then always turned on, or waiting for the state of being awake. The user transferred the operating habits of using the mobile phone to the use of the voice assistant.

The fourth most used features of the voice assistants are telling jokes, which account for 0.8%. Besides, the question of letting users tell stories is also in the top 20, accounting for 0.5%. Behind the joke is the question of needing additional information, "I want to listen," this can be listening to songs, jokes, or stories, so voice assistants need to ask the user, "What you want to listen." In the top 20, there are three other types, namely "Play," "Listen," and "Have listened," all of which are most relevant to music. More interesting is the ninth question, "Put a fart." It is another voice assistant in the Chinese market, the function of the T-mall Genie. Due to its popularity, the voice assistant users know this feature, so it will be used to test their voice assistant. Similarly, there is the final ranking, "Imitate cat saying," which is the name of a song that was very popular at the time. So the designers of voice assistants should keep up with the trend and seize the social hotspots and update the question and answer according to this.

The two questions that are listed later are related to the voice assistants, respectively, "What is your name," and "Who are you." The last thing to say is a func-

tion-related question, “One plus one equals?”, this is a question that can be used as a calculator function, but in our sampled NetEase Sing voice assistant it has not been developed yet.

As mentioned earlier, we have classified all conversations into 17 different categories. Below we will explain the number and proportion of these 17 categories. As shown in Table 2, in all categories, greetings accounted for the most, reaching 19.5% of the total. In the category of this accounted for 27.1%, which is the most common category of chat. The second most important question is the function. The percentage in the overall (15.5%) is slightly higher than 13.3% in the chat. The proportion of music is very high, ranking third, with a total of 10.9%. It is worth noting that the proportion of music in the overall situation is as high as 21.2%, which is the second most critical category in the overall situation. The meaningless proportion is 10%, 3.6% in chat, but itis the highest (22.6%) in the overall proportion, it makes sense as those utterances can not be processed by the current NetEase Sing NLP model will enter overall data. Because chat is not able to support multiple rounds of dialogue, the proportion of ask back is 8.7%. The user asks the voice assistant for 7.3% of the time about emotional problems.

We will not discuss more sub-categories here. The personality user about the voice assistant is very concerned, ranking the third in the chat, reaching 9.3%, and the personality in the overall situation is only 1.6%. The next is negative feedback, which is the user's direct expression of dissatisfaction with the voice assistant, with a total proportion of 5.7%. The next ridicule, jokes, and stories are all towards entertainment, accounting for 4.1%, 2.8%, and 2.5%, respectively, which together account for 9.4%. Encyclopedia's features account for 2.9% and users do not have many operations using voice assistants to search for knowledge. It is worth noting that the recognition is 2.9% in overall, which is higher than the 0.4% in the chat in both quantity and proportion. The total proportion of the company and the actual products is not very high, a total of 1.4%, but this part of the question and answer related to the company's image, public relations, and other complex external factors need to pay attention. The final ratio is the lowest intention, a total of 0.3%, and concentrated in the overall.

Table 2. 17 Categories of questions in chat and overall

Type	Chat		Overall		Total	
Category	*Percentage*	*Amount*	*Percentage*	*Amount*	*Percentage*	*Amount*
Greeting	27.1%	3,639	4.4%	302	19.5%	3,941
Function	13.3%	1,787	15.5%	1,054	14.0%	2,841
Music	5.7%	768	21.2%	1,440	10.9%	2,208
Meaningless	3.6%	488	22.6%	1,538	10.0%	2,026
Ask back	7.7%	1,040	10.7%	725	8.7%	1,765
Emotion	7.8%	1,046	6.4%	436	7.3%	1,482
Personality	9.3%	1,255	1.6%	108	6.7%	1,363
Negative feedback	7.7%	1,029	1.9%	129	5.7%	1,158
Ridicule	4.1%	556	4.1%	276	4.1%	832
Encyclopedia	2.7%	356	3.4%	234	2.9%	590
Joke	4.1%	548	0.3%	22	2.8%	570
Story	3.2%	431	1.1%	74	2.5%	505
Modal particle	1.5%	203	2.4%	160	1.8%	363
Recognition	0.4%	57	2.9%	198	1.3%	255
Competitor	0.9%	120	0.5%	34	0.8%	154
Company	0.7%	99	0.3%	23	0.6%	122
Multi-intent	0.0%	6	0.8%	53	0.3%	59
Total	100.0%	13,428	100.0%	6,806	100.0%	20,234

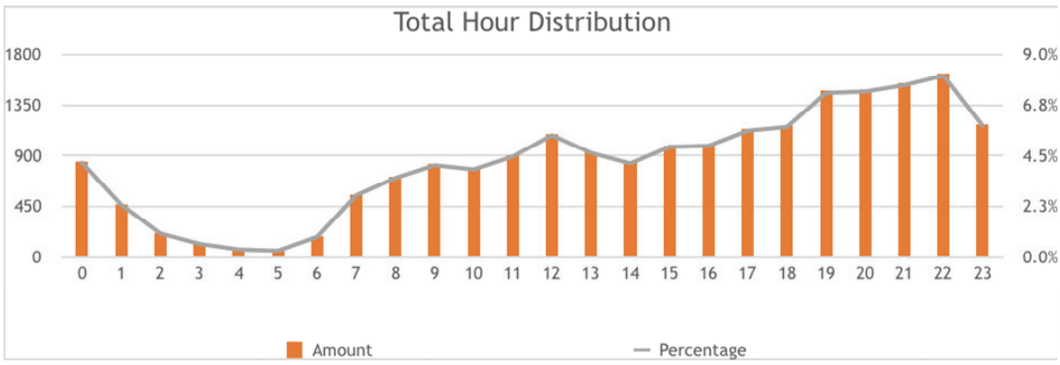

Fig. 2. The hour user distribution of all the utterances

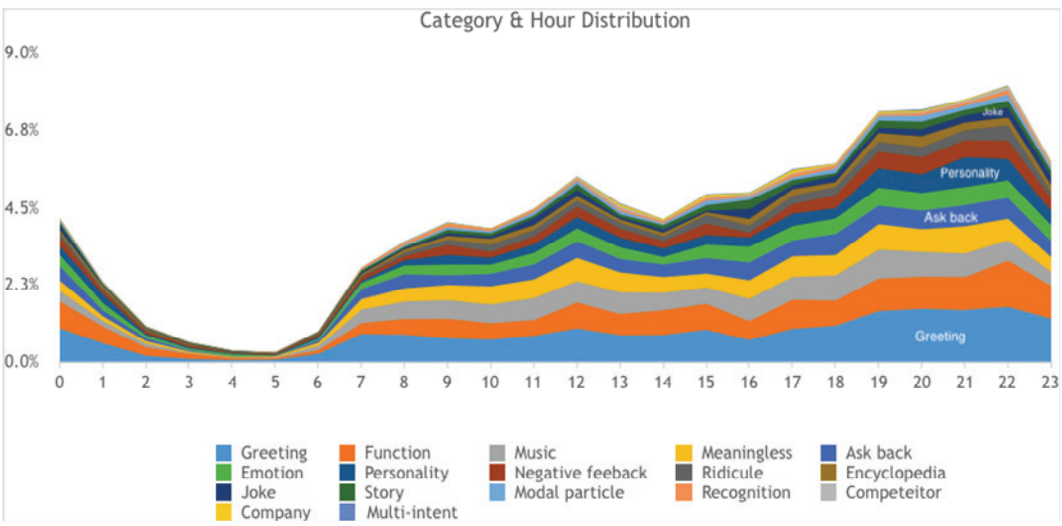

Figure 3. Category and hour distribution

Time of use refers to the time when users participate in a chat with the voice assistant. We consider 24 hours of one day using the local time. As shown in Fig. 2, the amount and percentage of user utterances during different hours of the day. We can see that early morning (4 AM and 5 AM) has the fewest number of user utterances, and the average number of user utterances increases steadily throughout the day and reaches the peak at noon then decreases to around 2 PM. One possible explanation is that, because most Chinese people have the habit of having a short afternoon sleep after dinner, they are sleeping during this time so they cannot chat with the voice assistant at the same time. The user started to use the voice assistant at 7 PM, which resulted in a peak continued until 10 PM. The reason is that most Chinese families have dinner between 6 PM and 7 PM. After dinner, they start the after-dinner activities. We also analyzed the proportion of users using different functions in different periods, as shown in the following Fig. 3.

In different periods, the proportion of greetings is the highest. 7 PM is the peak period for users to use music during the day. During 7 and 8 AM, the music used is the highest in the whole day. An interesting piece of data is that jokes and storytelling are the most demanded at 4 PM. At 0 AM and 1 AM, the proportion of emotions and people at the same time is higher than the proportion of the two classes at other periods. The negative feedback from users after 7 PM is also increasing, which is consistent with the increase in overall usage. At 10 PM, the proportion of users tuning voice assistants reached the highest.

6. Discussion

Our goal was to understand better what kind of chat users have with the voice assistant, how the user chats with them, and when they chat. Due to the protection of users' privacy, we can not get the users' personal information, so we cannot know the group characteristic of the users. However, we believe our research indicates that although ASR and NLP error obstacles are common, they are not the biggest threat to voice user interface design. The results showed that participants did have a preference for chats with the voice assistant. With our analysis of NetEase Sing activity data, we have a more concrete and accurate understanding of how users are chatting with their voice assistant (especially compared to self-reported usage). We found that the three main uses for the voice assistant are (1) greeting (2) function and (3) music. We also introduced some of the less frequently used chat categories.

The most talked-about by the user and the voice assistant is to say, “Hello.” The most common name in the greeting is the name of the voice assistant. The voice assistant used in the experiment requires the user to wake up first before entering the interaction process. However, if the intelligent hardware does not have an apparent sound effect or surface light after waking up, the user will mistakenly think that the voice assistant is not awake. This results in the user continuing to use the wake-up word to the voice assistant, even though the voice assistant is already activated at this moment. In the wake-up state, the NLP module will directly analyze the wake-up words, so the name of the voice assistant appears most frequently, which may be caused by improper design of the wake-up state feedback. Even if the number of voice assistant names is subtracted from the classification of greetings, its number is still close to the first, indicating that the percentage of greetings used is very high. Echoing the literature of Cheepen, one of the primary purposes of conversation is social (Cheepen, 1988). The conversation was as a way of establishing a social bond. In a human-human relationship, starting a conversation is imperative for getting to know somebody; usually, a conversation is started by saying “Hello” “Good morning” “How are you” or in Chinese way “Have you eaten” (Greeting). The second broad purpose of the conversation is transactional; people gather information or service they need by completing a conversation with the human or humanlike agent; we also call this is a task-oriented dialogue. That can explain the second and third category users chat with voice assistant are function and music. A study by Ammari et al. (2018) also found a similar result from their analysis. They identified music as the command categories most used by voice assistant users. Voice assistant provided users with the ability to play music. This music could be related to a particular genre (e.g., classical music), written by a particular artist (e.g., Jay Chou, one of the most pop singers in great China area) or a particular song (e.g., “Welcome to New York” by Taylor Swift). Playing music could also be related to users' daily routines. This finding echoes earlier results in Volokhin and Agichtein (2018), which show that contextual music recommendations depend on the activity the user is undertaking at home. For example, the music one plays when cooking might be different from what they played when they wanted to sleep, clean the house, or play with the children.

Meaningless has the highest percentage in Overall, indicating that it is more appropriate to define meaningless classifications. There are many sources of meaningless questioning. The core reason is that there is no common ground between the developer and the user behind the voice assistant, so that it will be classified into meaningless. The meaningless specific situation can be understood according to the user's original recording. At the end, when the system recognizes this sentence, the user uses the dialect, or there are several people talking at the same time, or just the developers do not understand what the user said (For example, The third kind people in Guangdong station may-

be a song or a story). However, there is controversy in the current technical ethics. Due to the pressure of public opinion, several technology giants are also suspending the monitoring of real-time voices of users.

Ask back is a supplement to the incomplete question, which is related to the current technical limitations. In the functional area, a specific function can design in a framework. Only when the user fills all the necessary slots in the frame, the voice assistant can find the corresponding information in the cloud. However, the chat is a non-structural, non-task dialogue, and there is no fixed purpose. It is the main problem in the current chat research. The proportion of nearly 9% in our research also reflects the necessity of solving this problem. At present, some solutions are modularized in deep learning and dialogue. Dialogue is no longer one question and one answer, but more questions and answers, the dialogues are written by editors are continuous. The shortcoming of this method is also very obvious. The labor cost is dozens of times, so there is still no mature in the commercial field to solve the problem.

As we mentioned above, the greeting is the embodiment of the social function of dialogue, and this part of emotion reflects the purpose of social function. In a conversation between people, people identified conversation as a fundamental tool used to transition towards friendship, allow them to know each other. Sharing personal information and emotion is seen as an essential step towards developing mutual trust and bond with others. Like emotions, personality also reflects the user's attitude towards the voice assistant. It is an attitude that they are interested in and willing to know the new “friend.” Through sharing personal information and discovering shared interests and traits, it is the critical to a relationship transition. For other researchers, they can use the proportion of emotional and personality issues in the chat as an indicator of how much the user trusts the voice assistant.

The opposite of the above two mentioned points is negative feedback, which is a direct response to the user's distrust of the voice assistant. However, the degree of distrust of the specific user can not only be seen by looking at the percentage but also needs to be combined with user interviews. After all, directly through the voice to express dissatisfaction is not the traditional way of Asians.

Ridicule is also part of the emotional role of dialogue. The purpose of ridicule is to get positive funny and exciting feedback. At this time, users are not just using voice assistants as computers; it plays the role of a social actor (Moon & Nass, 1996). Encyclopedias, stories, and jokes are the information needs of users at the information level, and they obtain information through voice. On the emotional level, the function of jokes is the same as that of ridicule. Many studies have shown that humor is an essential driver in dialogue, and people will deepen their trust and dependence on voice assistants because of interesting feedback (Tay et al., 2016; Ehrenbrink et al., 2017). The frequency of modal particles is not high. It is a kind of utterance between ridicule and meaningless.

The frequency of company and competitor is not high, accounting for 1.4%, but this part is essential. It is not a product or technology problem; it's a public relations issue. Voice assistant is an ambassador for the company. Its knowledge and evaluation of its own company and competing companies' product, service, founders, and culture will be the media hot point, which directly affects the user's evaluation of the voice assistant. Multi-intention only accounts for 0.3%, which is almost negligible.

The time users use the voice assistant is closely related to the Chinese people's work and life habits. The working period of most Chinese companies and government units is 9 AM to 6 PM and has one to two hours of mid-noon rest time. 7 PM to 0 AM is the primary period users chat with the voice assistant. In the daytime period, noon is the most frequently used time. This time is generally the time when people have lunch, so they can have time to interact with the voice assistant. From 7 PM, the interaction with the voice

assistant entered a peak period and continued until 10 PM. At 11 PM, usage has dropped rapidly, but it is still higher than the highest usage during the day. From 2 AM to 5 AM, the voice assistant is the least used. Voice assistant's developer should avoid the active communication between the voice assistant and the user at this time. In 2018, some users in the US had reported an unusual encounter with Amazon Alexa; the Echo device is emitting horrifyingly creepy laughs (Verma, 2018). They complained that it starts giggling randomly without any prompts at night; this brings a very terrible user experience.

Greetings account for the highest proportion of each period, 7 AM (wake up), noon (lunchtime), 6 PM to 7 PM (go back home) and 10 PM (sleeping preparation) is the period in which the peak of the day appears. 7 PM is the time which the user uses the music the most. The possible reason is that the user finished the day's work and returns to the home to play music. During 7 AM and 8 AM, the use of music is the highest in the whole day. The reason is that playing music after wake up can help people to wake up. In the future, this part of the conclusion needs to be combined with the user data of the music in the functional field.

At 0 o'clock and 1 o'clock in the middle of the night, the proportion of emotions and personality at the same time is higher than in other periods. We suspect that this may be because the Chinese prefer a relatively quiet, uninterrupted time to express their emotions and seek companionship from others. After 7 PM, the negative feedback from users is also increasing, which is consistent with the increase in overall usage. Because voice assistants are not mature enough, the more they are used, the easier to expose immature functions and problems, users express their negative emotions in a conversation with a voice assistant. 10 PM is the peak of the usage rate of the voice assistant for the whole day. This period is the end of the prime time TV program in China, and it is also the beginning of the sleep preparation state. So users turn their attention to the voice assistant, have more interaction with it.

7. Conclusion

As voice assistant becomes more widespread in these years and China became the biggest voice market, we need a better understanding of daily use of Chinese user with this technology. We identified different utterance categories based on previous studies, real user intent, and the commercial factors, annotated the utterances with 17 labels (categories). Drawing on 20,234 user utterances from NetEase Sing, we provide an exploratory study of what, when, and how Chinese user chat with voice assistant. We found that the three top chat usage for the voice assistant in China are (1) greeting (2) function and (3) music. Chinese user like to interact with the voice assistant mostly at night from 7 PM to 10 PM. They start to chat with voice assistant from 7 AM, and 12 AM is the usage peak of Chinese user during the day, as it is usually the lunch time for Chinese people. It is necessary for voice assistant to reduce the positive interaction activity with Chinese user at early morning, as few user will chat with voice assistant in this period. Last but not least, Chinese user is pretty polite, the whole percentage of negative feedback utterance is very low, less than 6%. The developers of voice assistant need not pay more extra attention, but they should teach the voice assistant get the ability to appease users' emotions with decent dialogue.

Though much studies remains to do in the area of interaction design in intelligent hardware, the current results can give some guidelines for voice user interface designers.

- **Build the first impression**

In the development of voice assistants, developers tend to focus on their functions and lose the overall view of voice assistants playing a role in user life. Meeting and greeting, greetings, etc are things that

people will do every day, and this is also a critical first impression. Developers, in the early days of development, will develop people's daily polite language to the voice assistant, so that users can leave the voice assistant they are using is a polite, friendly impression, more in the future. Interact with the voice assistant.

- **Show in different period**

In order to improve the activity of voice assistants, many companies will develop some features that are actively recommended. In the voice assistant, the user's usages changes significantly by time. The user communicates with the voice assistant at night and hopes to get a companion in the middle of the night. In the morning, they hope to excite themselves through music, and these discoveries, developers. In the corresponding period time, when the user wakes up, the voice assistant can actively push information and services to meet the user's expectations better.

- **Know your audience**

To know the user of a voice assistant as much as possible: What their background are? What do they think is useful ? Being aware of educational, gender, cultural differences is also important since audience's appreciation and knowledge often varies across these factors. Discover about misunderstandings can help designers get close to create a perfect voice assistant with good user experience.

REFERENCES

Akasaki, S., & Kaji, N. (2017). Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. arXiv preprint arXiv:1705.00746.
DOI: 10.18653/v1/p17-1120.

Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, Search, and IoT: How People (Really) Use Voice Assistants. ACM *Transactions on Computer-Human Interaction (TOCHI)*, *26*(3), 1-28.
DOI: 10.1145/3311956.

Austin, J. L. (1962). How to do things with words Oxford University Press.
DOI: 10.1093/acprof.oso/9780198245537.001.0001.

Akasaki, S., & Kaji, N. (2017). Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. arXiv preprint arXiv:1705.00746.
DOI: 10.18653/v1/p17-1120.

Canalys. (2019). China overtakes US in fast growing smart speaker market. https://www.canalys.com/newsroom/china-overtakes-us-in-fast-growing-smart-speaker-market.
DOI: 10.31857/s0869-5873897745-754-12467.

Cheepen, C. (1988). The predictability of informal conversation. Pinter Pub Ltd. DOI: 10.2307/414637.

Cho, Yu Suk., Eom, Kimin., & Joo, Hyo Min. (2009). The effect of the human voice that is consistent with context and the mechanical meloy on user's subjective experience in mobile phones. In 2009 Korean Society for Emontion and Sensibility (pp.531-544). DOI: 10.31274.

Duquette, A., Michaud, F., & Mercier, H. (2008). Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonomous Robots*, *24*(2), 147-157.
DOI: 10.1007/s10514-007-9056-5.Ehrenbrink

P., Osman, S., & Möller, S. (2017, November). Google now is for the extraverted, Cortana for the introverted: investigating the influence of personality on IPA preference. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction* (pp. 257-265). DOI: 10.1145/3152771.3152799.

Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. *Medical Reference Services Quarterly*, *37*(1), 81-88.
DOI: 10.1080/02763869.

Nass, C. I., & Brave, S. (2005). Wired for speech: How voice activates and advances the human-computer relationship (p. 9). Cambridge, MA: MIT press.
DOI: 10.1162/coli.2006.32.3.451.

Moldovan, C., Rus, V., & Graesser, A. C. (2011). Automated Speech Act Classification For Online Chat.

MAICS, *710*, 23-29.
DOI: 10.1007/978-3-642-67758-8_3.

Moon, Y., & Nass, C. (1996). How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication Research*, *23*(6), 651-674.
DOI: 10.1177/009365096023006002.

Price, R. (2016). Microsoft is deleting its AI chatbot's incredibly racist tweets. Business Insider.

Sato-Shimokawara, E., Shinoda, Y., Takatani, T., Lee, H., Wada, K., & Yamaguchi, T. (2016, August). Analysis of category estimation for cloud based chat robot. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 308-311). IEEE.
DOI: 10.1109/roman.2016.7745147.

Tay, B. T., Low, S. C., Ko, K. H., & Park, T. (2016). Types of humor that robots can play. *Computers in Human Behavior*, *60*, 19-28.
DOI: 10.1016/j.cnb.2011.08.011.

Verma Shubham (2018). Why so serious? Amazon's Alexa is 'laughing' at night and scaring users; here's what's happening. https://www.financialexpress.com/industry/technology/why-so-serious- amazons-alexa-is-laughing-at-night-and-scaring-users-heres-whats-happening/1091784/

Volokhin, S., & Agichtein, E. (2018, March). Understanding music listening intents during daily activities with implications for contextual music recommendation. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 313-316).
DOI: 10.1145/3176349.3176885.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. DOI: 10.1145/365153.365168.

Yoo, Cho-Rong., Kim Song-Hyun., & Kim, Jin-Woo. (2020). A Comparative Study of the Use of Intelligent Personal Assistant Services Experiences: Siri, Google Assistant, Bixby. *Science of Emotion and Sensibility*, *23*(1) 69-78.
DOI: 10.14695/KJSOS.2020.23.1.69

원고접수: 2020.03.18
수정접수: 1차 2020.04.14
2차 2020.04.29
게재확정: 2020.04.29