VICA, a Visual Counseling Agent for Emotional Distress

Yoshitaka Sakurai
School of Interdisciplinary Mathematical Sciences
Meiji University
Nakano, Japan
sakuraiy@meiji.ac.jp

Yukino Ikegami
IO Inc.
Tokyo, Japan
yknikgm@gmail.com

Motoki Sakai, Hiroshi Fujikawa, Setsuo Tsuruta
School of Information Environment
Tokyo Denki University
Inzai, Japan
{m-sakai,tsuruta}@mail.dendai.ac.jp
fujikawa@maruka.com

Avelino J. Gonzalez
School of Engineering and Computer Science
University of Central Florida
Orlando, Florida, USA
avelino.gonzalez@ucf.edu

Eriko Sakurai
Faculty of Service Management
Bunri University of Hospitality
Sayama, Japan
sakurai@bunri-c.ac.jp

Ernesto Damiani
Center for Cyber-Physical Systems
Khalifa University, UAE
Computer Science Department
Università degli Studi di Milano, Italy
ernesto.damiani@ku.ac.ae
ORCID: 0000-0002-9557-6496

Andrea Kutics
Dept. of Natural Sciences
International Christian University
Tokyo, Japan
matz@icu.ac.jp

Rainer Knauf
Dept. of Computer Science and Automation
Technische Universität Ilmenau
Ilmenau, Germany
rainer.knauf@tu-ilmenau.de

Fulvio Frati
Computer Science Department
Università degli Studi di Milano, Italy
fulvio.frati@unimi.it
ORCID: 0000-0001-8856-5150

***Abstract*— We present VICA, a Visual Counseling Agent designed to create an engaging multimedia face-to-face interaction. VICA is a human-friendly agent equipped with high-performance voice conversation designed to help psychologically stressed users, to offload their emotional burden. Such users specifically include non-computer-savvy elderly persons or clients. Our agent builds replies exploiting interlocutor's utterances expressing such as wishes, obstacles, emotions, etc. Statements asking for confirmation, details, emotional summary, or relations among such expressions are added to the utterances. We claim that VICA is suitable for positive counseling scenarios where multimedia specifically high-performance voice communication is instrumental for even the old or digital divided users to continue dialogue towards their self-awareness. To prove this claim, VICA's effect is evaluated with respect to a previous text-based counseling agent CRECA and ELIZA including its successors. An experiment involving 14 subjects shows VICA effects as follows: *i)* the dialogue continuation (CPS: Conversation-turns Per Session) of VICA for the older half (age >40) substantially improved 53% to CRECA and 71% to ELIZA. *ii)* VICA's capability to foster peace of mind and other positive feelings was assessed with a very high score of 5 or 6 mostly, out of 7 stages of the Likert scale, again by the older. Compared on average, such capability of VICA for the older is 5.14 while CRECA (all subjects are young students, age<25) is 4.50, ELIZA is 3.50, and the best of ELIZA's successors for the older (>25) is 4.41.**

***Keywords*— conversational agent, visual counseling agent, avatar, voice conversation, dialogue continuation, self-awareness**

***Corresponding author*:** Fulvio Frati
Computer Science Department – Università degli Studi di Milano
Via Bramante, 65 – 26013 Crema CR – Italy
Email: fulvio.frati@unimi.it
Tel. +39-0373-898032

I. INTRODUCTION

Worldwide, more people than ever before feel stressed, depressed and anxiety-ridden. According to recent studies (Weissman et al. 2017), an estimated 8.3 million American adults — about 3.4 percent of the U.S. population — suffer from serious psychological distress. In the UK a survey commissioned in 2018 by the Mental Health Foundation and undertaken by YouGov (Mental Health Foundation 2018) showed that almost three quarters (74%) of people feel overwhelmed or unable to cope with work-related issues. In Japan, already a decade ago governmental reports (Ministry of Health, Labor and Welfare 2008) showed that the majority (60%) of IT workers suffered from anxiety problems, psychological discomfort, and lack of effectiveness. Today, middle-aged people are often affected by job uncertainty or other forms of professional discomfort, while retired and/or elderly persons lament anxiety due to loneliness and physical decline. Young generations, and particularly millennials, are also suffering from stress related to unemployment or to low-income precarious jobs. Stressed, unhappy persons are increasing rapidly in number, becoming a significant social problem.

Unfortunately, the resource pool of professional counselors is very limited, and many people under psychological stress are not able to get professional help. As a result, some of them live with serious psychological distress, an umbrella term that runs from general hopelessness and nervousness all the way up to diagnosable conditions such as depression and anxiety (Weissman et al. 2017).

This situation has re-kindled interest in software counseling agents. Modern research on conversational software agents started as early as 1965 with ELIZA (Weizenbaum 1976) and its variant called DOCTOR (Weizenbaum 1966). DOCTOR mimicked a typical psychoanalyst's behavior, replying to clients with keywords taken from their input (a technique called *mirroring*). Many users felt as if DOCTOR was really listening to them. Despite considerable progress in this field and many implementations of conversational agents, most researchers concede that AI-based conversational agents cannot yet substitute human counselors in helping to find solutions to their clients'[1] real problems (Floridi et al. 2009).

However, a list of practical suggestions is not what many distressed people are looking for (Asai and Lambert 1966). Indeed, simply offloading their emotional burden to a trusted party has a *healing effect* that can improve peace of mind and decrease stress (Rogers 1957). Without mutual trust (Asai and Lambert 1966), humans do neither show their true feelings nor recount their sufferings. Mutual trust is not acquired once and for all; rather, counselors must continuously preserve it via trustworthy attitude. Alternating attentive listening to clients with such as unconditional positive regards (Rogers 1957) and context respectful prompts to narrow or clarify context (Powell 1966) are provided for clients to describe or be aware of their problems and solutions. Moreover, properly summarizing clients' emotional changes in the whole counseling dialogue is effective for promoting the client's clarification of the problem (Ivey et al. 2006).

As Artificial Intelligence (AI) products have gained acceptance, it is likely that some clients will forgo seeing human counselors altogether in favor of virtual ones. Others will interact with online "listeners" in addition to their human counselors. Also, counseling agents will find their way in the interfaces of complex cyber-physical systems (CPS) or IOT. *SimSensei Kiosk* (DeVault et al. 2014) is virtual human interviewer based on gestural (rather than verbal) responses for trust build-up. The AI agent *Ellie* (Rizzo and Talbot 2016) was designed to help treat patients with depression and post-traumatic stress disorder. Ellie's body language is designed to mirror that of an actual therapist. *Tess*, developed by X2AI Inc., is an interesting commercial verbal psychological AI project that administers highly personalized psychotherapy etc. instead of the mental health professional. To the best of our knowledge, no public disclosure has been made of the algorithmic engine behind Tess. Several projects have tried to create empathy between humans and software agents by using visual information rather than verbal interaction. Microsoft's Xiao Ice (Shum et al 2018) is a social chatbot that tries to establish an emotional connection showing empathy to its users. However, even such social chatbot has not emotional connection so much as our counseling agent at least in narrowing context towards clients' awareness of their problems or solutions.

The notion of empathy for trust building is also at the basis of our AI agent called Context-Respectful Counseling Agent: CRECA (Shinozaki et al. 2013), which was designed collaboratively by computer scientists and Japanese professional counsellors. CRECA replies with ELIZA-like paraphrases for empathy, followed by context respectful prompts to narrow the conversation's context towards clarifying/solving problems. CRECA's reasoning engine is based on a Context-Based Reasoning (CBR) mechanism (Gonzalez et al. 2010) supporting dynamic change of situations / goals. CRECA "listens" to clients and promotes their reflection and self-awareness by context preserving responses.

However, our original CRECA agent heavily relied on keyword typing. It had neither bi-directional voice/speech communication nor a cartoon-style avatar appearance. This created a "technology-oriented" context where clients who were elderly or affected by digitally divide found it difficult to spell out their problems.

To address emotionally distressed, non-computer-savvy and/or elderly users, we designed VICA, a novel multimedia (high-performance voice communication) counselling agent, suitable for cloud hosting. VICA's internal voice engine can provide

[1] Following customary practice in the counseling profession, in this paper we will use the term "client" rather than "patient" to designate people who seek professional counseling.

responses to emotionally loaded words in client utterances as well as comment on emerging relationship among clients' wishes (desires/ goals/ purposes), obstacles, and emotions. VICA is equipped with a two-dimensional image avatar and has a voice interface that uses Google's cloud speech API. However, the Google Speech API is vulnerable to various kinds of language traps that are typical of Japanese utterances including English words (Ashwell and Elam 2017). Therefore, VICA is equipped with its own application-level error recovery strategy to achieve high-quality interactions.

An early version of VICA equipped with an avatar and voice output was introduced in Yamamoto (2016). However, the input was still text. In this paper, VICA describes in detail the architecture, implementation and validation of the multimedia system. Experience has shown that, while younger clients can adapt to standard text-to-speech conversion hiccups, a seamless voice conversation is crucial for elderly or digitally divided clients to reach awareness of problems and solutions by VICA. Our experimentation shows that VICA's capability to foster peace of mind and other positive feelings is assessed as considerably high (over 5 out of 7 stages of the Likert scale) by the older 7 subjects (over forty years old) when compared with the assessment around 3 or 4 by the younger 7 subjects or with the assessment 4.5 by CRECA where all subjects are students less than twenty-five years old). Thus, effect of multimedia VICA equipped with high-performance voice communication is shown instrumental especially for the elderly clients.

The rest of the paper is organized as follows. Section 2 reviews some related work. In Section 3, we recap the design and implementation overview of our baseline CRECA agent called original CRECA, while Section 4 presents VICA's implementation in detail. Section 4 focuses on the human-like avatar and voice conversation facility, as well as details of the core using parts of CRECA especially the computation of responses exploiting utterances based on the relationship among the client's current and past ones. The experiment and its results are described in Section 5, while Section 6 draws our conclusions and discusses future work.

II. Related Work

We briefly review three major lines of related work: positive counseling, conversational agents and text summarization.

Within research on *counseling*, a significant notion is the one of *client-centered therapy* introduced by Rogers (Rogers 1957). It assumes that a therapist who provides unconditional positive regard will stimulate psychotherapeutic personality change in a vulnerable client, if the client perceives these attitudes (Rogers 1957). To help a client in finding solutions to his problems, human counselors need to firstly build up mutual trust with the client (Asai and Lambert 1966), otherwise clients will not express their true feelings. For instance, trust build-up is essential also in career counseling (Savickas 2011), where problems are often unclear and the client's motivation for solving them is low.

Our experience with CRECA (Shinozaki et al. 2013) has shown that trust build-up is not impaired by clients knowing that their interlocutors are software agents and therefore not "real persons". Indeed, humans feel less judged talking to software agents which provide unconditionally positive feedback and adopt a "I'm just here to listen" attitude (Shinozaki et al. 2013).

Regarding *conversational software agents*, work started more than fifty years ago with J. Weizenbaum's pioneering dialog agent ELIZA (Weizenbaum 1966) and its variation DOCTOR that mimicked interactions taking place in psychotherapy. ELIZA mirrors phrases provided by the client but does not maintain the context, so the average length of ELIZA interchanges, a.k.a. *continuation or CPS (Conversation-turn Per Session)*, was less than a half of the one achieved by today's conversational agents (Ashwell and Elam 2017) as well as our own CRECA (Shinozaki et al. 2013). The CRECA agent relies on ELIZA-like mirroring for empathy towards trust buildup and elaborately introduces context-respectfulness such as rather context-independent digging prompts, while using context-based reasoning (Hung et al. 2009) according to the dynamic change of situations and goals (Gonzalez et al. 2010).

Since ELIZA times, research has progressed significantly. In the early 1980's, ALICE (Wallace, 2003) was created, together with the Artificial Intelligence Markup Language (AIML). AIML has been used for more than two decades to represent the pattern-matching rules that link user-submitted words and phrases with topic categories.

Developers of modern conversation systems still face the same problem that Weizenbaum faced fifty years ago: they must initiate and maintain dialogue with humans based on past interactions with them. Many current dialogue systems use the same, psychotherapist questioning technique, and objectively measuring their performance is still an open issue. Ashwell and Elam (Ashwell and Elam 2017) proposed the experiment design we rely on in this paper. They compared five recent artificial dialogue systems with an online version of ELIZA used as a benchmark. In their study, more than 100 (one hundred) male and female participants interacted with the conversational agents over the Internet scoring each for conversation ability. Statistical significance of difference in length of interactions involving the same users shows that modern dialogue systems are a substantial, though not dramatic, improvement on their predecessor. At the end of the day, all conversational systems adapt their responses based on the way humans converse.

Recent work in automating conversational counseling has given rise to so-called 5W1H (who, what, when, where, why, how) approach (Han et al. 2013). The 5W1H information and four basic emotions (corresponding to happy, afraid, sad, and angry states) are extracted from user utterances. The counselor's communication is then tuned based on the current emotional state of the client. However, clients' problems may be too pressing for them to accept to undergo emotional profiling. Sometimes, clients feel the emotionally profiling 5W1H questions not to be worthy of an answer (Han et al. 2013).

Several projects have tried to create empathy between humans and software agents by using visual information rather than verbal interaction. The AI agent *Ellie* was designed by the Institute for Creative Technologies at the University of Southern California (Rizzo and Talbot 2016) to help treat patients with depression and post-traumatic stress disorder. Ellie's body language is designed to mirror that of an actual therapist. She responds to emotional cues, nods affirmatively when appropriate and adjusts in her seat. The underlying algorithm perceives 66 points on a person's face, uses a classic algorithm to map their positions to the client's emotional state (Anisetti et al. 2007) and moves accordingly.

An interesting commercial project in verbal AI counselling is *Tess*, developed by X2AI Inc., and described in the company's literature[2] as a "*psychological AI that administers highly personalized psychotherapy, psycho-education and health-related reminders, on-demand, when and where the mental health professional is not*." To the best of our knowledge, no public disclosure has been made of the algorithmic engine behind Tess.

SimSensei Kiosk (DeVault et al. 2014) is virtual human interviewer designed to create an engaging face-to-face interaction where the user feels comfortable talking and sharing information. SimSensei is designed to create interactional situations favorable to the automatic assessment of distress indicators, defined as verbal and nonverbal behaviors correlated with depression, anxiety or post-traumatic stress disorder (PTSD). A major difference between SimSensei and our VICA is that the former is based on gestural rather than verbal responses for trust build-up. For this reason, SimSensei has been validated comparing it to a "Wizard-of-Oz" baseline where two human operators and SimSensei competitively provided gestural responses.

Microsoft's XiaoIce (Shum et al 2018) is a social chatbot that tries to establish an emotional connection showing empathy to its users. The primary goal of XiaoIce is not necessarily to solve all the questions the users might have, but rather, to be a virtual companion to them. Its design takes not just intellectual quotient (IQ) but also emotional quotient (EQ) into account. To communicate effectively, XiaoIce interacts with users through multiple modalities, including text, speech, and vision. It uses deep learning technologies such as recurrent neural networks (RNN) for response generation and convolutional neural networks (CNN) for visual feature vector extraction. However, even such a state-of art social chatbot has rather just simple emotional connection with users. It is not so much as our counseling agent at least in CRECA's context respectful responses which includes narrowing contexts towards clients' awareness of their problems or solutions.

Assessment of conversational software agents by comparing them to humans in a variation of the classic Turing Test has been attempted many times in the framework of the Loebner Contest (Floridi et al 2009) and we believe that it falls well outside the scope of this paper. To prove our claim we have chosen to compare VICA mainly to its CRECA predecessor rather than to a human operator in disguise and supplementally use the conversational agent review (Shah et al. 2016) .

Most *summarization* methods (Gupta and Lehal 2010) focus on superficial text features such as word frequency and position of words, aiming only at reducing the reading or searching time. Meanwhile, in (Kennedy et al. 2012), a summarization method focusing just on emotions is proposed. In summarization for our context respectful client centered counseling, considering client's emotions and putting client's episodes in the chronological order is required. Furthermore, providing the overview of the past events leading to certain emotion is effective. However, those conventional works do not consider both emotion and time series at once. Many approaches to text summarization have been applied by intelligent tutoring systems for learning support (Guangbing et al. 2013) In our CA function, those solutions will be discussed when emotions or changes of emotions are summarized not in one sentence or one interaction, but in several interactions or sessions of a counseling dialogue. Moreover, our summarization is presented by voice output, which is even able to learn a real counsellor's natural voice tone.

III. TEXT-BASED SYSTEM

III.A. A text-based Counseling agent: CRECA

We use CRECA (Shinozaki et al. 2013) as a baseline or core of VICA. It exploits not only ELIZA-style mirroring of the current utterance but also generates its own responses for digging or narrowing the dialogue (counseling) context. Furthermore, it attempts summarization of the clients' previous utterances considering their emotional impact and its chronological order (emotional change history). Asking about the quality or effect of mirroring and/or summarization by such as confirmation including tag-question (e.g. "aren't you"), the agent can check how accurate or empathetic the listening (with unconditional positive regards) has been. This causes clients' trust in CRECA, keeping the dialogue open or continuing, which helps clients in seeing their own personal distortions or reviewing their own thinking as self-reflection or self-searching. Finally, these responses especially "digging" the context prevent the dialogue drifting away from the problem even when the client's emotional state or context has changed. Sentence patterns used here include problem-digging phrases (e.g. "please tell me more" or "give me more details"). Such digging

[2] https://x2.ai/info

responses give clients more chances of self-reflection or self-searching towards self-awareness of their problems and solutions. While original CRECA had a problem discovery phase asking clients vocation, affiliation, experience, and the problem field or subfields especially at the first counselling, this is not used as the core CRECA in VICA for generality and simplicity.

III.B. Implementation of CRECA

The architecture of the baseline system (original CRECA) is shown in Fig. 1. Terms and their structure are extracted by standard Natural Language Processing techniques with a dictionary for emotional words and then saved in context objects. The agent creates its dialogue responses using information stored in current context objects and in its counseling knowledge database (DB) including regular matching rules and procedural knowledge.

CRECA's *natural language dialogue processing module* consists of (1) an initialization and termination module that functions as an interface between humans and the agent, (2) a dialogue text analysis module; Analyzer in Fig. 1 and (3) a dialogue text input /output module. When the CA is launched, the initialization / termination module initializes the context objects and generates the dialogue-starting messages. At the end of the interaction, it saves all relevant information (including the conversation log) and generates the dialogue-end messages.

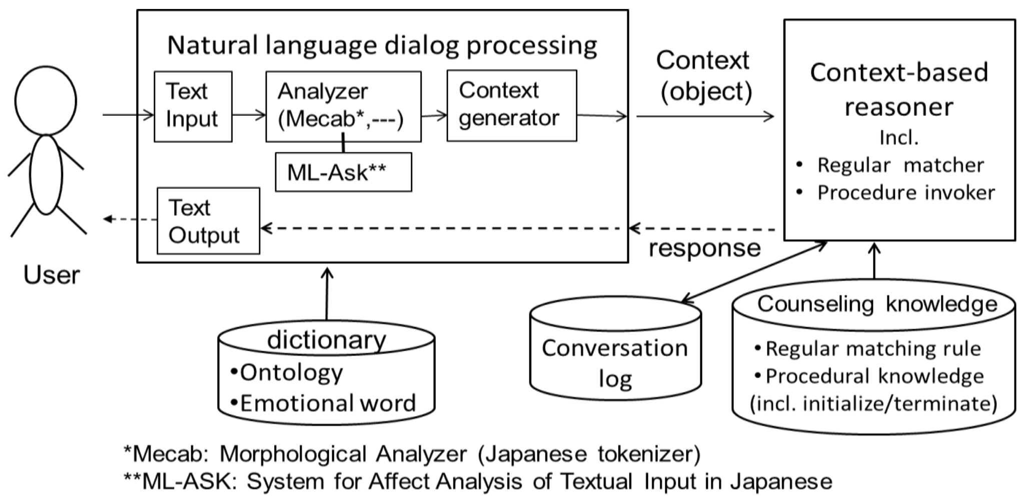

Fig. 1. Architecture of baseline CRECA (original CRECA)

The *dialogue text analysis module* tokenizes the dialogue texts using Mecab in fig. 1. Then, a context analysis is done. At each step of the dialogue, a new context object is generated according to the user input, the current state of the old context object, and the ontology dictionary. The ontology dictionary is used in the problem discovery phase of original CRECA, which VICA has not. CRECA also uses its internal counseling knowledge base to create responses according to the current context. The counseling knowledge contains regular matching rules for generating a digging prompt to narrow down the dialogue context towards the client's problems and enable responses that combine the identified context (object) with the prompt. The implementation detail is described in section IV (IV.A and IV.B using the work-out example).

III.C. Limitations

CRECA uses a few hundreds of rules or procedures (Shinozaki et al. 2013) for introducing experiences and vocations/ affiliations each other as well as asking clients their suffering problem fields/ subfields. This phase is called a problem discovery phase and discriminated from the main counseling phase called a problem solving phase. CRECA saves them as context to be used for response selection. However, there are too many vocations/ affiliations as well as too many such (sub) fields to manage them as context by rule or procedural knowledge including ontological one. As CRECA is extended, Therefore the problem discovery phase (an ontology dictionary also) has been removed.

Furthermore, a major limitation of this style of dialogue is the conversation to become boring, especially if the agent reacts always in the same way to the same input. CRECA uses fuzzy reasoning (as opposed to random selection) to avoid boring the client, alternating several digging prompts in case of no match of input dialogue with any context object (Shinozaki et al. 2013).

IV. VICA and its Implementation

The original CRECA agent was evaluated effective (Shinozaki et al. 2013). However, it has only text interface. It has neither bi-directional voice/speech communication medium nor even a cartoon-style avatar appearance. Its chat-style interface is not suitable for non-computer-savvy or elderly clients to spell out their problems. We newly designed VICA towards human-friendliness. We equipped it with a 2D avatar having voice input/output. High quality voice input followed by speech recognition is attained by using Google voice recognition tools such as Google cloud speech API as well as by our own software-level error correction tools.

IV.A. VICA's Processing flow

VICA's processing flow is as follows:

1. Input of the clients' voice data.
2. Activate voice recognition using Google speech recognizer (Cloud Speech API in Fig.2) to produce texts out of user utterances (client's voice data) and receive the recognized texts to hand them to CRECA by voice recognition handler as shown in Fig. 5.
3. Automatic correction of speech recognition results works, using our interactive software tool as follows:
    a. Activate Google speech recognizer again in case an error code indicating "insufficient precision" is returned.
    b. Execute automatic correction of the speech recognition results, using word/phrase recognition error correction dictionary dynamically improved by our interactive software tool.
        i. Perform the dictionary based automatic correction of speech recognition error using correction rules set (dictionary) stored as [recognition error (word or phrases), expected recognition (word or phrases)] pairs.
        ii. A [recognition error (word or phrase), expected recognition (word or phrase)] pair called "correction pair" is a tuple of frequently mistaken word/phrases and their corrections. For example, correction pairs = {("清掃部門 (cleaning department)", "製造部門 (production department)"), ..., ("博士 GO (doctor goes)", "博士号 (doctor's degree)) , ("やるじゃなくて (not doing)", "嫌じゃなくて (not dislike)). These correction pairs are manually entered and dynamically added to the dictionary by our interactive software tool for automatic correction on Google Chrome JavaScript.
        iii. For each of all correction pairs in the above-mentioned dictionary, the recognition error in recognized string is replaced with expected recognition.
4. Display the voice recognition result on a stripe window having deletion button. Both are located above the counselor avatar. If users think it is different from their utterances (specifically because they spoke in a noisy environment or they could not speak smoothly), they can pick the deletion button and input their utterance again.
5. Generation and logging of the reply response by CRECA's context-based reasoning engine that invokes regular matching rules and accompanied procedures generating such as tag questions, based on context, state or situation:
    a. the input text which was handed is taken in.
    b. Morphological analysis of the grammatical sentence structure.
    c. Generation of reply response and logs using regular matching rules and accompanied procedures for paraphrasing (incl. tag questions, personal pronoun/ verb conversion) and digging prompts.
6. Generation of response and its logging by CRECA's extended engine that invokes procedural knowledge based on context (state/ situation) :
    a. Generate a reply confirming emotional changes, if recognized by comparing current emotional words with previous ones in the conversation. Here, emotional words can be recognized using emotional word dictionary and emotional change can be detected using ML-ASK: System for affect analysis of textual input in Japanese (Ptaszynski et al., 2009).
    b. If improvement or change of feeling (emotion) is detected, add a corresponding remark or confirmation to the reply and log.
    c. If a desire (wish) is expressed, it is added along with the related obstacle and the resultant feeling/ emotion is added to the reply as well as to the log with the tag [wish], [obstacle], and [resultant emotion]. Desires (wishes), obstacles, and resultant emotion can be detected using regular matching rules as the example shown later in this section.
    d. In case no desire (wish) was expressed in the previous few dialogue cycles (two through five, decided by fuzzy reasoning), changes among emotional words are confirmed or their related factual or emotional words/ phrases/ sentences in the client's sayings (frequently occurring in dialogue session stored in the log, some with the tag such as [obstacle] or [resultant emotion]) are asked about the relation with the desire (a word or phrase tagged by [wish] and stored in the log) as the reply.
    e. If more than one of the above rules are applicable, detect the most appropriate one by fuzzy reasoning.
    f. At the beginning of the reply, insert output interjections such as "Hmm", calculating their repetition number and length.

g. Generate the reading of Japanese "kanji" for the derived output based on Kanji reading (Ruby) dictionaries.

7. Stop the conversation, if there is a final positive remark of the client such as "I feel refreshed" or "It was good" or there is no further clients input for the specific duration.

8. Analyze the current and former dialogs of this client (their length and their degree of success) for output generation accordingly.

IV.B. worked-out example

A worked-out example of the rule-based analysis of input sentences by pattern matching and of the creation of a response/log for paraphrasing, digging, and summarization is given below[3]:

1. Dialogue example (until a client is aware of themselves)

Client's input*: For me, software developer is a wish in life, but lack of programming skills is an obstacle. Job hunting makes me feel worried.*

VICA's reply: *Hum, Hum, for you, software development is a wish in life but lack of programming skills is an obstacle. Job hunting makes you feel worried, doesn't it? Please speak a little more in detail*."

Client's input*: I am always the last student that finishes given program problems in my programming class.*

VICA's reply: *Hum, Hum, you are always the last student that finishes given program problems in my programming class. Please say more about it.*

Client's input*: I have no skill of rapid programming.*

VICA's reply: *Hum, Hum, you have no skill of rapid programming. Please tell me the relation among no skill of rapid programming and your wish of software development.*

Client's input*: Oh, I found I can develop good quality programs and I would like to do research development of software.*

VICA's reply: *Hum, Hum, you found you can develop good quality programs and you would like to do research development of software, wouldn't you. Please tell me more.*

Client's input*: Now I am fine.*

2. Example of knowledge (mainly regular matching rules) and response generation

a. An example of context of starting input and its reply

Client's input*: For me, software developer is a wish in life, but lack of programming skills is an obstacle. Job hunting makes me feel worried.*

VICA's reply: *Hum, Hum, for you, software developer is a wish in life but lack of programming skills is an obstacle. Job hunting makes you feel worried, doesn't it? Please speak a little more in detail*.

b. An example of rules for the response generation in this starting context is as follows:

(1) Condition (Stimulus) part
[ur "For me, ((?:SE)| (?:PM)| (?: programmer)| (?: software developer) @1) is a wish (.*@2) but (.*@3) is an obstacle. ((?: job hunting)| (?: entrance)| (?: graduation)| (?: pass), @4) makes me feel (.*@5) ((?: worried)| (?: worry)| (?: Maybe, anxious) @ 6) (.*@7)"]

(2) Consequence part 1 (Response for reply) [u "Hum, Hum, %1 is a wish %2 but %3 is an obstacle. %4 makes you feel %5%6%7, doesn't it? Please speak a little more in detail."]

(3) Consequence part 2 (Response for logging)
[u" Hum, Hum, %1[wish] %2, but %3 [obstacle] is an obstacle. %4 makes you feel %5%6 [resultant emotion] %7, doesn't it? Please speak a little more in detail."]

Where: *u* indicates any Unicode sequence. *r* indicates a regular expression. Neither "*n*" nor "*m*" in such as "*@n*", "*?:@m*", %n, and %m does not exist in real (coding of) rules or knowledge representation but is just for explanation; "*@n*" in condition parts

[3] Here and in the remainder of the paper, English texts exchanged between VICA and human clients are manually back-translated from the original Japanese for English readers.

indicates matching and unification with *n* of *%n* in consequence parts of rules. "*.*@n*" indicates *n* can be unified with any words or phrases. "*?:@m*" indicates *m* is one word or phrase matching with that (context object) after "?:" which can unify with *m* of *%m* in consequence parts of rules. "|" indicates "or".

After inputting Client's utterance texts mentioned in the above example and analyzing them morphologically, generation of response by reasoning (regular matching) engine of CRECA is done using the above example rule as follows:

(1) In the condition part, after matching with "For me", "software developer" is unified with 1, and after matching with "is a wish", "in life" is unified with 2. After matching with "but", "*lack of programming skills"* is unified with 3. After matching with "is an obstacle", "*Job hunting"* is unified with 4, and after matching with "makes me feel", "worried" is unified with 6. Nothing is unified with 5 and 7.

(2) (Reply response generation): In consequence part 1, *1* of *%1* is "software developer". *2* of *%2* is "in life". *3* of *%3* is "*lack of programming skills*". *4* of *%4* is "job hunting". *6* of *%6* is "*worried*". Therefore, using the basic paraphrasing rule to substitute I, my, and me to you, your, and you respectively, the paraphrase response after "Hum, Hum," is generated, except the reason indicated by (…), as follows: For you, *software developer* (to which 1 is substituted) is a wish *in life* (to which 2 is substituted) but *lack of programming skills* (to which 3 is substituted) is the obstacle. *Job hunting* (to which 4 is substituted) makes you feel *worried* (to which 6 is substituted), doesn't it". Furthermore, using the last sentence in consequence part 1, "Please speak a little more in detail." is added as a prompt response for digging.
->
VICA's reply: *Hum, Hum, for you, software development is a wish in life but lack of programming skills is an obstacle. Job hunting makes you feel worried, doesn't it? Please speak a little more in detail.*

(3) (Log generation): Using consequence part 2, the following is logged ([…] such as [wish/goal], [obstacle], and [resultant emotion] is a tag representing each of these properties) since the matching and unification is the same in (2). "Hum, Hum, for you, *software developer* [wish] is a wish *in life*, but *lack of programming skills* [obstacle] is the obstacle. *Job hunting* makes you feel *worried* [resultant emotion], doesn't it? Please speak a little more in detail."

c. The knowledge for generating the remaining dialogue responses and the generated responses (omitting the log) are as follows:

Client's input: I am always the last student that finishes given program problems in my programming class.
Condition part of rule: [ur"I am (.*@1)"]
Consequence part of rule: [u"Hum, Hum, you %1", aren't you?]
->
VICA's reply: Hum, Hum, you are always the last student that finishes given program problems in your programming class, aren't you? Please say more about it.

Client's input: I have no skill of rapid programming.
Condition part: regular matching rule: [ur"I (.*@1)"] and procedural knowledge: if client said client's wish and current client utterance contains a word ("programming", in this case) frequently occurring in dialogue session (stored in the log, some with the tag such as [obstacle] or [resultant emotion])
Consequence part: [u"Hum, Hum, you have %1. Please tell me the relation among %1 and %2"]
NOTE: %2 is unified with client's wish namely *software developer* tagged with "[wish]" logged in (3) above.
->
VICA's reply: Hum, Hum, you have no skill of rapid programming. Please tell me the relation among no skill of rapid programming and your wish of software development.

Client's input: Oh, I found I can develop good quality programs and I would like to do research development of software.
Condition part of rule: [ur"I (.*@1)"]
Consequence part of rule: [u"Hum, Hum, you %1. Please tell me more."]
->
VICA's reply: Hum, Hum, you found you can develop good quality programs and you would like to do research development of software. Please tell me more.

Client's input: Now I am fine.

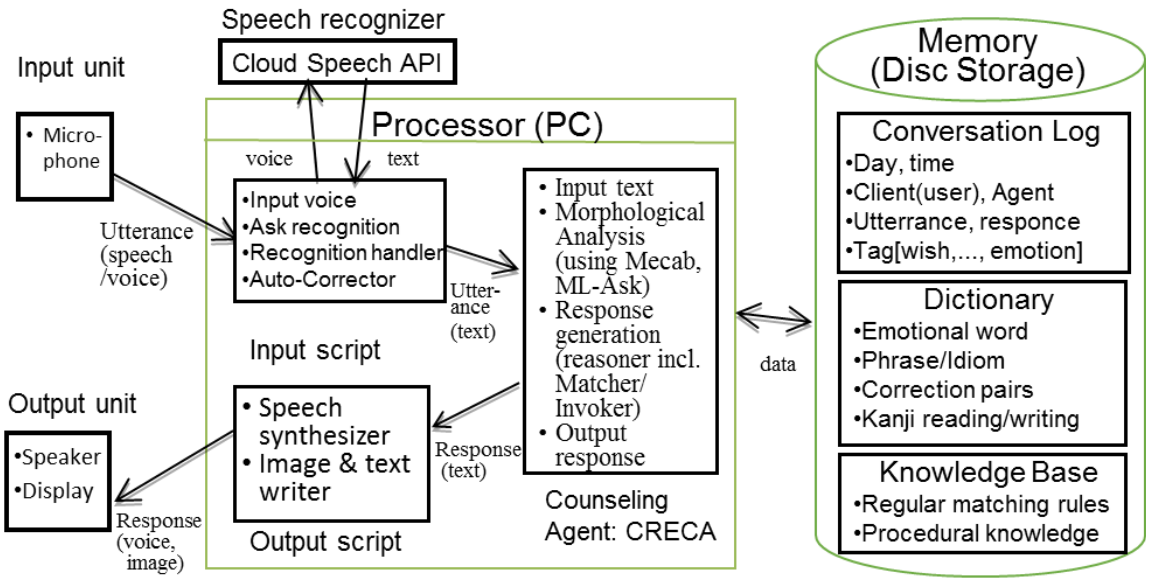

Fig. 2. Module structure for VICA's processing

IV.C. Module structure

Fig. 2 shows the module structure of VICA. Speech comes from the input unit/device to Google speech recognizer (Cloud Speech API). After recognition, the resultant text is transferred to CRECA for the reply/response generation, which has access to a memory storage that contains the dictionary and stores the conversation log/history. The generated reply is transferred to the sound generation equipment, which produces the reply sound/voice, which is transferred to the output unit.

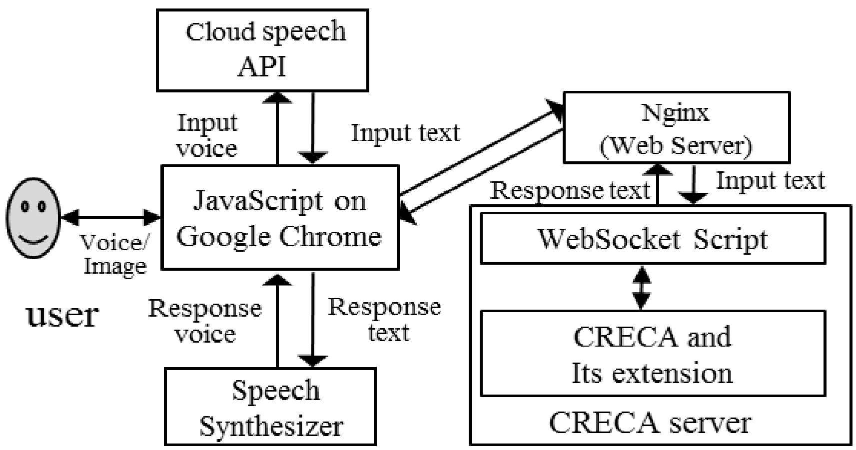

Fig. 3. Overall Structure of VICA using JavaScript on Google Chrome

Fig. 3 shows Google Chrome and its linked components used by VICA, for the multimedia (voice, image, text) interface. Clients carry out their speech (voice) conversation through JavaScript program on Google Chrome (client) connected to Google cloud speech API (server), Nginx (Web server) connecting CRECA (server) and Google speech synthesizer (server).

Speech (Voice) recognition, response generation, voice synthesis/output and 2D avatar display as shown in Fig. 4 are done using cloud speech API (cloud), CRECA (local) and speech synthesizer (local) through JavaScript on the client.

Fig. 4. Avatar of VICA in Google Chrome

The CA on the server is connected via a WebSocket open between JavaScript (client) and Python (server). Currently, Nginx is used as Web server, but Apache is also available. For speech recognition, we implemented a related HTML page, including JavaScript code for voice recognition handling shown in Fig. 5.

```
recognition.onresult = function(event){ //on receiving dialogue texts from voice recognizer
  var length = event.results.length; // length of received texts,received word (morpheme) length, or the number of received words
  var final_transcript = ""; //final (total) result buffer of recognition
    if (length > 0) { if there are received texts from Google speech recognizer, they are set in transcript buffer
            for (var i = 0; i < event.results.length; i++){ //increment index (initially 0 in this version) until received word length
            var recogStr = event.results[i][0].transcript;  //obtain a recognized text word (morphemes)
                final_transcript += recogStr; //concatenate the recognized text word to final_transcript (buffer)
                game.rootScene.removeChild(fukidashi); //clear (recognized) text display window
                fukidashi = create_fukidashi(final_transcript, 100, 0, 'white'); //create text display window
                game.rootScene.addChild(fukidashi); //show the intermediate text in text display window
                console.log(event.results[i][0].confidence); //log and display the degree of recognition confidence

                // final result of recognition
                if (event.results[i].isFinal || i == event.results.length){ // if result is the final one and index is final
                    if (event.results[i][0].confidence < 0.8){//if confidence is less than 80% (not OK)
                        msg.text = "もう一度聞かせてください。"; //Could you say that again
                        speechsynthesis.speak(msg); //VCA speaks the above message
                        return;
                    }
                }
            }
            if (final_transcript.length > 1 && final_transcript.replace(/ /g, '') != last_res){// received texts but not an echo
//last_res: a shared variable called echo indicator declared and set by get_response function to indicate the echo of CRECA's response
                fukidashi = create_fukidashi(final_transcript, 100, 0, 'white'); // create (text) display window
                game.rootScene.addChild(fukidashi); //All texts sent from Google are shown in (text) display window
                document.getElementById("info").innerText = "Google そうしんずみ... "; //All received from Google.
                get_response(final_transcript);// All texts were sent to CRECA
                document.getElementById("info").innerText = "VCA そうしんずみ";// All texts were obtained by VCA
                                                        // is notified on the right upper window
            }
            else {
                document.getElementById("info").innerText = "にんしきちゅう... ";//still recognizing
            }
    }
};
```

Fig. 5. Voice Recognition Handler

The above script handles the following tasks:

- Receive the recognized texts from Google Cloud speech API (server) for speech recognition.
- Return to activate auto/manual error recovery if an error code indicating “insufficient precision” or “not OK” is returned.
- Display the resultant texts as the speech text of the avatar if the result is OK, and
- Output the resultant texts in a stream to VICA’s CRECA (server) through Nginx (Web Server) and WebSocket Script.

The Google cloud speech API exploited in our VICA is known for its accuracy in speech recognition. Nevertheless, it was not sufficient, and our original error correction/ recovery mechanism is provided in the client side as mentioned above.

V. EXPERIMENTAL EVALUATION

Our experimental evaluation of VICA was modeled after the comparison of general-purpose conversational agents carried out in (Shah et al. 2016). We selected a sample of 14 persons. Our study size is consistent with the one of (Shah et al. 2016), as they included 100 subjects in total but divided them into groups to evaluate six different conversational agents. We balanced our sample w.r.t. the Japanese population with respect to age and gender, with a slight bias toward younger clients. Our sample’s median age is 40 years[4].

Like (Shah et al. 2016) and (Shinozaki et al. 2013), we used ELIZA as a baseline to compare CRECA and VICA. Our comparison metrics is also an adaptation of the one used in (Shinozaki et al. 2013): it is based on the dialogue continuation length (the average number of statements while a dialogue session is continuing) called CPS (Conversation-turns Per Session) in (Shum et al 2018). Again like (Shah et al. 2016), each of our subjects performed a single session with VICA and another one with CRECA (in addition to the baseline ELIZA). To avoid learning effects, sessions were in random order and each subject was free to terminate a session when she liked. Besides measuring continuation (CPS), we asked our subjects to rate VICA’s performance in terms of (1) trust/comfortability, (2) verbalization, (3) positive feeling on a classic Likert scale (Likert 1932) ranging from 1 to 7. Results are shown in Table I and Table II.

CPS of VICA was 18.9 ranging from 6 to 32 for the younger (=< 40) as shown in Table I. The CPS for the older age (>40) was 20 as shown in Table II. VICA’s average CPS 19.5 is 67% larger or longer than ELIZA’s CPS of 11.7, but VICA’s CPS 20 for the elderly people is further (71%) longer. This result almost matches CRECA’s CPS 23.7 on the previous evaluation described in (Shinozaki et al. 2013). However, CRECA’s CPS includes at least 10 conversation-turns or statements which CRECA forces in its initial phase called a problem discovery phase. These statements ask occupation, affiliation, experiences (activity, accomplishment), and problem (sub) fields. VICA has not such a phase as explicitly forcing dialogue. Removing this phase in measuring the CPS to reduce the inequality, CPS of CRECA whose subjects are all young students (<25) is around 13 or less. Compared with this CRECA’s CPS 13, VICA’s CPS 20 for the older subjects above the medium age (>40) is around 53% longer and the average CPS 18.9 is around 45% larger. As a result, VICA equipped with high-performance voice communication capability was shown to be more instrumental than CRECA and further more than ELIZA, especially for elderly persons.

TABLE I: EVALUATION RESULTS OF BELOW 40 YEARS OLD

Individual Effect	1	2	3	4	5	6	7	Ave.
dialogue continuation	18	18	20	30	8	6	32	18.9
trust / comfortability	5	3	5	5	4	1	4	3.86
verbalization/ awareness	4	4	4	6	4	2	5	4.14
positive feeling/ awareness	3	3	3	4	3	2	2	2.86

TABLE II: EVALUATION RESULTS OF OVER 40 YEARS OLD

individual effect	1	2	3	4	5	6	7	Ave.
dialogue continuation	20	22	18	18	20	22	20	20
trust / comfortability	4	2	3	3	3	5	4	3.43
verbalization/ awareness	5	5	5	4	6	5	5	5
positive feeling/ awareness	6	5	4	5	6	5	5	5.14

[4] The median of the Japanese population age has been steadily increasing from 40 toward 50 years in the last decades. According to the statistics portal www.statista.com, the median age of Japan's population reached 46.3 years in 2015.

The Likert score of the questions about self-awareness for VICA was consistently very high, again for older subjects above the medium age (>40). It was mostly 5 or 6 (average of 5.14) concerning verbalization/ awareness as well as positive feeling/ awareness as shown in table II. This Likert scale result showed the capability of VICA having high-performance voice communication to help users be aware of themselves through reflection or self-searching is as follows: VICA's capability in the scale is more than 5 positioned on the significantly "agree" side in average for the elderly (>40) subjects, compared with mostly around 3 or 4 for the younger (=<40) subjects. In our previous experimentation using the Likert scale (Shinozaki et al. 2013), CRECA for young (<23) undergraduate student subjects showed 4.5 as the average value of the questions about self-awareness regarding CRECA. That of ELIZA was 3.5.

According to the review in (Shah et al. 2016), the best of best current text-based dialogue systemsamong ELIZA's successors received scores (*Eugene Goostman* 63.56, *Cleverbot* 62.18) approximately 2½ times *conversationally better* than their predecessor *Eliza (web-version) 24.86. Here, the scale means conversational ability from 0 to 100* interrogated to human judges *where: 0='poor' machinelike 50= 'good' but still machinelike 100=humanlike.* Adjusting the measurement scale from Likert 1-7 to *conversational ability 0-100*, VICA of older age group is in line with such best of best conversational systems. Even in these comparison, the older age group is better and the younger age group is worse. More in detail, this review shows the younger age group (=<25) scored the machines higher than the older age group (>25); *Eugene Goostman* 65.13, 57.25 (4.44 in Likert), 63.56 (4.81 in Likert), *Cleverbot* 66.15, 51.00 (4.06 in Likert), 62.18 (4.73 in Likert) each for the younger, for the older, and in average. Meanwhile, in the awareness of VCA on both verbalization and positive feeling, Table I and Table II show the older age group (>40) assessed higher 67.8 (5.07 in Likert) than the younger (=<40) age group 41.7 (3.50 in Likert). Thus, VCA having high-performance speech communication interface is favorable to older or elderly people.

While our sample is too small to carry out percentile analysis, we collected some anecdotal evidence from three clients above the sample median age. They assessed VICA as follows:

A man of 59 years (non-computer savvy): *VICA made my mind significantly peaceful since I could confess my anxiety to a human-like avatar. Such voice conversation in front of a human-like avatar reminded me of Catholic confession.*

A woman of 41 years: *The human-like avatar listened my speech very respectfully. This brought me peace of mind.*

A man of 45 years: *Indeed, VICA made my mind peaceful although it sometimes responded before I finished the sentence which I wanted to confess.*

These experimental results/ experiences show that VICA equipped with high-performance voice communication capability is instrumental for elderly or non-computer savvy persons.

VI. CONCLUSIONS

We described VICA, a counseling agent equipped with an image avatar having a high-performance voice communication as well as an advanced reply function exploiting high-performances based on the emotional relationship among past and current ones. For high-performance voice communication, Google cloud-based speech API was not only exploited but also enhanced by an application-level error correction/ recovery mechanism. In percentile evaluation, the dialogue continuation (CPS) of VICA for the older age (>40) substantially improved 53% to CRECA and 71% to ELIZA. Evaluation on a Likert scale of 7 showed considerably high scores of average 5.14 again for old persons aged (>40) more than the sample median. By comparison on average, VICA for the older (age >40) was 5.14 while CRECA (students, age <25) was 4.50, ELIZA was 3.50, and the best of ELIZA's successors for the older (age >25) was 4.41. These experimental experiences show VICA equipped with high-performance voice communication capability is instrumental especially for elderly persons.

While these results are encouraging, several points remain open for future research. Here we briefly discuss two major opportunities for future work.

- *Improving user experience*: While we have been using Google standard avatars during the testing phase, we realize that customizing VICA's appearance (in the line of commercial products like *Tess*) may be critical to acceptance by more general (elderly or digital-divided) users. In the future, a lifelike image of the specific client will be taken by a photograph, and a lifelike avatar will be created out of it using FaceGen (2017). We also are considering using the Rapid Avatar Generator toolkit available at www.ict.usc.edu.

- *Achieving full language independence:* While technicalities of Japanese handling are outside the scope of this paper, it is important to remark that voice conversation in Japanese, especially on work-related matters, needs to handle homophony of Japanese and English terms, and success in conversion may depend on the clients' pronunciation.

In a recent paper, Ashwell and Elam (2017) were able to ascertain that the pronunciation of certain sounds is the single biggest reason for a fall in recognition accuracy when English and Japanese terms are used in the same sentence. We argue that pronunciation may not be an insurmountable barrier to using speech recognition system for counseling, if the agent is aware of the situation and openly addresses it in the dialogue. In all cases, our experience shows that progress in AI-based conversational agents requires general techniques for handling Asian and Middle Eastern languages that do not rely on the Roman alphabet.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP15K00349, JP15K00382 and by Artificial Research Promotion foundation. We thank the students of the Distributed Intelligent Systems Lab, Tokyo Denki University Japan and of the Machine Learning Lab, Meiji University Japan for their dedicated help to VICA implementation.

REFERENCES

Anisetti M, Bellandi V, Damiani E, Jeon G, and Jeong J (2007) Full Controllable Face Detection System Architecture for Robotic Vision. Proceedings 2007 Third International IEEE Conference on Signal-Image Technologies and Internet-Based System. pp. 748-756. https://doi.org/10.1109/SITIS.2007.89

Asay TP, Lambert MJ (1966) The Empirical Case for the Common Factors in Therapy: Quantitative Findings. The HEAT & SOUL of CHANGE, pp. 23-55, 1966. http://dx.doi.org/10.1037/11132-001

Ashwell T, and Elam J (2017) How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? Jalt Call Journal 13(1): 59–76

DeVault D et al. (2014) SimSensei Kiosk: A virtual human interviewer for healthcare decision support. Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, pp. 1061-1068

FaceGen (2017) FaceGen Modeller: 3D Face Generator (accessed 26 June, 2018). http://www.facegen.com/modeller.htm

Floridi L, Taddeo M, Turilli M (2009) Turing's Imitation Game: Still an Impossible Challenge for All Machines and Some Judges - An Evaluation of the 2008 Loebner Contest. Minds and Machines 19(1): 145-150. https://doi.org/10.1007/s11023-008-9130-6

Gonzalez AJ, Tsuruta S, Sakurai Y, Nguyen J, Takada K, and Uchida K (2010) Using Contexts to Supervise a Collaborative Process. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 25(1): 25-40. https://doi.org/10.1017/S0890060410000156

Guangbing Y, Nian-Shing C, Kinshuk SE, Anderson T, Wen D (2013) The effectiveness of automatic text summarization in mobile learning contexts. Comput. Educ. 68: 233-243. http://dx.doi.org/10.1016/j.compedu.2013.05.012

Gupta V, and Lehal GS (2010) A survey of text summarization extractive techniques. Journal of Emerging Technologies in Web Intelligence 2(3): 258-268. https://doi.org/10.4304/jetwi.2.3.258-268

Han S, Lee K, Lee D, Lee GG (2013) Counseling Dialog System with 5W1H Extraction. Proceeding of the SIGDIAL 2013 Conference, pp. 349-353. http://www.aclweb.org/anthology/W13-4054

Heung-Yeung Shum, Xiaodong He, Di Li (2018) From Eliza to XiaoIce: challenges and opportunities with social chatbots *Frontiers of Information Technology & Electronic Engineering* (ISSN 2095-9184, monthly) Chinese Academy of Engineering (CAE) and Zhejiang University, co-published co-published by Springer & Zhejiang University Press January 2018, Volume 19, Issue 1, pp 10–26

Hung V, Gonzalez A, De Mara R (2009) Towards a Context-Based Dialog Management Layer for Expert Systems. International Conference on Information, Process, and Knowledge Management, Cancun, 2009, pp. 60-65. https://doi.org/10.1109/eKNOW.2009.10

Ivey AE, Packard NG, Ivey MB (2006) Basic Attending Skills, Fourth Edition. Microtraning Associates, Ann Arbor, Michigan, pp. 93-97

Kennedy A, Kazantseva A, Inkpen D, Szpakowicz S (2012) Getting emotional about news summarization. Advances in Artificial Intelligence, pp. 121-132. https://doi.org/10.1007/978-3-642-30353-1_11.

Likert R (1932) A Technique for the Measurement of Attitudes. Archives of Psychology. 22(140):1-55. http://www.worldcat.org/oclc/812060

Mental Health Foundation (2018) Stress: Are we coping? London: Mental Health Foundation. (accessed 30 December, 2018). https://www.mentalhealth.org.uk/publications/stress-are-we-coping

Ministry of Health, Labour and Welfare, "General condition of mental health caring and smoking countermeasure consequence (Japanese)", 2008, http://www.mhlw.go.jp/toukei/itiran/roudou/saigai/anzen/kenkou07/index .html

Powell K (1966) Richard Rogers: Complete Works - Volume 3. Phaidon Press, London.

Ptaszynski Michal, et al. (2009) A system for affect analysis of utterances in Japanese supported with web mining. Journal of Japan Society for Fuzzy Theory and Intelligent Informatics 21.2: 194-213

Rizzo A, and Talbot T (2016) Chapter 18: Virtual Reality Standardized Patients for Clinical Training. The Digital Patient: Advancing Healthcare, Research, and Education, John Wiley & Sons, Hoboken, New Jersey, United States. https://doi.org/10.1002/9781118952788.ch18

Rogers CR (1957) The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology 21(2): 95-103. http://dx.doi.org/10.1037/h0045357

Savickas ML (2011) Career Counseling. American Psychological Association, Washington

Shah H, Warwick K, Vallverde J, Wu D (2016) Can Machines Talk? Comparison of ELIZA with Modern Dialogue Systems. Computers and Human Behavior 58(C): 278-295. https://doi.org/10.1016/j.chb.2016.01.004

Shinozaki T, Yamamoto Y, Tsuruta S (2013) Context-based counselor agent for software development ecosystem. Computing 97(1): 3-28. https://doi.org/10.1007/s00607-013-0352-y

Yamamoto Y, Shinozaki T, Tsuruta S, Damiani E, and Knauf R (2016) Equipping a Context Respectful Counseling Agent with a Human-like Voice Synthesizer. Proceedings of 2016 World Automation Congress (WAC), pp. 1-6. https://doi.org/10.1109/WAC.2016.7583031

Wallace RS (2003) The elements of AIML style. ALICE AI Foundation. Retrieved on July 15, 2018 from https://files.ifi.uzh.ch/cl/hess/classes/seminare/chatbots/style.pdf

Weissman J, Russell D, Jay M, Beasley JM, Malaspina D, Pegus C (2017) Disparities in Health Care Utilization and Functional Limitations Among Adults With Serious Psychological Distress, 2006–2014. Psychiatric Services 68(7): 653-659. https://doi.org/10.1176/appi.ps.201600260

Weizenbaum J (1966) ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine. Commun. ACM 9(1):36-45. https://doi.org/10.1145/365153.365168

Weizenbaum J (1976) Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman & Co. New York