New Dimensions in Testimony Demonstration

Ron Artstein and **Alesia Gainer** and **Kallirroi Georgila**
Anton Leuski and **Ari Shapiro** and **David Traum**
University of Southern California Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista CA 90094, USA
{artstein|gainer|kgeorgila|leuski|shapiro|traum}@ict.usc.edu

Abstract

New Dimensions in Testimony is a prototype dialogue system that allows users to conduct a conversation with a real person who is not available for conversation in real time. Users talk to a persistent representation of Holocaust survivor Pinchas Gutter on a screen, while a dialogue agent selects appropriate responses to user utterances from a set of pre-recorded video statements, simulating a live conversation. The technology is similar to existing conversational agents, but to our knowledge this is the first system to portray a real person. The demonstration will show the system on a range of screens (from mobile phones to large TVs), and allow users to have individual conversations with Mr. Gutter.

1 Introduction

This demonstration presents *New Dimensions in Testimony*, the first dialogue system prototype to enable a conversation with a real person who is not available for conversation in real time. Technology such as the telegraph, telephone and videoconferencing allowed people to communicate with each other across long distances with increasing fidelity, but required that the participants make themselves available for conversation at the same time. Other technologies such as writing, audio recording and video recording allowed people to send messages across time, but did not allow synchronous conversation. In the past two decades, embodied conversational agents – that is, artificial characters controlled by computer programs – have been able to converse with users with increased complexity and naturalness. Our system demonstrates how conversational agent technology can be used with recorded video statements from a real person to create a conversation that is offset in time: the speaker recorded his statements in the past as a message to the future, and users now can interact with him and hold a conversation as if the speaker were present.

The *New Dimensions in Testimony* prototype is intended to emulate a conversation with Holocaust survivor Pinchas Gutter. Holocaust education today relies to a great extent on survivors talking to audiences in museums and classrooms, relating their experiences directly and creating an intimate connection with their audiences (Bar-On, 2003). However, the youngest survivors are in their seventies today, and in a few years there will be no more survivors left to tell the story in person. The prototype will afford future generations the opportunity to engage in such conversation, talking to Pinchas Gutter and asking him questions about his life before, during and after the Holocaust. What makes our project unique is the ability to connect on a personal level with a survivor, and the history, even when that survivor is not present.

The technology can have a wide range of applications, such as preserving the memory of a person for the future (historical figures as well as ordinary people); enabling conversation with family and friends who are temporarily unavailable (traveling, deployed overseas, or incarcerated); allowing popular speakers (leaders, celebrities) to engage with multiple people at the same time; and enabling access to expert knowledge and customer service.

Proceedings of NAACL-HLT 2016 (Demonstrations), pages 32–36,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

2 Technical details

In the *New Dimensions in Testimony* prototype, users talk to a persistent representation of a Holocaust survivor presented on a video screen, and a computer algorithm selects and plays individual video clips of the survivor in response to user utterances. The result is much like an ordinary conversation between the user and the survivor. The system has been described in detail in previous publications, covering the proof of concept (Artstein et al., 2014), the content elicitation process (Artstein et al., 2015), the language processing (Traum et al., 2015a), the full prototype (Traum et al., 2015b), and ethical considerations (Artstein and Silver, 2016). Here we give a brief description of the language processing technology and the system's runtime components.

2.1 Language processing

At the heart of the runtime computer system is a response classifier and dialogue management component called NPCEditor (Leuski and Traum, 2011), which selects a response to each user utterance. NPCEditor combines the functions of Natural Language Understanding (NLU) and Dialogue Management – understanding the utterance text and selecting an appropriate response. The NLU functionality is a classifier trained on linked question-response pairs, which identifies the most appropriate response to new (unseen) user input. The dialogue management logic is designed to deal with instances where the classifier cannot identify a good direct response. During training, NPCEditor calculates a response threshold based on the classifier's confidence in the appropriateness of selected responses; at runtime, if the confidence for a selected response falls below the predetermined threshold, that response is replaced with an "off-topic" utterance that asks the user to repeat the question or takes initiative and changes the topic (Leuski et al., 2006); such failure to return a direct response, also called non-understanding (Bohus and Rudnicky, 2005), is usually preferred over returning an inappropriate one (misunderstanding). The current system uses a five-stage off-topic selection algorithm which is an extension of that presented in Artstein et al. (2009). Figure 1 shows a sample dialogue illustrating the handling of non-understanding.

User Hello Pinchas, how are you?
Las Vegas how are you

Pinchas Can you just repeat that?

User Hello Pinchas, can you hear me?
how thick is can you hear me

Pinchas I can hear you, yeah.

User Pinchas, can you tell me how old you are?
Vegas can you tell me how old you are

Pinchas I was born in nineteen thirty two and we are now in two thousand and fourteen and you can make your own reckoning.

User Can you tell me a little bit about your childhood?

Pinchas My life before the war, I was born in Lodz in Poland, and I was eight years old when the war started. So the first eight years I lived in a home, a very religious one, a Hasidic home. I had a twin sister, my. . .

User Have you ever met Hitler?
mermaid Hitler

Pinchas Please repeat that.

User Did you ever meet Hitler?

Pinchas Did I ever meet Hitler? Well, Hitler was a Führer of Germany and I was born in Poland and I was a little Jewish boy, so the chances of somebody like me meeting Hitler was impossible, because Hitler was. . .

User What would you do if you met Hitler?

Pinchas Geyn drerd arein, that's Yiddish and what it means go and bury yourself.

Figure 1: Sample dialogue from system testing, illustrating correct responses to noisy speech input as well as the handling of non-understanding using off-topic responses. Speech recognition output is shown in italics when it departs substantially from the spoken text. The final utterance was originally recorded as a response to the question "If you could ask one question or say anything to Adolf Hitler, what would that be?"

The system has over 1700 recorded responses (a total of almost 18 hours of video), allowing it to give appropriate direct responses to about 64% of the user questions, with 20% off-topic responses and the remaining 16% being errors. This is sufficient to enable a reasonable conversation flow (Traum et al., 2015a). Between responses the system loops through short videos of idle behavior by the survivor, giving the feeling of live presence. When the user starts speaking, this changes to concentrated listening behavior, adding to the feeling of engagement.

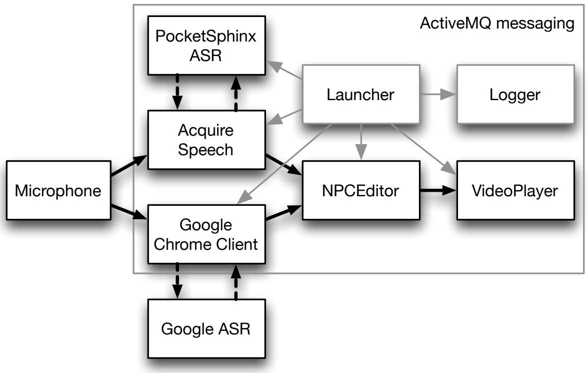


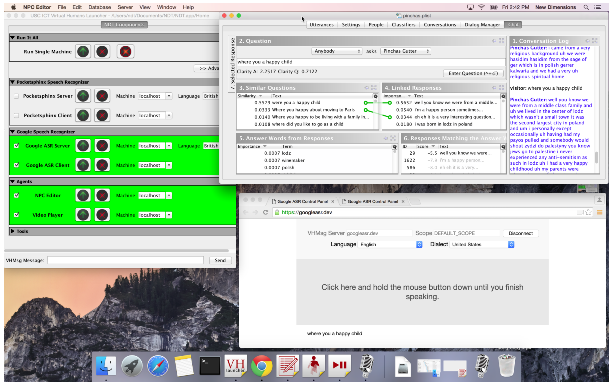
Figure 2: System architecture: Black lines show the data flow through the system, while gray arrows indicate the control messages from the Launcher interface. Solid arrows represent messages passed via ActiveMQ, and dotted lines represent data going over TCP/IP.

2.2 Software components

The system is built on top of the components from the USC ICT Virtual Human Toolkit, which is publicly available (Hartholt et al., 2013).[1] Specifically, we use the AcquireSpeech tool for capturing the user's speech, CMU PocketSphinx[2] and Google Chrome ASR[3] tools for speech recognition, NPCEditor (Leuski and Traum, 2011) for classifying the utterance text and selecting the appropriate response, and a video player to deliver the selected video response. The individual components run as separate applications and are linked together by VHMsg[4] messaging over ActiveMQ: each component connects to the broker server and sends and receives messages to other components via the broker. The system setup also uses the JLogger tool for recording the messages, and the Launcher tool that controls starting and stopping of individual tools. Figure 2 shows the overall system architecture. A typical session on a Mac is shown in Figure 3.

2.3 System hardware

A typical installation is run on a 15-inch MacBook Pro with Retina display, connected via HDMI to an external monitor or television. We have used displays ranging from a basic 22-inch desktop monitor for personal interaction to a large theatre projector screen, though our preferred display is an 80-inch high definition television in vertical orientation (Figure 4). This allows showing the speaker at approximately life size, making it appropriate for one-on-one and small group interaction, as well as large group interaction in a theatre setting.



USC ICT Virtual Humans Launcher - /Users/ndt/Documents/NDT/NDT.app/Home

□ Pocketsphinx Server	Machine localhost	Language British
□ Pocketsphinx Client	Machine localhost	
☑ Google ASR Server	Machine localhost	Language British
☑ Google ASR Client	Machine localhost	
☑ NPC Editor	Machine localhost	
☑ Video Player	Machine localhost	

VHMsg Message:

pinchas.plist

2. Question

Anybody asks Pinchas Gutter

where you a happy child

Clarity A: 2.2517 Clarity Q: 0.7122

3. Similar Questions

Similarity	Text
0.5579	were you a happy child
0.0333	Where you happy about moving to Paris
0.0140	Where you happy to be living with a family in...
0.0108	where did you like to go as a child

4. Linked Responses

Importan...	Text
0.5652	well you know we were from a middle...
0.0540	i'm a happy person sometimes...
0.0344	eh eh it is a very interesting question...
0.0180	i was born in lodz in poland

5. Answer Words from Responses

Importance	Term
0.0007	lodz
0.0007	winemaker
0.0007	polish

6. Responses Matching the Answer W

ID	Score	Text
29	-5.5	well you know we were...
1622	-7.9	i'm a happy person...
586	-8.0	eh eh it is a very...

1. Conversation Log

Pinchas Gutter: i came from a very religious background uh we were hasidim hasidim from the sage of ger which is in polish gerrer kalwaria and we had a very uh religious spiritual home

visitor: where you a happy child

Pinchas Gutter: well you know we were from a middle class family and uh we lived in the center of lodz which wasn't a small town it was the second largest city in poland and um i personally except occasionally uh having had my payos pulled and somebody would shout zydzi do palestyny you know jews go to palestine i never experienced any anti-semitism as such in lodz uh i had a very happy childhood uh my parents were

Google ASR Control Panel | Google ASR Control Panel

https://googleasr.dev

VHMsg Server googleasr.dev Scope DEFAULT_SCOPE

Language English Dialect United States

Click here and hold the mouse button down until you finish speaking.

where you a happy child

Figure 3: A typical desktop runtime environment: Launcher on the left, NPCEditor top right, Google Chrome ASR bottom right. Video player is displayed (maximized) on an external screen, and JLogger is minimized.

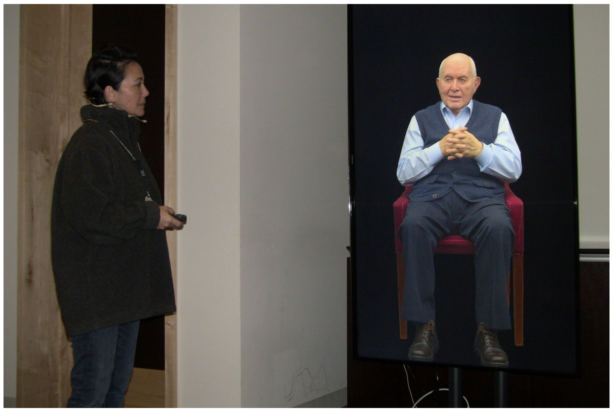


Figure 4: User talking to Mr. Gutter on a large TV.

For small, informal demonstrations in a quiet setting, we have had good results using the MacBook Pro's built-in microphone for audio capture, and the built-in trackpad as a push-to-talk button. In more challenging environments we use a Sennheiser HSP-4 headworn microphone, which works well to isolate the user's speech from the background noise. The microphone is connected to a wireless transmitter-receiver pair and sent to the computer through a Focusrite Scarlett 2i2 USB recording audio interface. Push-to-talk functionality is provided

[1] http://vhtoolkit.ict.usc.edu
[2] http://cmusphinx.sourceforge.net
[3] https://www.google.com/intl/en/chrome/demos/speech.html
[4] https://sourceforge.net/projects/vhmsg/

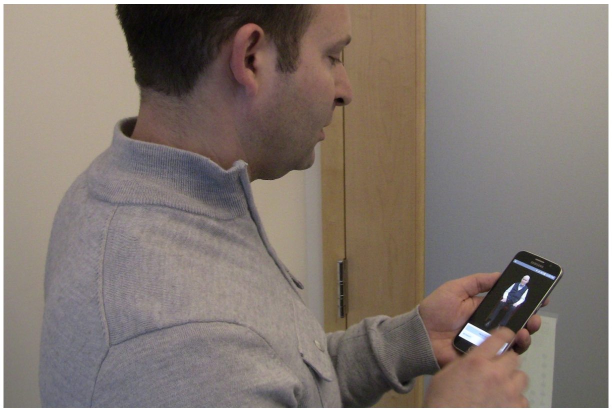


Figure 5: User talking to Mr. Gutter on a mobile phone.

by a wireless mouse, removing any physical connection with the computer and allowing the user full freedom of movement. The speaker's audio is normally transmitted over HDMI together with the video, but can be routed through the Focusrite interface to external speakers when needed.

2.4 Mobile version

The mobile version (Figure 5) is built using an Android-based virtual human software platform (Feng et al., 2015). This platform allows script-based access to speech recognition, video playback, and dialogue management services via Jerome, an implementation of the NPCEditor algorithm.

In order to accommodate the smaller display and mobile nature of a handheld device, the videos were reduced from 1080×1920 to 270×480, effectively reducing the size of the videos by a factor of 16. This results in a change in video file size from approximately 1.7 gb per hour (28 mb per minute) of content to 110 mb per hour (1.75 mb per minute) of content. Frequently used videos, such as those for listening and off-topic responses are stored locally on the mobile device, while the rest are stored on a video-streaming cloud service and are retrieved on demand. Streaming videos of such size via wifi connection yields similar response times to playing the videos locally on the device, and greatly reduces the size of the mobile app. An additional button on the app allows the user to indicate explicitly that a given response is inappropriate to the question asked; this information is used for future classifier training.

The classification algorithm and data are processed locally on the device. Speech recognition is handled via the Android's interface to Google ASR. Thus, there are three network messages for each user utterance: one to obtain the results of the ASR, another to retrieve the desired video if found, and a third to store the recognized question and response in a cloud-based database, for later analysis. The classifier data can be replaced through an update to the mobile app, thus allowing for easy propagation of improvements in the question/answer interaction as larger amounts of data are captured and analyzed.

3 Demonstration outline

The demonstration will feature a live interaction between participants and Pinchas Gutter, on both desktop and mobile platforms. Depending on the participant's preference, interaction will be either moderated (speech relayed by a demonstrator) or direct (participant operating the push-to-talk and talking into the microphone). The live conversations will highlight Mr. Gutter's understanding, his ability to deal with non-understanding of user utterances, and the overall coherence of the conversation. It will also showcase many of Mr. Gutter's moving personal stories, and illustrate the sense of closeness and bonding that can form when talking to a person through a system of time-offset interaction.

Acknowledgments

The New Dimensions in Testimony prototype is a collaboration between the USC Institute for Creative Technologies, the USC Shoah Foundation, and Conscience Display. It was made possible by generous donations from private foundations and individuals. We are extremely grateful to The Pears Foundation, Louis F. Smith, and two anonymous donors for their support. The Los Angeles Museum of the Holocaust, the Museum of Tolerance, New Roads School in Santa Monica, and the Illinois Holocaust Museum and Education Center offered their facilities for data collection and testing. We owe special thanks to Pinchas Gutter for sharing his story, and for his tireless efforts to educate the world about the Holocaust.

This work was supported in part by the U.S. Army; statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Ron Artstein and Kenneth Silver. 2016. Ethics for a combined human-machine dialogue agent. In *AAAI Spring Symposium SS-16-04: Ethical and Moral Considerations in Non-Human Agents*, pages 184–189, Stanford, California, March. AAAI Press.

Ron Artstein, Sudeep Gandhe, Jillian Gerten, Anton Leuski, and David Traum. 2009. Semi-formal evaluation of conversational characters. In Orna Grumberg, Michael Kaminski, Shmuel Katz, and Shuly Wintner, editors, *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, volume 5533 of *Lecture Notes in Computer Science*, pages 22–35. Springer, Heidelberg, May.

Ron Artstein, David Traum, Oleg Alexander, Anton Leuski, Andrew Jones, Kallirroi Georgila, Paul Debevec, William Swartout, Heather Maio, and Stephen Smith. 2014. Time-offset interaction with a Holocaust survivor. In *Proceedings of IUI*, pages 163–168, Haifa, Israel, February.

Ron Artstein, Anton Leuski, Heather Maio, Tomer Mor-Barak, Carla Gordon, and David Traum. 2015. How many utterances are needed to support time-offset interaction? In *Proceedings of FLAIRS-28*, pages 144–149, Hollywood, Florida, May.

Dan Bar-On. 2003. Importance of testimonies in Holocaust education. *Dimensions Online: A Journal of Holocaust Studies*, 17(1).

Dan Bohus and Alexander I. Rudnicky. 2005. Sorry, I didn't catch that! – An investigation of non-understanding errors and recovery strategies. In *Proceedings of SIGDIAL*, pages 128–143, Lisbon, Portugal, September.

Andrew W. Feng, Anton Leuski, Stacy Marsella, Dan Casas, Sin-Hwa Kang, and Ari Shapiro. 2015. A platform for building mobile virtual humans. In Willem-Paul Brinkman, Joost Broekens, and Dirk Heylen, editors, *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26–28, 2015 Proceedings*, volume 9238 of *Lecture Notes in Computer Science*, pages 310–319. Springer, August.

Arno Hartholt, David Traum, Stacy C. Marsella, Ari Shapiro, Giota Stratou, Anton Leuski, Louis-Philippe Morency, and Jonathan Gratch. 2013. All together now: Introducing the virtual human toolkit. In Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira, editors, *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29–31, 2013 Proceedings*, volume 8108 of *Lecture Notes in Computer Science*, pages 368–381. Springer, August.

Anton Leuski and David Traum. 2011. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine*, 32(2):42–56.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of SIGDIAL*, Sydney, Australia, July.

David Traum, Kallirroi Georgila, Ron Artstein, and Anton Leuski. 2015a. Evaluating spoken dialogue processing for time-offset interaction. In *Proceedings of SIGDIAL*, pages 199–208, Prague, Czech Republic, September.

David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. 2015b. New Dimensions in Testimony: Digitally preserving a Holocaust survivor's interactive storytelling. In *Interactive Storytelling: 8th International Conference on Interactive Digital Storytelling, ICIDS 2015, Copenhagen, Denmark, November 30 – December 4, 2015 Proceedings*, volume 9445 of *Lecture Notes in Computer Science*, pages 269–281. Springer, December.