

Maintaining Credible Dialogs in a VideoBot System with Special Audio Techniques

Doug DeGroot
Leiden Institute of Advanced Computer Science
Leiden University, The Netherlands
degroot@liacs.nl

Abstract

An approach to using actual audio and video clips of real humans to construct artificial, conversational agents and bots is presented. This approach differs from other schemes focusing on believable, emotional, intelligent agents and bots in that it begins with real human subjects but constructs artificial behaviors and interactions with human users as opposed to beginning with artificial characters and trying to construct real interactions. The approach presents several production challenges during the filming, postproduction, and scripting phases of bot creation that make it difficult for human users to sustain suspension of disbelief during interaction with the bot. Various approaches to solving these problems are presented and described.

1. Introduction

The Internet enables the deployment of autonomous, semi-intelligent agents and bots (Web robots) to assist users, companies, groups, devices, or even other bots in carrying out a wide variety of tedious, time-consuming, difficult, or expert-level tasks. Simple examples include bots that hunt for the best price or terms for some item you want to buy or that hunt down and gather specific news that you might be interested in. The great majority of bots are "worker" bots, examples of which include mailbots, spambots, warbots, shopping bots, auction bots, site indexing bots, and the like [Leonard]. Most of these are designed to recede into the background of an application or system -- to do their work silently, with little or no human intervention, and without being noticed.

Other types of bots, however, are emerging that are instead designed specifically to be noticed, to perform tasks other than worker-bee-level tasks, and to affect not only the specific task assignment but also the emotional experiences of the humans involved in their use. Example application domains for this type bot include entertainment, shopping consultations, training, companionship,

artificial pets, and virtual cloning, to name a few [Hayes-Roth, Maes]. To be successful, these types of bots must exhibit — to varying degrees — intelligence, believability, and emotions. Early examples of these bots include chatterbots, date bots, game bots, personal advisers, customer service bots, knowbots, etc. [BotSpot, AgentLand].

A wide variety of traditional computer and communication technologies are involved in the creation, control, communication, and evolution of these bots, including artificial intelligence, knowledge databases, distributed computing, natural language recognition, communication protocols, fuzzy logic, computer animation, audio and video, etc. Further, though, other scientific and social fields are proving to be essential to the success of these bots, including personality profiling, psychology, cognitive science, linguistics, philosophy, models of the mind, and social anthropology, to name a few [Nass].

These additional fields reach pre-eminence in those application domains that exploit "human like" bots — bots that exhibit believable personality, emotions, and intelligence. In these domains, success is often maximized only when the human users of the applications that deploy these bots are either "fooled" into believing the bot is a human or when they are at least *willing* to suspend their disbelief of the bot's humanity long enough to effectuate the experience or task at hand. (This "suspension of disbelief" is actually a stronger condition than we currently believe necessary for our VideoBots approach; however, we continue to adopt it as a key requirement of our research for the sake of sufficiency.)

Because VideoBots are "human like" in appearance, sound, and behavior to begin with, it is possible to construct bots that for all practical purposes are perceived as being humans (not just being "human like") with whom interactions are occurring in real time. This perception should continue until such time as any technology glitches provide sufficient clues to the human participant for the belief to become disbelief. Thus while animated, graphical approaches to bots must focus on achieving believable humanness, VideoBots must also focus on keeping from destroying it.

2. VideoBots as Synthetic Humans

Many visual-bot paradigms attempt to present bot interfaces that appear “human-like”. However, the majority of these bot paradigms rely on graphically rendered artificial characters (either animated or static) for the visual interface. Usually, these synthetic characters merely appear to be human or perhaps even only human-like. Others use images of actual human faces as textures and then develop 3-dimensional, graphically rendered animation sequences based on these faces, hoping that these texture-mapped, synthetic characters will appear to be more “human-like.” Almost all of these approaches rely on simple textual output or speech synthesis for the audible interface [MSAgent, Massaro, Ananova, etc.].

Being “human-like” is a different requirement than being “believable,” and depending on the bot paradigm, it can be either a stronger or a weaker requirement. Also, the requirement to appear to be human differs from that of appearing to be “human-like”; and appearing to be “human” differs considerably between having to appear to be any human (a generic human) versus a specific human, including possibly well known humans. Furthermore, being “well known” can differ between being broadly well known (e.g., John Wayne or Marilyn Monroe [Thalmann]) and well known to only a few (e.g., by family members).

We are researching and developing a class of “human-like” bots whose primary mode of interaction with the user is based on real video footage and audio clips of actual humans. Thus when the user interacts with the bot, the user sees and hears a real human. The actions, statements, facial expressions, and responses seen and heard by the user are all recordings of their actual enactment by the human on which the bot is based (or at least largely so; even though certain elements of the human-computer interface may exploit computer-generated graphics, animation, and speech synthesis, these are incidental to experiencing the bot.)

As the human user converses with the VideoBot, the system’s underlying dialog manager attempts to maintain a coherent, satisfying, and credible dialog with the user by selecting an appropriate set of audio-video clips to be played in response. The audio-video clips can be replayed in an indeterminate order. Because the audio responses by the VideoBot are accompanied by the corresponding video, the dialog manager must also take into consideration the content and nature of the video clips when making its audio selection; this challenge is tougher than that faced by text-only or audio-only chatterbots. In addition, the dialog manager must select appropriate video sequences to replay when moments of silence (non-conversation) are reached between the human user and the VideoBot. Any failure to achieve and sustain believably human actions here would quickly harm the user’s suspension of disbelief, whether willing or coerced.

We refer to this class of bots as “VideoBots” in order to distinguish them and the consequent approach to their development from other work on animated, emotional

persona, especially those whose main design goal is also achieving “believability”, as defined in [Loyall]. Our research is currently focusing more on sustaining both believability and human-like qualities rather than on achieving them.

3. Audio Characteristics of a VideoBot

We have developed several prototypes of VideoBots to date in an attempt to help validate the utility of the VideoBot paradigm. In particular, we are exploring the specific use of VideoBots as “companions” in next-generation TVs and smart homes, but these are simply representative domains. In these prototypes, one or more household members can interact with the VideoBot(s) employed by the household. Our current bots are able to carry on a fairly limited conversation with a user, perform limited planning tasks on behalf of a user, turn on or off or otherwise control a number of consumer appliances within the home, process emails, respond to doorbells and phone calls, control room lighting, change TV channels, and the like.

The scope of this paper does not allow for a complete description of the VideoBot paradigm (further details can be found in [DeGroot]). Instead, we discuss several fundamental characteristics of VideoBots that we believe are most pertinent and problematic to maintaining credible dialogs. Among these are:

1. VideoBots speak to and listen to the user, usually in a conversational manner (although certain VideoBot application domains may involve announcements only rather than conversations). Accordingly, they can be considered to be relatives of Chatterbots [Laven, Suereth], but relatives that use video and speech (audio) rather than text (and in our case, real speech rather than synthesized speech).
2. VideoBots speak to the user using primarily pre-recorded speech segments consisting of one or more entire sentences (although both multi-modal hypermedia and computer manipulation of the utterances is exploited when desirable and/or necessary).
3. Each audio segment has at least one corresponding video segment that can be played back simultaneously with the audio. The time durations of these must be closely matched.
4. VideoBots listen to the user and, to a large degree, employ speaker-independent speech recognition to do so. (Text-based input is provided, but we expect it will be seldom used.)
5. A VideoBot is generally expected to be cast as a “live person,” and the video and audio clips that are presented to the viewer are expected to be “believably” the person depicted in the video. (The viewer may or

may not know the identity of the person on whom the VideoBot is based.)

6. A VideoBot can participate in 2-way communication with the user in either visible or non-visible states, and the video used to portray the VideoBot can be turned off and on without affecting the communication.
7. A VideoBot can (optionally) be observed doing its work, even if much of that work occurs behind the scenes through a set of computer agents that cooperate with the bot and present it the information required to perform its job. Further, between tasks, the bot must be visible, exhibiting believable “idle” or “busy work” activity.
8. Although the interactions, the sequences of interactions, and the various responses to input from the user or external world are possibly random, the responses presented to the user are nevertheless ordered, actual responses. If the VideoBot becomes “lost” or “confused”, it must both recover and explain away the problem.
9. VideoBots must be able to converse in real time; any unexpected or unnatural delays caused by internal system state or Internet communication delays must be able to be identified and “explained away” by the bot, in a believable manner.

4. Creating Credible Conversations

As a fundamental requirement of our research, a VideoBot must first and foremost satisfy the requirements of “believability.” As in [Loyall], we require the definition of “believable” to include the ability to allow (sic) the user to suspend disbelief and to provide a convincing portrayal of the personality they expect or come to expect. As Loyall points out, these goals differ from most work in autonomous agents in that “the focus is on building agents that have distinct, specific personalities” [Loyall].

However, we go further in that we also require the definition of “believable” to include the ability to both allow and persuade the user to suspend disbelief about whether the user is viewing and interacting with a generic human or the specific, real person on whom the VideoBot is based. In addition, our work also focuses on building agents that have specific, known personalities. The re-creation of the audio and video interface is thus critically important to achieving and sustaining the suspension of disbelief on the part of the user. These requirements are similar in nature to those faced by the creators of Max Headroom and other synthetic actors [Thalman]. VideoBots, however, have to support dynamic, interactive communication.

Our VideoBot architecture currently supports conversation between two people (entities) and enables primitive multi-modal discourse. On the user’s side, we support

speech as the dominant input mode; on the VideoBot’s side, we utilize playback of prerecorded audio segments (coupled with the accompanying video, under normal circumstances) as the dominant output mode. We are currently using Apple’s Speech Recognition Kit for input [ASRKit] and Apple’s QuickTime system for both audio and video playback [Stern].

Using the word “conversation” with respect to our current VideoBots is more than a bit presumptuous on our part, as the current domains of discourse are purposefully limited at this point. Our current research thrust for providing conversations focuses more on the system architecture requirements for maintaining credibility in the dialogs more so than on supporting multiple modes of discourse and high degrees of flexibility during a conversation [Allen].

Typically, we expect a conversational part of the discourse to be a somewhat orderly, purposeful, and directed exchange of sentences (or short sequences of sentences), queries, and responses. As typified by many current ChatterBots, we expect a high degree of conversational “turn taking”, i.e., either the user or the bot says something, the other party says something in return, and so on.

However, unlike most chatterbots, we specifically allow for:

1. Multiple, asynchronous threads of conversation between the two parties.
2. Interruptions of one party by the other, during a speech segment, so-called “barges-ins”. (Note that in particular, we do allow the VideoBot to interrupt the human on certain occasions.)
3. Suspensions and resumptions of extant conversational threads.
4. The use of external control devices to provide input and to set system state (e.g., radio frequency or infrared remote control units, light switches, etc.)

As the conversation takes place between the user and the bot, the system plays back a dynamic sequence of highly coupled audio and video clips. During pauses, or while the user is speaking, the system must continue to play back video sequences, even if these are “idling” sequences. Our VideoBot architecture allows video clips to be arbitrarily arranged and dynamically played back so that multiple emotional responses can be simulated. The choice of video and audio clips to be played depends on the internal state of the bot’s knowledge base, the state of the system it is embedded within, and the state of any conversational threads in progress.

5. Problems with Real-Time Interactions

Because VideoBots are created from sets of prerecorded audio and video clips, several specific problems arise from

the requirement to maintain credible dialogs between a VideoBot and a human user. Many of these problems are unique to the VideoBot paradigm, but others are shared with other conversational bot paradigms [see e.g., Hutchens, Lucente]. In addition, certain audio and discourse related problems arise from several of the particular application domains we are exploring. (For example, in the smart house arena, the bot must be able to follow household members throughout the house and interact with them in multiple rooms, regardless of whether there is a TV in the room, and even if so, regardless of whether it is currently turned on.)

In the remainder of this paper, we describe and discuss several of these problems and some of our current approaches to solving them. In particular, we focus on several “tricks” such as those commonly used in various well-known chatterbots. However, these tricks are not as easily introduced into VideoBots, and many require special, additional audio and video “tricks” as well.

6. Fixed Collections of Audio/Video Clips

As part of the VideoBot production process, the audio and video clips are filmed during the preproduction stage. Generally, the actor or actress must be filmed wearing the same outfit, in the same setting, with the same lighting, same miking, etc. Once filmed, digitized, and captured on the computer, these audio and video assets constitute the entire set of available character movements, facial expressions, emotions, interactions, speech segments, etc. If this set proves insufficient during bot construction, it is possible but difficult to augment this set at a later date by filming additional clips, as long as the same actor/actress, setting, clothing, etc. are used. However, when it comes to constructing the bot, the set of assets is fixed, and it is only those assets used during the bot construction process that can be used to animate the playback and interactivity aspects of the VideoBot.

This fixed-asset aspect of a VideoBot limits the set of behaviors and utterances that can be used to construct believable behaviors and credible dialogs on the part of the bot. Because original speech (encoded in digital audio files) is used instead of text-to-speech synthesis, if there is no relevant audio response that can be used by the VideoBot’s Dialog Manager, it is impossible to generate one on the fly, as can be done with bots that utilize speech synthesis. Thus it is possible for the user to make a query or comment to the bot for which the bot has no ability to audibly reply in a highly meaningful way. In such events it is not so much a matter of not knowing what to say as it is of not having a pre-recorded audio clip to utilize.

This problem is somewhat related to that faced by text-only chatterbots, including well-known chatterbots such as Eliza [Weizenbaum], Julia [Mauldin], and Hex [Hutchens]. These chatterbots resort to several “tricks” when the program cannot find a suitable pattern match with which to base a canned response. One of the most widely used tricks is creating a reply that includes part of

the user’s input, thereby pretending to have understood the user’s input. Other examples include changing the subject, asking for more information, and — for more brazen chatterbots — questioning the user’s intelligence (e.g., Hutchens’ Hex bot). While the use of such tricks has been resoundingly criticized [Shieber], in practice they have proven highly successful in inducing credible dialogs.

7.1 Non-committal Audio/Video Sequences

We have attempted to adapt several of these “tricks” to the VideoBot approach by utilizing several audio/video production and postproduction techniques. The simplest of these is prerecording a set of generic, non-committal responses, such as “OK”, “I certainly will (can) do that”, “That won’t be a problem,” etc. Additionally, we film each of these responses under different emotional states assumed by the actor (e.g., bored, reluctant, happy, obedient). Both the audio and the video reflect the bot’s emotional state since the actor/actress changes facial expressions, voice characteristics, and physical gestures throughout. This allows us to easily and naturally exploit the wide range of emotion-laden voicings possible in human speech. The more of these emotional, non-committal responses there are, the more credible we can make the dialog, as the VideoBot’s Dialog Manager can control how frequently this type of reply is used and which specific replies are given and when.

7.2 Diversionsary Clips

Another “trick” we use is prerecording multiple audio responses (similar to those above) but overlaying them on different video clips. This way, if there are n non-committal audio clips and m non-committal video clips, we can present $n \times m$ different combinations of these particular audio-video clips, thereby providing an even richer set of responses to the user to cover for the fact that the bot cannot directly reply with a highly meaningful, specific response.

This technique presents another problem, though, in that the audio will not be lip-synched to the video. To solve this problem, we can resort to another set of “tricks” that make lip-synching unnecessary. For example, one method is to have the bot essentially conceal or cover its mouth during the comment by scratching its chin during filming, placing its hand over its mouth as if in deep contemplation, etc. Alternatively, the bot can turn around or to the side before beginning to talk, as if to retrieve a piece of paper with notes on it, thereby occluding its face during the time the audio clip is played back. Given a set of such “diversionary” video clips (where the bot’s lips are not clearly visible or prominent), it is generally possible to overlay any desired audio clip to create a seemingly appropriate (yet diversionary) response by the bot, thereby maintaining credibility of the dialog.

7.3 Time-Stretched Video Clips

Apple's QuickTime architecture allows the choice of video clip and audio clip to be determined dynamically, at playback time. The only precaution that must be taken is to time-stretch the video to match the time duration of the audio when necessary. Whereas time-stretching the audio clip can easily lead to readily observable audio problems, including raising or lowering the pitch of the bot's voice, time-stretching the video rarely leads to observable problems, since the video clips in question are relatively short and the audible experience compensates for any visual miscues.

7. Directed vs. Non-Directed Utterances

Another significant problem is the need to be able to differentiate utterances that are directed to the VideoBot from those that are not. Examples of the latter can include the user speaking to himself or to someone else in the room, or audio emanating from a TV or radio in the same room as the user. Failure to adequately differentiate between directed and undirected utterances can quickly lead to frustration on the part of the user, error states within the bot, or out-of-sequence dialogs between the bot and the user. This latter occurrence would be a dead give-away that the bot was non-human, and thus believability and/or suspension of disbelief would be quickly shattered.

Currently, we are following the simple escape mechanism supplied with Apple's Speech Recognition Kit of calling out the bot's name before each utterance, or at least before any utterance that has been more than 15 seconds since the last utterance addressed to the bot. While this works to a high degree of satisfaction, it can occasionally feel tedious. Further, the arbitrary 15-second limit is not always easily judged by the user; if the limit has expired without the user realizing it, she/he may speak to the bot, without including it's name, only to have the bot ignore the user due to the time limit expiration. When and if the user realizes that the bot has failed to hear and/or respond, the user will have to repeat the statement, but this time, speaking the bot's name.

To help overcome this problem, there are several additional "tricks" we can employ. The simplest would be to resort to some sort of command-and-control interface (such as a graphical dashboard); there could be a light that glows green when the bot is listening and red when it is not (just for an example). While perhaps suitable for graphically rendered, artificial bots, this approach would quickly destroy any suspension of disbelief that a VideoBot was a real human. (We do, however, safely resort to this solution when the bot has receded into the background and remains non-visible, e.g., while a television program is playing, the bot has "gone away", but the bot is still able to listen and reply to spoken commands.) In this case, the psychological reaction is similar to that of having a green light glow on a telephone receiver while the user is connected to a party on the other end of the line.

For another approach, we have the VideoBot turn its attention away from the screen (i.e., look away, turn to the side, start reading a book, etc.). Then, when the user next wants to speak with the bot, the user can see that the bot is not paying attention and is not awaiting a command. Having to call out the bot's name before the rest of the utterance, then, feels more natural and less tedious. If the bot is seen to be still looking straight at the screen, the user can assume the bot is still listening, and so the bot's name can be omitted from the utterance [ASRK].

We also are experimenting with various control devices that the user is required to trigger before each utterance; examples of these include keyed-microphones (as used in walkie-talkies and CB radios) that are embedded into typical consumer electronic remote controls, or wearable versions, such as the communicators used in the Star Trek television series. One experiment that embedded a voice transmitter in the TV remote control has proven quite successful so far. However, the approaches that will prove most natural and successful and lead to easier suspension of disbelief remains to be seen.

8. Face-to-Face vs. In-the-Same-Room

Most conversational bot paradigms adopt either a "faceless" approach or a direct, "face-to-face" conversational approach, even when the conversation is achieved using only text. The result is a graphical persona that appears on the computer screen and seems to be talking to and listening to the user with whom it is carrying on a conversation. We have found that this leads to the psychological tendency to converse with and manage the discourse expectations in the same way one would with a real human that was physically present with the user, and in the same room. Because these psychological expectations are critical to achieving and maintaining believability and suspension of disbelief [Nass, Loyall], we believe it may prove advantageous to introduce simulated, artificial distance between the bot and the user, in order to encounter fewer chances of "blowing it" with respect to maintaining credibility in the dialog.

Clearly, supporting face-to-face conversation does not require creating the impression of being in the same room; videophones, for example, do not give this impression. Consequently, we are experimenting with a variety of distance-inducing visual paradigms to explore their effectiveness in increasing the credibility of the dialog. For example, one approach we are considering would have the VideoBot shown at all times using a telephone on its end to converse with the user. Only when the bot has the phone to its ear can it hear and respond to the user. When it has been given a task to perform, it can temporarily lay the phone down on the desk (or hang up, depending), perform the task, and then pick the phone back up to report on the results or to receive further instructions. The effect is similar to that seen in movies when speaking to a prisoner behind a glass screen but with a phone. To heighten the illusion of distance, elevating the camera and

angling down on the human actor during filming should prove useful.

We hope to additionally experiment with a setting similar to a remote control station (e.g., an airplane cockpit) but with a prominently displayed microphone front and center that is required to be keyed by the bot before he/she can listen or talk. Both these models also benefit from being able to use the communication device (telephone or microphone) to obscure the actor's mouth at key points, to assist with artificial voice-overs.

9. Barge-Ins, Overlapped Speech, and Interruptions by the User

Experiments with users have shown that they frequently begin saying something new to the bot while the bot is busy performing some task or while in the process of replying to a previous input from the user. In general, this poses no problem, for the speech input processor runs asynchronously with the audio playback routines. The system can thus capture and pre-process the new utterance while the bot continues to speak or act.

If these utterances were simply stacked requests or commands by the user that added to a previous utterance (e.g., "Oh yeah, and please set the lighting to theater style."), the problem would be relatively simple to handle from a "believability" point of view. In the example just given, no audible reply is even necessary, and the bot can simply issue the relevant control commands to set the lighting appropriately. However, since the interruptions can be commands to cancel a previous command or to change some part of a previous command (e.g., "Sorry, I meant channel 5, not 9!" or "Never mind, I'd rather watch the news."), it would certainly diminish believability if the bot continued responding to the current utterance and gave no indication that it had understood, much less even heard, the newer utterance until it had completed responding to the previous utterance.

Handling these types of interruptions requires that the speech recognition manager be able to interrupt the system's audio and video playback manager and request a check to see if the user has issued a "contradiction". If not, the new utterance is stacked for subsequent processing, and the current video and audio segment are allowed to play to completion. Then and only then is the stacked utterance fully processed, and the appropriate audio and video clips played back in response to that utterance.

More difficult problems arise if the new utterance is a contradiction or cancellation-type command, such as "Never mind, Buster." In this event, the currently playing audio/video segment must be aborted and a new one — one that presents a response consistent with the contradictory utterance — must be initiated.

To maintain believability, the bot's responsiveness to interruptions must be limited to less than 2 seconds — the less the better. Since many of the bot's video clips or clip sequences will be significantly longer than 2 seconds, it becomes necessary to create artificial segues between the

playing video clip and the appropriate response clip. We overlay prerecorded audio clips (e.g., "Huh? Oh, sure.") with these very short video segues and then start the prerecorded response clip appropriate to the interruption. Creating all these artificial segues can be one of the most tedious parts of producing a VideoBot's assets, yet their utility in maintaining credible dialogs is undeniable.

10. Personalizing the Dialog

As part of the VideoBot construction process, the human subject is filmed while performing a set of scripted actions, making a variety of pre-scripted responses with a prespecified set of emotions [DeGroot]. During this process, the human is likely to have no idea to whom he or she will be speaking. Thus all utterances directed to the listener must either be nameless (e.g., "Sure," rather than, "Sure, Bob.") or generic (e.g., "Yes, Ma'am" or "10-4, Good Buddy.")

Under certain circumstances, this type of response may actually prove more desirable than personalized responses. However, to ensure that both are provided, we can record the human subject speaking a number of names and then insert these spoken-name audio segments into the fuller audio clip. For example, changing "Sure thing, I'd be happy to do that," into "Sure thing, I'd be happy to do that, *Sally*," is a simple matter of appending the audio clip for "Sally" onto the end of the nameless response clip (similar to techniques used in Interactive Voice Response (IVR) systems). Since all audio clips are dynamically appended to each other during execution of the VideoBot, this is easily performed, although at a slight performance cost. (Alternatively, the audio clip can be preprocessed to have the name segment permanently pasted to the end of the response clip. This would be most appropriate when the number of users is expected to be small.)

Inserting prerecorded names into audio clips, such as changing, "Sure thing, I'd be happy to do that," into "Sure thing, *Sally*, I'd be happy to do that," is slightly more complicated. Doing this requires cutting the response audio clip right after the pause following "thing", inserting the clip for "Sally", and then appending the remainder of the original audio clip. Unlike the previous solution, this modification is not one that easily lends itself to dynamic execution, and it would thus be best performed during preprocessing.

In either case, the editing operations are fairly simple, and they can easily be mechanized with asset processing scripts, thereby providing inexpensive audio personalization of a large number of audio/video clips.

While both of these solutions work fine for the audio, they both create a separate problem for the video clips. Consider that the VideoBot will be shown talking to the user from a variety of positions (e.g., straight on, turned to the side, head down) and during a variety of acts (straightforward looking, scratching it's head, eye's rolling, and the like).

It would clearly prove impractical to film each of the many names being said from all the required camera positions and action sequences in order to provide corresponding video that could be inserted along with the inserted, prerecorded audio. Accordingly, there are no video frames to accompany the inserted audio, and hence the additional video must be synthesized. Personalizing an audio/video clip this way requires replicating some single, relevant video frame a sufficient number of times to cover the length of time added by the additional audio clip. This is usually the last frame of the video associated with the end of the first part of the audio clip. The time duration of the inserted audio clip is usually one second or less, and thus only 30 or so frame replications are required.

It should be noted though, that this approach to personalization does not provide actual, realistic video that is lip-synched to the audio. Instead, the video sequence is artificially extended with static video to cover the extra audio. We have not yet been able to determine the degree to which believability is jeopardized by failing to provide for synthesized lip-synching in the video while playing back the audio for the user's name, but we expect (hope?) that the disruption to "reality" will be nominal.

11. Conclusions

Just as various dialog "tricks" have been shown to be effective in heightening believability of chatterbots [Loebner], we have found that similar types of tricks can also be deployed in a VideoBot to heighten believability. But even more, they have proven successful in helping maintain the illusion on the part of a user that the user is actually engaged in a conversation with a real human, and often, with a specific human. This paper has described several issues and tricks that can be employed in a VideoBot to help maintain a credible dialog with a human user.

Once it becomes practical to create speech-synthesis engines that accurately recreate a given human's voice, the need for and value of some of these tricks will be partially reduced. For example, it will be possible then to dynamically create utterances that are both 1) highly relevant responses and 2) believably the voice of the specific human on which the bot is based. This will make possible chatterbots that use both speech recognition and speech synthesis that are thoroughly believable. Even then, though, the newly created audio clips will have no actual video clips to accompany them, and even if an existing video clip is used, the audio will not be lip-synched to the video.

It is tempting to assume that the solution would be to leap to fully synthesized, graphical characters, wherein the bot system could presumably create both dynamic video and accurately synthesized speech together, providing the necessary facial expressions and lip-synching, as in [Thalman]. This leads to the challenge of having to create the character, its movements, facial expressions, visual setting, and the like in a fully believable manner, if the

user is to feel as if she/he is really talking to a specific, real human. This approach requires crossing the chasm from *disbelief* to *belief*; we believe the VideoBot paradigm more easily begins with *belief* and crosses over to *disbelief* only when the audio/video playback and dialog fail to *sustain* believability; this presents entirely different challenges to the designer. We therefore believe that certain of the tricks presented herein will likely remain valuable.

12. Bibliography

- [AgentLand] One of the more recent sites dedicated to agents and bots; this one is based in France; <http://www.agentland.com>.
- [Alice] The A.L.I.C.E. AI Foundation, <http://www.alicebot.org>.
- [Allen] "An architecture for more realistic conversational systems," James Allen, George Ferguson and Amanda Stent, *Proceedings on the International Conference on Intelligent User Interfaces*, Jan. 14-17, 2001, Santa Fe, NM USA, ACM, New York, pp. 1-8.
- [Ananova] "Ananova, a face for the web," D. Ian Hooper, CNN.com, Jan. 18, 2000.
- [ASRK] "The Speech Recognition Manager Revealed," Matt Pallakoff and Arlo Reeves, *develop: The Apple Technical Journal*, Issue 27, September 1996, pp. 6-12.
- [BotSpot] One of the oldest and largest collections of items related to bots; <http://www.botspot.com>.
- [DeGroot] "An Architecture for the Support of Interactive Software VideoBots", in preparation, 2001; please send email if interested.
- [Hayes-Roth] "Staffing the Web with Interactive Characters: Intelligent agents in the form of personable characters interact with and serve human customers," Barbara Hayes-Roth, Vaughan Johnson, Robert van Gent, and Keith Wescourt, *Comm. of the ACM*, Vol. 43, No. 3, March, 1999, pp. 103-105.
- [Hutchens] "How to Pass the Turing Test by Cheating," Jason Hutchens, Centre for Intelligent Information Processing Systems, Dept. of E & EE, The University of Western Australia, TR97-05, December, 1996.
- [Laven] The Simon Laven Page: Chatterbot Central, Simon Laven, <http://www.simonlaven.com>.
- [Leonard] *Bots: The Origin of New Species*, Andre Leonard, Hardwired, 1997.
- [Loebner] "In Response," Hugh Gene Loebner, *Comm. of the ACM*, Vol. 37, No. 6, 1994, pp. 79-82. Also see, "Home page of the Loebner Prize - 'The First Turing Test,'" <http://www.loebner.net/Prize/loebner-prize.html>.

- [**Loyall**] *Believable Agents: Building Interactive Personalities*, A. Bryan Loyall, Ph.D. dissertation, Carnegie Mellon, CS Dept., CMU-CS-97-123, May, 1997.
- [**Lucente**] "Conversational interfaces for e-commerce applications," Mark Lucente, *Comm. of the ACM*, Vol. 43, 2000, pp. 59-61.
- [**Maes**] "Artificial Life Meets Entertainment: Lifelike Autonomous Agents," Pattie Maes, *Comm. of the ACM*, Vol. 58, No. 11, Nov. 1995, pp. 108-114.
- [**Massaro**] *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, Dominic W. Massaro, MIT Press/Bradford Books Series in Cognitive Psychology, 1997.
- [**Mauldin**] "Chatterbots, Tonymuds, and the Turing Test: Entering the Loebner Prize Competition," Michael Mauldin, *Proceedings of AAAI-94*.
- [**MSAgent**] *Developing for Microsoft Agent: Microsoft ActiveX Technology for Interactive Characters*, Microsoft Press, Redmond, Washington, 1998.
- [**Nass**] "Speech interfaces from an evolutionary perspective," Clifford Nass and Li Gong, *Comm. of the ACM*, Volume 43, 2000, pp. 36 - 43
- [**Shieber**] "Lessons from a restricted Turing test," Stuart M. Shieber, *Communications of the ACM*, Volume 37, Issue 6, 1994, pp. 70-78.
- [**Stern**] *QuickTime 5 for Macintosh and Windows: Visual Quick Start Guide*, Judith Stern, Robert Lettieri, Peachpit Press, 2001.
- [**Suereth**] *Developing Natural Language Interfaces: Processing Human Conversations* (with CD-ROM), Russell Suereth, McGraw-Hill, 1997.
- [**Thalman**] *Synthetic Actors in Computer Generated 3D Films*, Nadia Magnenat Thalmann and Daniel Thalmann, Springer Verlag, 1990.
- [**Weizenbaum**] *Computer Power and Human Reason*, J. Weizenbaum, W.H. Freeman and Co., New York, 1976.