A Toolkit for Exploring the Role of Voice in Human-Robot Interaction

Zachary Henkel
Center for Robot-Assisted Search and Rescue
Texas A&M University
College Station, Texas, USA
zmhenkel@tamu.edu

Vasant Srinivasan
Center for Robot-Assisted Search and Rescue
Texas A&M University
College Station, Texas, USA
vasant.s@gmail.com

Robin R. Murphy
Center for Robot-Assisted Search and Rescue
Texas A&M University
College Station, Texas, USA
murphy@cse.tamu.edu

Victoria Groom
CHIME Lab
Stanford University
Stanford, CA, USA
vgroom@stanford.edu

Clifford Nass
CHIME Lab
Stanford University
Stanford, CA, USA
nass@stanford.edu

ABSTRACT

This paper describes an open source speech translator toolkit created as part of the "Survivor Buddy" project which allows written or spoken word from multiple independent controllers to be translated into either a *single synthetic voice*, *synthetic voices for each controller*, or *unchanged natural voice of each controller*. The human controllers can work over the internet or be physically co-located with the Survivor Buddy. The toolkit is expected to be of use for exploring voice in general human-robot interaction.

Categories and Subject Descriptors

I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

General Terms

Human Factors, Experimentation

1. INTRODUCTION

The Survivor Buddy project is motivated by our prior work which suggests that a dependent victim will treat the robot as a *social medium*, that is, the robot will be both a medium to the "outside" world and a local, independent entity devoted to the victim (e.g., a buddy). One function of the medium is to provide two-way audio communication between the survivor and the emergency response personnel, but more interesting capabilities emerge by fully exploiting web applications. For example, responders could play therapeutic music with a beat designed to regulate heartbeats or breathing or the dependent could make requests- consider that two trapped Australian miners requested MP3 players with a Foo Fighters Album while waiting for rescue [3] . A web-enabled robot with a LCD screen could permit the survivor to videoconference with responders (or family), watch live TV (e.g., to get validation of the magnitude of the disaster or to be entertained), movies, or listen to music. The idea is that a web-enabled, multi-media robot allows i) the survivor to take some control over the situation and find a soothing activity while waiting for extrication and ii) responders to support and influence the state of mind of the victim.

Figure 1: View from the Survivor Buddy webcam showing test subjects looking at the robot despite the voice originating from a different location. Also a top view of the room in the sub-picture.

The potential significance of choice of robotic voice in a social medium was highlighted by our recent experiences with "SciGirls." Survivor Buddy was used by four middle-school girls to illustrate the scientific process for an episode of "SciGirls," a Public Broadcast System science reality show. The SciGirls worked with the robot for a week. They hypothesized that an extroverted Survivor Buddy would be better liked than introverted Survivor Buddy, created intro- and extroverted versions of affective behaviors and voice, then tested their hypothesis by having 12 of their friends play tic-tac-toe with the robot. Two of the SciGirls used the text to speech interface in the Center for Spoken Language (CSLU) Toolkit [1] to generate two synthetic voices for the

HRI'11, March 6–9, 2011, Lausanne, Switzerland.
ACM 978-1-4503-0561-7/11/03.

robot with the pitches and frequencies associated with introversion and extroversion.

An examination of the sessions captured by the Survivor Buddy's webcam showed that the SciGirls' friends fixated on the stationary robot head rather than the external loud speakers, with the exception of the first group which could see the SciGirls. While the SciGirl experimentation was limited, it does raise the question *is voice more important than non-verbal cues for certain situations?* If voice is more important than nonverbal cues, it may be a more cost-effective way to make existing non-anthropomorphic robots socially consistent or to reduce costs in new robots. Fig. 1 shows one group clearly looking at robot but as can be seen from the top view this is almost 90° from the actual source of the voice. The fixation on the robot head versus the source of the voice is not surprising given the primacy of the face over other communication channels in coordinating the give-and-take of conversation [2]. That the friends liked the extroverted voice better is also consistent with the literature, as extroverts are perceived as more sociable, warm, and engaging[4]. But the relative contribution of voice versus nonverbal cues remains unclear. On one hand, the information being conveyed in game playing may be simpler and the environmental conditions favorable. Less structured conditions such as a disaster with background noise and high stress may require non-verbal affect to amplify the intent of the agent.

2. SPEECH TRANSLATION TOOLKIT

In order to explore questions about the type of voice for a social medium that will minimize or eliminate distrust, dislike or cognitive dissonance, we have built and continue to expand a speech translation toolkit. The speech translation toolkit is a framework for extending already existing voice recognition and text to speech systems into the realm of rescue robotics. By wrapping recognition and generation abilities as network services, we allow users to choose the best tools available, regardless of their platform or installed libraries. For example, our base implementation provides users with recognition services based on Microsoft's Speech APIs and generation services based on FreeTTS or Apple's built in operating system voices. The toolkit is also able to use recognition technologies like CMU's Sphinx 4 or the Center for Spoken Language (CSLU) Toolkit [1], provided a network wrapper is written for these packages. The network wrapper allows connected clients to input not only raw data for translation, but also tuning parameters, like speed or pitch; this affords the controller more options and techniques for communication.

The toolkit is simple to deploy, using peer-to-peer communication between the involved parties and access to a central server for a distribution list of connected clients and services. The client software, written in Java, is divided into three distinct clients distinguished by their roles, is platform independent and easily distributable. It has the following functions:

Natural Voice Pass through between all parties connected: the number of connected parties is limited by bandwidth, usually up to three communicate easily.

Voice alteration via real time effects on voice stream: because the toolkit provides byte level access to the audio streams, we are able to employ digital signal processing techniques to control attributes like reverb or echo.

Voice standardization via voice recognition then synthesis: a controller can choose a synthetic voice to use and then speak in that voice via a speech recognition process followed by a speech generation process.

Text To Speech with many choices for voices and options: for example the FreeTTS package provides the voices kevin16, mbrola1, mbrola2, and mbrola3, as well the option to create voices with FestVox. Other packages, as simple as built in operating system voices, provide varying voices as well. Packages like the CSLU Toolkit, which allows tweaking of voice parameters to create new and unique voices, may also be supported through the tunning parameters feature.

The three software clients are distinguished by the roles they play in the disaster situation. One client is managed remotely and supports simple listening and responding. This client runs on the rescue robot itself, and is the interface that the victim encounters. It requires no interaction from the victim, other than speaking and listening. The second interface, designed for remote guests, allows a remote user to communicate with facilitators and the victim, at controlled intervals. This client is ideal for a remote expert, or family members of the victim. The victim and the guest are able to communicate, but only as allowed by the facilitator. The third client, the facilitator client, acts as a gatekeeper between the other clients, as well as participates in the communications. The facilitator client allows management of which clients can speak with the victim and who can login to the system in general. Additionally, the facilitator client is able to decide who may access specific data on the central server. For example, the central server may collect and distribute information about the robot's proximity to a victim. This data may be relevant and useful to a facilitator, but not to a guest.

3. CURRENT AND FUTURE WORK

The toolkit will be used in experiments starting in Spring 2011 which will explore the appropriate identity (loyalty, expression, locus of control) and type of voice (controller, synthetic for each controller, single synthetic) for the robot serving as a social medium. The toolkit will be part of a larger Social Medium toolkit that will go beyond speech translation and enable the exploration of role identity mapping and collision resolution between multiple controllers. The expanded toolkit will also permit the association of scripts, policies, priorities, and decision-making aids with different controller roles. Code is available upon request.

Acknowledgments

This work was supported in part by NSF Grant IIS-0905485 "The Social Medium is the Message" and by a HRI grant from Microsoft External Research.

4. REFERENCES

[1] Center for spoken language (cslu) toolkit. WEB, 2009. http://cslu.cse.ogi.edu/toolkit/index.html.

[2] M. Knapp and J. Hall. *Nonverbal Communication in Human Interaction*. Holt, Rinehart, Winston, 1978.

[3] B. I. News. Dave grohl to meet trapped miners http://news.bbc.co.uk/2/hi/entertainment/5301658.stm, Aug. 2006.

[4] D. Watson and L. A. Clark. Handbook of personality psychology, 1997.