Services over VoIP for Robot Companion

Thierry Simonnet[1] and Gérard Chollet[2]

[1] ESIEE-Paris,
2 boulevard Blaise Pascal, Cité DESCARTES,BP 99,
93162 Noisy le Grand cedex, France
t.simonnet@esiee.fr
http://www.esiee.fr/
[2] CNRS-LTCI, TELECOM-ParisTech,
46 rue Barrault, 75634 Paris cedex 13, France
{gerard.chollet}@telecom-paristech.fr,
http://www.telecom-paristech.fr/

Abstract. An increasing number of people are in need of help at home (elderly, isolated and/or disabled persons, people with mild cognitive impairment). Several solutions can be considered to provide tele-care to these people and to maintain them as long as possible at home. In recent years, we are witnessing a gradual migration of traditional telephone service network i.e. Public Switched Telephone Network (PSTN) world to the Internet Voice over Internet Protocol (VoIP). This phenomenon is playing a major role in telecommunication developments due to its advantages, operational infrastructure simplification and significant reduction in the cost of communication services that are linked to them. The likelihood of implementing a Voice and Video over IP solution will be helpful for elderly people to have connection with their families, Robot-Companion and medical professionals especially in the field of trust relationship concerning their health condition. This paper describes how VoIP solution can be used to advance communication process by means of controlling a robot machine and especially with the use of ASR system like Julius. *abstract* environment.

Keywords: VoIP, ASR, Robot Companion

1 Introduction

European Scientists have found three new major genetic links to pre-Alzheimer, affecting up to 20% of people with brain-wasting disease and it was the most significant such discovery in 15 years. Alzheimer disease affects more than 26 million people globally and it has no cure with any good treatment and the need for effective remedies is pressing on, with the number of cases estimated to go beyond 100 million by 2050 [7]. In Europe, there is also an increasing demand for maintaining dependent people at home, to reduce hospital load, improve their quality of life and strengthen their social link [6], [16]. To this extent, a need for communication systems and telecare technologies arose. Maintaining such

people at home often requires medical assistance, excellent and reliable communication tools, used by their relatives and the caregivers.As part of ongoing projects, research has been conducted towards efficient solutions for audio/video communication between people and system control, using an unified channel. All these projects have a common objective, to support the elderly in daily life by integrating the existing technologies for managing and interacting, mostly using speech recognition [5], with their domestic ambient environment in order to increase their autonomy, safety and improve their quality life.

This article is organized as follows. We present environment, platform overview in Section II, technicals descriptions in Sections III, IV and V. A detail of technical integration is given in Section V before concluding in Section VII.

2 Environment, platform overview

The current platform is composed of three parts: a master server, a smart home and a remote client. A master server, handling:

- Lightweight Directory Access Protocol (LDAP) server (registry and authentication).
- Mail server.
- Collaborative environment based on Horde Groupware.
- Asterisk Internet Protocol Private Branch eXchange (IP-PBX, or IPBX) for voice/video communications routing.
- Julius ASR server (can be hosted by another server)

A Smart Home, equipped with:

- A robot companion featuring a camera, a display and a VoIP client.
- Various sensors for person monitoring.
- Internet gateway (local IPBX).

A remote client system, basically a Personal Computer with

- A web browser.
- A VoIP Client

3 VoIP architecture and services

3.1 A Unified and standardized communication solution

We need to address different kind of media for different equipments. A VoIP solution offers a complete and unified communication infrastructure and then we can use it to develop various services. This infrastructure can be use for a closed group of users but also to be plugged on public VoIP network. This solutions has all criteria for medical/paramedical usage :

- Supports various Internet infrastructure (public IP, private IP, ADSL box...)

- Interoperability with public and private (ekiga.net, google talk, skype...) telecommunication networks
- Low latency (less than100ms with H263 video) mandatory for remote control (robot, home automation)
- Automatic internet bandwidth adjustment.
- Single solution for videoconferencing, robot relay and the Smart Home control.
- Support for various clients (softphones, IP phones, mobile phones, specialized softphones for remote control...)
- Choice of audio and video codecs
- Communication robustness
- Compatibility with IPBX call centers
- Ability to set up centralized services (low cost of deployment) as IVR, ASR, multi-conferencing, voice and video messaging
- Unique identifier (phone number).
- Centralization of data (voice, video)
- Internationalization with customization of the language user.

3.2 Communication infrastructure

Internet is the main communication media for the project. VoIP solutions imply the use of a PBX. We will use Asterisk PBX from DIGIUM Company [8], [3] for the first version of the infrastructure. Other products can be used like Kamailio (OpenSER) or IPXSecs. Asterisk PBX has standard configuration for classical communications but needs new and modified communication module for our purposes. Patient network will use private IP addresses, then it will be necessary to have a local PBX to manage local communications and to act as a gateway to make or receive a call from public or private domains.

When a call is started, a SIP request is sent to the PBX, which transmit it to the other client. When this signalling communication is done a direct one is established using RTP (Real Time protocol) (See Fig 1). This protocol, over UDP, keeps the packet order and drops old ones. To establish a communication between 2 private networks, it is necessary to use trunking services to allow all communications use a path through PBXs.

Codecs A codec (Code-DECode) is a module that can Code and DECode an analog or a digital signal. For VoIP codec is used for norm but also for the module itself. X264 codec code and decode streams that use MPEG-4 AVC/H264 norm. PBXs are not designed for stream translation. A direct RTP communication is set between 2 clients and then clients must have compatible codecs that respect norms.

Asterisk can handle for signalling purposes:

- Voice : law, alaw, gsm, ilbc, speex, g726, adpcm, lpc10, g729, g723
- Video : h261 [10], h263 [11], h263+, h264 [12]

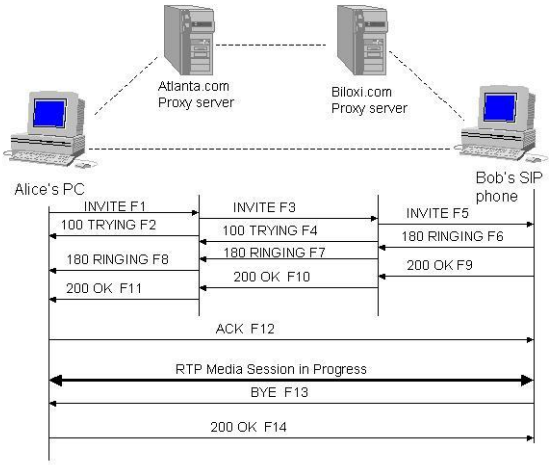

Fig. 1. Call dialog

For the project, we can use: ulaw, alaw, speex (8KHz narrowbad and 16KHz wideband) for voice and H261, H263 and H264 for video. The key point is using a well balanced setup between "compression", "delay" and "video quality". Increasing compression will also increase delay due to buffer use.

Alarms It is possible to transmit alarms using SIP MESSAGE method (see Table 1 for SIP Methods). Asterisk doesn't fully handle this method and it was necessary to implement properly RFC 3428 [19] for SIP channel. It is also possible to use T.140 (RFC 4103) [20] method for Instant Messaging/Alarms communications. Both solutions are implemented.

3.3 Services

Central PBX can be interconnected to a call center with economical and security features:

- Direct connection with an auto-connected client and using specific dial number.
- Security in the possibility of encryption (VPN or stream encryption ...), OSP, closed group of subscribers for the confidentiality. Management by phone number and not by identity.
- Dialplan will transfer any specific local calls to call centre one.
- Depending on partners' needs, we will develop dedicated modules for PBX (MP4, Interactive Voice/Video Response - IVR - solution, RTSP, Speech to text - ASR...).

SIP Method	Description	RFC
ACK	Acknowledge final response to Invite	3261
BYE	Terminate a session	3261
CANCEL	Cancel a previous	3261
INFO	Mid-session signaling	2976
INVITE	Initiate a session	3261
MESSAGE	Allows the transfer of IMs	3428
NOTIFY	Event notification	3265
OPTIONS	Query to find the capabilities	3261
PRACK	Acknowledgement for Provisional responses	3262
PUBLISH	Publish event state	3903
REFER	Transfer user to a 3rd party	3515
REGISTER	Register with a SIP network	3261
SUBSCRIBE	Request asynchronous event notification	3265
UPDATE	Update parameters of a session	3311

Table 1. SIP Methods

- No data duplication
- Ability to centralize all the data for exploitation or study purposes.
- Patient home PBX is setup on a Plug computer (SheevaPlug for example) for patient home use.
- A PBX module will be developed to handle a Speech to Text tool. This will allow when needed a direct transcription of calls for medical use. This will also allow to handle voice orders. Such a centralized ASR will handle multi-language tools and avoid unitary installation. A great exploitation benefit.

4 Robot Control

For some applications, such as telecare, it is necessary to pilot a robot companion usng the video stream it provides. Several options were evaluated based on the following environmental constraints: Commands sent to the robot must be synchronized with the video feed to account for latency and driving the robot must be a straightforward operation. The first constraint quickly prompted toward sending robot commands over the same carrier as the voice and video were. Two primary options were studied:

4.1 Sending commands over the voice channel

This could be achieved with e.g. inband Dual-Tone Multi-Frequency (DTMF). Synchronisation would be guaranteed (the VoIP client handles synchronisation between voice/video feeds and between each ends feeds). This would have had the drawback of introducing a new complexity as the audio channel would have had to be processed in order to filter out DTMF command signals and translate

them to the robot controls. Also, this approach would rely on good audio quality, and introduced some latency (duration of the DTMF code and processing). Also, a loss of audio signal (in case of lag for instance) would also mean a loss of control over the robot, which might be a security risk.

4.2 Sending commands over a third channel

A simpler approach was then devised: instead of encapsulating commands into the voice channel, we thought about using a dedicated text channel. The reasons for this are numerous:

- No more signal processing required.
- Does not rely on audio channel (more resilient).
- Groundwork for the text channel already exist in Ekiga [15], and simple to implement in Asterisk.
- No latency: complexity of commands virtually unlimited (any amount of text is sent almost instantaneously).
- Assigning missions to the robot, or even completely reprogramming it, could be done this way.

5 Remote automatic speech recognition interface

Speech is probably the most natural way that human beings employ to communicate between themselves, also being one of the most impressive system interfaces for human-computer interaction. The use of Automatic Speech Recognition (ASR) technologies becomes even more interesting when applied to the case of users who are not familiar with (or are physically/mentally unable to manage) the traditional computer interfaces. The potential of ASR Research englobes from vocal commands to complex dialog systems, capable to identify one's state of mind and to detect distress situations.

" "Julius" is a high-performance, two-pass large vocabulary continuous speech recognition (LVCSR) decoder software for speech-related researchers and developers. Based on word N-gram and context-dependent HMM, it can perform almost real-time decoding on most current PCs in 60k word dictation task. Major search techniques are fully incorporated such as tree lexicon, N-gram factoring, cross-word context dependency handling, enveloped beam search, Gaussian pruning, Gaussian selection, etc. Besides search efficiency, it is also modularized carefully to be independent from model structures, and various HMM types are supported such as shared-state triphones and tied-mixture models, with any number of mixtures, states, or phones. Standard formats are adopted to cope with other free modeling toolkit such as HTK, CMU-Cam SLM toolkit, etc." [13]

Julius in combination with Asterisk is used for different purposes. First one is to act like prompter to get acoustic models for different languages (dutch, french, spanish), using the VoIP infrastructure with its limitation (codecs act

as filters, narrowband/wideband...). It is then possible to have a live learning processus for each language. A phone number is setup for each combination of language and codec configuration. The second use is the most basic one : speech controlled IVR (Instant Voice Response) for disabled of to avoid key press action for elderly people. The third use is speech driven remote control for robot mainly, but also for smart home equipment.

6 ASR and VoIP

With such a centralized platform, ASR can be centralized using phone technologies facilities. Asterisk provides various speech tools but no embedded ASR tool. A proper commercial model, trained and annotaded by professionals costs a lot. Open Source project Julius offers the services we need. Asterisk offers a a generic speech API that can be used with Julius. Julius can have input and output redirected to any socket. The aim was to have an Asterisk module that can manage Asterisk speech functions, usable in a dialplan. The dialplan API is based around a single speech utilities application file, which exports many applications to be used for speech recognition. These include an application to prepare for ASR, to activate a grammar and to play back a sound file while waiting for the person to speak.

We started with app_julius module [14] developed by Danijel Korzinek and Dikshit Thapar. Dialplan Flow:

1. Create an ASR object using SpeechCreate()
2. Activate your grammars using SpeechActivateGrammar(Grammar Name)
3. Call SpeechStart() to indicate you are going to do recognize speech immediately
4. Play back your audio and wait for recognition using SpeechBackground(Sound File|Timeout)
5. Check the results and do things based on them
6. Deactivate your grammars using SpeechDeactivateGrammar(Grammar Name)
7. Destroy your speech recognition object using SpeechDestroy()

A simple macro is used in the dialplan to confirm word recognition. ARG1 is equal to the file to play back after "I heard..." is played.

7 Conclusion

It is certainly great to achieve such flexible level of communication using open source softwares. Although the need to work towards the modelization of more robust acoustic models for ASR (in order to increase the recognition rates), all the needed infra-structure is currently available and ready to make progress towards the multiple kinds of applications, in its many types of contexts (e.g. telemedicine, security, vocal commands etc.) that it is capable to handle.

Acknowledgment

Some research leading to these results has received funding from the European Communitys seventh Framework Programme (FP7/2007-2013) under grant agreement n216487.

References

1. J. Apostolopoulos, "Reliable Video Communication over Lossy Packet Networks using Multiple State Encoding and Path Diversity", Visual Communications and Image Processing (VCIP), 392,409, 2001.
2. M. Armstrong, D. Flynn, M. Hammond, S. Jolly and R. Salmon, High Frame-Rate Television, BBC Research White Paper WHP 169, 2008.
3. Asterisk, The future of telephony, Jim van Meggelen, Leif Madsen, Jared Smith, O'Reilly
4. B. Burger, I. Ferran, and F. Lerasle, Multimodal Interaction Abilities for a Robot Companion, Conference on Computer Vision Systems (ICVS'08), Santorini, Greece, 2008.
5. G. Chollet,D. R. S. Caon, T. Simonnet, J. Boudy, vAssist: Le Majordome des personnes dépendantes. In: 2e Conférence Internationale Sur laccessibilité et les systèmes de suppléance aux personnes en situations de handicap, Paris - France (2011).
6. N. Clement, C. Tennant, C. Muwanga, Polytrauma in the elderly: predictors of the cause and time of death. Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine, v. 18, n. 1, p. 26, 2010. ISSN 1757-7241,http://www.sjtrem.com/content/18/1/26
7. Commission of the European Communities, 2009. The Health Report of European initiative on Alzheimers disease and other dementias. Brussels N 380.
8. Digium, The Open Source PBX & Telephony Platform, http://www.asterisk.org/.
9. G. Herlein et al., RTP Payload Format for the Speex Codec, draft-ietf-avt-rtp-speex-07, http://tools.ietf.org/html/draft-ietf-avt-rtp-speex-07, 2009.
10. International Telecommunication Union, "H.261: Video codec for audiovisual services at p x 64 kbit/s", Line Transmission of Non-Telephone Signals, 1993.
11. International Telecommunication Union, "H263: Video coding for low bit rate communication", SERIES H: Audiovisual and Multimedia Systems Infrastructure of audiovisual services, Coding of moving Video, 2005.
12. International Telecommunication Union, "H264: Advanced video coding for generic audiovisual services", SERIES H: Audiovisual and Multimedia Systems Infrastructure of audiovisual services, Coding of moving Video, 2003.
13. Julius ASR, http://julius.sourceforge.jp/en_index.php
14. D. Korzinek, module app_julius, http://forge.asterisk.org/gf/project/julius/
15. D. Sandras, Ekiga, http://ekiga.org.
16. World Health Organization, 2002, The European Health Report, European Series, #97.
17. Xiph.Org Foundation, "Speex: A Free Codec For Free Speech", http://speex.org/.
18. SIP protocol, RFC 3261, http://www.ietf.org/rfc/rfc3261.txt
19. UDP, RFC 3428, http://www.ietf.org/rfc/rfc3428.txt
20. T.140, RFC 4103, http://www.ietf.org/rfc/rfc4103.txt