Automated Audio-visual Dialogs over Internet to Assist Dependant People

Thierry Simonnet, Samuel Ben Hamou
R&D department
ESIEE-Paris
Noisy le Grand, France
{t.simonnet, s.benhamou}@esiee.fr

***Abstract*—With today's advancements in medical treatments and care fields, an increasing number of people need help or assistance at home. Concerned categories are mostly elderly, isolated or disabled persons even though persons with mild cognitive impairment are also considered. One of the main issues these categories are facing is the lack of constant communication to help maintaining some kind of social link, with families, friends and eventually caregivers. Moreover, the fact that current hardware accessories and technologies are not always suited to allow for such functionalities, in a generic situation, tend to increase that particular gap. Many studies suggest the use of Automatic Speech Recognition (ASR) to have a global control over devices and communication means. The resulting architecture allows for a pseudo butler to be put in place, serving as the main entry point to manage daily common communication tasks as well as giving access to different web services via a voice controlled environment. Once put in application, it could stand as a complete and integrated solution for remote communication and monitoring of the designated population target.**

Keywords-ASR; Voice over IP (VoIP).

I. INTRODUCTION

Our ongoing research projects are focusing on creating a unified solution/channel to be used for daily tasks management but also audio/video communications between people. The main reason for this unified solution comes from an increasing demand for maintaining dependent people at home [5][9]. In [20], the World Health Organization assessed the restructuration of hospital services, with an increased role for substitution between different levels of care, strengthening primary health care services, increasing patient choice and participation in health services and improving outcomes through technology assessment and quality development initiatives. According to these recommendations, the number of telesurveillance implementations and pilot experimentations has been growing in Europe, especially within the Northern countries.

To reduce financial costs, hospitals load but also improve the patients quality of life, it has been recently considered to keep them at home, thus a need for suitable communication and telecare technologies has arisen. For such a specific panel, we also often need medical assistance. This has a direct implication on the quality of services as we need reliable communication tools and a really easy learning curve for the end users. We are relying on IP technologies as Asymetric Digital subscriber Lines (ADSL [28]) are available all across Europe for affordable prices. It has a small drawback as upload bandwidth is generally limited and thus has to be taken into account for data/video communications. Also, considering that mobile networks are relying on the same technology for data exchange, IP solutions represent the obvious choice as they will probably be suited for future network evolutions. At the moment, in order to offer a good service, we have to focus on video and audio quality, which are directly linked to the available bandwidth and compression algorithms.

We will present the environment with a platform overview and an explanation of our choices in Section 2, while Section 3 will focus on technical descriptions. Integration will be covered in Section 4 and results in Section 5. The conclusion will focus on identifying current and remaining issues, but also put this in perspective.

II. REMOTE AUTOMATIC SPEECH RECOGNITION INTERFACE

A. Speech recognition

The most natural and obvious way for humans to interact and communicate nowadays is speech. Considering our end-users, who may not always be acquainted with traditional computer interfaces, it makes sense to focus on Automatic Speech Recognition technologies as the primary way of interaction with the system. It allows for vocal commands to be passed, but is also able to eventually identify mood states or for example detect particular/distress situation.

B. Usage scenarios

The system has to pose as a virtual butler with access to a centralized and collective memory database, with either audio and or visual representations. Some ways to interact with it are as follow:

- Find his way: the butler, as a service, can be used on a gps enabled smartphone allowing some guidance.
- Manage a diary, appointments, bills payment...
- Answer the phone, messaging, mail...
- Find information on the web.
- Detect abnormal situations, behaviors through a wearable vital/actimetric device [2].

 ISBN: 978-1-61208-256-1

- Provide recipes; keep history of menus prepared for friends/family.
- Remember faces/names/information through the phone camera.

Some of these features are already available on smartphones, others are being developed such as the Microsoft MyLifeBits project [7].

III. VOIP ARCHITECTURE AND SERVICES

A. Existing Platform

As part of different projects (CompanionAble, vAssist) the current platform can take many shapes even though some areas are commonly shared:

- Asterisk Internet Protocol Private Branch eXchange (IPBX) for all video/audio communications.
- Julius ASR server.

At the users' homes:

- At least one platform featuring a camera, display and VoIP client (computer/phone/tablet/tvbox).
- Sensors for person monitoring.

B. A Unified and standardized communication solution

As a result of the devices heterogeneity, we needed to be able to handle different kinds of media. The VoIP solution allows us not only to take care of that aspect, but gives us the possibility to extend functionalities via plugin developments. Moreover, this infrastructure has the ability to be inserted into a public VoIP network for cross domains/technologies communications. Current advantages are:

- Support for various Internet infrastructures (e.g., public/private IP, ADSL box).
- Interoperability with public and private telecommunication networks (e.g., google talk, skype).
- Low latency (less than 100ms with H263 video)
- Automatic bandwidth adjustment for Quality Of Service.
- Support for various clients (e.g., softphones -phone softwares-, IP phones, mobile phones).
- Large choice of audio and video codecs.
- Ability to set up centralized services (low cost of deployment) as IVR (Interactive Voice Responce), ASR, multi-conferencing, voice and video messaging.
- The user is linked to a unique identifier (a phone number).
- Centralization of data (voice, video).
- Out of the box internationalization.

C. Communication infrastructure

As VoIP solutions imply the use of a PBX, we decided to use Asterisk from DIGIUM. It has standard configurations for regular calls but allows us to tweak it extensively for our purposes. Regarding the fact that patient networks will use private IP addresses, we initially believed a local PBX was needed not only for local communications, but also for call forwardings/translation from the public to private domains. After some tests, we would probably need this in some restricted cases for PSTN translations, but for the average user, it will probably not be necessary (depending on the id-number allocation).

When a call is started, a SIP [23] request is sent to the PBX, which transmits it to the end-client. When this signalling communication is done, a direct tunnel is established using RTP (Real Time protocol) [24]; (see Fig. 1). This protocol, over UDP [25], keeps the packet order and drops old ones. Fig. 2 shows how different components are set on ISO layers. Depending on the service definition, it might be necessary to use trunking services to allow all communications through PBXs (see Fig. 3).

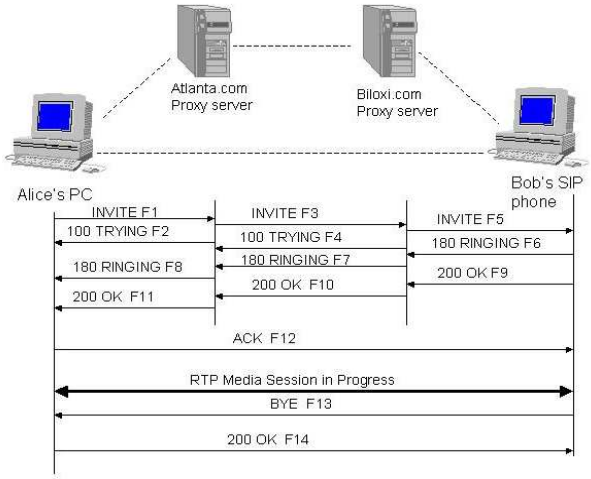

Figure 1. Call Dialog.

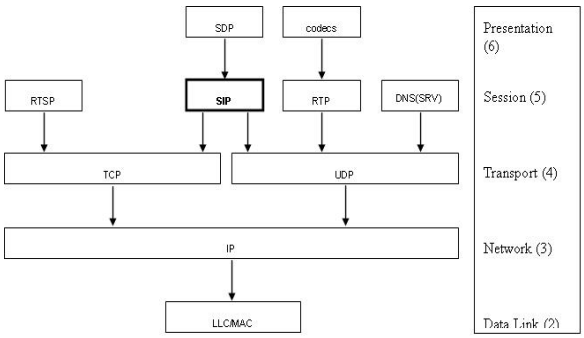

Figure 2. SIP and OSI.

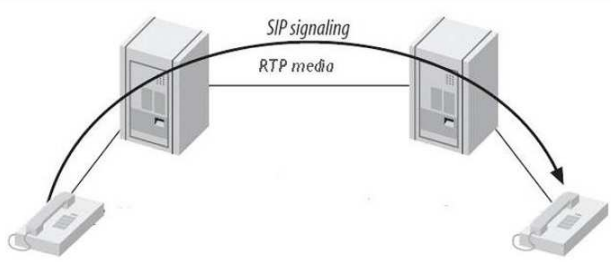

Figure 3. SIP trunking architecture.

1) A codec, is a module that can COde and DECode an analog or a digital signal. For VoIP communications, many codecs are available. As PBXs are not designed for stream translation, we initially needed to make sure both clients used the same normalized codecs. Later on we worked with a transcode plugin to eventually reduce this impact.

Asterisk can handle at least:

- Voice: ulaw, alaw, gsm, ilbc, speex [10][22], g726, adpcm, lpc10, g729, g723;
- Video: h261 [11], h263 [12], h263+, h264 [13][19], MPEG-4, VP8.

For our systems, we decided to use: µlaw, alaw, speex for the voice encoding and H261, H263 and H264 for the video part. The key point is finding a good fit between "compression", "delay" and "video quality" as increasing the compression rate increases the delay due to buffer use and a higher processing load per time unit.

2) Alarms: There are multiple ways to handle and transmit alarm signals he goal of the vAssist project is to provide specific voice controlled Home Care and Communication Services for two target groups of older persons: Seniors suffering from chronic diseases and persons suffering from (fine) motor skills impairments. The main goal is the development of simplified and adapted interface variants for tele-medical and communication applications using multilingual natural speech and voice interaction (and supportive graphical user interfaces where necessary).and all could be implemented in parallel. The first one is to use the SIP MESSAGE method. (see Table 1 for SIP Methods). As Asterisk does not handle it, it is necessary to implement RFC 3428 [26] for SIP channel. We also could use T.140 (RFC 4103 [27]) method for Instant Messaging/Alarms communications. Last would be to simply automate calls/messages/mails to emergency services when a specific signal has been sent from the monitoring device.

TABLE I. SIP METHODS

SIP Method	*Description*	*RFC*
ACK	Acknowledge final response to Invite	3261
BYE	Terminate a session	3261
CANCEL	Cancel a previous call	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

D. Performances

Two different VoIP clients are currently used for performances and codecs compatibility: ekiga [29] for the PC platform and linphone [30] for either PC, Android, iOS platforms. These clients are customized for HD and low delays communication. We currently use wideband Speex audio codec and H263 or H264 video codecs depending on the platform with a specific bandwidth adaptation module. It makes sure instant messaging and voice delays are being kept as low as possible by reducing video resolution in case of congestion. This ensures low delays communication over internet (less than 100ms for a PC to PC communication over internet, less than 200ms for a PC to smartphone communication using WiFi). Tests with other standard VoIP clients and skype gave delays between 200ms and 500ms for long term communication (more than 3 hours long). All these tests were done between two private networks with their own Asterisk IPBX.

IV. VOICE-BASED SYSTEM INTERFACE

A. ASR and VoIP

The main advantage of such a centralized platform is that services and tools can be accessed with all connected devices. Regarding ASR, there is no embedded ASR tool into Asterisk. Julius, an Open source project, offered all the services we needed and has the ability to redirect either input and/or output streams to any ip socket. Its speed and real time speech recognition for large vocabulary made it a perfect candidate for this purpose. We found the app_julius module [15], developed by Danijel Korzinek and Dikshit Thapar, which allowed us to make a direct connection between Asterisk and Julius which follows this flow :

 ISBN: 978-1-61208-256-1

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

```
[macro-speech-confirm]
exten => s,1,SpeechActivateGrammar(yes_no)
exten => s,2,Set(OLDTEXT0= ${SPEECH_TEXT(0)})
exten => s,3,Playback(heard)
exten => s,4,Playback(${ARG1})
exten => s,5,SpeechStart()
exten => s,6,SpeechBackground(correct)
exten => s,7,Set(CONFIRM=${SPEECH_TEXT(0)})
exten => s,8,GotoIf($["${SPEECH_TEXT(0)}" = "1"]?9:10)
exten => s,9,Set(CONFIRM=yes)
exten => s,10,Set(CONFIRMED=${OLDTEXT0})
exten => s,11,SpeechDeactivateGrammar(yes_no)
```

The voice-based MMI (Maximum Mutual Information) functionality uses a voice recognition module based on Julius and HTK (Hidden Markov Toolkit) softwares (Julius for recognition, HTK for training) with adaptation facilities to customize the system to our speakers' constraints and needs..

B. *Julius, HTK*

The voice recognition module is based on the use of conventional Hidden Markov Models (HMM) to model statistically the acoustic models of phonemes and / or words in the vocabulary. We use software tools such as HTK [21] and Julius [16]. Language models (linguistic probabilities, which are complementary to acoustic probabilities) are implicitly addressed in the use of such models to make robust word recognition in a given sentence (use of statistical N-grams and rules of grammar).

V. CURRENT PROJECTS

The following projects came to life after realizing that no particular solution was fitted to offer a unified solution for this matter. Of course we could find some products purely specialized in the video or audio communication or more recently some virtual “butlers” like SIRI started to appear though with limited functionalities of the medical context, but regarding the unified user experience, the simplicity of use, a lot of work could have been done.

A. *CompanionAble*

The main idea behind the CompanionAble project was to provide a synergy between Robotics and Ambient Intelligence technologies and integrate them into a fully assistive environment. A robotic companion was working collaboratively with a smart home environment. CompanionAble addressed the issues of social inclusion and homecare of persons suffering from chronic cognitive disabilities prevalent among the increasing European older population. This is obviously a subsection of a more generic group of persons with elevated requirements and constraints. ASR has been used for service managements and SIP technologies have been put in place for audio/video communications, integrated into a standardized GUI. Yet, the two technologies were not linked and ran concurrently. Also, SIP had been proposed to handle commands for the robot movements via the messaging service. Usage of the robot proved to be quite accepted but costs and infrastructure requirements (a smarthome environment) unfortunately reserved it to a very few: It would be unable to fit in a small flat as the ones found in the big european cities and at the same time would not be able to handle stairs in the case of a house with multiple floors.

B. *vAssist*

As for CompanionAble, the goal of the vAssist project is to provide specific voice controlled Home Care and Communication Services for two target groups of older persons: Seniors suffering from chronic diseases and persons suffering from (fine) motor skills impairments. The main idea is the development of simplified and adapted interface variants for tele-medical and communication applications. The main difference with the previous project stands in the two following points:

- To target a wider audience, standard equipment is preferred. It reduces the development costs and existing hardwares and interfaces in the home of the users can be used such as PC, TV, mobile phone or tablets.
- Every service of the system must be defined not only to use graphical user interfaces but also multilingual natural speech and voice interaction. In a sense, existing services can be adapted/enhanced to support these aspects.

In this aspect, Asterisk and Julius represent the first accesspoints to such a service.

VI. CONCLUSION AND FUTURE WORK

The infrastructure for testing a physical or virtual butler is in place. Open source software components were selected for telecommunications (PBX - Asterisk) and for automatic processing of speech (Julius). Experimental results were

 ISBN: 978-1-61208-256-1

obtained during the CompanionAble project. Under vAssist, more common devices (smartphones, tablets...) will be preferred [1]. The Asterisk server is ready for testing services related to usage scenarios listed in Section 2.

So far, telephony signals have been sampled at 8kHz but our experimentations showed that we would probably need to work with higher-rates codecs (e.g., Speex 16kHz), better acoustic models and then finally to improve the platform from Narrowband to Wideband.

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

ACKNOWLEDGMENT

Some parts of this research leading to these results have received funding from the European Commission Seventh Framework Program (FP7/2007-2013) under grant agreement number 216487.

REFERENCES

[1] N. Armstrong, C. Nugent, G. Moore, and D. Finlay, “Using smartphones to address the needs of persons with Alzheimer's disease,” Annales des Télécommunications, vol. 65, pp. 485-495, 2010.

[2] J.L. Baldinger, et al., “Tele-surveillance System for Patient at Home: the MEDIVILLE system,” 9th International Conference, ICCHP 2004, Paris France, Series : Lecture Notes in Computer Science, Ed. Springer, 2006.

[3] R. Bayeh, “Reconnaissance de la Parole Multilingue: Adaptation de Modeles Acoustiques Multilingues vers une langue cible,” Thèse (Doctorat) TELECOM Paristech, 2009.

[4] D.R.S Caon, et al., “Experiments on acoustic model supervised adaptation and evaluation by k-fold cross validation technique,” In: ISIVC. 5th International Symposium on I/V Communications and Mobile Networks. Rabat, Morocco: IEEE, 2010.

[5] N. Clement, C. Tennant, and 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, [retrieved: February, 2013]

[6] A. Constantinescu and G. Chollet, “On cross-language experiments and data-driven units for alisp (automatic language independent speech processing),” In: IEEE Workshop on Automatic Speech Recognition and Understanding. Santa Barbara, CA, USA, 1997, pp. 606-613.

[7] Digium, The Open Source PBX & Telephony Platform, http://www.asterisk.org/, [retrieved: February, 2013].

[8] J. Gemmell, G. Bell, and R. Lueder, “MyLifeBits: a personal database for everything,” Communications of the ACM, vol. 49, Issue 1, pp. 88-95, 2006. http://research.microsoft.com/en-us/projects/mylifebits/, [retrieved: February, 2013].

[9] L.N. Gitlin and T. Vause Earland, “Améliorer la qualité de vie des personnes atteintes de démence: le rôle de l'approche non pharmacologique en réadaptation,” J.H. Stone, M. Blouin, editors. International Encyclopedia of Rehabilitation, 2011. http://cirrie.buffalo.edu/encyclopedia/fr/article/28/, [retrieved: February, 2013].

[10] G. Herlein, J. Valin, A. Heggestad, and A. Moizard, “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, [retrieved: February, 2013].

[11] International Telecommunication Union, “H.261: Video codec for audiovisual services at p x 64 kbit/s,” Line Transmission of Non-Telephone Signals, 1993.

[12] 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.

[13] 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.

[14] Julius ASR, http://julius.sourceforge.jp/en_index.php, [retrieved: February, 2013].

[15] D. Korzinek, module app_julius, http://forge.asterisk.org/gf/project/julius/, [retrieved: May, 2012].

[16] A. Lee, T. Kawahara, and K. Shikano, “Julius - an open source real-time large vocabulary recognition engine,” EUROSPEECH, pp. 1691-1694, 2001.

[17] A.S. Rigaud, et al., “Un exemple d'aide informatisé á domicile pour l'accompagnement de la maladie d'Alzheimer : le projet TANDEM”, NPG Neurologie - Psychiatrie - Gériatrie. N°6, Vol.10, ISSN :1627-4830, LDAM édition/Elsevier, ScienceDirect, April 2010, pp 71-76.

[18] T. Schultz and K. Katrin, “Multilingual Speech Processing,” Elsevier, 2006.

[19] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the H.264/AVC Video Coding Standard,” IEEE Transactions on Circuits and Systems for Video Technology, 2003.

[20] World Health Organization, The European Health Report, European Series, #97, 2002.

[21] S. Young, et al., “The HTK Book”, version 3.4. Cambridge, UK: Cambridge University Engineering Department, 2006.

[22] Xiph.Org Foundation, “Speex: A Free Codec For Free Speech”, http://speex.org/, [retrieved: February, 2013].

[23] SIP protocol, RFC 3261, http://www.ietf.org/rfc/rfc3261.txt, [retrieved: February, 2013].

[24] RTP , RFC 3550, http://www.ietf.org/rfc/rfc3550.txt, [retrieved: February, 2013].

[25] UDP, RFC 0768, http://www.ietf.org/rfc/rfc0768.txt, [retrieved: February, 2013].

[26] SIP Message Extension, RFC 3428, http://www.ietf.org/rfc/rfc3428.txt, [retrieved: February, 2013].

[27] RTP Payload for Text Conversation, RFC 4103, http://www.ietf.org/rfc/rfc4103.txt, [retrieved: February, 2013].

[28] http://en.wikipedia.org/wiki/Asymmetric_Digital_Subscriber_Line, [retrieved: February, 2013].

[29] http://www.ekiga.org, [retrieved: February, 2013].

[30] http://www.linphone.org, [retrieved: February, 2013].

 ISBN: 978-1-61208-256-1