Automated Audio-visual Dialogs over Internet to Assist Dependant People

Thierry Simonnet
R&D department
ESIEE-Paris
Noisy le Grand, France
t.simonnet@esiee.fr

Gérard Chollet, Daniel Caon
CNRS-LTCI
TELECOM-ParisTech
Paris, France
{*chollet, caon*}*@telecom-paristech.fr*

Jérôme Boudy
TELECOM-SudParis
Evry, France
jerome.boudy@it-sudparis.eu

***Abstract*—An increasing number of people are in need of help at home (elderly, isolated and/or disabled persons and people with mild cognitive impairment). Several solutions can be considered to maintain social links while providing tele-care to these people. Many proposals suggest the use of automatic speech recognition (ASR) to control equipments and to maintain social link. In this paper, we will look at an environment constrained solution, its drawbacks (such as latency) and its advantages (e.g.flexibility, integration). A key design choice is to control equipments using a Voice over Internet Protocol (VoIP) solution with an ASR system, while addressing bandwidth limitations, providing good communication quality to obtain the best results for speech recognition. The resulting platform offers a powerful framework for setting up a virtual butler but also different services as voice controlled equipments and text transcriptions of medical talk.**

***Keywords*-ASR; Voice over IP (VoIP);**

I. INTRODUCTION

As part of ongoing projects, research has been conducted towards the implementation of efficient solutions for audio/video communications between people and system control, through an unified channel.

In Europe, there is an increasing demand [5], [9], [20], for maintaining dependent people at home, to reduce hospitals load, improve their quality of life and strengthen their social links. To this extent, a need for suitable communication systems and telecare technologies has arisen. Maintaining such people at home often requires medical assistance, excellent and reliable communication tools, handled by their relatives and the caregivers [17].

Nonetheless, several constraints have to be taken into account: These systems make use of an internet connection and as such they rely on bandwidth availability. Most European personal internet accesses use Asymmetric Digital Subscriber Line (ADSL) technology, offering a limited upload bandwidth. Still, in order to offer accurate and exploitable communication, image and sound quality must be maintained. Video quality is usually highly dependent on available bandwidth, and how different compression algorithms perform with limited bandwidth.

This article is organized as follows. We present environment, platform overview and explanation of our choices in Section 2, technicals descriptions in Section 3. A detail of technical integration is given in Section 4. Results are shown in section 5. Then the identified remaining issues and perspectives are discussed before concluding in Section 6.

II. REMOTE AUTOMATIC SPEECH RECOGNITION INTERFACE

A. Speech recognition

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 includes domains from vocal commands to complex dialog systems, capable to identify one's state of mind and to detect distress situations.

B. Usage scenarios

Suppose that one always has access to a personal and collective memory-prosthesis, relayed by an audio-visual Butler.

There are different way to interact with the Butler. Here are samples of actions :

- find his way as the Butler embodied by a Smartphone, connected to GPS, knows our location and can guide us,
- record short video or photos to an album of his recent past,
- remember the name and other information about a person that we met, the Butler equipped with a camera takes a picture and finds this information,
- shopping, shopping list, prices, ...
- provide recipes, remember which menu were prepared for his friends, family, ...
- view and update their diary, appointments, bills to pay, planned parties,
- answer the phone, messaging, ...
- find informations on the web

 ISBN: 978-1-61208-186-1

- detect situations of distress, abnormal behavior through a wearable vital/actimetric sensors-based device [2], ...

Some of these features are already available on smartphones, others are being developed such as the Microsoft MyLifeBits project [8].

III. VoIP architecture and services

A. Existing Platform

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

- Asterisk Internet Protocol Private Branch eX change (IP-PBX, or IPBX) for voice/video communications routing; [7]
- Julius ASR server (can be hosted by another server) [14].

Home, equipped with:

- a platform 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 VoIP Client or a smartphone.

B. A Unified and standardized communication solution

Different kinds of media for different equipments need to be addressed. A VoIP solution offers a complete and unified communication infrastructure and then various services can be developed with it. This infrastructure can be used for a closed group of users but also be plugged on public VoIP network. This solution meets all criteria for medical/paramedical usage :

- supports various Internet infrastructures (e.g., public IP, private IP, ADSL box);
- interoperability with public and private (e.g., ekiga.net, google talk, skype) telecommunication networks;
- low latency (less than 100ms 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 (e.g., 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 (Interactive Voice Responce), ASR, multi-conferencing, voice and video messaging;
- unique identifier (phone number);
- centralization of data (voice, video);
- internationalization with customization of user language.

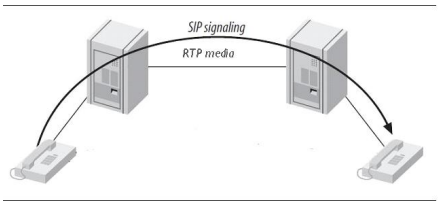

Figure 1. SIP trunking architecture

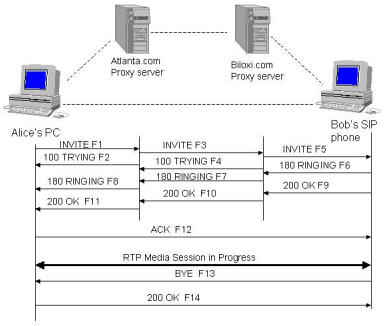

Figure 2. Call dialog

C. Communication infrastructure

Internet is the main communication media for the project. VoIP solutions imply the use of a PBX. The Asterisk PBX from DIGIUM Company will be used for the first version of the infrastructure. Asterisk PBX has standard configuration for classical communications but needs new and modified communication module for our purposes, repectively for voice and video transmissions. 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 [23] 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) [24]; (See Fig 2). This protocol, over UDP [25], keeps the packet order and drops old ones. Fig 3 shows how different components are set on ISO layers. To establish a communication between 2 private networks, it is necessary to use trunking services to allow all communications through PBXs (see Fig 1).

1) 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 thus clients must have compatible codecs that respect norms.

 ISBN: 978-1-61208-186-1

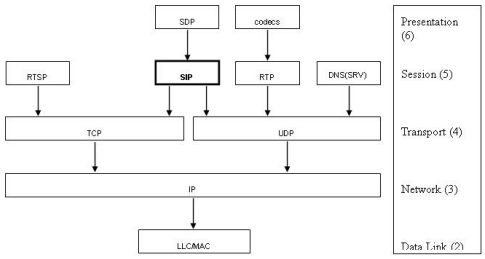

Figure 3. SIP and OSI

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 I
SIP METHODS

Asterisk can handle :

- voice : ulaw, alaw, gsm, ilbc, speex [10], [22], g726, adpcm, lpc10, g729, g723;
- video : h261 [11], h263 [12], h263+, h264 [13], [19].

For our vAssist system, we can use: law, alaw, speex 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 rate indeed increases the delay due to buffer use and a higher processing load per time unit.

2) Alarms: It is also possible to transmit alarms using SIP MESSAGE method. (see Table 1 for SIP Methods). Asterisk doesn't handle this method and it is necessary to implement RFC 3428 [26] for SIP channel. It is also possible to use T.140 (RFC 4103 [27]) method for Instant Messaging/Alarms communications. Both solutions could be implemented.

D. Services

1) Central PBX server: It is necessary to interconnect the central PBX to a call center with economical and security parameters. The following features have to be implemented:

- direct connection with an auto-connected client and using specific dial number;
- security in the possibility of encryption (e.g., VPN or stream encryption), OSP, closed group of subscribers for confidentiality. Management by phone number and not by identity;
- dialplan will transfer any specific local calls to the right call centre ;
- depending on partners' needs, we will develop dedicated modules for PBX (e.g., MP4, Interactive voice/video respond solution, RTSP (Real-Time Stream Protocol), Speech to text - ASR);
- no data duplication;
- ability to centralize all the data for exploitation or study purposes;
- patient home PBX is setup on a Plug computer 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 including voice order handling. Such a centralized ASR (a great exploitation benefit) will handle multi-language tools and avoid unitary installation.

E. Performances

This unified platform is used for communication and videoconferencing purposes but also for robot remote control. Two different VoIP clients are used for performances and codecs compatibility : ekiga (www.ekiga.org) for PC platform and linphone (www.linphone.org) for PC and Android platform. These two clients are customized for HD and low delays communication. We use wideband Speex audio codec and H264 video codecs with a specific bandwidth adaptation module that reduces videoresolution in case of bandwidth congestion, keeping instant messaging and voice delays as low as possible. This solution allows 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) and far less reliability (unexpected end of communication, freezing...). All these tests were done between 2 private networks with their own Asterisk IPBX, using trunking facilities to go through internet to demonstrate that it is possible to remote control a robot using standard Internet infrastructure.

IV. VOICE-BASED SYSTEM INTERFACE

A. ASR and VoIP

With such a centralized platform, ASR can be accessed using phone technologies facilities. Asterisk provides various speech tools but no embedded ASR tool. A proper commercial model, trained and annotaded by professionals costs. Open Source project Julius offers the services we need. Asterisk offers a generic speech API that is 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 the ASR system, to activate the relevant grammar according to the context application context or menu and to play back a sound file while waiting for the person to speak.

We use app_julius module [15] 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.

```
[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 softwares (Julius for recognition, HTK for training) with adaptation facilities to customize the system to the Elderly person use 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).

Sentence (repeated 10x by the same speaker)	Sentence totally correct(%)	Semantically correct(%)
hellep	100	100
help me	50	90
kom naar de keuken	100	100
kom eens naar de keuken	100	100
wil je naar de keuken komen	60	60
...	...	...
hector ik ga lunchen met kennisen	100	100
hector ik ga lunchen met buren	100	100
wolly ik ga uit eten	100	100
wolly ik ga uit eten met vrienden	100	100
wolly ik ga uit eten met kennisen	100	100
ja graag	40	80
Average percentage	86.39	94.44
Lowest percentage	40	60
Highest percentage	100	100

Table II
RECOGNITION RATES

C. Data, Adaptation

Adaptation strategies are implemented, especially those based on crossed multilingual adaptation between languages with rich phonetic materials.

Research of Tania Schultz [18], Rania Bayeh [3] and Gerard Chollet [6] on language independent and multilingual speech recognition, serve as a starting point.

V. EVALUATION

A first validation of the automatic speech recognition system was performed on data recorded in the European CompanionAble project (www.companionable.net). 22 Elderly Dutch speakers were recorded for one hour each in an experimental house (SmartHome) in Eindhoven. They repeated the phrases uttered by a "prompter', person reading at slow acoustic level the Prompt texts. Table II (containing 37 different phrases) gives results for one of these speakers, after adaptation of acoustic models by MLLR (classical technique of adaptation also studied in [4]) to its voice.

The rate of " semantically correct" is a way to describe that two sentences are at the same level of meaning (e.g., "help me" and "hellep"), so if "help me" is recognized instead of "hellep" the sentence is 100% correct (semantically) and voice dialogue can take place without problems.

The second software validation was conducted on 20 speakers (all seniors) without repetition of phrases, and the measurement is shown at the rate of word recognition. A classical MAP adaptation technique (also studied in [4]) was applied from a set of 10 adaptation sentences for each speaker. Figure 4 shows the results by language models and n-grams (2-gram and 6-gram), with and without adaptation of acoustic models.

Improvements are achieved through the hidden Markov models adaptation and the language model precision. We

 ISBN: 978-1-61208-186-1

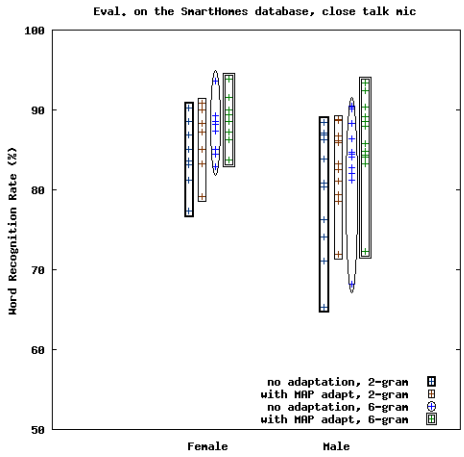

Figure 4. Evaluations of 20 elderly speakers (lavalier microphone).

also notice that not all users test the recognition system with the same success rate.

VI. Conclusion

The infrastructure for testing a mobile butler is in place. It uses both free software components for telecommunications (PBX - Asterisk) and for automatic processing of speech (Julius). Experimental results were obtained by automatic speech recognition of recorded data in the project CompanionAble. Under vAssist, a smartphone (Android) will be used [1]. The Asterisk server is ready for testing services related to usage scenarios listed in Section 2.

Today, telephony speech signals are generally sampled at 8kHz. First experimentations gave us good results but we 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 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 multiple kinds of applications, in many types of contexts (e.g., telemedicine, security, vocal commands, etc) that it is capable to handle.

Acknowledgment

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

References

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

[2] Baldinger, J.L. 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] Bayeh, R. Reconnaissance de la Parole Multilingue: Adaptation de Modeles Acoustiques Multilingues vers une langue cible. Thèse (Doctorat) TELECOM Paristech, (2009);

[4] Caon, D.R.S. 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] Clement, N., Tennant, C., and Muwanga, C.: 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;

[6] Constantinescu, A. and Chollet, G. 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: p. 606-613, (1997);

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

[8] Gemmell, J., Bell, G. and Lueder, R., MyLifeBits: a personal database for everything, Communications of the ACM, vol. 49, Issue 1 (Jan 2006), pp. 88.95. http://research.microsoft.com/en-us/projects/mylifebits/

[9] Gitlin LN and Vause Earland T.. Améliorer la qualité de vie des personnes atteintes de démence: le rôle de l'approche non pharmacologique en réadaptation. In: JH Stone, M Blouin, editors. International Encyclopedia of Rehabilitation, (2011). Available online: http://cirrie.buffalo.edu/encyclopedia/fr/article/28/;

[10] Herlein G. 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.

[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

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

[16] Lee, A., Kawahara, T., and Shikano, K. In: EUROSPEECH. Julius - an open source real-time large vocabulary recognition engine. p. 1691-1694, (2001);

 ISBN: 978-1-61208-186-1

[17] Rigaud, A.S. et al. “Un exemple d’aide informatisé á domicile pour l’accompagnement de la maladie d’Alzheimer : le projet TANDEM”, NPG Neurologie - Psychiatrie - Gériatrie. N6, Vol.10, pp. 71-76, ISSN :1627-4830, LDAM édition/Elsevier, ScienceDirect, (April 2010).

[18] Schultz, T. and Katrin, K. Multilingual Speech Processing. Elsevier, (2006);

[19] Wiegand, T., Sullivan, G.J., Bjontegaard, G., and Luthra, A., Overview of the H.264/AVC Video Coding Standard, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2003.

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

[21] YOUNG, S. J. 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/.

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

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

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

[26] UDP, RFC 3428, http://www.ietf.org/rfc/rfc3428.txt

[27] T.140, RFC 4103, http://www.ietf.org/rfc/rfc4103.txt

 ISBN: 978-1-61208-186-1