vAssist:
The Virtual Interactive Assistant for Daily Home-Care.

Daniel R.S.Caon*, Thierry Simonnet†, Pierre Sendorek*, Jérôme Boudy‡ and Gérard Chollet*
*CNRS-LTCI, TELECOM-ParisTech,
46, rue Barrault, 75634 Paris cedex 13, France
Email: daniel.caon, pierre.sendorek, gerard.chollet@telecom-paristech.fr
http://www.telecom-paristech.fr/
†ESIEE,
ESIEE , 2 boulevard Blaise Pascal, Cité DESCARTES,BP 99,
93162 Noisy le Grand cedex, France
Email: t.simonnet@esiee.fr
http://www.esiee.fr/
‡TELECOM-SudParis,
9, rue Charles Fourier, 91011 EVRY cedex, France
Email: jerome.boudy@it-sudparis.eu
wwww.it-sudparis.eu

***Abstract*—The "Ambient Assisted Living" program supports the vAssist project for the European community. The portable devices (smartphones) are the keys to integrate open-source systems of Voice/Video over IP and Automatic Speech Recognition in order to produce a centralised and interactive virtual assistant for daily activities. To increase the life quality of the users (which can be called as care recipients) is part of the objectives. The assistant in this behavior is here called 'the Butler' and provides personalised health care and security, being able to handle the detection of distress expressions and vocal commands. The target population are handicaps and elderly people living alone at home.**

I. Introduction

Our often faulty memory is not rarely supplemented by the internet, which can be considered the "world's memory", directly accessible through search engines in smartphones. This concept is not so recent, as in 1945, V. Bush wrote an article [7] describing an electronic device (the memory extender, or "memex") linked to a library and capable to display books and movies.

The collection and the storage of personal information is now possible [15]. Hence, local and remote servers are the founding blocks of our "memorial prosthesis" [14]. Patients with Alzheimer's disease, including its early forms ("Mild Cognitive Impairment") [2], [4], [13] and also handcaps are considered as the main target population of such device. Any kind of support (cognitive stimulation, daily life support and also surveillance) being provided by the "Butler" can greatly increase the life quality of the care recipients. The "Butler" is designed to fit into an ecosystem of communicating objects, having two parts:

- A fixed entertainment platform including tasks such as cognitive stimulation exercises, telephony, Internet access, television, video-conferencing (relationship with the caregiver, family and carers), management records, medical needs, banking, taxes and invoices;
- A portable communicating system, the Butler allowing the patient to alert the caregiver if dropped or with abnormal behavior, knowing its location outside home, remaining in telephone contact with her caregiver and reminding tasks.

This article is structured as follows:

- Some usage scenarios are proposed in Section II;
- Architecture for Voice/Video over IP is detailed in Section III;
- The speech dialogue system is described in Section IV;
- The evaluation of the automatic speech recognition system is provided in Section V;
- Followed by conclusions and perspectives.

Some parts of this paper have been presented recently [9], focusing the handicap public, which encourages this kind of research with great expectations.

II. Usage scenarios

Suppose that one always has access to a personal and collective memory-prosthesis, relayed by an audio-visual Butler. What are the uses we could do? Here is a list:

- 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 meet, the Butler is 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 information on the web;
- Detect situations of distress, abnormal behavior through a wearable vital/actimetric sensors-based device [6].

Some of these features are already available on smartphones, others are being developed as the project MyLifeBits Microsoft [15].

III. Voice and Video over IP

The Butler (smartphone) communicates with the server in accordance with the Session Initiation Protocol (SIP).

A. SIP (Session Initiation Protocol)

Using a VoIP infrastructure involves setting up a PBX (Private Branch Exchange, also called PABX) software. We selected the product from Digium [1]. It is configured to handle voice communications, but also to support the video ones.

When a call is started, a SIP request is sent to the PBX, which transmits it (using dial plan, or for our case to the central PBX) to the other client. When this signalling communication is done a direct one is established using RTP (Real Time protocol). This protocol, over UDP (User Datagram Protocol), keeps the packet order and drops old ones. A self-local switch must be installed in each patient and connected to the central PABX (Private Automatic Branch Exchange). This architecture will use the SIP trapezoid topology (see Fig 1).

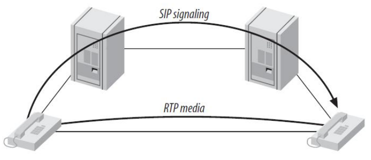

Fig. 1. SIP trapezoid architecture.

B. PABX

The Asterisk PBX is configured to provide all telephony standard features and to offer video transmission services. Voicemail services is also setup to provide services as close as possible to the conventional phone.

An essential component for transmitting video is CODEC. A codec (Code-DECode) is a module that can Code and DECode an analog or a digital signal. For VoIP the codec is used to normalize also the module itself. The X264 (codec) handles streams that use MPEG-4 AVC/H264 normalisation.

A direct RTP communication is set between 2 clients and then clients must have normalised and also compatible codecs. 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. High definition codecs such as H264 are not the best ones for Real Time video calls. They are used for HD television, or web application, where the transmission delay is not critical (unidirectional).

For first trials, the use of H261 offers a good compromise between video definition and transmission delays.

C. Client SIP

A SIP client must have, for our application of video telephony, some specifications: low latency and good image quality.

A QoS (quality of service) system should be established to ensure a constant flow. The SIP client must be able to dynamically balance the compression ratio and quality of different channels (video and audio) to fit the bandwidth. It must also provide arbitration between the video and audio streams. An application of such monitoring does not need a great sound quality, but low latency, while the priority for application of video telephony is voice.

The Asterisk PBX doesn't provide codec translation. Therefore the choice of codecs and softphone is important. The SIP enables a self-negotiation between the two clients to get a couple of codecs with a compatible and optimal method.

We chose the Ekiga Softphone for several reasons:

- It is open-source, thus it can be changed;
- Portability (Linux, Windows);
- Multiple built codecs (H261, H264, Theora, PCMU, GSM,...);
- Graphical interface can be easily modified using Glade (the Gnome GUI builder).

We worked on the client Ekiga, integrating the open source project and maintaining the Windows version of the project. An automatic answer function was added, which may be of interest to the call of the elderly. First experiments showed a transmission delay on the local network of about 400ms. These delays have been reduced to 80ms using the internet, without QoS.

D. Limitations and Developments

- We have found an inconsistency in the use of Windows Vista. We concentrate on a target Windows XP.
- Much work has been done to reduce transmission time (between 30ms and 80ms)
- The Asterisk PBX is public (provided you have an account) and operational with any SIP client (Ekiga, X-Lite, linphone, ...): wagram.esiee.fr.
- The services offered are:
  - Conference: Changing channels video via Dual-Tone Multi-Frequency (DTMF) (use of dial button);
  - Instant Messaging during a conversation;
  - Interact with gtalk, msn, aim and yahoo messenger;
  - Interfacing with Julius, an open-source large vocabulary continuous speech recognition (LVCSR) system;
  - Interactive Voice and Video Response (IVVR).

IV. Speech-based system interface

The speech-based interface functionality uses a speech recognition module based on Julius LVCSR (Large Vocabulary Speech Recognition) and HTK (Hidden Markov Model Toolkit) softwares.

A. 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 [19] and Julius [16]. Julius is used for recognition and HTK for training.

Acoustic and language model adaptation are conducted to customize the system to the users (specially elderly) accessibility constraints. Thus, the 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).

B. Data, Adaptation

Adaptation strategies will be implemented, especially those based on crossed multilingual adaptation between languages with rich phonetic materials. Research of Tania Schultz [18], Rania Bayeh [5] and Gérard Chollet [11] on language independent and multi-lingual speech recognition, serve as a starting point. Classical HMM (acoustic models) adaptation methods (MAP and MLLR) are described below and further evaluated at section V.

1) Maximum a Posteriori: Hidden Markov models adaptation is here accomplished using a classical maximum a posteriori (MAP [12]) approach (also known as Bayesian adaptation). The informative priors contain the speaker independent model parameters. Given a state *j* and a mixture component *m*, the adaptation formula becomes:

$$\hat{\mu}_{jm} = \frac{N_{jm}}{N_{jm}+\tau}\bar{\mu}_{jm} + \frac{\tau}{N_{jm}+\tau}\mu_{jm} \qquad (1)$$

where the weighting of the a priori knowledge to the adaptation speech data is represented by τ and the occupation likelihood of the adaptation data is represented by N, defined as,

$$N_{jm} = \sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t) \qquad (2)$$

where $L_{jm}^{r}(t)$ is the occupancy probability for a given state *j*, a mixture component *m* and an observation sequence *r* at the time t.

μ_{jm} was the speaker independent mean (a priori knowledge) and $\bar{\mu}_{jm}$ is the mean of the observed adaptation data, defined as,

$$\bar{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{r}(t)} \qquad (3)$$

Hence MAP has as drawback the need of large adaptation data in order to build speaker dependent models, as it is defined at component level.

2) Maximum Likelihood Linear Regression: The Maximum Likelihood Linear Regression adaptation technique [17] computes a set (or one global) transformation(s) in order to adapt the Hidden Markov Models, reducing the mismatch between the initial model set (speaker independent) and the adaptation data (speaker dependent).

This technique makes use of transformation matrices to reestimate the model parameters. Its main advantage is the ability to overcome the adaptation accuracy rates obtained with the MAP technique when low adaptation data is available.

V. Evaluation

A first validation of the automatic speech recognition system was performed on data recorded in the CompanionAble European project (www.companionable.net). 22 Dutch speakers were recorded for one hour each in an experimental house (SmartHome) in Eindhoven. They repeated the phrases uttered by a “prompter”.

This section is divided into Database Acquisition (subsection V-A), Model Adaptation (subsection V-B) and Results (subsection V-C)

A. Database acquisition

The sound recordings are conducted at SmartHomes (www.smart-homes.nl). Adaptation sentences (a set of 10 phonetically balanced) have also been recorded for each speaker.

Specific features concerning this database are shown in Table I.

TABLE I
SmartHomes database specifications

Feature	Value
Duration	40 to 60 min per person
Sampling rate	16 kHz
Syncronised microphone channels	10
Speakers	20 elderly + 1 adult woman
Adaptation sentences per speaker	10
Sessions per speaker	1
Scenario sentences per session	96, in the standard scenario (played by all speakers) 55, in the review scenario (played by adult women)
Speakers recorded	>60 years old: 12 males and 8 females, in the standard scenario <60 years old: 1 young female speaker, in the review scenario

B. Model Adaptation

In this paper we conduct experiments of adaptation (using MAP and MLLR techniques) with phonetically balanced sets of speech data trying to cover as much as possible all acoustic space (part of the SmartHomes database has been built exactly for speaker adaptation research).

C. Results

Table II describes the number of synchronised channels for each microphone.

TABLE II
RECORDED MICROPHONE CHANNELS

Microphone	Channels
Close-talking mic. (lavalier fixed on the speaker)	1
Simple omnidirectionnal mic.	1
Directionnal mic. A	3 cardioids + omnidirectionnal = 4
Directionnal mic. B	3 cardioids + omnidirectionnal = 4
	Total = 10

The Table III (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 [8]) to its voice.

TABLE III
RESULTS ON THE DUTCH SPEECH RECOGNITION SCENARIO

Sentence (repeated 10 times by the same speaker)	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

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 recognised 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 [8]) was applied from a set of 10 adaptation sentences for each speaker. Figure 2 shows the results by language models and n-grams (2-gram and 6-gram), with and without adaptation of acoustic models.

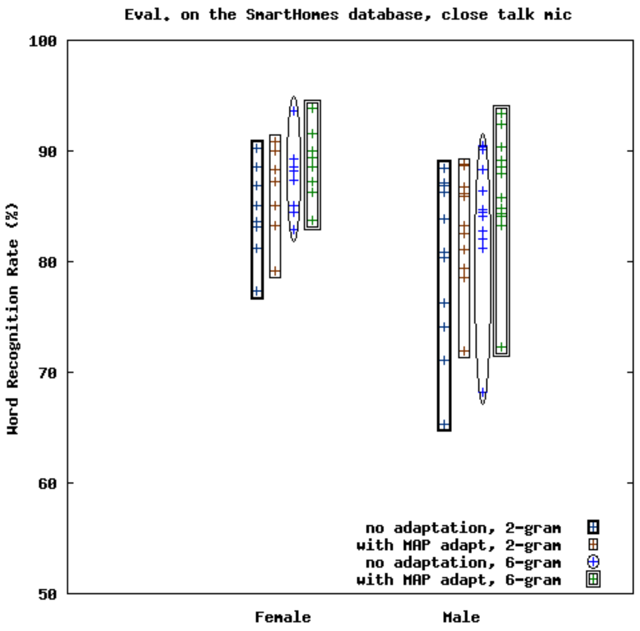

Fig. 2. Evaluations of 20 elderly speakers (lavalier microphone).

Improvements are achieved through the hidden Markov models adaptation and the language model precision. We also see that not all users test the recognition system with the same success rate. Female speakers tend to provide higher recognition rates than male speakers, this is a phenomena which has been previously studied by [3] with French and English databases. The female speech is considered to be more difficult to process (the higher fundamental frequency and the typical shorter vocal tracts of female voices leads to higher formant frequencies and therefore to shorter useful bandwidth, making pitch extraction and formant measurements harder to extract), however the male speakers produce more disfluencies (e.g. filled pauses, sloppy pronunciations and repetitions) while speaking.

VI. CONCLUSIONS AND PERSPECTIVES

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. The Asterisk server is ready for testing services related to usage scenarios listed in Section 2.

ACKNOWLEDGMENT

The thesis of Daniel R. S. Caon is partly funded by the European project CompanionAble (FP7/2007-2013, Grant Agreement No. 216487).

REFERENCES

[1] *Asterisk - The Open Source Telephony Projects.* Available online: https://www.asterisk.org/

[2] AS. Rigaud et al. *Un exemple d’aide informatisée 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 dition/Elsevier, ScienceDirect, (April 2010).

[3] Adda-Decker, M.; Lamel, L. *Do speech recognizers prefer female speakers?* In Proceedings of the InterSpeech Conference (Interspeech ’05), pp. 2205-2208, Lisbon, Portugal, (September 2005);

[4] 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);

[5] Bayeh, R. *Reconnaissance de la Parole Multilingue: Adaptation de Modèles Acoustiques Multilingues vers une langue cible.* Thése (Doctorat) TELECOM Paristech, (2009);

[6] Baldinger, J.-L. et al. *Tele-surveillance system for patient at home: The mediville system.* In: MIESENBERGER, K. et al. (Ed.). Computers Helping People with Special Needs. Springer Berlin / Heidelberg, 2004, (Lecture Notes in Computer Science, v. 3118). p. 623-623. 10.1007/978-3-540-27817-7_59. http://dx.doi.org/10.1007/978-3-540-27817-7_59

[7] Bush, Vannevar *As We May Think.* The Atlantic Monthly. Volume 176, No. 1; pages 101-108. (July, 1945);

[8] 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);

[9] Chollet, G.; Caon, D. R. S.; Simonnet, T.; Boudy, J.. vAssist: Le Majordome des personnes dépendantes. In: 2e Conférence Internationale Sur l'accessibilité et les systèmes de suppléance aux personnes en situations de handicap, Paris - France (2011).

[10] Clement, N.; Tennant, C.; 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

[11] Constantinescu, A.; 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);

[12] Jean-Luc Gauvain and Chin-Hui Lee. *MAP estimation of continuous density HMM: theory and applications.* In Proceedings of the workshop on Speech and Natural Language (HLT '91). Association for Computational Linguistics, Stroudsburg, PA, USA, 185-190. DOI=10.3115/1075527.1075568 http://dx.doi.org/10.3115/1075527.1075568 (1992)

[13] Gitlin LN, 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/

[14] http://nomemoryspace.wordpress.com/2008/03/17/la-prothese-memorielle-limpact-de-la-publication-des-historiques-sur-la-societe-de-linternet/

[15] Jim Gemmell, Gordon Bell and Roger Lueder, *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/

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

[17] Mokbel, C. *Reconnaissance de la parole dans le bruit: bruitage/débruitage.* PhD Thesis - Ecole Nationale Supérieure des Télécommunications, 1992.

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

[19] Young, S. J. et al. *The HTK Book*, version 3.4. Cambridge, UK: Cambridge University Engineering Department, (2006).