Proceedings of

eNTERFACE'14

10th International Summer Workshop on Multimodal Interfaces

Training School

Bilbao, June 9th – July 4th 2014

Editors:

Daniel Erro, Inma Hernáez

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

AHOLAB
Signal Processing Laboratory

Forewords

After nine successful editions – Mons 2005, Zagreb 2006, Istanbul 2007, Paris 2008, Genoa 2009, Amsterdam 2010, Pilsen 2011, Metz 2012, and Lisbon 2013 – eNTERFACE came to Spain for the first time in 2014. Organized by Aholab research group and hosted by the Faculty of Technical Engineering of the University of the Basque Country in Bilbao, the 10th International Summer Workshop on Multimodal Interfaces – eNTERFACE'14 – took place from June 9th to July 4th 2014. Also for the first time, given the special support from the International Speech Communication Association (ISCA), it was distinguished as ISCA Training School.

Once again, eNTERFACE was a unique opportunity for students and experts all over the world to meet and effectively work together, so as to foster the development of tomorrow's multimodal research community. It gathered about 60 researchers coming from institutions in 13 different countries, not only from Europe but also from Singapore, China, Iran, etc. Thanks to the sponsorship of ISCA and the European Association for Signal Processing (EURASIP), the program included five keynote presentations by Nick Campbell, Anton Nijholt, Juan M. Montero, Rubén San-Segundo and Mark J. F. Gales.

The projects undertaken by the eNTERFACE'14 attendees were the following:

1. Auracle - How are the salient cues situated in audiovisual content?
2. A character animation authoring platform for the 99%
3. Exploring stylistics of expert gestures and singing voice with motion capture and reactive statistical mapping
4. MILLA - Multimodal interactive language learning agent
5. ZureTTS - Online platform for obtaining personalized synthetic voices

After four weeks of intense collaborative work seasoned with diverse social activities, all these projects resulted in promising results, showy demonstrations and/or fully operative systems, most of which are reported later in this document.

The organizers of eNTERFACE’14 would like to express their gratitude to all the people who made this event possible and fruitful: the project leaders and co-leaders, for their valuable proposals and their undeniable commitment to carry them out; all the attendees and their funding institutions, for their participation and their fantastic attitude; the official sponsors, for their generous support; the invited lecturers, for their willingness to come to Bilbao in June; the members of the steering committee (Albert, Antonio, Benoit, Bülent, Christophe, Gualtiero, Igor, Milos, Olivier, Thierry and Yves), for their advice and collaboration in the project selection process; Alicia, David, Loreto, Paula and Sofía, for taking care of the social dimension of this event. It was a great pleasure to meet you all and build together this 10th edition of eNTERFACE.

Daniel Erro

Chairman of eNTERFACE’14

eNTERFACE’14 Sponsors

EURASIP

eman ta zabal zazu

Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

Escuela
Universitaria de
Ingenieria Técnica
Industrial de Bilbao

Bilboko Industria
Ingeniaritza
Teknikoko
Unibertsitate Eskola

AHOLAB

Signal Processing Laboratory

BFA
DFB

Bizkaiko Foru Aldundia
Diputación Foral de Bizkaia

RTH

ikerbasque

Basque Foundation for Science

metro bilbao

BilbaoTurismo
& Convention Bureau

eNTERFACE Steering Committee

Albert Ali Salah, University of Amsterdam

Antonio Camurri, University of Genova

Benoit Macq, Université Catholique de Louvain

Bülent Sankur, Bogazici University

Christophe d'Alessandro, CNRS-LIMSI

Daniel Erro, University of the Basque Country - Ikerbasque

Gualtiero Volpe, University of Genova

Igor Pandzic, Zagreb University

Inma Hernáez, University of the Basque Country

Milos Zelezny, University of West Bohemia

Olivier Pietquin, Metz Supélec

Thierry Dutoit, Faculté Polytechnique de Mons

Yves Rybarczyk, New University of Lisbon

eNTERFACE'14 Local Committee

General chair: Daniel Erro

Co-chair: Inma Hernáez

Administration: Eva Navas

Technical support: Ibon Saratxaga

Infrastructure: Begoña Blanco

Web management: Agustín Alonso

Outline

Project Title *Page*

Auracle: how are salient cues situated in audiovisual content?

Christian Frisson, Nicolas Riche, Antoine Coutrot, Charles-Alexandre Delestage,
Stéphane Dupont, Onur Ferhat, Nathalie Guyader, Sidi Ahmed Mahmoudi, Matei Mancas,
Parag K. Mital, Alicia Prieto Echániz, François Rocca, Alexis Rochette, Willy Yvart

***Abstract*—The Auracle project aimed at investigating how sound alters the gaze behavior of people watching moving images, using low-cost opensource systems and copyleft databases to conduct experiments.**

We created a database of audiovisual content comprising: several fragments of movies with stereo audio released under Creative Commons licenses, shorts with multitrack audio shot by participants, a comic book augmented with sound (events and music). We set up a low-cost experimental system for gaze and head tracking synchronized with audiovisual stimuli using the Social Signal interpretation (SSI) framework.

We ran an eye tracking experiment on the comic book augmented with sound with 25 participants. We visualized the resulting recordings using a tool overlaying heatmaps and eye positions/saccades on the stimuli: CARPE. The heatmaps qualitatively observed don't show a significant influence of sound on eye gaze.

We proceeded with another pre-existing database of audiovisual stimuli plus related gaze tracking to perform audiovisual content analysis in order to find correlations between audiovisual and gaze features. We visualized this exploratory analysis by importing CARPE heatmap videos and audiovisual/gaze features resampled as audio files into a multitrack visualization tool originally aimed at documenting digital performances: Rekall.

We also improved a webcam-only eye tracking system, CVC Eye Tracker, by porting half of its processing stages on the GPU, a promising work to create applications relying on gaze interaction.

***Index Terms*—Eye tracking, multimodal recording, saliency, computational attention.**

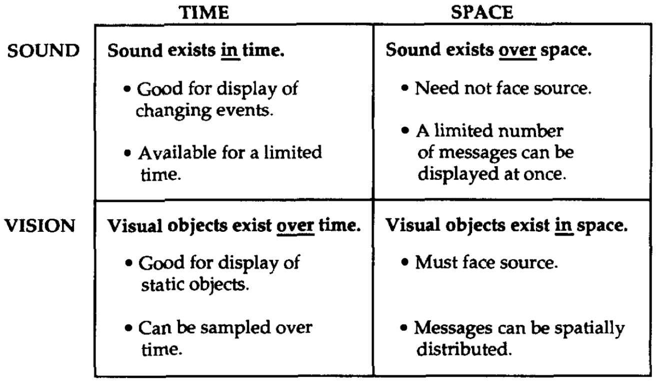
I. Introduction

Our research question is to find if gaze tracking analysis can help to understand whether or not sound influences vision when watching audiovisual content. We will first introduce facts about the sound/image relation (I-A), eye tracking (I-B) and other signals that can help to answer this question (I-C). In Section II we describe the two gaze recording setups that we put together or optimized. In Section III we illustrate the databases we used for the experiments and for the analysis. In Section IV we explain our experimental protocol for gaze recordings using an audiovisual comic book as stimulus. In Section V we report the techniques we employed to perform the analysis. In Section VI we illustrate the tools we used to produce visualization that supports our analysis.

A. The sound/image relation

The sound/image relation can be addressed by different standpoints: from the observations of the data analysts that study audio and visuals as media content or interaction channels; and from the perspective of audiovisual content creators.

Figure 1 compares the complementary modes of sound and vision across time and space as defined by Gaver for the design of human-computer interfaces [1]. It underlines the differences between audition and vision in human perception. In our case, studying the relation between sound and image implies to perform spatiotemporal analysis.



	TIME	SPACE
SOUND	**Sound exists <u>in</u> time.** • Good for display of changing events. • Available for a limited time.	**Sound exists <u>over</u> space.** • Need not face source. • A limited number of messages can be displayed at once.
VISION	**Visual objects exist <u>over</u> time.** • Good for display of static objects. • Can be sampled over time.	**Visual objects exist <u>in</u> space.** • Must face source. • Messages can be spatially distributed.

Fig. 1. Complementary modes of sound and vision by Gaver [1]

Electroacoustic music composer and professor Michel Chion has been studying the relationship between sound and image [2]. For movie analysis he coined new terms, for instance one that defines how space and time can be composed together:

Synchresis, an acronym formed by the telescoping together of the two words synchronism and synthesis: "The spontaneous and irresistible mental fusion, completely free of any logic, that happens between a sound and a visual when these occur at exactly the same time."

B. Eye tracking

Eye tracking is the measurement of the eye movements made by a viewer on a visual array, such as a real-world scene, a computer or a smartphone screen. This technique provides an unobtrusive, sensitive, real-time behavioral index of ongoing visual and cognitive processing. Eye tracking research has been ongoing for several decades, effervescent in the last, discovered a couple of centuries ago. Some predictions foresee it as the new trend in video game controllers, "hands-free". The price and form-factor size of these devices is plunging, about soon to be part of consumer-grade video game hardware and to be integrated in mobile devices, most probably through infrared pupil tracking [3].

Companies such as SensoMotoric Instruments (SMI) and Tobii provide research-grade commercial solutions that are very precise (less that 0.5° of error) and fast (samplerate in kHz) but however are very expensive (tens of thousands of Euros). Low-cost solutions with specific hardware are emerging, such as the Eye Tribe in 2003 (see Section II-B), less precise (between 0.5 and 1° of error) and fast (30 frames per second) but affordable for about hundred Euros. Low-cost solutions without specific hardware, solely based on webcams, have been developed in the research communities for a decade, such as CVC Eye-Tracker (see Section II-A) and ITU GazeTracker (evaluated by its authors [4] for usability testing versus higher-end solutions).

Besides the qualitative specification offered by dedicated hardware, another major advantage brought by expensive commercial systems is the comprehensive software tools that are provided with their product. Such tools usually cover the whole workflow for eye tracking experiments: device support, test and stimuli preparation, test recording, tests analysis with statistics. A few opensource tools providing such a fully-fledged solution exist, recently surveyed in [5], for instance OGAMA (OpenGazeAndMouseAnalyzer)[1] [6] released under a GPL license unfortunately only for Windows platforms and currently not supporting video stimuli.

C. Other signals

While the purpose of this project is to record a audiovisual database to analyze the correlation between user gaze and sound, we chose to record other signals to go deeper into the analysis. We opted for recording processed data from the Microsoft Kinect as the video signal and depth map. We have also recorded the head pose estimation and the facial "Action Units" [7] which are data for coding facial movements, for example, to know whether the user is smiling or neutral.

II. Setups

Motivated by exploring open source and low-cost solutions, we decided to investigate two eye-tracking systems. The first relying solely on a webcam is of interest to build interactive applications on mobile devices using built-in hardware (Section II-A). The second is as well low-cost but provides greater precision than the first setup for our experiments (Section II-B).

A. Setup 1: webcam-based eye tracker

Although infrared (IR) based eye-tracking techniques are used in almost all commercial eye-tracker devices and software, visible light methods pose an alternative that has the advantage of working with common hardware such as webcams. CVC Eye-Tracker [8] is an open source eye-tracking software[2] which is an improved version of Opengazer [3] and which requires no special equipment such as IR lights, IR cameras, etc.

The pipeline of the eye-tracker is shown in Fig. II-A. The application uses the images captured from the webcam to calibrate a gaze point estimator and then use this to output gaze estimations. The components of the system are:

- **Point Selection:** 8 facial feature points are selected automatically on the subject's face. A combination of Haar cascades, geometrical heuristics and a novel eye-corner detection technique is used to choose the points.
- **Point Tracking:** The points selected in the first step are tracked over the subsequent webcam frames, using a combination of optical flow and 3D head pose based estimation.
- **Calibration:** The regions containing the left and right eyes are extracted and used to calibrate a Gaussian Process (GP) estimator for the gaze point.
- **Gaze Estimation:** The left and right eye images extracted from the latest webcam frame are mapped to the screen coordinates using the GP estimator.

A key variable that has to be taken into account in real time video processing applications is the computation time, which can be so elevated in case of processing high definition (HD, Full HD, etc.) videos. Therefore, we developed a version of the eye-tracker that exploits the high computing power of graphic processing units (GPUs). These components dispose of a large number of computing units, which can be very well adapted for parallel computation. Our GPU implementation is applied on the computationally most intensive steps of the webcam-based eye tracker (Fig. II-A):

- Eyes, nose and frontal face detection
- Optical flow-based points tracking

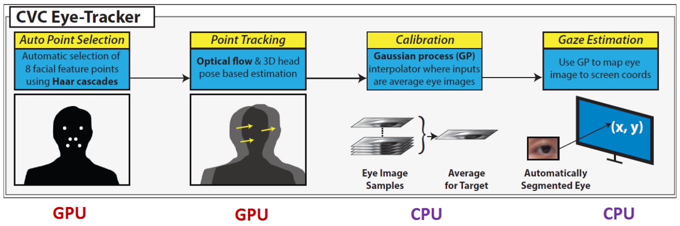

CVC Eye-Tracker
Auto Point Selection
Automatic selection of 8 facial feature points using Haar cascades
Point Tracking
Optical flow & 3D head pose based estimation
Calibration
Gaussian process (GP) interpolator where inputs are average eye images
Gaze Estimation
Use GP to map eye image to screen coords
Eye Image Samples
Average for Target
Automatically Segmented Eye
(x, y)
GPU
GPU
CPU
CPU

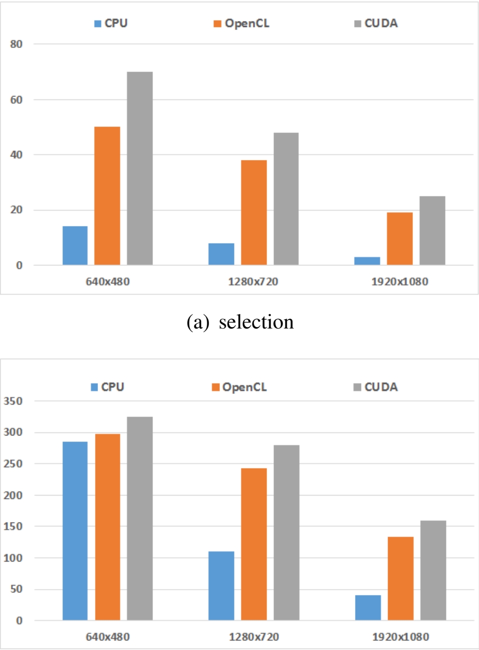
Fig. 2. GPU-based Eye tracker

The proposed implementation can exploit both NVIDIA and ATI graphic cards, based on CUDA[4] and OpenCL[5]. The CUDA version consists of selecting a number of GPU threads so that each thread can perform its processing on one or a group of pixels in parallel. More details about this process and the applied optimization techniques can be found in [9], [10].

Otherwise, the OpenCL implementation is based on the same process, but using a specific syntax related to OpenCL. The main advantage of OpenCL is its compatibility with both NVIDIA and ATI graphic cards, as it was proposed as a standard for GPU programming. However, CUDA, which allows to program NVIDIA cards only, offers better performances (Fig. 3) thanks to its adapted programming architecture.

[1] OGAMA: http://www.ogama.net
[2] CVC Eye-Tracker: http://mv.cvc.uab.es/projects/eye-tracker
[3] OpenGazer: http://www.inference.phy.cam.ac.uk/opengazer/
[4] CUDA. http://www.nvidia.com/cuda
[5] OpenCL.http://www.khronos.org/opencl

Fig. 3 compares performance between CPU, CUDA and OpenCL implementations (in terms of fps) of points (eyes, nose and frontal face) detection and optical flow-based tracking. These accelerations allowed to improve the process of webcam-based eye tracker with a factor of 3x. As result, our GPU-based method allows real time eyes tracking with high definition videos.



(a) selection

(b) tracking

Fig. 3. CUDA and OpenCL-based points selection and tracking performance

This setup does not yet provide a sufficient accuracy for performing experiments with gaze recording. In next section we describe a low-cost system that we used for our experiments.

B. Setup 2: low-cost eye tracker and depth camera

The goal of the second setup is to capture different data from a user analysis with synchronized stimuli. The recording of this data was done using the Social Signal interpretation (SSI) framework[6] developed at Augsburg University. The SSI framework gives different tools to record and analyze human behavior through sound, video, and a large range of commercial sensors [11].

In this setup, we combined a web camera, the Eye Tribe [7](a low cost eye tracking device) and the Microsoft Kinect. The utility of the webcam is to apply the CVC Eye-Tracker on the video. CVC Eye-Tracker can also be applied on the video stream from the Kinect, but with a smaller resolution. As stimuli, a video player and a comics book player are linked to the framework to obtain sensor data correctly synchronized. SSI also contains processing modules to filter and extract features from the recording signals. For example, it allows to obtain the head pose estimation and the Animated Units from the Kinect through the Kinect SDK.

The use of SSI requires writing a pipeline in XML. This pipeline contains all the structure of the system to obtain the synchronized date at the end. The XML pipeline contains 4 different parts. The first part is the sensor declaration. In this part all the sensors with the raw data extract from each sensors are declared. The second part contains filter to prepare and combine data for the visualization of the third part. The last part contains all parameters for the data storage.

Figure II-B shows the monitoring display of our pipeline recording all the sensors in sync. The next sections provide more information about the plugins implementing access to sensors and display of stimuli.

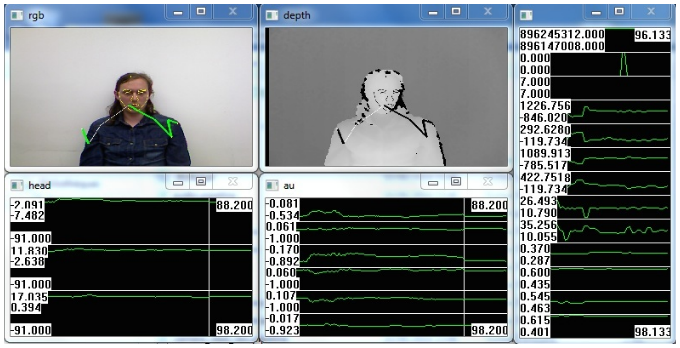


Fig. 4. Visualization of multimodal recording in SSI: Kinect RGB video (top-left), depth map video (top-middle), head pose (bottom-left), action units (bottom-middle) and Eye Tribe signals (right)

1) *Webcam:* The aim is to know the reactions of the user facing an audio-visual stimuli, it is normal shoot and record the user. This can easily be done using a webcam. In this case we used a 720p webcam at 30 FPS. The video was recorded using the FFMPEG Plugin from SSI

2) *Kinect:* To go further in the user behavior analysis, there should be an analysis of the user's face. The Kinect SSI plugin gives access to several metrics available from the Kinect SDK. Data extracted from the Kinect in this setup are the RGB Image with depth map, the head pose estimation, the user skeleton in seated mode and the facial animation unit. The Kinect sensor contains two CMOS sensors, one for the RGB image (640 x 480 pixels at 30 fps) and another for the infrared image from which the depth map is calculated, based on the deformation of an infrared projected pattern [12].

The main use of the Kinect is the user skeleton tracking. Skeletal Tracking is able to recognize users sitting. To be correctly tracked, users need to be in front of the sensor, making sure their head and upper body are visible (see Figure II-B2). The tracking quality may be affected by the image quality of these input frames (that is, darker or fuzzier frames track worse than brighter or sharp frames).

The Kinect head pose estimation method returns the Euler rotation angles in degrees for the pitch, roll and yaw as described in Figure II-B2, and the head position in meters relatively to the sensor being the origin for the coordinates.

From the face are extracted: the neutral position of the mouth, brows, eyes, and so on. The Action Units (AU) represent the difference between the actual face and the neutral face. Each AU is expressed as a weight between -1 and +1.

[6] SSI: http://www.openssi.net

[7] The Eye Tribe: http://theeyetribe.com

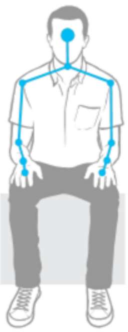


Fig. 5. User tracking in seated mode

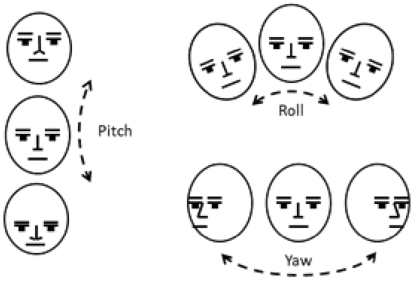


Fig. 6. Three different degrees of freedom: pitch, roll and yaw [13]

3) Eye Tribe: gaze + calibration: At the beginning of the project the data streamed by the Eye Tribe was partially recorded by SSI. We enhanced the existing Eye Tribe plugin to allow SSI to record all the gaze calibration and data provided by the Eye Tribe. The complete Eye Tribe's data description is available on their website[8]. Each participant had to perform the Eye Tribe calibration process before each stimuli type. During the calibration process SSI was recording the gaze data frame stream and the calibration data frame event. We recorded the gaze stream to have practical calibration gaze points and eventually recheck the calibration afterwards. After the launch of the calibration recording SSI pipeline we used the Eye Tribe server for the calibration which comes with Eye Tribe and consists of circular targets appearing on random grid positions on the screen. The announced Eye Tribe precision is around 0,5° to 1° which gives an average precision of 0,5cm to 1cm on the screen in practical. Practically, if the participant moves her/his head, the calibration is lost. The participant relative position to the screen and tracker is also important for a good tracking quality. The Eye Tribe and the Kinect are both using infrared illumination, these have to be properly placed to limit their interference. We used the Eye Tribe at 30 frames per second, our reference framerate for the recordings.

4) ffmpeg for movie stimuli: We wanted to get the sensors recorded in sync with the movie stimuli with two constraints: frame-accuracy and the ability to randomly sequence movies.

The SSI core plugin for video reading which uses the ffmpeg framework didn't provide the frame timing information. We modified the SSI plugin to log the frame numbers.

The ffmpeg plugin is not suitable for sequencing many video stimuli inside one pipeline. SSI provides a tool for stimuli named ModelUI that resembles a slideshow authoring tool, but not allowing frame-accurate synchronization between video stimuli (when embedded as slideshow elements) and the sensors pipeline. We had to manually generate random sequences of video stimuli by concatenating fragments (using the ffmpeg library). We experienced frame dropping through this process.

5) UDP socket client logging events from the comic book player: For the SSI synchronization of the comic book player, we simply recorded the comic book's logs through a local UDP connection. SSI provides a core plugin that reads UDP frames as events and record them in a .events xml format file. The recorded logs were: comic pages and their timestamp; sounds status (played or stopped) and type (music/effect/ambience) and sequencing option (looped or triggered once).

6) Discussion: SSI is a fully-fledged solution beyond synchronized recording providing realtime machine learning capabilities on multimodal signals.

The strengths of SSI are its availability as opensource project under a GPL license and its vast number of plugins for supporting many devices.

SSI also comes up with several drawbacks, regarding calibration and video stimuli synchronization. One calibration for both trackers (eye and kinect) was required per SSI pipeline, so we wanted to be able to launch a single pipeline per participant to facilitate the tests. From our experience with the ffmpeg plugin (see Section II-B4), we decided to drop video stimuli for the experiments (see Section IV) and choose only the interactive comic book as stimulus.

The SSI distribution (binaries and source code) that we adapter (with modified ffmpeg and Eye Tribe plugins) and used for the experiments is available on a github repository[9].

III. Database

A. State-of-the-art of audiovisual and gaze datasets

During the last decade, an exponentionally increasing number of computational visual saliency model has been proposed [14]. To be able to fairly evaluate and compare their performance ([15], [16]), over a dozen of videos datasets annotated with eye tracking data has been publicly released[10] [17]. Yet, aside from a few papers [18], [19], [20], [21], [22] authors never mention the soundtracks or explicitly remove them, making participants look at silent movies which is far from natural situations.

Here, we propose a freely available database of audiovisual stimuli and related human fixations, allowing to validate models on the relation between image and sound, with diverse genres beyond movies, documentaries and broadcasts, such as animation [23] and video games, focusing on special directing and sound design techniques [24].

[8] Eye Tribe API: http://dev.theeyetribe.com/api/

[9] SSI Auracle: http://github.com/eNTERFACE14Auracle/AuracleSSI

[10] Eye-Tracking databases repertory: http://stefan.winkler.net/resources.html

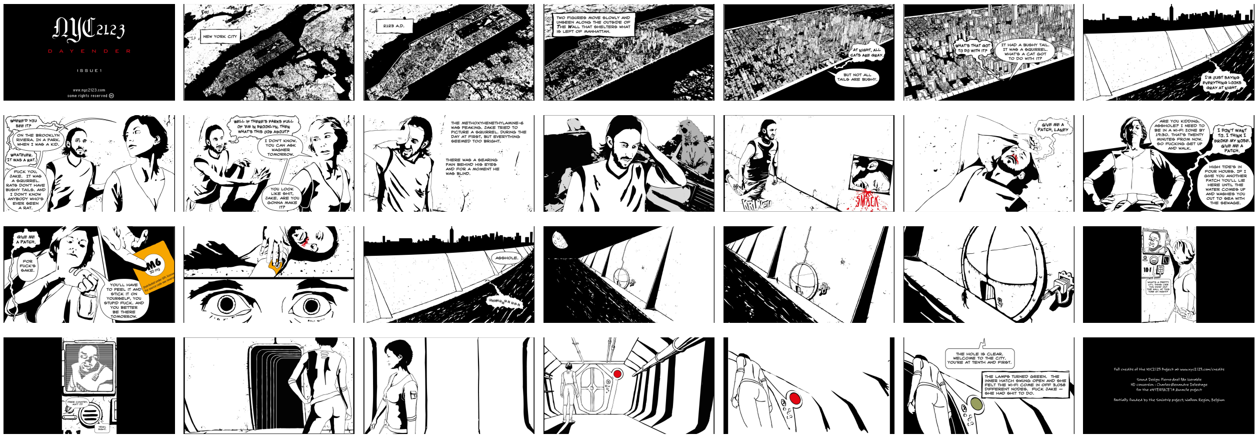

NYC2123
DAYENDER
ISSUE1
www.nyc2123.com
some rights reserved
NEW YORK CITY
2123 A.D.
TWO FIGURES MOVE SLOWLY AND UNSEEN ALONG THE OUTSIDE OF THE WALL THAT SHELTERS WHAT IS LEFT OF MANHATTAN.
AT NIGHT, ALL CATS ARE GRAY.
BUT NOT ALL TAILS ARE BUSHY.
WHAT'S THAT GOT TO DO WITH IT?
IT HAD A BUSHY TAIL. IT WAS A SQUIRREL. WHAT'S A CAT GOT TO DO WITH IT?
I'M JUST SAYING EVERYTHING LOOKS GRAY AT NIGHT.
WHERE'D YOU SEE IT?
ON THE BROOKLYN RIVIERA. IN A PARK. WHEN I WAS A KID.
WHATEVER. IT WAS A RAT.
FUCK YOU, JAKE. IT WAS A SQUIRREL. RATS DON'T HAVE BUSHY TAILS. AND I DON'T KNOW ANYBODY WHO'S EVER SEEN A RAT.
WELL IF THERE'S PARKS FULL OF 'EM IN BROOKLYN, THEN WHAT'S THIS JOB ABOUT?
I DON'T KNOW. YOU CAN ASK WAGNER TOMORROW.
YOU LOOK LIKE SHIT, JAKE. ARE YOU GONNA MAKE IT?
THE METHOXYHENETHYLAMINE-6 WAS PEAKING. JAKE TRIED TO PICTURE A SQUIRREL. DURING THE DAY AT FIRST, BUT EVERYTHING SEEMED TOO BRIGHT.
THERE WAS A SEARING PAIN BEHIND HIS EYES AND FOR A MOMENT HE WAS BLIND.
KRUNCH
SMACK
GIVE ME A PATCH, LANEY
ARE YOU KIDDING, ASSHOLE? I NEED TO BE IN A WI-FI ZONE BY 19:30. THAT'S TWENTY MINUTES FROM NOW. SO FUCKING GET UP AND WALK.
I DON'T WANT TO. I THINK I BROKE MY NOSE. GIVE ME A PATCH.
HIGH TIDE'S IN FOUR HOURS. IF I GIVE YOU ANOTHER PATCH YOU'LL LIE HERE UNTIL THE WATER COMES UP AND WASHES YOU OUT TO SEA WITH THE SEWAGE.
GIVE ME A PATCH.
FOR FUCK'S SAKE.
M6
30 mg
YOU'LL HAVE TO PEEL IT AND STICK IT ON YOURSELF, YOU STUPID FUCK. AND YOU BETTER BE THERE TOMORROW.
ASSHOLE.
UHHHHHHHHHH
10-1
WHAT'S A PRETTY LIT'L THING LIKE YOU DOIN' OUT THE WALL AT THIS TIME UH NIGHT?
FREE COUNTRY, AIN'T IT?
YEAH, RIGHT.
THE HOLE IS CLEAR. WELCOME TO THE CITY. YOU'RE AT TENTH AND FIRST.
THE LAMPS TURNED GREEN. THE INNER HATCH SWUNG OPEN AND SHE FELT THE WI-FI COME IN OFF 3,058 DIFFERENT NODES. FUCK JAKE — SHE HAD SHIT TO DO.
Full credits of the NYC2123 Project at www.nyc2123.com/credits
Sound Design: Pierre-Axel PAx Izerable
HD conversion : Charles-Alexandre Delestage
for the eNTERFACE'14 Auracle project
Partially funded by the Sonixtrip project, Walloon Region, Belgium

Fig. 7. Frames of the NYC2123 comic book reworked with less text, as audiovisual stimulus for the experiments

B. The Auracle database

1) Movies with multitrack audio: To provide material for the experiments following our initial plans to include movie stimuli, 8 excerpts from Creative Commons movies were selected for the added values sound offered against audio, with an access to the source files in order to modify the sound channels. We also wrote and shot 5 original shorts with a cinema-like aspect. Each of these had separate audio channels for music, ambience and effects, so the sound mixing could be changed during the experiments using a control surface.

These movies could unfortunately not be used as stimuli for experiments due to our issues with implementing a suitable SSI pipeline with frame-accurate synchronization between video stimuli and sensor recording (see Section II-B4).

2) A comic book augmented with sound: We used a digital comic book with sound effects and music as a stimuli. We wanted that all the material (comic book, sound effects, music) would be released under a Creative Commons license. We adapted the source files of the NYC2123 comic book[11] to make all pages match a single orientation and resolution. Pierre-Axel Izerable produced the sound design and provided the audio files and a description of their sequencing. We assembled the visuals and the sounds together in a prototype with which sounds can be looped or played once, in transition between two frames. Participants can read the comic book at their own pace, the prototype records the slide transitions as well as if an audio file is played, stopped, or looped; the type of audio content of the file (ambient, sound effect or music). This information is also sent to the SSI recorder through an UDP connection so it can be synchronized with the sensors data. After some testing we decided to remove some large framed text from the original comic because the participants were taking a huge proportion of the experiment duration to read those and the reading path and time were not the gaze features we wanted to extract. Fig. 7 illustrates these reworked frames containing less text. An HTML5 version of the NYC2123 comic book augmented with sound is available online[12].

IV. Experiments

A. Participants

25 participants agreed to take an experiment consisting in watching the comic book while their gaze was being monitored. Participants were split into two groups: one group experienced the comic book augmented with sound, the other group read the comic book without soundtrack. Figure IV-A shows the apparatus used for the experiments.

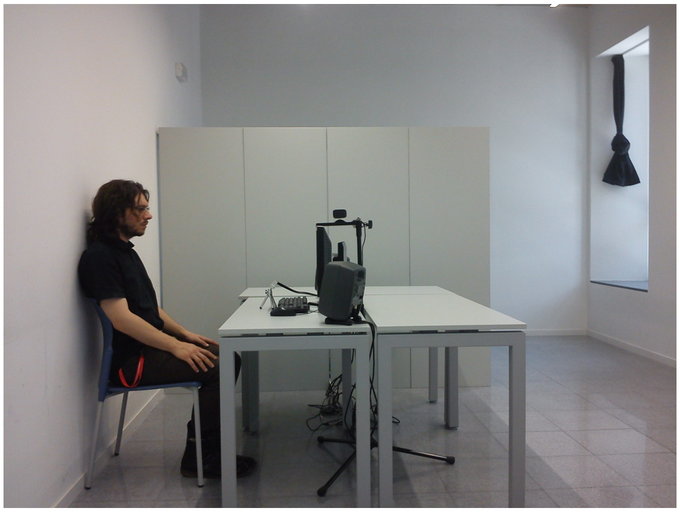


Fig. 8. Apparatus used for the experiments

Participants didn't receive financial reward but were offered Belgian chocolate.

We used the opensource survey management application LimeSurvey[13] to implement and collect pre- and post-experiment questionnaires. The pre-experiment questionnaire allowed to collect qualitative data: sex, spoken language(s). Among the 25 participants, 6 were females and 19 males. 6 reported English as their native language. 12 reported to suffer from visual impairment: 11 included left/right eye myopia/presbyopia or strabismus; and 1 color blindness. 3 reported to suffer from hearing impairment, including slight loss in bass/medium/treble frequency range or tinnitus.

[11] NYC2123 comic book: http://nyc2123.com

[12] Auracle NYC2123 Issue 1 augmented with sound: http://github.com/eNTERFACE14Auracle/AuracleNYC2123Issue1

[13] LimeSurvey: http://www.limesurvey.org

B. Collected data

For each participant, data was collected in files containing:

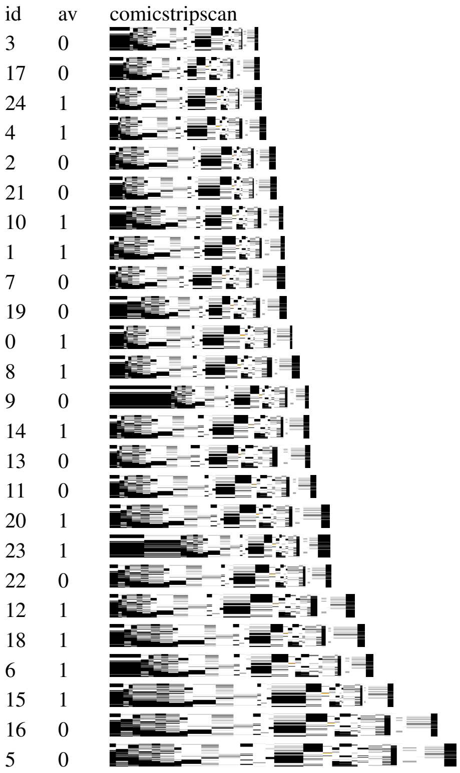
- events from the comic book player with timestamps: next comic book frame, sound start/loop/end;
- streams from the trackers: Kinect RGB/depthmap videos and head pose and action units, EyeTribe eye fixation status and raw/average eye positions;
- calibration status from the Eye Tribe including an estimation of its quality on a [0;5] scale.

C. Preliminary analysis on reading times

In this Section, we perform analysis on temporal aspects: reading times that can be grouped by participant, by comic book frame, and by condition (with or without sound).

1) User-defined page switching pace complicates experimental analysis: Participants could read the comic book at their own pace, they could decide when to switch to the next page manually by pressing the right arrow key on a keyboard.

Table I visualizes the variations in reading time between participants and on each comic book frame. For each participant, a replay video was generated with a fixed framerate, along the time spent reading each comic book frame. Each of these replay videos was summarized into a slit-scan (by concatenating horizontally a 1-pixel-wide central vertical column extracted in each frame).



id	av	comicstripscan
3	0	
17	0	
24	1	
4	1	
2	0	
21	0	
10	1	
1	1	
7	0	
19	0	
0	1	
8	1	
9	0	
14	1	
13	0	
11	0	
20	1	
23	1	
22	0	
12	1	
18	1	
6	1	
15	1	
16	0	
5	0	

TABLE I
COMICSTRIPSCANS ORDERED PER DURATION AGAINST PARTICIPANT ID AND AV CONDITION (WITH OR WITHOUT SOUND)

This illustrates how our stimuli have a variable duration depending on the browsing pace desired by each participant, what makes the inter-subject analysis more difficult.

Figure 9 visualizes times per frame grouped by user. Besides users 5, 6, 15 and 16 (at the bottom of Table I), most users read most frames in between 2 and 10 seconds, except for some frames considered as outliers. Analyzing the effect of frame content on reading times may help to further understand this.

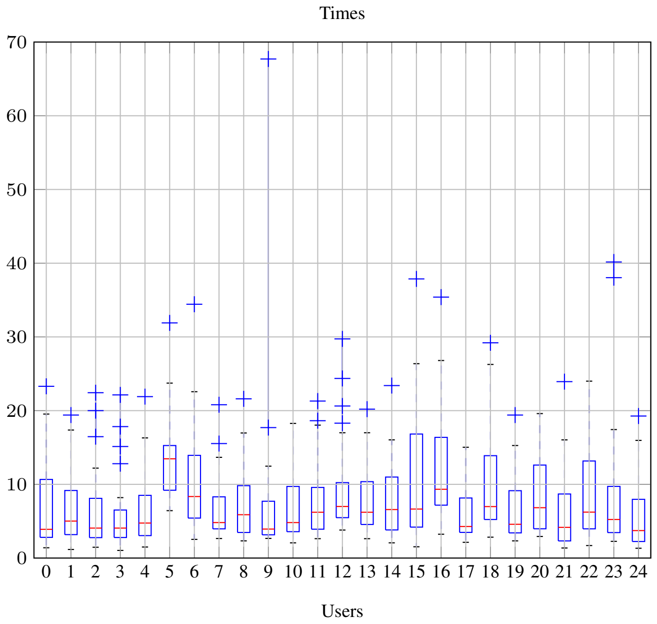


Fig. 9. Times (s) per frame grouped by user

2) Influence of frame content on reading times: Figure 10 visualizes times per user grouped by frame. Reading times are clearly greater for frames containing much text (7-9, 13-14, 26). Further analysis may be undertaken to correlate text length and complexity with reading times.

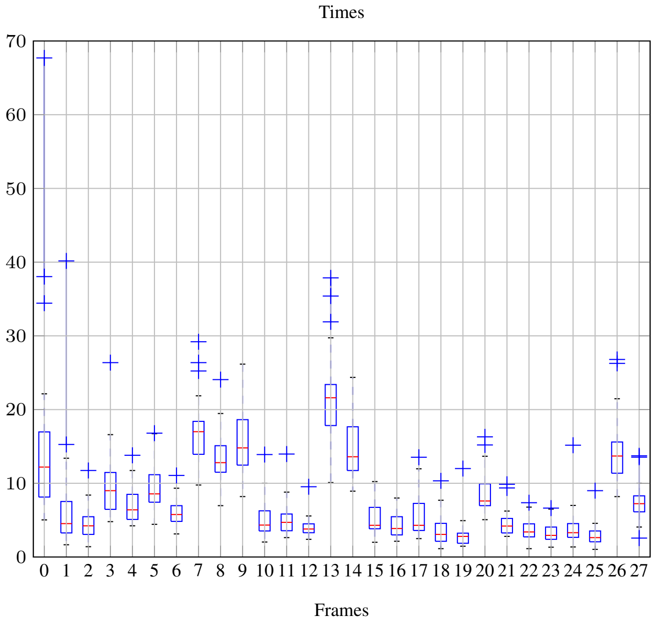


Fig. 10. Times (s) per user grouped by frame

3) Influence of the audio condition (with/without soundtrack) on reading times: The Shapiro-Wilk test statistic (W) and its p-value (p) can be computed to to test the null hypothesis whether reading times (for all frames and participants) grouped by both conditions (with and without sound) are taken from a normal distribution [25]. Times "with sound" (W=0.82, p=1.15e-19) and "without sound" (W=0.77, p=1.55e-21) are not from a normal distribution (the null hypothesis is rejected).

We want to compare two unpaired groups (different amount of participants and varying times per frame) whose distribution normality is not assumed, therefore we choose to compute the Mann-Whitney test statistic (u) and its p-value (p) [26]. With u=6.11e5 and p=0.49, we can assume that there is no signifiant difference in reading times per frame per participant between both groups (with/without audio), what Figure 11 illustrates.

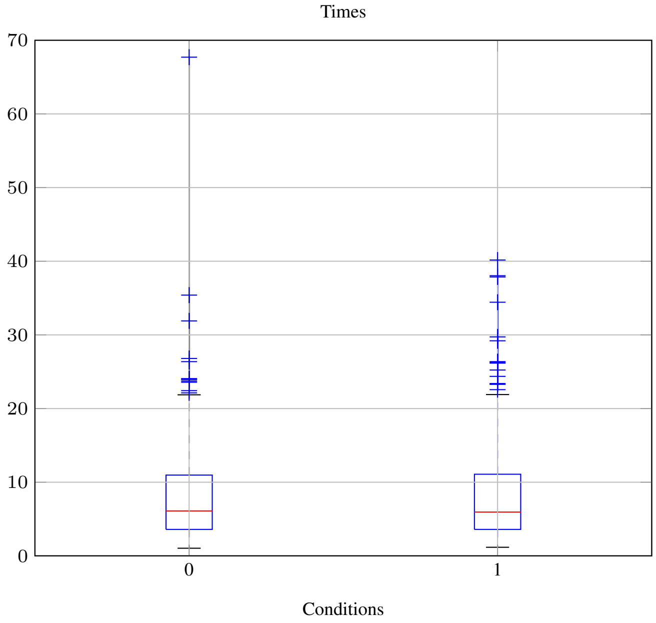


Fig. 11. Times (s) per user per frame grouped by condition (0: without sound, 1: with sound)

This experiment could only be undertaken during the last week of the workshop, after facing technical difficulties on synchronized recording and thus deciding to simplify the test protocol discarding video stimuli. We used a pre-existing database of video stimuli and related gaze recordings by Coutrot el al. [19] to test and choose techniques for analysis (Section V) and visualization (Section VI) .

V. Analysis

In this section, we report how audiovisual content and gaze recordings can be analyzed; and how the analyses of both can be correlated.

A. Audiovisual content analysis

Audio content can be described by a collection of commonly employed musical information retrieval features: the first 4 central moments of the real portion of a fast Fourier transform, spectral flux, entropy, first 13 MFCCs, and their flux. These are computed using the default parameters from MIRtoolbox [27]. Alternative methods could be employed which attempt to describe the content by its latent feature dimensions [28].

Regarding visual features, we compute optical flow, shown to correlate significantly with the entropy of eye-movements from multiple participants during dynamic scene viewing [29], [30]. This feature describes a vector of motion for each pixel for every frame. The resulting vectors are binned in a 360 degree histogram for each frame, creating a 360-element feature vector.

B. Gaze recordings analysis

We decided to describe gaze by the dispersion of the eye-movements, following the tracks of Coutrot et al [19]. Dispersion is defined as the mean of the Euclidean distance between the eye positions of different observers on a given frame.

C. Correlations between audiovisual/gaze features

Correlating two multivariate features can be done by a canonical correlation analysis. This method attempts to describe either multivariate feature as a linear combination of the other. By reducing the data to single linear combination, we can produce a single metric describing how well one domain is described by the other. The resulting feature can be used in a program such as Rekall (see Section VI-B) for alternative methods of linear editing, producing meaningful thumbnails, or simply providing the editor another cue for understanding their media. When dealing with larger collections, this feature could possibly be used for align/sync audiovisual content.

VI. Visualization supporting analysis

Techniques for visualizing eye tracking have been elaborated since the 1950, as surveyed in the state-of-the-art report by Blascheck et al.[31]. Some of these visualization techniques are integrated into tools that allow exploratory analysis and interactive browsing, such as the recent and comprehensive tool ISeeCube by Kurzhals et al.[32]. Unfortunately, a majority of these tools are neither open source nor cross-platform. We opted for combining two opensource solutions: CARPE (Section VI-A) that computes heatmaps and scanpaths, and Rekall (Section VI-B) that allows to browse audio and video content through a multitrack timeline.

A. Gaze heatmaps and scanpaths with CARPE

CARPE [29] produces heatmaps overlaid on video and imports eye-tracking data stored in text files. Each eye-movement is convolved with a 2-degree isotropic Gaussian distribution. The resulting distribution is normalized and converted to a jet-colormap. The CARPE source code is available on a github repository[14]. We use CARPE for producing example visualizations of our data. A heatmap video rendered with CARPE with the gaze data recorded through the experiments on the interactive version of the NYC2123 comic book is available online[15].

[14] CARPE: https://github.com/pkmital/NSCARPE

[15] CARPE replay videos of Auracle NYC2123 Issue 1: https://archive.org/details/AuracleNYC2123CARPE

B. Multitrack audiovisual/gaze visualization with Rekall

Tools that allow exploratory browsing of audiovisual content are multimodal annotation tools (surveyed by some of the participants of the current project through a previous eNTERFACE project [33]) and non-linear audio/video editing tools and sequencers. Few of these tools are cross-platform, opensource and easy to adapt to our needs. Rekall [16] is a new opensource tool for browsing and annotating audiovisual content, designed for the documentation of live artistic performances [34]. It is released under an opensource license (CeCILL, similar to GPL). Until the end of the eNTERFACE'14 workshop, its first version was developed as a desktop application solely relying on the Qt framework[17], pre-computing and visualizing simple thumbnails such as audio waveforms and video keyframes. Rekall was thus a good candidate for our requirements (opensource, cross-platform, easy to adapt).

Figure 12 illustrates how Rekall can be adapted to visualize altogether a stimulus video with a heatmap of gaze fixations overlaid using CARPE and multitrack thumbnails of features (dispersion and audiovisual correlation) and video (keyframes). Such a prototype has been obtained throughout the workshop with a limitation: some of the files recorded through the experiments are converted to files supported by Rekall, rather than having implemented the computation of their visualization directly into Rekall. These are the heatmap videos pre-computed using CARPE, importable as video content by Rekall; and the audiovisual/gaze features resampled at a minimal standard audio sampling rate (22050 Hz) to be directly importable as audio content from which Rekall computes waveforms.

Next steps would include to adapt Rekall to support various and modular thumbnail techniques through a plugin architecture, so that a single standalone application can be used without external conversions.

VII. Conclusion and Future Work

Throughout this project we have learned by trial and error the current state of the art on: low-cost eye tracking, synchronized multimodal recording, opensource eye tracking visualization and analysis.

We applied this experimental pipeline to an experiment with 25 participants whose gaze was tracked while watching an interactive comic book augmented with sound. The collected data is not relevant for further analysis, therefore we decided not to distribute it. However the experiment can be replicated on different datasets using our system.

Further work is required to allow the analysis of the correlation between the direction of the head and eye tracking.

To better understand the zones of interest in each comic book frame, clues could be mined from visual analysis, such as *objectness* to track protagonists and objects in the scene [35], and text detection and recognition [36].

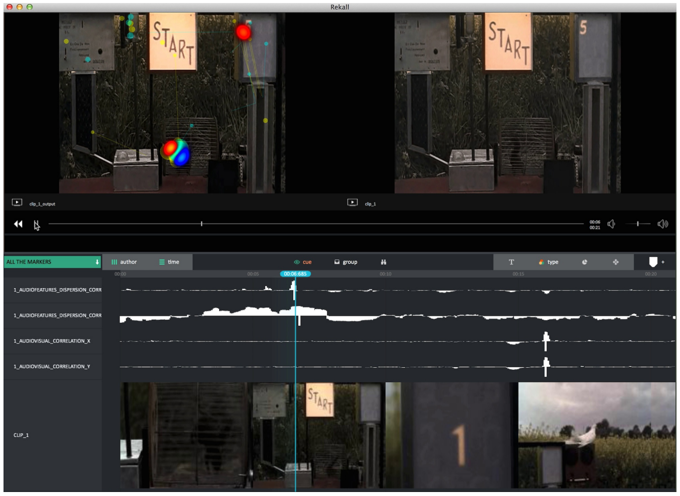

Rekall
clip_1_output
clip_1
00:06
00:21
ALL THE MARKERS
author
time
cue
group
type
00:00
00:05
00:06:685
00:10
00:15
00:20
1_AUDIOFEATURES_DISPERSION_CORR
1_AUDIOFEATURES_DISPERSION_CORR
1_AUDIOVISUAL_CORRELATION_X
1_AUDIOVISUAL_CORRELATION_Y
CLIP_1

Fig. 12. Visualization in Rekall of a stimulus video (top-right) with a heatmap of gaze fixations overlaid using CARPE (top-left) and multitrack thumbnails (below) of features (dispersion and audiovisual correlation) and video (keyframes)

Acknowledgment

Christian Frisson, Alexis Rochette and Stéphane Dupont have been partially funded by a grant from the Walloon Region of Belgium, GREENTIC program, entitled SonixTrip (number 1317970) (http://www.sonixtrip.be) for the preparation of this project and their participation to eNTERFACE'14.

Christian Frisson received a grant from the International Speech Communication Association (ISCA) to partially cover his expenses during his participation to the eNTERFACE'14 workshop (http://www.isca-speech.org).

Onur Ferhat's work was supported in part by the Spanish Gov. grants MICINN TIN2009-10435 and Consolider 2010 MIPRCV, and the UAB grants.

Sidi Ahmed Mahmoudi has been partially funded by the computer science department of the faculty of Engineering of UMONS. Sidi Ahmed Mahmoudi would like to thank the PPP Slowdio project for supporting his research activities.

François Rocca is supported by the Integrated Project LinkedTV (www.linkedtv.eu) funded by the European Commission through the 7th Framework Programme (FP7-287911).

The authors would like to thank the local organizers of eNTEFACE'14 for their kindness, their local assistance regarding logistics, their nice schedule of social events, their successful call of ISCA grants to support participating students.

The NYC2123 comic book (http://nyc2123.com) and its Adobe Illustrator source files created by Chad Allen and Paco Allen are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License.

The authors would like to thank Pierre-Axel Izerable for having produced the sound design for the comic book and accepted to release it under a Creative Commons Attribution-NonCommercial-ShareAlike 2.5 License.

The authors would like to thank Johannes Wagner for having granted them access to their bleeding-edge repository for the SSI framework [11].

[16] Rekall: http://www.rekall.fr – http://github.com/Rekall/Rekall

[17] Qt: http://qt-project.org

The authors would like to thank Jerome Urbain for his introduction and help on SSI pipeline creation.

The authors would like to thank Marc Thunissen from Océ Software Laboratories for the HTML5/CSS/JS code that allowed us to replicate the NYC2123 Issue 1 comic book augmented with sound.

Christian Frisson would like to thank the creators of Rekall for the insightful discussions they shared about their tool: Clarisse Bardiot, Guillaume Jacquemin, Guillaume Marais and Thierry Coduys.

Christian Frisson would like to thank Serheyev Pavlo for having trustfully shipped his for sale EyeTribe eye tracker before he could receive the payment, while official sources were out of stock. This helped us getting started.

References

[1] W. W. Gaver, "The sonic finder: An interface that uses auditory icons," *Human-Computer Interaction*, vol. 4, pp. 67–94, 1989.

[2] M. Chion, *Audio-Vision: Sound on Screen*. Columbia University Press, 1994.

[3] C. Holland and O. Komogortsev, "Eye tracking on unmodified common tablets: Challenges and solutions," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '12. ACM, 2012, pp. 277–280.

[4] S. A. Johansen, J. San Agustin, H. Skovsgaard, J. P. Hansen, and M. Tall, "Low cost vs. high-end eye tracking for usability testing," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '11. New York, NY, USA: ACM, 2011, pp. 1177–1182.

[5] V. Krassanakis, V. Filippakopoulou, and B. Nakos, "Eyemmv toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification," *Journal of Eye Movement Research*, vol. 7, no. 1, pp. 1–10, 2014.

[6] A. Vokhler, V. Nordmeier, L. Kuchinke, and A. M. Jacobs, "Ogama - OpenGazeAndMouseAnalyzer: Open source software designed to analyze eye and mouse movements in slideshow study designs," *Behavior Research Methods*, vol. 40, no. 4, pp. 1150–1162, 2008.

[7] P. Ekman, W. Friesen, and J. Hager, *The Facial Action Coding System*. The MIT Press, 2002.

[8] O. Ferhat, F. Vilariño, and F. J. Sanchez, "A cheap portable eye-tracker solution for common setups," *Journal of Eye Movement Research*, vol. 7, no. 3, pp. 1 – 10, 2014.

[9] S. A. Mahmoudi, M. Kierzynka, P. Manneback, and K. Kurowski, "Real-time motion tracking using optical flow on multiple gpus," *Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 62, pp. 139–150, 2014.

[10] S. A. Mahmoudi, M. Kierzynka, and P. Manneback, "Real-time gpu-based motion detection and tracking using full hd videos," in *Intelligent Technologies for Interactive Entertainment*, 2013, pp. 12–21.

[11] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 831–834.

[12] H. Gonzalez-Jorge, B. Riveiro, E. Vazquez-Fernandez, J. Martínez-Sánchez, and P. Arias, "Metrological evaluation of microsoft kinect and asus xtion sensors," *Measurement*, vol. 46, no. 6, pp. 1800–1806, 2013.

[13] "Face tracking. microsoft developper network." [Online]. Available: http://msdn.microsoft.com/en-us/library/jj130970.aspx

[14] A. Borji and L. Itti, "State-of-the-art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[15] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and Human Fixations: State-of-the-art and Study of Comparison Metrics," in *Proceedings of the 14th International Conference on Computer Vision (ICCV 2013)*, Sydney, Australia, 2013, pp. 1–8.

[16] ——, "A study of parameters affecting visual saliency assessment," in *Proceedings of the 6th International Symposium on Attention in Cognitive Systems (ISACS'13)*, Beijing, China, 2013.

[17] S. Winkler and R. Subramanian, "Overview of eye tracking datasets," in *Proceedings of the 5th International Workshop on Quality of Multimedia Experience*, ser. QoMEX, 2013.

[18] C. Quigley, S. Onat, S. Harding, M. Cooke, and P. König, "Audio-visual integration during overt visual attention," *Journal of Eye Movement Research*, vol. 1, no. 2, pp. 1–17, 2008.

[19] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier, "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, pp. 1–10, 2012.

[20] M. L. H. Võ, T. J. Smith, P. K. Mital, and J. M. Henderson, "Do the eyes really have it? Dynamic allocation of attention when viewing moving faces," *Journal of Vision*, vol. 12, no. 13, pp. 1–14, 2012.

[21] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of Vision*, vol. 14, no. 8, pp. 1–17, 2014.

[22] ——, "An Audiovisual Attention Model for Natural Conversation Scenes," in *IEEE International Conference on Image Processing (ICIP)*, Paris, France, 2014.

[23] R. Beauchamp, *Designing Sound for Animation*. Focal Press, 2005.

[24] V. T. Ament, *The Foley Grail: The Art of Performing Sound for Film, Games, and Animation*. Focal Press, 2009.

[25] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[26] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, vol. 18, no. 1, p. 5060, 1947.

[27] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," *International Conference on Digital Audio Effects*, pp. 237–244, 2007.

[28] P. Mital and M. Grierson, "Mining Unlabeled Electronic Music Databases through 3D Interactive Visualization of Latent Component Relationships," in *NIME 2013: New Interfaces for Musical Expression*, Seoul, Korea, 2013.

[29] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion," *Cognitive Computation*, vol. 3, no. 1, pp. 5–24, Oct. 2010.

[30] T. J. Smith and P. K. Mital, "Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes," *Journal of Vision*, vol. 13, pp. 1–24, 2013.

[31] T. Blascheck, K. Kurzhals, M. Raschke, M. Burch, D. Weiskopf, and T. Ertl, "State-of-the-Art of Visualization for Eye Tracking Data," in *State of The Art Report (STAR) from the Eurographics Conference on Visualization (EuroVis)*, R. Borgo, R. Maciejewski, and I. Viola, Eds. The Eurographics Association, 2014.

[32] K. Kurzhals, F. Heimerl, and D. Weiskopf, "Iseecube: Visual analysis of gaze data for video," in *Proceedings of the Symposium on Eye Tracking Research and Applications*, ser. ETRA '14. New York, NY, USA: ACM, 2014, pp. 351–358.

[33] C. Frisson, S. Alaçam, E. Coşkun, D. Ertl, C. Kayalar, L. Lawson, F. Lingenfelser, and J. Wagner, "Comediannotate: towards more usable multimedia content annotation by adapting the user interface," in *Proceedings of the eNTERFACE'10 Summer Workshop on Multimodal Interfaces*, Amsterdam, Netherlands, July 12 - August 6 2010.

[34] C. Bardiot, T. Coduys, G. Jacquemin, and G. Marais, "Rekall: un environnement open source pour documenter, analyser les processus de creation et faciliter la reprise des uvres sceniques," in *Actes des Journées d'Informatique Musicale*, 2014.

[35] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *IEEE CVPR*, 2014.

[36] L. Neumann and M. J, "Real-time scene text localization and recognition," in *IEEE CVPR*, 2012.

Christian Frisson graduated a MSc. in "Art, Science, Technology (AST)" from Institut National Polytechnique de Grenoble (INPG) and the Association for the Creation and Research on Expression Tools (ACROE), France, in 2006. In 2015, he obtained his PhD with Profs Thierry Dutoit (UMONS) and Jean Vanderdonckt (UCL) on designing interaction for organizating media collections (by content-based similarity). He has been a fulltime contributor to the numediart Institute since 2008.
http://christian.frisson.re

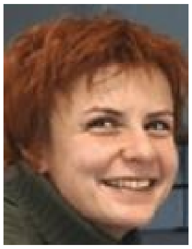
Nicolas Riche holds an Electrical Engineering degree from the University of Mons, Engineering Faculty (since June 2010). His master thesis was performed at the University of Montreal (UdM) and dealt with automatic analysis of the articulatory parameters for the production of piano timbre. He obtained a FRIA grant for pursuing a PhD thesis about the implementation of a multimodal model of attention for real time applications.
http://tcts.fpms.ac.be/attention

Antoine Coutrot holds an Engineering degree and a PhD in Cognitive Science from Grenoble University (France). During his PhD, he studied the influence of sound on the visual exploration of dynamic scenes. He currently holds a Research Associate position at CoMPLEX, University College London (UK), where he develops statistical model to understand how high and low level features guide our gaze.
http://www.gipsa-lab.fr/~antoine.coutrot

Charles-Alexandre Delestage is an Msc student in Audiovisual Communication's Management in Valenciennes. He starting a doctorate degree fall 2014 within DeVisu (France) and TCTS (Belgium) research laboratories. His research is oriented on the automation in the audiovisual processes of production and the impact on the audiences.
http://www.univ-valenciennes.fr/DEVISU/membres/delestage_charles_alexandre

Stéphane Dupont received the PhD degree in EE at the Faculty of Engineering of Mons (Belgium) in 2000. He was post-doctoral associate at ICSI (California) in 2001-2002 where he participated to the ETSI standardisation activity on robust speech recognition. In 2002, he joined Multitel (Belgium) to coordinate the speech recognition group and several industrial and EU-funded projects. In 2008, he joined UMONS (Belgium). Dr. Dupont interests are in speech, music, audio and multimedia processing, machine learning, and multimodal human-computer interaction. He holds 3 international patents and has authored/co-authored over 100 papers in these areas.
http://www.tcts.fpms.ac.be/~dupont

Onur Ferhat holds a BSc degree in Computer Engineering from Boğaziçi University, Turkey and an MSc degree in Computer Vision from Universitat Autònoma de Barcelona (UAB). He is currently a postgraduate student in Computer Vision Center, Barcelona and an assistant teacher in UAB. His research interests include computer vision, eye-tracking and human-computer interaction.
http://onurferhat.com



Nathalie Guyader obtained a PhD degree in Cognitive Sciences at the University Joseph Fourier of Grenoble (France) in 2004 under the supervision of J. Hérault and C. Marendaz on a biologically inspired model of human visual perception to categorize natural scene images. Since 2006, she has been an associate professor at the University Joseph Fourier of Grenoble and at the Grenoble Image Signal and Automatism laboratory (GIPSA-lab). Her research mainly concerns human visual attention through two approaches : eye-tracking experiment analysis and computational modeling.
http://www.gipsa-lab.grenoble-inp.fr/page_pro.php?vid=98

Sidi Ahmed Mahmoudi received the graduate engineering degree in computer science from the University of Tlemcen, Algeria, the masters degree in multimedia processing from the Faculty of Engineering in Tours, France, and the PhD degree in engineering science from the University of Mons, Belgium, in 2006, 2008, and 2013, respectively. Currently, he is a postdoc researcher at the University of Mons, Belgium. His research interests are focused on real time audio and video processing for watching slow motions, efficient exploitation of parallel (GPU) and heterogeneous (multi-CPU/multi-GPU) architectures. He also participated in national (ARC-OLIMP, Numdiart, Slowdio PPP) projects and European actions (COST IC 805).
http://www.ig.fpms.ac.be/fr/users/mahmoudisi

Matei Mancas holds an ESIGETEL Audiovisual Systems and Networks engineering degree (Ir.), and a Orsay Univ. D.E.A. degree (MSc.) in Information Processing. He also holds a PhD in applied sciences from the FPMs on computational attention since 2007. His research deals with signal saliency and understanding. Part of this research is done for artisitic purposes within the numediart institute. Matei is now the SmartSpaces Group leader @ the Numediart Research Institute (UMONS Institute for Creative Technologies).
http://tcts.fpms.ac.be/~mancas

Parag K. Mital B.Sc, M.Sc., Ph.D, is a computational media artist investigating how film together with eye-movements, EEG, and fMRI can help to explore how people attend to and represent audiovisual scenes. His arts practice builds this understanding into audiovisual scene synthesis, computational mashups, and expressive control of audiovisual content. He is currently working on audiovisual decoding with fMRI at Dartmouth College, Bregman Music and Audio Research Studio.
http://pkmital.com

Alicia Prieto Echániz graduated in Advertising and Public Relations and has a master's degree in Corporate Communications from the University of the Basque Country. Specialized on photography, graphic design and online marketing, has taken part in three short film contests and has juggled her studies with voluntary work in brain injury daycare centers and as a instructor in urban summer camps. She's currently working for the Cultural Association Tarasu as the communication's manager, which dedicates to the spread of the Japanese culture.
http://es.linkedin.com/pub/alicia-prieto-echniz/33/955/3/en

François Rocca holds an Electrical Engineering degree from the Faculty of Engineering of Mons (UMONS) since June 2011. He did his master's thesis in the field of emotional speech analysis, and more especially on laughter frequencies estimation. He is currently pursuing a PhD thesis on Real-time 2d/3D head pose estimation, face tracking and analysis by markerless motion tracking for expression recognition.
http://tcts.fpms.ac.be/~rocca

Alexis Rochette is currently researcher on the SonixTrip project for the art and science laboratory (Laras) of Institut de Recherche de l'Institut Supérieur Industriel de Bruxelles (IrIsib). He holds an Industrial Electronic Engineering degree from Institut Supérieur Industriel de Bruxelles (Isib) since 2013. He did a Final internship for the Master Degree and Master thesis at Klavis Technologies S.A from Brussels where he developed of a CopperLan application on a embedded platform. He also holds a Bachelor Degree in Applied Electronics since 2011 for which he did a Final internship at the Royal Observatory of Belgium.

Willy Yvart graduated a Master Degree in Multimedia, Audiovisual, Information and Communication Sciences from DREAM departement of the University of Valenciennes (France) in 2011. Since 2013, he is a PhD candidate under the joint supervision of Thierry Dutoit (UMONS, Belgium) and Sylvie Leleu-Merviel (UVHC, France) on the study of semantics metadata in massive music library in order to improve indexing and searching techniques.
http://www.univ-valenciennes.fr/DEVISU/membres/yvart_willy

MILLA – Multimodal Interactive Language Learning Agent

João P. Cabral, Nick Campbell, Shree Ganesh, Emer Gilmartin, Fasih Haider, Eamonn Kenny, Mina Kheirkhah, Andrew Murphy, Neasa Ní Chiaráin, Thomas Pellegrini and Odei Rey Orozko

***Abstract*—The goal of this project was to create a multimodal dialogue system which provides some of the advantages of a human tutor, not normally encountered in self-study material and systems. A human tutor aids learners by:**

- **Providing a framework of tasks suitable to the learner's needs**
- **Continuously monitoring learner progress and adapting task content and delivery style**
- **Providing a source of speaking practice and motivation**

MILLA is a prototype language tuition system comprising tuition management, learner state monitoring, and an adaptable curriculum, all mediated through speech. The system enrols and monitors learners via a spoken dialogue interface, provides pronunciation practice and automatic error correction in two modalities, grammar exercises, and two custom speech-to-speech chatbots for spoken interaction practice. The focus on speech in the tutor's output and in the learning modules addresses the current deficit in spoken interaction practice in Computer Aided Language Learning (CALL) applications, with different text-to-speech (TTS) voices used to provide a variety of speech models across the different modules. The system monitors learner engagement using Kinect sensors and checks pronunciation and responds to dialogue using automatic speech recognition (ASR). A learner record is used in conjunction with the curriculum to provide activities relevant to the learner's current abilities and first language, and to monitor and record progress.

***Index Terms*—language learning, CALL, spoken dialogue system.**

I. Introduction

Language learning is an increasingly important area of human and commercial endeavour as increasing globalisation and migration coupled with the explosion in personal technology ownership expand the need for well designed, pedagogically oriented language learning applications.

While second languages have long been learned conversationally with negotiation of meaning between speakers of different languages sharing living or working environments, these methods did not figure in formal settings. In contrast, traditional formal language instruction followed a grammar-translation paradigm, based largely on the written word. The advent of more communicative methodologies in tandem with increased access to audio-visual media in the target language had led to much greater emphasis on use of the language in both the spoken and written forms. The Common European Framework of Reference for Language Learning and Teaching (CEFR) recently added a more integrative fifth skill – spoken interaction – to the traditional four skills – reading and listening, and writing and speaking [1]. The nature of language curricula is also undergoing change as, with increased mobility and globalisation, many learners now need language as a practical tool rather than simply as an academic achievement [2].

The language learning sector has been an early adopter of various technologies, with video and audio courses available since the early days of audiovisual technology, and developments in Computer Assisted Language Learning (CALL) have resulted in freely available and commercial language learning material for autonomous study. Much of this material provides learners with reading practice and listening comprehension to improve accuracy in syntax and vocabulary, rather like exercises in a textbook with speech added. These resources greatly help develop discrete skills, but the challenge of providing tuition and practice in the ever more vital "fifth skill", spoken interaction, remains.

Much effort has been put into creating speech activities which allow learners to engage in spoken interaction with a conversational partner, the most difficult competence for a learner to acquire independently, with attempts to provide practice in spoken conversation (or texted chat) using chatbot systems based on pattern matching (e.g. Pandorabots) [3] or statistically driven (e.g. Cleverbot) [4] architectures.

An excellent overview of uses of speech technology in language education is given by [5], covering the use of ASR and TTS to address specific tasks and implementations of complete tutoring systems. Ellis and Bogart [9] outline theories of language education / second language acquisition (SLA) from the perspective of speech technology, while Chapelle provides an overview of speech technology in language learning from the perspective of language educators [10]. Simple commercial pronunciation tutoring applications range from "listen and repeat" exercises without feedback or with auto-feedback. In more sophisticated systems the learner's utterance is recorded and compared with a target or model, and then feedback is given on errors and strategies to correct those errors. Interesting examples of spoken production training systems based on speech technology, where phoneme recognition is used to provide corrective feedback on learner input, include CMU's Fluency [6], KTH's Arthur [7] and MySpeech [8].

Dialog systems using text and later speech have been

João P. Cabral, Nick campbell, Emer Gilmartin, Fasih Haider, Eamonn Kenny, Andrew Murphy and Neasa Ní Chiaráin are with Trinity College Dublin, Ireland

Mina Kheirkhah is with Institute for Advanced Studies in Basic Sciences, Zanjan, Iran

Shree Ganesh is with University of Goettingen

Thomas Pellegrini is with Université Toulouse, France

Odei Rey Orozko is with University of the Basque Country, Bilbao, Spain

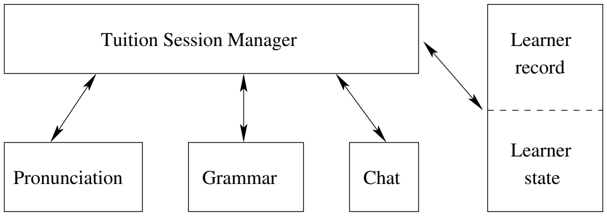


Fig. 1: General architecture of the MILLA system.

successfully used to tutor learners through a natural language interface in science and mathematics subjects. For example, relevant paradigms are the AutoTutor [11] and ITSPOKE [12] systems. In language learning, early systems such as VILTS [13] presented tasks and activities based on different themes which were chosen by the user, while other systems concentrated on pronunciation training via a conversational interface [7].

The MILLA system developed in this project is a multimodal spoken dialogue system combining custom language learning modules with other existing web resources in a balanced curriculum, and offering some of the advantages of a human tutor by integrating spoken dialogue both in the user interface and within specific modules.

II. MILLA System Overview

Figure 1 shows the general architecture of the MILLA system. MILLA's spoken dialogue Tuition Manager (Figure 1) consults a curriculum of language learning tasks, a learner record and learner state module to greet and enrol learners. It also offers language learning submodules, provides feedback, and monitors user state. The tasks comprise spoken dialogue practice with two chatbots, general and focussed pronunciation practice, grammar and vocabulary exercises. All of the tuition manager's interaction with the user can be performed using speech and gestures.

The tuition manager and all interfaces are written in Python 2.6, with additional C#, Javascript, Java, and Bash in the Kinect, chat, Sphinx4, and pronunciation elements respectively.

III. Tuition Management

MILLA's spoken dialogue Tuition Session Manager is a Spoken Dialog System (SDS) that guides the user through the system. The SDS is rule-based, i.e. depending on the answer of the user, the system provides one answer or another. As shown in Figure 2 the Tuition Session Manager first welcomes the user and checks if they already have an account. If the user does not have an account, the system offers to create one. Then, the system consults the curriculum of language learning tasks, the learner record and learner state associated to the user. The way the Tuition Manager updates the learner record is explained in SectionV. The user is asked to choose a language learning submodule and she is redirected to the selected learning module. Meanwhile, the Tuition Manager monitors the user state so that it can offer another alternative tasks ifsignals of frustration or lack of interest are detected by the system (as planned for a future version). The way the Tuition Manager monitors the user state is explained in Section IV.

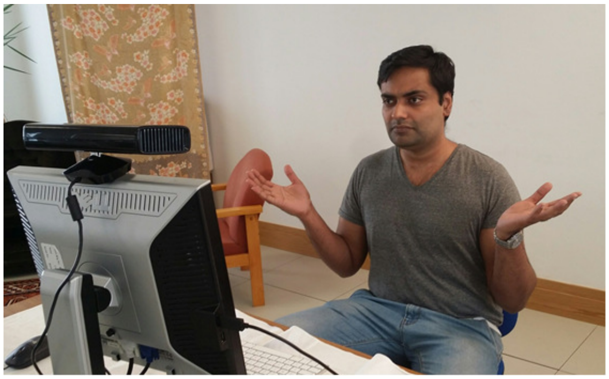


Fig. 3: Example of a gesture ("I don't know") which is detected by kinect in the learner state monitor module.

Spoken interaction with the user is performed through TTS and ASR. The first is implemented using the Cereproc's Python SDK [15], while the second is based on the CMU's Sphinx4 ASR [16] through custom Python bindings using W3C compliant Java Speech Format Grammars. During the design phase the dialogue modules were first written in VoiceXML for rapid prototyping purposes, and then ported to Python.

IV. Learner State Monitor

Microsoft's Kinect SDK [17] is used for gesture recognition. MILLA includes a learner state module to eventually infer learner boredom or involvement. As a first pass, gestures indicating various commands were designed and incorporated into the system using Microsoft's Kinect SDK. The current implementation comprises four gestures: "Stop", "I don't know", "Swipe Left" and "Swipe Right". Figure 3 shows a snapshot of the "I don't know" gesture. They were modelled by tracking the skeletal movements associated with these gestures and extracting joint coordinates on the x, y, and z planes to train the gesture classifier. Python's socket programming modules were used to communicate between the Windows machine running the Kinect and the Mac laptop hosting MILLA.

V. Learner progression - Curriculum and Learner Record

MILLA creates a learner record for each user which is used in conjunction with the curriculum and user state model to guide users to relevant activities, monitor in-session performance and learner state through the predefined gestures or any other infered information such as boredom and frustration. It also record the learner's progress along the curriculum. The curriculum consists of activities for each of the modules tagged with level, first language suitability, duration, and any other information needed to run the activity. As an example, there is a curriculum entry for a focussed pronunciation activity based on the difference between the [ɪ] and [iː] sounds in "fit" and "feet" respectively. It contains information on the

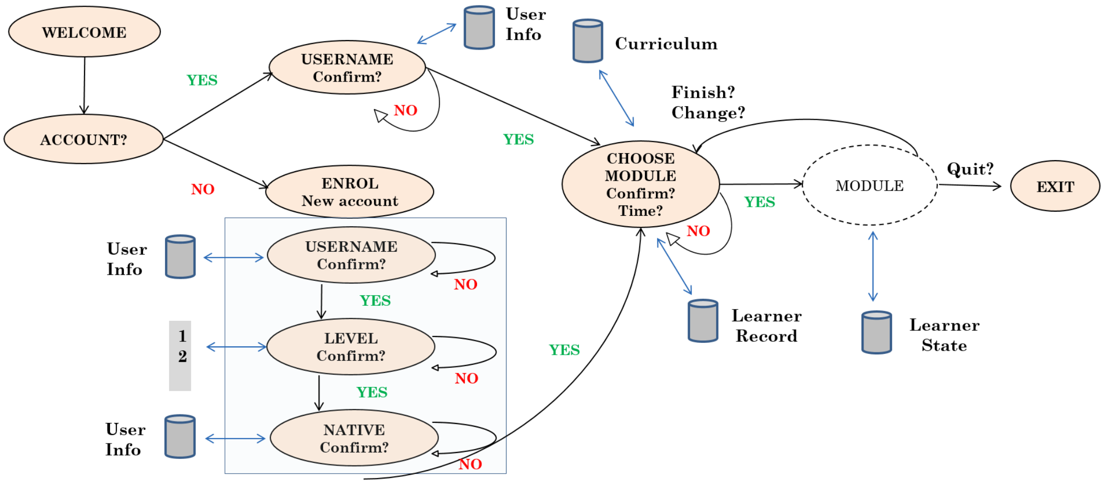


Fig. 2: Diagram of the tuition manager component.

sentence used in this exercise ("These shoes fit my feet"), including suitable explanatory graphics and tips to display on the user interface, level, and the first languages where these sounds are frequently confused. It also contains automatically extracted phonetic renderings of the target phrase and an "erroneous" version for use as parameters in the focussed pronunciation module. For the general pronunciation, chat and grammar modules, which are web-based, the system stores the relevant *urls* for different levels and activities plus the score or accumulated time needed to progress to the next activity or level. This abstraction will facilitate curriculum authoring and editing in the future.

When a learner logs on to the system and chooses their learning module, the learner record is queried to check the user's level, first language, and which activities have been completed. The curriculum is then searched for the next relevant activities in the module, and the learner is directed to suitable activities.

When the module progression is based on time accumulated, the system allows the user to choose how long they will stay on the activity. On completion of the activity the system updates the learner record and prompts the user to choose a new activity or quit. The curriculum and learner records are currently stored as JSON lists. The plan is to port them to an SQL database as the system develops.

VI. Pronunciation Tuition

MILLA incorporates two pronunciation modules. They are both based on comparison of learner production with model production using the Goodness of Pronunciation (GOP) algorithm [18]. However one is first language (L1) focused by taking into account common pronunciation errors from L1 learners, whereas the other provides general pronunciation error feedback independently of L1.

GOP scoring involves two phases: 1) a free phone loop recognition phase which determines the most likely phone sequence given the input speech without giving the ASR any information about the target sentence, and 2) a forced alignment phase which provides the ASR with the phonetic transcription and force aligns the speech signal with the expected phone sequence. Then, the GOP score is computed by comparison of the log-likelihoods obtained from the forced alignment and free recognition phases.

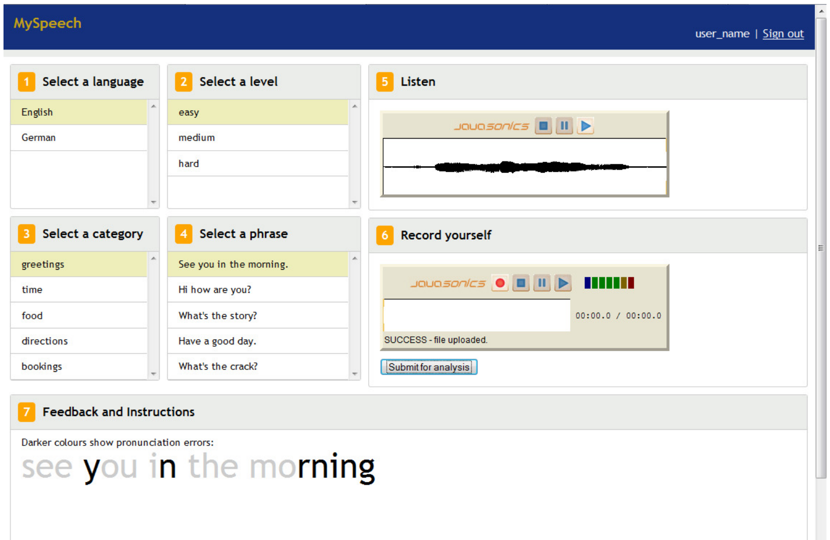
A. Focussed tuition based on common L1 specific errors

The first module was implemented using the HTK toolkit [19] and is defined by five-state 32 Gaussian mixture monophone acoustic models provided with the Penn Aligner toolkit [20], [21]. In this module, phone specific threshold scores were derived by artificially inserting errors in the pronunciation lexicon and running the algorithm on native recordings, as in [22]. After preliminary tests, we constrained the free phone loop recogniser for more robust behaviour, using phone confusions common in specific L1's to define constrained phone grammars. A database of utterances with common errors in several L1's was built into the curriculum (for the learner to practice), so that the system offers relevant pronunciation training based on the learner's first language, which is obtained from the learner record.

B. General Phrase Level Tuition

The second pronunciation training module is a phrase level trainer which is accessed by MILLA via the MySpeech web service [8]. It tests pronunciation at several difficulty levels as described in [23]. Difficulty levels are introduced by incorporating Broad Phonetic Groups (BPGs) to cluster similar phones. A BFG consists of phones that share similar articulatory feature information, for example plosives and fricatives. There are three difficulty levels in the MySpeech system: easy, medium and hard. The easiest level includes a greater number of BPGs in comparison to the harder levels.

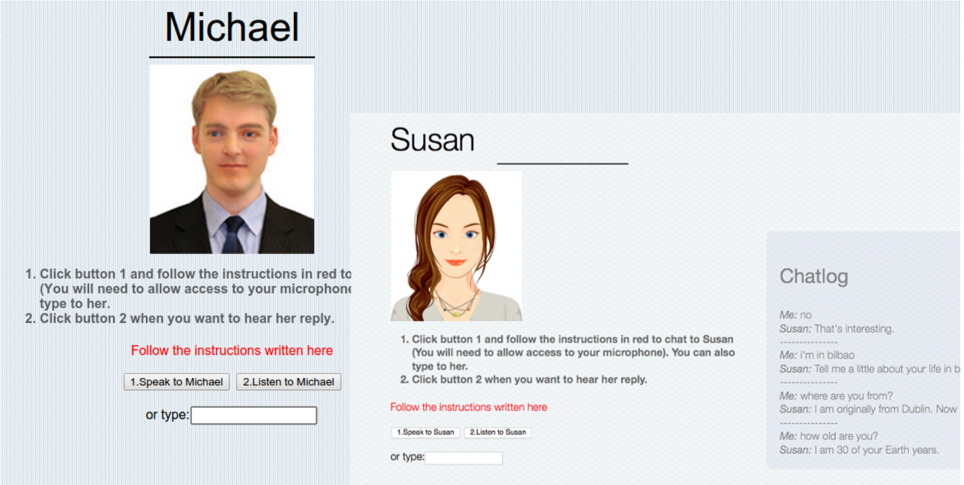
Figure 4 shows the web interface of MySpeech. It consists of several numbered panels for the users to select sentences and practice their pronunciation by listening to the selected sentence spoken by a native speaker and record their own version of the same sentence. Finally, the results panel shows the detected mispronunciation errors of a submitted utterance using darker colours.



Darker colours show pronunciation errors:

see you in the morning

Fig. 4: Web interface of the MySpeech system for pronunciation training.



1. Click button 1 and follow the instructions in red to (You will need to allow access to your microphone type to her.
2. Click button 2 when you want to hear her reply.

1. Click button 1 and follow the instructions in red to chat to Susan (You will need to allow access to your microphone). You can also type to her.
2. Click button 2 when you want to hear her reply.

Me: no
Susan: That's interesting.

Me: i'm in bilbao
Susan: Tell me a little about your life in b

Me: where are you from?
Susan: I am originally from Dublin. Now

Me: how old are you?
Susan: I am 30 of your Earth years.

Fig. 5: Web interfaces for the "Michael" and "Susan" chatbots.

VII. Spoken Interaction Tuition (Chat)

In order to provide spoken interaction practice, MILLA sends the user either to Michael (Level 1) or to Susan (Level 2), two chatbots created using the Pandorabots web-based chatbot hosting service [24]. Figure 5 shows the web interface for the chatbot service with Michael and Susan.

In this work, these bots were first implemented in text-to-text form in AIML (Artificial Intelligence Markup Language). Then, TTS and ASR were added through the Web Speech API, conforming to W3C standards [25]. The system design is based on previous research in the field of CALL as well as consultation with language teachers and learners [26]. The system allows users either to speak to the bot, or to type chat questions/responses. The user receives text feedback from the bot and can also listen to these utterances pronounced in the accent of the bot (Michael: British-English and Susan: American-English). A chat log was also implemented in the interface, allowing the user to read back or replay previous interactions.

VIII. Grammar, Vocabulary and External Resources

MILLA's curriculum includes a number of graded activities from the OUP's English File and the British Council's Learn English websites. Wherever possible the system scrapes any scores returned by these web services for exercises and incorporates them into the learner's record, while in other cases the progression and scoring system includes a time required to be spent on the exercises before the user progresses to the next exercises (as explained in Section V). During the project custom morphology and syntax exercises created using VoiceXML, which will be ported to MILLA.

IX. Future Work

MILLA is an ongoing project. In particular work in progress includes the development of a Graphical User Interface and avatar to provide a more immersive version. We also have a plan to incorporate several new modules into MILLA. Finally, user trials are planned in several centres providing language training to immigrants in Ireland.

Acknowledgment

João P. Cabral and Eamonn Kenny are supported by the Science Foundation Ireland (Grant 12/CE/I2267) as part of CNGL (www.cngl.ie) at Trinity College Dublin. Sree Ganesh is supported by GCDH University of Goettingen. Emer Gilmartin is supported by the Science Foundation Ireland Fastnet Project, grant 09/IN.1/I2631. Thomas Pellegrinis participation was partially supported by an IDEX "Nouveaux Entrants" fellowship. Fasih Haider is supported by the METALOGUE (Multiperspective Multimodal Dialogue: dialogue system with metacognitive abilities) project under FP7-ICT, grant 611073. For details on METALOGUE visit http://cordis.europa.eu/project/rcn/110655_en.html. Neasa Ní Chiaráin and Andrew Murphy are supported by the ABAIR Project, funded by the Irish Government's Department of Arts, Heritage, and the Gaeltacht. We would like to thank the eNTERFACE organisers for providing a richly collaborative setting for this work.

References

[1] Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. Language Teaching, 39(03), 167–190.

[2] Gilmartin, E. (2008). Language Training for Adult Refugees: The Integrate Ireland Experience. Adult Learner: The Irish Journal of Adult and Community Education, 97, 110.

[3] "Pandorabots - A Multilingual Chatbot Hosting Service". [Online]. Available at http://www.pandorabots.com/botmaster/en/home. [Accessed: 14-Jun-2011].

[4] "Cleverbot.com - a clever bot - speak to an AI with some Actual Intelligence?". [Online]. Available at http://www.cleverbot.com/. [Accessed: 18-Apr-2013].

[5] Eskenazi, M. (2009). An overview of spoken language technology for education. Speech Communication, vol. 51, no. 10, 832–844.

[6] Eskenazi, M. and Hansma, S. (1998). The fluency pronunciation trainer, in Proc. of the STiLL Workshop.

[7] B. Granström. (2004). Towards a virtual language tutor. in Proc. of InSTIL/ICALL Symposium 2004.

[8] Cabral, J. P., Kane, M., Ahmed, Z., Abou-Zleikha, M., Szkely, E., Zahra, A., Ogbureke, K., Cahill, P., Carson-Berndsen, J. and Schlögl, S. (2012). Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz. In Proc. of International Conference on Language Resources and Evaluation (LREC), Istanbul.

[9] Ellis, N. C. and Bogart, P. S. (2007). Speech and Language Technology in Education: the perspective from SLA research and practice, in Proc. of ISCA ITRW SLaTE Farmington PA.

[10] Chapelle, C. A. (2009). The Relationship Between Second Language Acquisition Theory and Computer-Assisted Language Learning, in Mod. Lang. J., vol. 93, no. s1, 741–753.

[11] Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A. and Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. Behav. Res. Methods Instrum. Comput., vol. 36, no. 2, 180–192.

[12] Litman D. J. and Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. in Demonstration Papers at HLT-NAACL 2004, pp. 5–8.

[13] Rypa M. E. and Price, P. (1999). VILTS: A tale of two technologies. in Calico J., vol. 16, no. 3, 385–404.

[14] Jokinen, K. and McTear, M. (2009). Spoken Dialogue Systems. in Synth. Lect. Hum. Lang. Technol., vol. 2, no. 1, 1–151.

[15] CereVoice Engine Text-to-Speech SDK |CereProc Text-to-Speech. (2014). Retrieved 7 July 2014, from https://www.cereproc.com/en/products/sdk

[16] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P. and Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition.

[17] Kinect for Windows SDK. (2014). Retrieved 7 July 2014, from http://msdn.microsoft.com/enus/library/hh855347.aspx

[18] Witt, S. M. and Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 30(2), 95–108.

[19] Young, S., Evermann, G., Gales, M., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). The HTK book version 3.4. http://htk.eng.cam.ac.uk/.

[20] Young, S. (n.d.). HTK Speech Recognition Toolkit. Retrieved 7 July 2014, from http://htk.eng.cam.ac.uk/

[21] Yuan, J. and Liberman, M. (2008). Speaker identification on the SCOTUS corpus. Journal of the Acoustical Society of America, 123(5), 3878.

[22] Kanters, S., Cucchiarini, C. and Strik, H. (2009). The goodness of pronunciation algorithm: a detailed performance study. In SLaTE (pp. 49–52).

[23] Kane, M. and Carson-Berndsen, J. (2011). Multiple source phoneme recognition aided by articulatory features. In Modern Approaches in Applied Intelligence, 426–435. Springer.

[24] Wallace, R. S. (2003). Be Your Own Botmaster: The Step By Step Guide to Creating, Hosting and Selling Your Own AI Chat Bot On Pandorabots. ALICE AI foundations, Incorporated.

[25] W3C. (2014). Web Speech API Specification. Retrieved 7 July 2014, from https://dvcs.w3.org/hg/speech-api/rawfile/tip/speechapi.html

[26] Ní Chiaráin, N. (2014). Text-to-Speech Synthesis in Computer-Assisted Language Learning for Irish: Development and Evaluation. Unpublished doctoral dissertation, Trinity College, Dublin.

ZureTTS: Online Platform for Obtaining Personalized Synthetic Voices

Daniel Erro*, Inma Hernáez*, Eva Navas*, Agustín Alonso, Haritz Arzelus, Igor Jauk, Nguyen Quy Hy,
Carmen Magariños, Rubén Pérez-Ramón, Martin Sulír, Xiaohai Tian, Xin Wang and Jianpei Ye

Abstract—**The primary goal of the ZureTTS project was the design and development of a web interface that allows non-expert users to get their own personalized synthetic voice with minimal effort. Based on the increasingly popular statistical parametric speech synthesis paradigm, the system was developed simultaneously for various languages: English, Spanish, Basque, Catalan, Galician, Slovak, Chinese, and German.**

Index Terms—**Statistical parametric speech synthesis, speaker adaptation.**

I. Introduction

SPEECH synthesis technologies have evolved during the last decade from selection and concatenation based paradigms [1] to statistical parametric ones [2], [3]. The main advantages of hidden Markov model (HMM) based speech synthesis are its enormous flexibility for speaker/style adaptation [4], the low footprint of the voice models (those of a high-quality voice can be stored in less than 5 MB in some cases!), the ability of generating smooth synthetic signals without annoying discontinuities, etc. Importantly, the availability of an open source statistical parametric speech synthesis system, HTS [5], has played a key role in this technological revolution. Statistical parametric speech synthesis has enabled many new applications that were not possible in the previous technological frameworks, such as the design of aids for people with severe speech impairments [6], personalized speech-to-speech translation [7], and noise-robust speech synthesis [8].

In parallel, the market of hand-held devices (smartphones, tablet PC's, etc.) has grown substantially, together with their capabilities. As a result, communication barriers are diminishing while new ways of entertainment are emerging. In these two contexts, low-footprint personalized text-to-speech (TTS) synthesizers arise as a very interesting outcome of years of research. Nevertheless, it is still hard for a non-expert user to get a personalized TTS. The ZureTTS project bridges the gap between users and speaker-adaptive speech synthesis technology by providing a web interface that helps them obtaining personalized synthetic voices in an intuitive and automatic manner. This interface can be used, for instance, by people suffering from degenerative pathologies to create a "backup" of their voice before surgery or before the symptoms are noticeable, or by anyone wishing a GPS to speak in his/her own voice or that of a relative.

The ZureTTS project was undertaken by an international team of researchers during eNTERFACE'14, covering up to 8 languages: English, Spanish, Basque, Catalan, Galician, Slovak, Chinese, and German. This article contains a detailed description of the project and the way it was developed. Section II presents a general overview of the technology and roughly describes the system in relation to it; Sections III, IV and V go into the details of the different tasks accomplished during the development; Section VI discusses some open issues and future perspectives, and it summarizes the main conclusions.

II. General framework

A. Technological framework

Fig. 1 shows the general diagram of a speaker adaptive HMM-based speech synthesis system such as HTS [5], which is the core of ZureTTS. Basically, HTS provides the tools to (i) learn a global statistical correspondence between labels extracted from text and features extracted from the acoustic realization of that text, (ii) adapt the trained statistical models to new incoming data from a particular speaker [4] and (iii) generate the most likely sequence of acoustic features given a specific set of labels [9] (these tools correspond to the blue blocks in Fig. 1).

The so-called labels are phonetic, linguistic or prosodic descriptors of the text, which are stored in a specific format that the HTS engine is able to understand. The information they contain is typically related to phones (code of the current one and those in the surroundings, time location, position in syllable, position in word, etc.), syllables (accent, position in word, position in phrase, etc.), words, phrases, pauses... The label extraction process (green block in Fig. 1) is obviously

*Project leaders. The remaining co-authors are equal contributors and have been listed in strict alphabetical order.

D. Erro, I. Hernáez, E. Navas, A. Alonso and J. Ye are with Aholab, University of the Basque Country, Alda. Urquijo s/n, 48013 Bilbao, Spain (contact e-mail: derro@aholab.ehu.es). D. Erro is also with IKERBASQUE, Basque Foundation for Science, 48013 Bilbao, Spain.

H. Arzelus is with VicomTech-IK4, Paseo Mikeletegi 57, Parques Tecnológicos de Euskadi - Gipuzkoa, 20009 San Sebastian, Spain.

I. Jauk is with the VEU group, Teoria Senyal i Comunicacions, Technical University of Catalonia, C/ Jordi Girona 1-3, 08034 Barcelona, Spain.

N.Q. Hy and X. Tian are with NTU-UBC Research Centre of Excellence in Active Living for the Elderly, Nanyang Technological University, 639798 Singapore.

C. Magariños is with Multimedia Technology Group, AtlantTIC, University of Vigo, Campus Lagoas-Marcosende s/n, 36310 Vigo, Spain.

R. Pérez-Ramón is with Laslab, University of the Basque Country, Facultad de Letras, Paseo de la Universidad 5, 01006 Vitoria, Spain.

M. Sulír is with Department of Electronics and Multimedia Communications, Technical University of Košice, Letná 9, 040 11 Košice, Slovakia.

X. Wang is with the National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China.

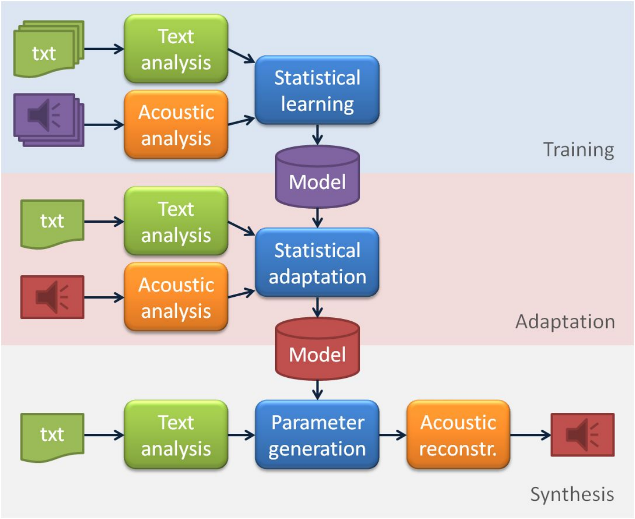


Fig. 1. General diagram of a speaker-adaptive statistical parametric speech synthesis system.

language-dependent. In principle, the front-end of any existing TTS system can be used to accomplish this task as long as the information it handles can be translated into the appropriate label format. Since this project deals with many languages, many language-specific text analyzers were used as will be explained in Section IV.

Vocoders are used to extract acoustic features from audio signals and also to reconstruct audio signals from acoustic features (orange blocks in Fig. 1). The statistical learning process implies the use of appropriate acoustic parameterizations exhibiting not only good speech analysis/reconstruction performance but also adequate mathematical properties. A typical state-of-the-art vocoder extracts acoustic features at three different levels: logarithm of the fundamental frequency (f_0), Mel-cepstral (MCEP) or linear prediction related representation of the spectral envelope, and degree of harmonicity of different spectral bands.

The statistical model that results from the training stage of a speaker-adaptive system (purple model in the top part of Fig. 1) is often an "average voice model" [4]. In other words, it is learned from data (recordings + texts) from a large number of speakers so that it covers the variability not only of the language but also of its many possible acoustic realizations. These kind of models are more easily adapted to a few recordings of a new unknown speaker than those obtained from one only speaker.

B. Performance flow

The overall structure and performance of the system is graphically described in Fig. 2 (for clarity, we have used the same colors as in Fig. 1). The users interact remotely with the system through a website, and the system itself is hosted on a server that carries out the core tasks and contains the necessary data.

1) The client side: Roughly speaking (see next section for more detailed information), the ZureTTS website allows registered users to record their voice in any of the available languages, visualize the recordings and listen to them for validation, transmit the recordings to the server, and tell the server to start the adaptation process. When the training process finishes, the website gives the user access to the artificial voice so that he/she can validate it or manually modify some intuitive aspects of it (average pitch, vocal tract length, speaking rate, loudness...) with an appropriate acoustic feedback. When the user approves the voice with the desired manual modifications, the website presents the final models of the synthetic voice for download.

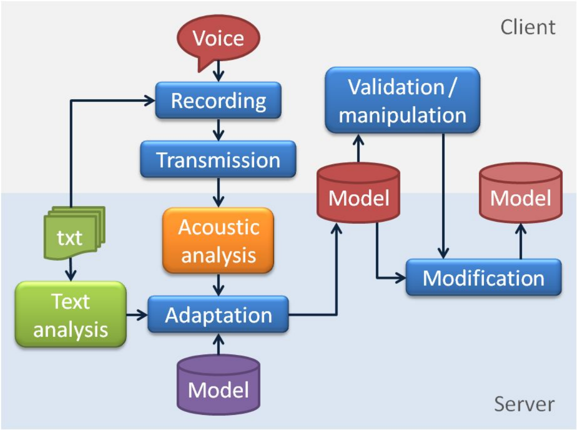
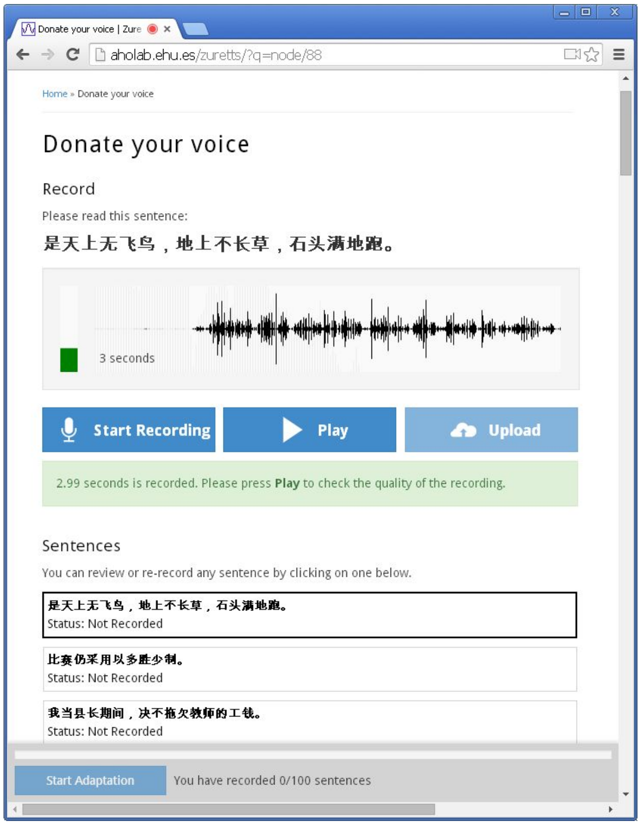


Fig. 2. General diagram of the system.

2) The server side: For each available language, the server contains an initial statistical model (either the model of a generic voice or, preferably, an average voice model [4]), the text analyzer and the vocoder that were used during the training of the initial voice, the scripts for adaptation and modification of models, and a phonetically balanced set of sentences to be recorded by the user. When a particular user selects a language, the server sends him/her a set of sentences (about 100) so that the recording process can be started. Then, when the user transmits the recordings to the server, the server carries out the adaptation (similarly as in the mid part of Fig. 1) and yields a "personalized" model. Since the adaptation process may take some time (depending on the size of the initial model and also on the concentration of user requests), the server alerts the user via e-mail when it has finished. If the user asks for any voice modification through the website after having previewed the result, the server embeds such modification into the model itself, making it permanent. Finally, it stores the "modified personalized" model in the user's internal zone so that it can be either downloaded or used within the ZureTTS portal. Last but not least, the server hosts a web service that allows an external application to connect and synthesize speech from a particular model.

The next three sections describe the work we conducted to develop each part of the ZureTTS system, namely the website itself, the initial voice models and text analyzers for each language under consideration, and the synthetic voice manipulation procedures.



Donate your voice

Record

Please read this sentence:

是天上无飞鸟，地上不长草，石头满地跑。

2.99 seconds is recorded. Please press **Play** to check the quality of the recording.

Sentences

You can review or re-record any sentence by clicking on one below.

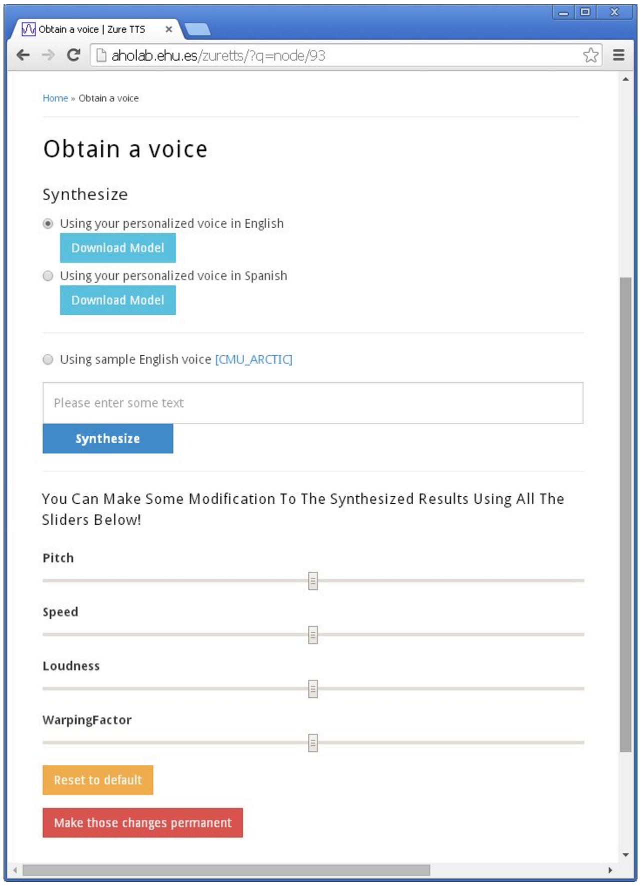
是天上无飞鸟，地上不长草，石头满地跑。
Status: Not Recorded

比赛仍采用以多胜少制。
Status: Not Recorded

我当县长期间，决不拖欠教师的工钱。
Status: Not Recorded

You have recorded 0/100 sentences

Fig. 3. The recording interface for Chinese language.



Obtain a voice

Synthesize

● Using your personalized voice in English

○ Using your personalized voice in Spanish

○ Using sample English voice [CMU_ARCTIC]

Please enter some text

You Can Make Some Modification To The Synthesized Results Using All The Sliders Below!

Fig. 4. Getting and trying personalized or generic artificial voices in different languages.

III. The ZureTTS Website

A. Functional description

The website includes both a public access area, providing information about the project and registration forms, and a private area available only to registered users. In this private area, two functional parts can be distinguished: the *"Donate your voice"* area (depicted in Fig. 3) and the *"Obtain a voice"* area (depicted in Fig. 4).

1) Donation area: When a registered user accesses the voice donation area (Fig. 3), he/she is firstly asked to choose the target language. Then, a speech recorder and a list of about 100 sentences are displayed on the screen. The user is expected to read aloud the sentences and record them in a silent environment. The recorder was designed to give the user useful visual feedback to help him/her control the level of the signals, thus avoiding saturation. In the current implementation, the user is forced to listen to each recorded sentence for validation before transmitting it to the server. By communicating with the central database, the system controls the number of uploaded/remaining recordings. When all of them have been received by the server, the user is given the chance to start the adaptation of the underlying voice models to the samples of his/her voice by pressing the *"Start adaptation"* button.

2) Voice obtaining area: This area is normally accessed when a user receives via e-mail the confirmation that the adaptation process has finished. As can be seen in Fig. 4, the user can then synthesize speech using his/her personalized synthetic voice in any of the languages for which the necessary utterances have been recorded and the adaptation has finished successfully (as an example, in Fig. 4 there are two personalized voices available: English and Spanish). Some generic synthetic voices are also available for registered users who do not want to have a personalized one (in Fig. 4 there is one generic voice in English). A number of modifications are also available in case the user wants to tune some intuitive aspects of the synthetic voice (more details are given in section V). These modifications are applied through sliders, and for moderate factors they do not alter the perceived identity of the voice significantly. Each time the *"Synthesize"* button is pressed, the system synthesizes the message written in the text area using the selected voice model with the modifications given by the current positions of the sliders. Once the modifications are made permanent by pressing the *"Make those changes permanent"* button, the user's voice models are overwritten and replaced by the modified ones.

B. Technical description

The frontend website was written in HTML, Javascript and CSS. It leverages the new Web Audio API of HTML5 to implement a plugin-free voice recorder directly on the webpage; currently this API has been adopted by several major web browsers such as Google Chrome and Mozilla Firefox and will be supported in the next versions of Microsoft

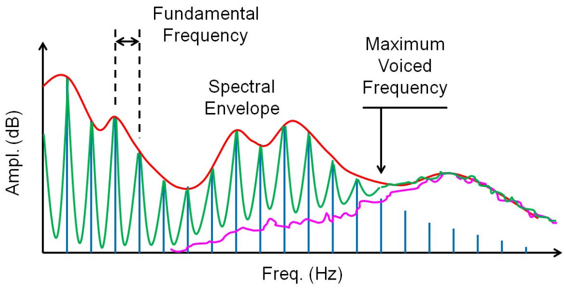


Fig. 5. Assuming that a given speech frame is the sum of a harmonic component (blue) and a noise component (magenta), which occupy the lower and the upper band of speech respectively, Ahocoder measures the log-f_0, the MCEP representation of the spectral envelope (red), and the maximum voiced frequency, defined as the boundary between harmonic and noisy bands.

Internet Explorer. The synthesis page also leverages several capabilities of HTML5 such as audio tag, range input for better compatibility with new browsers and mobile devices. Several popular Javascript and CSS libraries are also used in the frontend such as Jquery, Knockoutjs and Bootstrap.

The backend webservice was mainly written in PHP, so as to be integrated with the mature user management system of Drupal. The webservice is responsible for generating JSON data (recording status, etc.) for the web interface, accepting inputs from the web interface and coordinating different other processing scripts for data collection, adaptation and synthesis.

With an aim towards extending the usability of the results of ZureTTS, we also implemented a multilingual-TTS web service based on the XML-RPC protocol by exploiting the potential of the Drupal framework. Thus, any XML-RPC client (an external application, for instance) can synthesize speech in the offered languages and even manipulate its parameters (as in Fig. 4) just by sending messages in the appropriate XML format to the XML-RPC sever of ZureTTS, which in turn sends XML responses containing encoded audio signals.

IV. Language-Specific Voices Models and Tools

This task of ZureTTS involved the development of initial (average when possible) voice models and text processing tools for all the languages under consideration. A common framework was assumed in terms of HMM topology and acoustic analyzer (vocoder). This approach enabled a more homogeneous implementation of the different functionalities of the system and facilitated some operations, such as those involved in the voice post-editing process (which will be explained in Section V). Excluding these acoustic modeling aspects, voices and tools are language-dependent, so the subsections below describe how the problem was tackled for each language.

The model topology we used is the "standard" one in HTS v2.2, namely 5-state context-dependent left-to-right hidden semi Markov models (HSMMs = HMMs with explicit state duration distributions) [10] where each state is characterized by a multivariate Gaussian emission distribution given by a mean vector and a diagonal covariance matrix. When dealing with parameters that can take discontinuous values, such as the local fundamental frequency of the signal, multi-space distributions (MSD) [11] were used. The vectors used to train the model were appended 1^{st} and 2^{nd}-order dynamics, and the global variance of the acoustic parameters was modeled together with the parameters themselves [9].

The vocoder we used in ZureTTS is called Ahocoder [12]. As depicted in Fig. 5, it handles three streams of acoustic information: log-f_0, MCEP coefficients, and the so-called maximum voiced frequency, which stands for the local degree of harmonicity of the signal. The relationship between MCEP coefficients $\{c_i\}_{i=0...p}$ and spectral envelope $S(\omega)$ can be formulated as follows[1]:

$$\log S(\omega) = \sum_{i=0}^{p} c_i \cos(i \cdot \mathrm{mel}(\omega)) \tag{1}$$

where $\mathrm{mel}(\omega)$ is the Mel-scaled version of ω. All speech signals were digitized at 16 kHz sampling rate and were analyzed/reconstructed at 5 ms frame shift. The order of the MCEP parameterization was always equal to 39, which resulted in 42 static parameters per frame (plus their dynamics).

A. English

An average voice model for English was trained from the 7 voices (2 female + 5 male, 1132 utterances each) in the CMU ARCTIC database [13], similarly as in the HTS demo scripts [14] (except for the vocoder).

The text analysis tools were taken from the well known Festival speech synthesis system [15], developed by the University of Edinburgh.

B. Spanish

In order to train an average voice model for Spanish, we used the "phonetic" subset of the Albayzin speech database [16]. It contains a total of 6800 utterances and 204 different speakers, each of which recorded either 160, 50 or 25 utterances. The phonetic segmentation of the database was not available, so it was carried out automatically via forced alignment of HMMs using HTK [17].

The text analysis tool we employed was taken from AhoTTS, the open-source TTS system developed by Aholab [18]. The structure of the HTS labels was similar to the one described in [19].

C. Basque

The average voice model was trained from the databases described in [20]. They contain a total of 9 voices (5 female, 4 male), all of which include 1 hour of speech except for two (female and male) which include 6 hours each. It is noteworthy that due to time constraints, only half of the voices were used to train the initial model.

Similarly as in Spanish, the front-end of AhoTTS [18] was used for text analysis and label extraction according to the specifications in [19].

[1] Some authors include a multiplicative factor 2 for $i > 0$. In this case such factor is omitted to make the MCEP coefficients compatible with some of the transformations described in Section V.

D. Catalan

Thanks to the Festcat project [21], an open database of 10 different voices (5 female, 5 male) was available for Catalan language [22]. The amount of speech material was 10 hours for two of the voices therein (female and male) and 1 hour for the remaining ones. The speakers were all professional. As well as for Basque, due to time constraints, only half of the voices were used during training.

For text analysis we used a Catalan front-end compatible with Festival [15] that had been released in the framework of the Festcat project [21] too. It included word normalization, a lexicon, a letter-to-sound (L2S) converter and a part-of-speech (POS) tagger.

E. Galician

Due to the lack of freely available speech synthesis databases for Galician languages, we used a single voice provided by the University of Vigo to train the initial model of the system. It contained 1316 utterances (1 hour 13 minutes of speech) recorded by a professional male speaker.

For text analysis, we integrated the front-end of Cotovia [23], the TTS system developed by GTM, University of Vigo.

F. Slovak

The initial average voice model was trained from a total of 17903 utterances (more than 36 hours of speech). This material contained 18 different voices taken from several databases:

- A big Slovak speech database composed by 4526 phonetically balanced sentences, all of them spoken by two different speakers, female and male (about 6 hours of speech per speaker), and recorded in a professional studio. This database had been specifically designed for synthesis purposes [24].
- A smaller database containing 330 phonetically balanced sentences recorded by both a female and a male speaker (40-50 minutes of speech each).
- A speech recognition database with 14 different voices (7 female and 7 male) and a variable number of utterances per speaker (between 469 and 810, 80-140 minutes).

Similarly as for English and Catalan, the Slovak text analyzer used in this project was based on Festival [15]. It included a big pronunciation dictionary containing about 150k problematic Slovak words (a word is considered problematic when there is a mismatch between the written form and its corresponding canonical pronunciation), a rule-based L2S conversion system and a set of token-to-word rules for basic numerals (from zero to several billion).

G. Chinese

Two databases generously provided by iFLYTEK Co. Ltd. were utilized to train a Chinese average voice model. Both databases had been recorded in neutral reading style with 16 kHz sampling rate and 16 bits per sample. The first one contained 1000 utterances (110 minutes) from a male speaker, the average utterance duration being 6.6 seconds. The second one contained 1000 utterances (186 minutes) from a female speaker, with average duration 11.2 seconds. The texts of the two databases were different.

For a correct performance of the TTS system, the Mandarin Chinese text analyzer must parse the input text into a composite structure where not only the phoneme and tone for every syllable but also the prosodic structure of the whole sentence is specified. Unfortunately, such text analyzer is not currently available; thus we built a Mandarin Chinese text analyzer for this project. The text analysis procedure includes: (i) word segmentation, (ii) POS tagging, (iii) grammatical parsing, (iv) L2S conversion, and (v) prosodic prediction. The last two steps are parallel but they both depend on the results of the first three steps.

To implement steps (i)-(iii), an open-source parser called ctbparser [25] was utilized. This parser is based on conditional random fields (CRF). The provided CRF models for word segmentation, POS tagging, and grammatical parsing were trained on The Chinese TreeBank [26] and have been reported to achieve good results on the three tasks [25].

For the L2S conversion, every Chinese character (we assume it as one syllable) must be converted into the composition of phonemes and tone, or pinyin. This conversion can be implemented through a search in a pronunciation dictionary. However, Chinese is well known for polyphones: the pronunciation of one syllable may differ in different contexts. Therefore, we built a hierarchical dictionary and adopted a simple search strategy: if the current grammatical unit, such as a word or phrase, can be found in the dictionary, the corresponding pinyin sequence is used; otherwise, the pinyin of all the syllables of the grammatical unit are retrieved and concatenated into a pinyin sequence.

Getting the pinyin sequence is not enough for high-quality speech synthesis in Mandarin Chinese. Another necessary component is the prosodic hierarchy [27]: some adjacent characters should be pronounced as a single prosodic word and several prosodic words form one single prosodic phrase. Such hierarchy resembles the grammatical structure of a sentence; thus it is possible to derive the prosodic structure based on the results of the grammatical parsing mentioned above. In this project, we adopted grammatical features similar to those mentioned in [28] and used decision trees [29] to build the prosodic prediction model. The model was trained from 1900 sentences in the aforementioned corpus. Tests performed using the remaining 100 sentences showed that the performance of the prosodic prediction model achieved similar results as those reported in [28].

H. German

The German initial voice has been created by using only one voice called “Petra” [30]. The Petra corpus is an extract of the German BITS database [31]. It contains a total of 399 sentences and it was originally recorded, transcribed and segmented for the use in the BOSS speech synthesis system [32]. Thus, a set of scripts had to be implemented in order to translate the original files of the corpus (in BOSS-XML format) into an HTS-compatible label format. Basically we extracted the same type of information as in [19], with very few differences. The same scripts are used in combination with the BOSS text analyzer at synthesis time. This part of the work was not yet fully operational at the time of writing this paper.

V. User-Driven Post-Edition of Synthetic Voices

The purpose of this part of the project was to provide the algorithms to manually manipulate intuitive aspects of synthetic speech. In accordance with the interface described in Section III, the modifications must be applicable: a) on the acoustic parameters of a given signal, and b) on the models that generate such parameters. When performing modifications on the acoustic parameters, the user can listen to the modified version of the signal and tune the parameters according to his/her desires until the results are satisfactory. At that moment, the system will "print" the new values on the HMM, thus making them permanent. In this regard, linear transforms are optimal: given a set of acoustic vectors $\{\mathbf{x}_t\}$ generated by an HMM, we can either transform directly the vectors, $\hat{\mathbf{x}}_t = \mathbf{A}\mathbf{x}_t + \mathbf{b}$, or transform the mean vectors and covariance matrices of all the HMM states:

$$\hat{\boldsymbol{\mu}} = \check{\mathbf{A}}\boldsymbol{\mu} + \check{\mathbf{b}} \ , \quad \hat{\boldsymbol{\Sigma}} = \check{\mathbf{A}}\boldsymbol{\Sigma}\check{\mathbf{A}}^{\top} \tag{2}$$

where

$$\check{\mathbf{A}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A} \end{bmatrix} \ , \quad \check{\mathbf{b}} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \tag{3}$$

Note that the block structure of $\check{\mathbf{A}}$ and $\check{\mathbf{b}}$ allows transforming both the static and the dynamic parts of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For consistency, the global variance models are also transformed through eq. (2) for null bias vector $\check{\mathbf{b}}$ and for $\check{\mathbf{A}}$ equal to the element-wise product of $\mathbf{A}$ by itself.

Focusing on spectral modifications, the choice of MCEP coefficients as acoustic parameterization is particularly useful in this context because it has been shown that, in the MCEP domain, frequency warping[2] and amplitude scaling[3] operations can be formulated as a product by a matrix and an additive bias term, respectively [33], [34].

As depicted in Fig. 4, users are shown a textbox where they can type any sentence to test the trained synthetic voice. They are also presented with a set of sliders to manipulate the four aspects of voice that we describe next, and a button to make the modifications permanent.

1) Speech rate: The effect of the modification can be previewed by reconstructing the waveform from the generated parameters at a different frame rate, i.e. the frame shift is multiplied by the lengthening factor d. Modifications are imposed to the duration model by multiplying its means and covariances by d and d^2, respectively. In addition, the dynamic parts of the acoustic models are multiplied by $1/d$ (deltas) or $1/d^2$ (delta-deltas) for consistency.

2) Average pitch: This modification was implemented through a linear transform of the $\log f_0$ stream where $\mathbf{A} = [1]$ and $\mathbf{b} = [\log \kappa]$, κ being the selected pitch factor.

3) Vocal tract length: Among the existing types of parametric frequency warping curves [33], we chose the one based on the bilinear function as in [34]. In the MCEP domain, the corresponding transformation matrix $\mathbf{A}$ takes only one parameter α as input (see [34] and [35] for details):

$$\mathbf{A} = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots \\ 0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \dots \\ 0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4}$$

Positive values of α produce a vocal tract length reduction, i.e. move the spectral events to higher frequencies[4], and vice versa.

4) Loudness: Even without increasing the power of a signal, its perceptual loudness (and also its intelligibility) can be increased by reallocating energy in some specific bands. In [8] a filter enhancing the band 1–4 kHz was used for this exact purpose. In this work, loudness modification was implemented through a bias vector $\mathbf{b}$ that contains the MCEP representation of the mentioned filter, multiplied by a weighting factor ℓ tuned by the user.

Apart from these four simple modifications, algorithms were developed to allow more expert users to modify the spectrum of the synthetic voice via arbitrary piecewise linear frequency warping and amplitude scaling curves (for simplicity, their corresponding graphical controls have not been included in Fig. 4). These curves will be referred to as $W(\omega)$ and $A(\omega)$ respectively. Basically, the user "draws" them by choosing the position of a number of reference points in a 2D surface. To translate these curves onto the MCEP domain, they are first resampled at a sufficient resolution ($K = 257$ frequency bins between $\omega = 0$ and $\omega = \pi$). The warping matrix $\mathbf{A}$ is given by $\mathbf{C} \cdot \mathbf{M} \cdot \mathbf{S}$, where $\mathbf{M}$ is a $K \times K$ sparse matrix that contains the correspondence between source and target frequency bins (similarly as in [36]), $\mathbf{S}$ is the matrix that transforms a MCEP vector into K log-amplitude spectral bins, and $\mathbf{C}$ is the matrix that converts bins into MCEP coefficients. In accordance with eq. (1), the elements of $\mathbf{S}$ are given by

$$s_{k,i} = \cos\left(i \cdot \mathrm{mel}\left(\pi k/K\right)\right) \ , \quad 0 \le k \le K \ , \quad 0 \le i \le p \tag{5}$$

and $\mathbf{C}$ can be expressed as

$$\mathbf{C} = \left(\mathbf{S}^{\top}\mathbf{S} + \lambda \mathbf{R}\right)^{-1} \mathbf{S}^{\top} \tag{6}$$

where $\lambda = 2 \cdot 10^{-4}$ and $\mathbf{R}$ is a diagonal perturbation matrix whose i^{th} element is $r_{i,i} = 8\pi^2 i^2$ (see [36] for more details). The amplitude scaling vector $\mathbf{b}$ that corresponds to $A(\omega)$ can be obtained[5] by multiplying $\mathbf{C}$ by a K-dimensional vector containing the log-amplitude spectral bins of $A(\omega)$.

VI. Discussion and Conclusion

In this paper we have described the ZureTTS system, a system than includes a set of tools and algorithms covering

[2] Frequency warping modifies the frequency axis of speech spectrum according to a specific mapping function.

[3] Amplitude scaling can be understood as a filtering process that modifies the log-amplitude spectrum of speech.

[4] A similar formulation is typically used to implement the $\mathrm{mel}(\cdot)$ function of eq. (1).

[5] The MCEP representation of the aforementioned loudness filter was calculated exactly this way (though for efficiency it was calculated offline only once, then stored at code level).

the main goals of our project, i.e., the easy creation by non-expert users of a personalized synthetic voice. There are, however, some aspects of the performance that deserve an explicit discussion.

As we have mentioned in Section II-B, users record their voice while reading a set of phonetically-balanced sentences displayed on the screen. Since the texts of these sentences are known, the context labels needed by HTS to perform adaptation are easy to obtain. Nevertheless, it is necessary to align such labels with the input recordings. As labels are defined at phone level, phone segmentation is needed whenever a user transmits a set of recordings to the server and presses the *"Start adaptation"* button. This was not an issue for those languages where Festival-based text analyzers were being used (namely English, Catalan and Slovak), since Festival already provides the necessary tools to do this (it is even capable of detecting and considering the speaker's short pauses that are not consistent with the punctuation marks of the text). For the remaining languages, a solution had to be investigated. The problem of using speech recognition technology was that it required either the availability of appropriate pre-trained speaker-independent recognition models for each language, which would hinder the incorporation of new languages, or the ability to train them exclusively from the user's recorded material, which could be too scarce for an accurate modeling. Hence, we decided to use the initial voice models (synthesis HSMMs), which are obviously available for all languages and are supposed to be of high accuracy, in forced alignment mode. To overcome the possible spectral mismatch between the initial models and the input voice, we followed the strategy described in [35], which consists of the following steps:

1) Get phone durations via forced alignment between the input acoustic vectors and the HSMMs (the mathematical formulation can be found in [35]).
2) Calculate the vocal tract length factor α (see section V-3) that makes the acoustic vectors maximally closer to a similar sentence generated from the HSMM with the current durations.
3) Transform the input vectors through the matrix given by eq. (4) and go back to the first step until convergence is reached.

This procedure resulted in highly satisfactory phone segmentations while getting some interesting information about the vocal tract length of the input voice[6]. The main limitation of this method is that it does not consider the inclusion of short pauses. This issue should be addressed for a more accurate overall performance of the system for the involved languages.

We also studied the possible use of the output likelihoods of HSMM forced alignment for utterance verification. Note that in the current implementation of the system the user him/herself is asked to validate the recordings before transmission, which makes the system prone to undesired problems. Unfortunately, we were not able to implement an accurate detector of mispronounced sentences within the duration of the ZureTTS project. Future works will possibly address this issue.

Another limitation of the system is the lack of control over the quality of the input (presumably home-made) recordings. Given the profiles of the project participants, this aspect was not explicitly tackled, but informal tests indicate that passing the recorded signals through a Wiener filter before acoustic analysis avoids gross noise-related artifacts and does not substantially harm the performance of the system in terms of adaptation. In any case, it is logical to suppose that users will take care of the recording conditions as long as they want to get a high-quality personalized voice.

As detailed in Section IV, language-specific initial voice models were trained from databases that were available to the project participants with adequate permissions. In some cases the databases were not optimal for this task; in some others the amount of training material was not sufficiently large; also, time constraints imposed the need of using only a subset of some relatively large databases. Consequently, there are some significant performance differences between languages. Most of the models are currently being retrained or properly trained, which is likely to result into significant performance improvements in the short term.

The acquisition of a powerful dedicated server is undoubtedly one of the necessary changes to be carried out in the near future. Currently, the system runs in a mid-performance server and requests from different users are processed sequentially on a first-in first-out basis, which sometimes results in unnecessary delays.

Finally, beyond its immediate practical goals, ZureTTS provides a framework to investigate very interesting challenges such as dynamically incorporating the acquired knowledge (basically from the recordings) into the average voice models, or designing new adaptation strategies that reduce the number of sentences that the user has to record for a successful adaptation. When properly advertised, the ZureTTS web portal is expected to facilitate the communication and interaction between researchers and users. In accordance with the market trends it is still necessary, however, to provide the users not only with web services such as those described in Section III, but also with user-friendly "apps" that are compatible with the personalized voices yielded by ZureTTS. With regard to this, an Android version of AhoTTS [18] will be launched soon.

Acknowledgment

The authors would like to thank the support of the Basque Government (ZURETTS+, S-PE11UN081), the Spanish Ministry of Economy and Competitiveness (SpeechTech4All, TEC2012-38939-C03-03, and FPI program), the European Regional Development Fund, the Ministry of Education of the Slovak Republic and Slovak Academy of Sciences (VEGA, 1/0386/12), the Media Development Authority of Singapore (Grant MDA/IDM/2012/8/8-2 VOL 01), the sponsors of eNTERFACE'14 (ISCA, UPV/EHU, EURASIP, RTTH, IKERBASQUE, the Province and City Council of Bilbao, and Metro Bilbao), iFLYTEK Co. Ltd. for providing the Mandarin speech synthesis corpus, and the developers and maintainers of HTS for making its code publicly accessible.

[6] This privileged information is not yet exploited in the current implementation of the system.

References

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[2] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[4] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, 2009.

[5] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Speech Synthesis Workshop*, 2007, pp. 294–299.

[6] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," *Acoustical Science & Tech.*, vol. 33, no. 1, pp. 1–5, 2012.

[7] J. Dines, H. Liang, L. Saheer, M. Gibson, W. Byrne, K. Oura, K. Tokuda, J. Yamagishi, S. King, M. Wester, T. Hirsimki, R. Karhila, and M. Kurimo, "Personalising speech-to-speech translation: unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis," *Computer Speech & Lang.*, vol. 27, no. 2, pp. 420 – 437, 2013.

[8] D. Erro, T. Zorilă, Y. Stylianou, E. Navas, and I. Hernáez, "Statistical synthesizer with embedded prosodic and spectral modifications to generate highly intelligible speech in noise," in *Proc. Interspeech*, 2013, pp. 3557–3561.

[9] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[11] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[12] D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal Sel. Topics in Signal Process.*, vol. 8, no. 2, pp. 184–194, 2014.

[13] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[14] [Online]. Available: http://hts.sp.nitech.ac.jp

[15] [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival

[16] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, and J. B. M. no, "Albayzin speech database: design of the phonetic corpus," in *Proc. 3rd European Conf. on Speech Commun. and Tech.*, 1993, pp. 175–178.

[17] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2006.

[18] A. Alonso, I. Sainz, D. Erro, E. Navas, and I. Hernáez, "Sistema de conversión texto a voz de código abierto para lenguas ibéricas," *Procesamiento del Lenguaje Natural*, vol. 51, pp. 169–175, 2013.

[19] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sánchez, I. Saratxaga, E. Navas, and I. Hernáez, "HMM-based speech synthesis in Basque language using HTS," in *Proc. FALA*, 2010, pp. 67–70.

[20] I. Sainz, D. Erro, E. Navas, I. Hernáez, J. Sánchez, I. Saratxaga, and I. Odriozola, "Versatile speech databases for high quality synthesis for Basque," in *Proc. 8th Int. Conf. on Language Resources and Eval.*, 2012, pp. 3308–3312.

[21] [Online]. Available: http://festcat.talp.cat

[22] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Pérez, "Corpus and voices for Catalan speech synthesis," in *Proc. LREC*, 2008, pp. 3325–3329.

[23] E. Rodríguez-Banga, C. García-Mateo, F. J. Méndez-Pazó, M. González-González, and C. Magariños-Iglesias, "Cotovia: an open source TTS for Galician and Spanish," in *Proc. IberSpeech*, 2012.

[24] M. Sulír and J. Juhár, "Design of an optimal male and female slovak speech database for HMM-based speech synthesis," in *Proc. 7th Int. Workshop on Multimedia and Signal Process.*, 2013, pp. 5–8.

[25] [Online]. Available: http://sourceforge.net/projects/ctbparser/

[26] N. Xue, F. Xia, F.-D. Chiou, and M. Palmer, "The penn chinese treebank: Phrase structure annotation of a large corpus," *Nat. Lang. Eng.*, vol. 11, no. 2, pp. 207–238, 2005.

[27] A. Li, "Chinese prosody and prosodic labeling of spontaneous speech," in *Speech Prosody*, 2002.

[28] Y.-Q. Shao, Z.-F. Sui, J.-Q. Han, and Y.-F. Wu, "A study on chinese prosodic hierarchy prediction based on dependency grammar analysis," *Journal of Chinese Information Process.*, vol. 2, p. 020, 2008.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[30] D. Moers and P. Wagner, "The TTS voice "Petra"," Bielefeld University, Tech. Rep., 2010.

[31] [Online]. Available: http://www.bas.uni-muenchen.de/Forschung/BITS

[32] [Online]. Available: http://sourceforge.net/projects/boss-synth

[33] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech & Audio Processing*, vol. 13, pp. 930–944, 2005.

[34] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. Audio, Speech, & Lang. Process.*, vol. 21, no. 3, pp. 556–566, 2013.

[35] D. Erro, A. Alonso, L. Serrano, E. Navas, and I. Hernáez, "New method for rapid vocal tract length adaptation in HMM-based speech synthesis," in *8th ISCA Speech Synthesis Workshop*, 2013, pp. 125–128.

[36] T.-C. Zorilă, D. Erro, and I. Hernaez, "Improving the quality of standard GMM-based voice conversion systems by considering physically motivated linear transformations," *Commun. in Computer & Inf. Science*, vol. 328, pp. 30–39, 2012.

Daniel Erro$^{(6)}$ received his Telecommunication Eng. degree from Public University of Navarre, Pamplona, Spain, in 2003 and his Ph.D. degree from Technical University of Catalonia, Barcelona, Spain, in 2008. Currently he holds an Ikerbasque Research Fellowship at Aholab, University of the Basque country, Bilbao, Spain. His research interests include speech analysis, modeling, modification, conversion, reconstruction and synthesis. He was the general chair of eNTERFACE'14.

Inma Hernáez received the Telecommunication Eng. degree from the Technical University of Catalonia, Barcelona, Spain, and the Ph.D. degree from the University of the Basque Country, Bilbao, Spain, in 1987 and 1995, respectively. She is a Full Professor in the Faculty of Engineering, University of the Basque Country. She is founding member of the Aholab Signal Processing Research Group. Her research interests are signal processing and all aspects related to speech processing. She is highly involved in the development of speech resources and technologies for the Basque language. She was the general co-chair of eNTERFACE'14.

Eva Navas received the Telecommunication Eng. degree and the Ph.D. degree from the University of the Basque Country, Bilbao, Spain. Since 1999, she has been a researcher at AhoLab and an associate professor at the Faculty of Industrial and Telecommunication Engineering in Bilbao. Her research is focused on expressive speech characterization, recognition, and generation.

Agustín Alonso$^{(5)}$ received his Telecommunication Eng. degree in 2010 and his M.Sc. degree in 2013, both from the University of the Basque Country, Bilbao, Spain. Currently he is a PhD student at Aholab Signal Processing Laboratory, University of the Basque Country, focusing on speech synthesis, transformation and conversion.

Haritz Arzelus(1) received his Computer Eng. degree from the University of the Basque Country, San Sebastian, in 2009. Since October 2009, he has been working in Vicomtech-IK4 as a researcher in speech and language processing technologies on local, national and European projects such BERBATEK, SAVAS and SUMAT. He participated actively in the creation of Ubertitles S.L. (2013), a company oriented to provide automatic subtitling, developing and integrating the technology on which it is based.

Igor Jauk(4) received his Master degree in Phonetics and Computational Linguistics at the University of Bonn, Germany, in 2010. After a research period at the Bielefeld University, Germany, in artificial intelligence and dialogue systems and at the Pompeu Fabra University, Spain, in audiovisual prosody, he now holds an FPU grant for a Ph.D. degree at the Technical University of Catalonia, Barcelona, Spain. His research interests are expressive speech synthesis and information retrieval.

Nguyen Quy Hy(10) received his bachelor in Computer Science from Nanyang Technological University, Singapore, in 2003. He is currently a Master student in the same university. His research interests include all aspects of software engineering and speaker-adaptive expressive speech synthesis.

Carmen Magariños(3) received her Telecommunication Eng. degree in 2011, and her M.Sc. degree in 2014, both from the University of Vigo, Spain. Currently she works as a PhD student at the Multimedia Technology Group of the University of Vigo. Her research interests are focused on speech technology, mainly on HMM-based speech synthesis, hybrid models and speaker adaptation.

Rubén Pérez-Ramón(2) finished his studies on Spanish Language and Literature at Universidad Autónoma de Madrid, Spain, in 2010. He completed a Masters degree on Speech Technology and another one on Judicial Phonetics, and has collaborated with CSIC in several punditries. He is currently preparing his Ph.D. thesis at the University of the Basque Country, Vitoria, Spain.

Martin Sulír(9) received his M.Sc. (Eng.) degree in the field of Telecommunications in 2012 at the Department of Electronics and Multimedia Communications of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice. He is currently PhD student at the same department. His research interests include text-to-speech synthesis systems.

Xiaohai Tian(11) received his Computer Application and Technology degree from Northwestern Polytechnical University, Xian, China, in 2011. Currently he pursues his Ph.D degree at School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include speech signal processing, voice conversion and speech synthesis.

Xin Wang(8) received his B.Eng. degree from University of Electronic Science and Technology of China, Chengdu, China, in 2012. Currently he pursues a master's degree in University of Science and Technology of China, Hefei, China. His research interests include speech synthesis in Text-to-Speech and Concept-to-Speech.

Jianpei Ye(7) received the Telecommunication Eng. degree and the M.Sc. degree in Space Science and Technology from the University of the Basque Country, Bilbao, Spain, in 2013 and 2014, respectively. He is currently a junior researcher in Aholab, University of the Basque Country. His research interests include statistical approaches to speech processing such as voice conversion, speech analysis and speech synthesis.