King's Speech: Pronounce a Foreign Language with Style

Georgios Athanasopoulos
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
georgios.athanasopoulos@uclouvain.be

Céline Lucas
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
celine.lucas@uclouvain.be

Alessandro Cierro
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
alessandro.cierro@uclouvain.be

Robin Guérit
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
robin.guerit@uclouvain.be

Kaori Hagihara
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
kaori.hagihara@uclouvain.be

Julie Chatelain
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
julie.chatelain@uclouvain.be

Sébastien Lugan
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
sebastien.lugan@uclouvain.be

Benoît Macq
ICTEAM-ELEN, Université Catholique de Louvain, Belgium
---
benoit.macq@uclouvain.be

ABSTRACT

Computer assisted pronunciation training requires strategies that capture the attention of the learners and guide them along the learning pathway. In this paper, we introduce an immersive storytelling scenario for creating appropriate learning conditions. The proposed learning interaction is orchestrated by a spoken karaoke. We motivate the concept of the spoken karaoke and describe our design. Driven by the requirements of the proposed scenario, we suggest a modular architecture designed for immersive learning applications. We present our prototype system and our approach for the processing of spoken and visual interaction modalities. Finally, we discuss how technological challenges can be addressed in order to enable the learner's self-evaluation.

KEYWORDS

Immersive Language Learning; L2 Pronunciation; Spoken Karaoke; Computer Assisted Pronunciation Training; Gamification; Audiovisual Speech Technology.

ARTICLE INFO

Received: 26 October 2017
Accepted: 23 July 2018
Published: 02 August 2018

https://dx.doi.org/10.7559/citarj.v10i2.414

1 | INTRODUCTION

The GRAAL [1] project is concerned with developing a set of tools to facilitate self-training on foreign (second) language (L2) pronunciation, with the first target being learning French. This set of tools includes a method for aligning a phrase uttered by a native speaker and the phrase repeated by a learner. Following the alignment, the speech utterance is decomposed in parts and different features are extracted. After analysis of the identified phonemes (vowels and consonants), prosody, intonation, rhythm, etc., an evaluation feedback is given to the learner. The GRAAL project also aims to provide a multimodal interface for rendering the native speaker's discourse and for returning feedback to the learner.

In this paper, we describe our approach during the eNTERFACE 2017 workshop [2] towards developing new interaction modalities aiming to better personalize GRAAL to the preferences and specificities of each learner. The proposed prototype system consists of a narrative story whose role is to capture the attention and stimulate the motivation of the learner while guiding him throughout the interaction with the spoken karaoke. The spoken karaoke is designed to facilitate the self-assessment of the learner, similarly to the traditional "mirror exercise", yet enhanced with modern audiovisual capabilities. The choice of an audiovisual feedback type is motivated by the strong correlation that is known to exist between face motion, vocal tract shape and speech acoustics. The important role of audiovisual speech in second (non-native) language comprehension has also been highlighted in behavioral studies (Barrós-Loscertales, 2013).

The core idea in the developed prototype is to replace the song of a traditional karaoke by native utterances within the context of a narrative story. The learner can adjust the utterance's playback speed and, using the microphone and camera of the portable device (e.g., smartphone), can register his pronunciation. At the end of each story chapter, the learner's audiovisual recordings are displayed side-by-side to those of a native speaker. For each recording, the learner can choose which voice to hear (i.e., none, both, only the native speaker's voice, or only the learner's voice). Differences in the production of the karaoke utterance can be identified by the learner by comparing the audio and visual events in the two versions. This comparison stimulates the self-evaluation feedback effect.

From the signal processing point of view, different elements are necessary for supporting the intended functionalities. The proposed type of feedback visualization is useful only when the audio and video events of the native speaker and the learner are in synchrony and therefore easily comparable. Hence, the main technological challenges lie in time alignment and audiovisual synchronization through time-scaling. As feedback is provided to the learner at the end of each story chapter, the signal processing tasks can be performed offline while the learner continues his interaction with the narrative story. An overview of the proposed system and its components is shown in Figure 1.

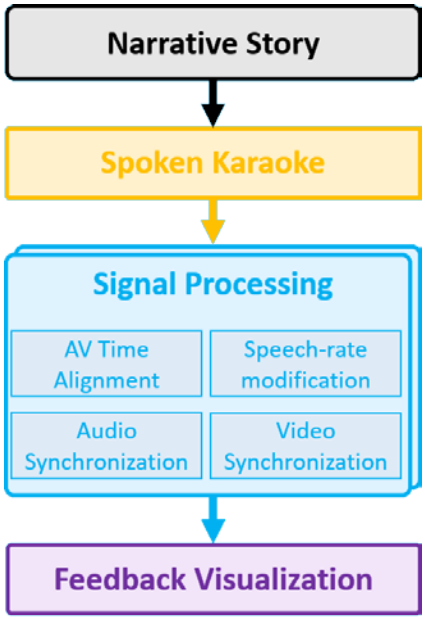

Figure 1 | Overview of the system and its components.

The remainder of this paper is organized as follows. Section 2 motivates the chosen scenario and describes its design with emphasis being given to immersive and gamification aspects. Section 3 presents the modular architecture adopted for the development of the prototype. In Section 4, we detail the audio and visual signal processing components that were implemented to support the scenario. We also highlight some practical implementation issues and describe our approach. Finally, in Section 5 we conclude and present a future perspective.

2 | SCENARIO DESIGN

In this paper, we propose a spoken karaoke concept powered by immersive conditions and storytelling which aims to empower learners in an interactive pronunciation training. In the spoken karaoke, the song is replaced by appropriately selected native utterances of the target language. In the following subsections we motivate the spoken karaoke concept, present the narrative story and discuss empowerment and gamification aspects of our design.

2.1 SPOKEN KARAOKE

One of the most well-known exercises in pronunciation training is the "mirror exercise". Like a dancer who learns a new choreography by working in front of a mirror, in pronunciation learning the mirror is also a fundamental tool for perceiving what is happening when a learner pronounces the target language. Speaking a foreign language leads to produce mouth and tongue movements that might

not exist in the learner's mother tongue. Beyond the face, it is the whole body that will move differently when pronouncing the rhythm and melody of the new language to be learned. There is therefore a whole work of imitation and mime to be achieved. The mirror is the oldest tool and one of the most effective to imitate these new movements of the face and the body.

Nowadays, portable devices such as smartphones, tablets and laptops are part of our everyday life and immediate environment. These devices have an integrated frontal camera, so a type of mirror, the essential and ancestral tool to work these new mouth, tongue, head and body gestures specific to the pronunciation of each language. The spoken karaoke is designed so that it takes advantage of modern technology, while it incorporates all the benefits of the "mirror exercise". It makes use of an audiovisual interface allowing the learner's pronunciation to be automatically synchronized and displayed side-by-side to that of a native speaker.

The concept of the spoken karaoke is designed to involve all the actions that are essential to the learning and improvement of the foreign language pronunciation. These actions can be summarized as follows: listen (potentially more than once), discriminate, identify, manipulate (pausing the utterance, adjusting its playback speed), repeat (what was perceived), compare (what was produced with what was perceived), restart.

Finally, the spoken karaoke concept requires a selection of utterances in accordance to the learning objectives and the level of the learner. In our design, this selection has been validated according to a specific pedagogical progression for the learning of French pronunciation (Briet, 2014).

2.2 NARRATIVE STORY

The design of the spoken karaoke is particularly interesting on different levels. Foremost, it necessitates an element that captures the attention of the user, stimulates intrinsic motivation and guides him throughout the interaction by creating immersive conditions. Previous studies have shown that storytelling is characterized by its capability to support learning processes and create more engaging and exciting learning environments (Kalogeras, 2013). In our approach, the use of a story allows learners to become part of the narrative

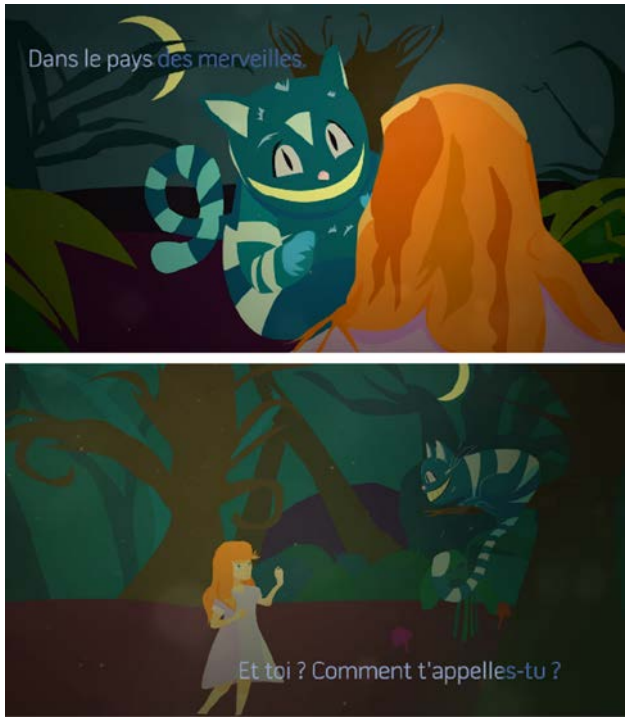

Figure 2 | Sample captures of the animated narrative story. The dynamically highlighted subtitles assist the learner in comprehending the speech dialogue.

plot and be emotionally linked to virtual characters. The goal of this immersive experience is to ease learners to dare to speak the targeted foreign language.

Each karaoke utterance is well entwined in the immersive story and is the key for moving the story forward. The learner discovers the story as a viewer and interacts with it by acting on behalf of the main character. To this aim, he is requested to repeat the utterances that are proposed and demonstrated by the spoken karaoke. The learner can choose among the given options how the main character responds to the narrative environment and thus decides the flow of the story. These alternative scenes serve as a way to engage the learner into practicing sequences more than once, each time potentially choosing a different flow in the plot.

In the work presented in this paper, the non-linear story is built to be a Virtual Reality (VR), fully immersive experience in 180°. Example captions of the narrative story are shown in Figure 2. The storyline is implemented in the Unity game engine [3] which acts as the scenario's scheduler by calling up the necessary components such as e.g., the audiovisual synchronization. This modular approach is further discussed in Section 3.

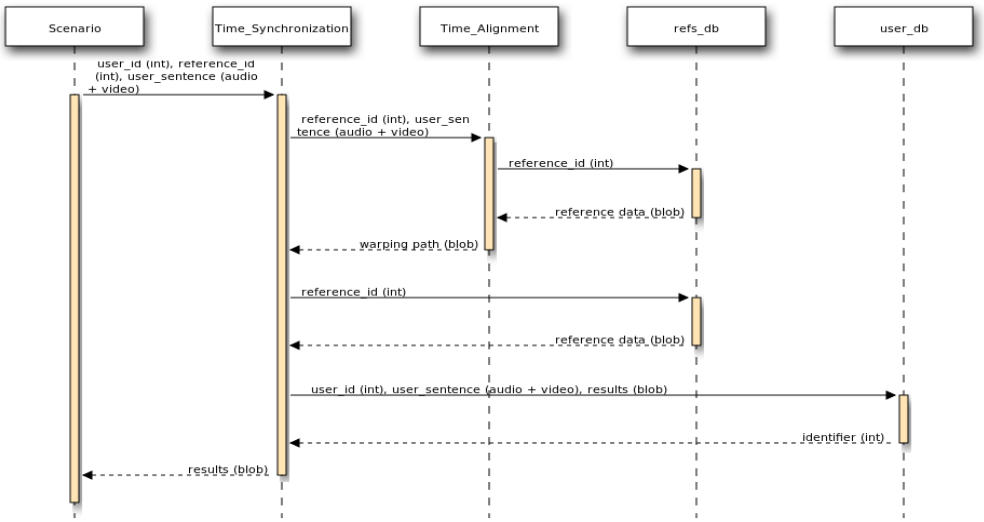

Figure 3 | A sequence diagram depicting the communication between the developed components according to the modular architecture.

2.3 GAMIFICATION

In recent years, gamification, which can be defined as the application of game-design elements and game principles in non-game contexts, has become a trending topic in many fields. Gamification is particularly suitable in the case of education, where it can be integrated effectively to motivate students and enhance learning (Hamari, 2014). There are some obvious overlaps between games and the classroom that make the gamification of curriculum a logical approach: game players work to achieve specific goals and win, while students in the classroom work to achieve specific learning objectives and succeed in exams; game players progress from level to level based on performance, while in the classroom students must pass prerequisite courses and show some level of understanding before addressing an upper academic level. Based on a review of popular gamification taxonomies, 6 inseparable gamification persuasive strategies can be enumerated as follows (Cugelman, 2013): Clear goals and challenges setting; Constant feedback on performance; Reinforcement through rewards (not punishments); Progress monitoring and comparison with self and others; Social connectivity; Fun and playfulness.

These principles that make gamification addictive, along with other popular methods, see e.g., (Miller, 2014), have been considered throughout the design of the narrative immersive story. As an example, the storyline incorporates a point gaining system that increases with the learner's progress in the story, and an experience level that increases based on the pronunciation assessments undertaken by the learner.

3 | MODULAR ARCHITECTURE

Within an immersive pronunciation learning application, it is envisioned that different pedagogical scenarios, learning exercises, pronunciation analysis methods, even new languages should be easily integrated. Hence, the overall architecture should anticipate a modular and flexible design.

In the scope of this work, we have defined a modular architecture for immersive learning applications. Each component has a well-defined set of inputs and outputs and can be evoked by any other component in an asynchronous event-based manner. Different components can share reoccurring functionalities by reusing the subcomponents that implement it. A component can be concurrently instantiated by different scenarios, components, or users. Therefore, no central component acting as a single orchestrator is required. The execution of a scenario can advance while the processing of preceding calls is taking place. The communication between the different components is realized through WebSocket protocol (Fette, 2011), allowing for full-duplex communication channels over a single TCP connection. Hence, the components can be implemented using different programming languages and run on different platforms or operating systems.

A view of the modular architecture for the spoken karaoke is illustrated in Figure 3. In this sequence diagram, the learning scenario calls up the time

synchronization component and provides it with the appropriate inputs. In this example, the time alignment functionality is decoupled from the time synchronization and, hence, it can be readily reused by a different scenario or component that requires it. It is worth noting that each component (that is described in more detail in Section 4) is responsible of retrieving the required reference information and data from a reference database. Therefore, the calling scenario is not concerned with what sub-components are involved and their specificities. Once the time synchronization processing has been completed, the results are stored in the user database for future reference. Finally, the results are returned to the calling scenario.

4 | AUDIOVISUAL SIGNAL PROCESSING

Displaying side-by-side the utterance of a native speaker with that recorded by a learner necessitates the use of a common time reference. As the speech rate of each recording is inevitably different, even in karaoke settings where the learner is expected to imitate the native's speaking style, the use of an alignment method is necessary. The information carried in the visual dynamics of speech complements the information of the acoustic speech signal and has the potential to enhance its comprehension. Therefore, it is important that the signal processing takes into account both audio and visual modalities. Our implementation follows a two steps approach. In the first step, the relative timing differences between the corresponding audiovisual utterances are identified using a timing analysis technique. The result of this step is the estimation of the timing relationship (time-warping path) between the two input sequences. The goal of the second step is to appropriately time-scale the original sequence(s) in accordance with the time-warping path so that the sequences are synchronous with the timing reference. The following subsections further discuss these two steps and present the details of our implementation.

4.1 TIME ALIGNMENT

For the time alignment of the two audiovisual signals, we have implemented a Dynamic Time Warping (DTW) algorithm (Müller, 2007). This choice is motivated by the fact that DTW does not require a phonetic transcription of the karaoke utterances, nor training of an acoustic model on the learner's voice as it is the case in HMM-based forced alignment methods (Brognaux, 2016). In addition, integrating visual features in the DTW feature vector is a relatively straightforward task.

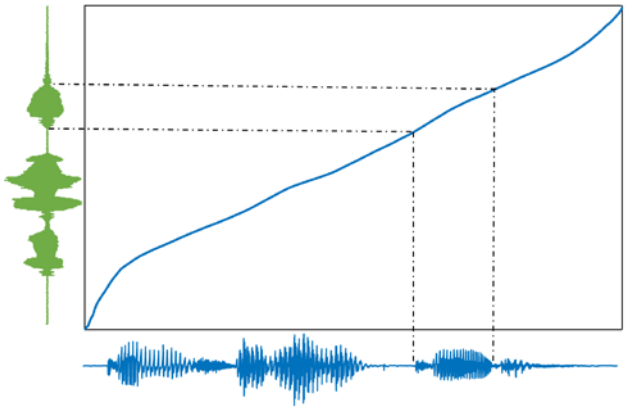

Figure 4 | Example of time-warping path between a reference (vertical axis) and repeated (horizontal axis) speech signals. The dashed lines connect the boundaries of the same distinct sound in both utterances.

Essentially, DTW algorithms aim at measuring the similarity between two temporal sequences which may vary in speed. The two sequences are aligned by warping the time axis of their feature vectors (e.g., Mel-frequency Cepstral Coefficients (MFCCs) of speech, or visual cues such as normalized lips distance) iteratively until an optimal match, in terms of similarity, between the two sequences is reached. In our implementation, the similarity matrix is defined as the cosine distance between the feature vectors (Turetsky, 2003). Figure 4 shows an example of the computed time-warping path between two speech utterances produced by two speakers with different voice characteristics (i.e., one male, one female).

In order to achieve a robust alignment performance, we have incorporated two main improvements in the DTW implementation. First, to compensate for the fact that different speakers have vocal tracts of different characteristics, resulting to speaker depended speech signals, a Vocal Tract Length Normalization (VTLN) has been adopted. Similarly to the approach of (Soens, 2012b), the VTLN is performed prior to the DTW by applying different frequency warping functions to the learner's speech signal (Stadniczuk, 2013). The parameters of each function are iteratively estimated. The VTLN warping function and parameters that maximize the total similarity measure over the resulting time-warping path are selected for the given utterance.

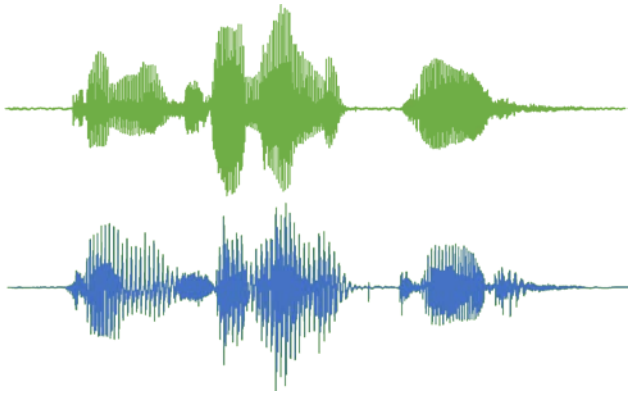

Figure 5 | The two speech signals after synchronization where each distinct sound is located at the same time instant (horizontal axis) in both utterances.

Furthermore, the computation of the time-warping path through dynamic programming is performed in forward and backward fashion (i.e., starting from the last processing frame). The two time-warping paths obtained from this process are then averaged in a single final time-warping path estimate. To the best of our knowledge, this technique has never been applied in previous DTW implementations and, in our experience, it results in smoother time-warping paths, a favorable property when performing time-scaling of speech signals (Soens, 2012).

4.2 TIME SYNCHRONIZATION

Typically, in the case of speech, a time-scaled modified version of the signal is constructed by simply overlap-adding windowed segments from the original speech signal, a method known as Overlap-Add Synthesis method (OLA). Different algorithms can be found in literature. A variant known to result in high quality output for speech signals, is the Waveform Similarity based Overlap-Add (WSOLA) (Verhelst, 1993). Besides producing high quality output when applied to speech (i.e., prosodic aspects of speech, such as timbre and pitch, stay unaffected), WSOLA has low computational cost, it is robust against background noises, and does not require any additional preprocessing (such as pitch-marks estimation that is necessary in pitch-synchronous OLA methods).

The automatic temporal synchronization of two speech utterances has been previously investigated in the context of different applications such as e.g., automatic dialogue replacement systems (Soens, 2012). In these applications, similarly to the spoken karaoke, due to the dynamic nature of the time-scaling, the scaling factor of WSOLA is not constant

Figure 6 | Example of two speakers before (upper panel) and after (lower panel) visual synchronization. In the latter, both speakers pronounce the same phoneme and hence their realizations (e.g., lips rounding) can be efficiently compared.

but proportional to the inverse of the time-warping path as described in (Verhelst, 1997).

A previous analysis (Verhelst, 2003) has shown that distortions can occur with neither the time-warping path estimation, nor the time-scaling procedure being responsible. These distortions are related to acoustic-phonetic differences e.g., due to strong co-articulation in one of the utterances. In order to reduce potential audible distortions, our implementation proposes the modification of both utterances, an approach that to the best of our knowledge has never been considered in previous studies. The spoken karaoke, unlike other applications where the timing of the reference signal needs to be strictly respected (e.g., as in voice replacement), allows some flexibility under the condition that the produced utterances are synchronized. Hence, an artificial time reference can be used for time-scaling the signals. This artificial time reference is chosen so that it minimizes the joint time-scaling amount that is required for modifying both utterances. As a result, distortions due to excessive time-scaling of only one of the utterances are avoided. Informal testing suggests that distributing distortions in both utterances results in less audible artifacts when compared to the traditional approach where only one of the two utterances is modified.

In addition, in our implementation, speech pauses exceeding 150ms are detected and removed before time-scaling is performed using an energy-based Voice Activity Detection (VAD). Long silent segments in speech signals are known to result in horizontal and vertical stretches (singularities) in the time-warping path and cause distortions during the WSOLA synthesis process (Soens, 2012). After removal of these segments, a moving average filter

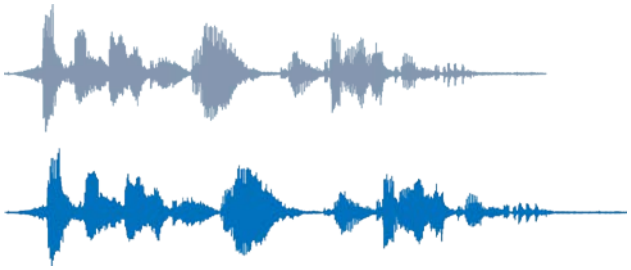

Figure 7 | The original utterance (up) and its slowed down version (bottom) after time-scaling with a constant factor of 1.2.

is applied for further smoothing the time-warping path. Finally, the removed pauses are appropriately reintroduced in the obtained time-scaled signals. A visual example of the time-scaling output considering the above enhancements for the same two utterances as in Figure 4 is shown in Figure 5.

Concerning the video sequences, their time-scaling is more straight-forward. It can be achieved through re-sampling of the original frames, either by duplicating and dropping video frames, or by using a variable frame rate in accordance to the estimated time-warping paths. It is worth noting that in the scope of this paper, we assume that the audio and video input streams are synchronous and hence we do not address "lip sync" issues where audio events occur before or after the associated video frames. Figure 6 shows an example of visual synchronization implemented by duplicating and dropping video frames.

4.3 SPEECH-RATE MODIFICATION

A distinct application of the time-scale modification is the speech-rate modification. A key issue in speech-rate modification is to speed up or down the speech signal without affecting its frequency content (e.g., the perceived pitch), nor introducing any artifacts. In the context of the interactive spoken karaoke, speech-rate modification is applied for automatically adapting the perceived intelligibility of the native speech to the level and preferences of each learner. Previous research has not only identified speech-rate as a major factor affecting second language comprehension, but it has also highlighted the importance of giving listeners control of the degree of its modification (Zhao, 1997).

In our system, the speech-rate modification is performed before each utterance's playback and it is based on the same time-scaling algorithm that is used for the time synchronization. Besides its robust outcome, the scaling factor of WSOLA can be easily controlled by the learner. An example of speech-rate modification is shown in Figure 7. In this example the speech is slowed down with a constant factor of 1.2, hence the duration of the modified utterance is scaling factor times the length of the original utterance, and so the speech rate is inversely proportionally decreased.

4.4 SOUND SPATIALIZATION

The immersive aspects of the narrative story as described in Section 2.2 are reinforced by the use of binaural technology for rendering 3D audio and sound effects (Møller, 1992). The special 3D audio is presented to the learner via headphones. To model the relationship between the intended virtual acoustic source position and the signals received by the listener, we have employed the binaural Head-Related Impulse Responses (HRIR). The spatial 3D audio was efficiently synthesized at playback time using frequency domain convolution of the story's audio dialogues and sound effects with the appropriate HRIR. To this purpose we have utilized an available HRIR database (Algazi, 2001) which offers a wide range of measurements with satisfactory spatial resolution.

5 | CONCLUSION & FUTURE WORK

In this paper, we presented our approach in developing a foreign language pronunciation training prototype system. A key aspect of our approach is the use of storytelling and gamification elements for supporting a novel spoken karaoke concept in combination with a new self-evaluation paradigm which is powered by audiovisual signal processing techniques. We discussed the system components, their integration, and highlighted implementation issues. We also presented a modular architecture for immersive language learning applications.

Technical validation of the intended functionality has been conducted in component level as well as for the system as a whole. We plan to perform user experience testing for formally evaluating the performance and effectiveness of our system. Regarding the system's impact on foreign language pronunciation learning, our intention is to evaluate the learners' progress after long term use of the system taking into account various learning conditions. Moreover, we intend to evaluate the new interaction modalities in combination with existing automatic pronunciation assessment functionality developed within GRAAL project. Besides feedback

on usability, these experiments will also serve data collection purposes for further system development.

Initial informal feedback during the development of the system indicates that creating immersive conditions could have a positive impact on pronunciation learning. Moreover, an automatic pronunciation assessment could provide input to higher level visualizations related to the mouth opening, the position of the tongue, the lips rounding, etc. These generated visual patterns (e.g., a 3D avatar and 2D sagittal planes) could deliver a complementary stimulating evaluation feedback to the learner. The pronunciation analysis results could be used in the future to feed the empowerment and personalization modules, as well as to contribute to immersive aspects. To this extend, pronunciation exercises adapted to each learner's needs and preferences could be proposed based on the assessment outcome. The spoken karaoke concept offers a unique opportunity to experiment with various pedagogical conditions as, e.g., the adaptation of the system to the modality that is preferred by each learner. Finally, from signal processing point of view, we believe that the fusion of audio and visual modalities could further improve the perception of the corrective feedback by the learner.

ACKNOWLEDGEMENTS

This research work was carried out in the scope of the eNTERFACE 2017 workshop organized at the Centre of Digital Creativity (CCD), Escola das Artes, Universidade Católica Portuguesa. Parts of the research reported in this paper were performed in the context of the Région Wallonne FSO project GRAAL (#1510519), and the FEDER project UserMEDIA (#501907-379156) co-funded by the European Union and the Région Wallonne.

ENDNOTES

[1] GRAAL: Guidage en Réalité Augmentée pour l'Apprentissage des Langues.

[2] Information regarding eNTERFACE workshops can be found at http://www.enterface.net

[3] Unity website: https://unity3d.com

REFERENCES

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF Database. Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics

Barrós-Loscertales, A., Ventura-Campos, N., Visser, M., Alsius, A., Pallier, C., Rivera, C. Á., & Soto-Faraco, S. (2013). Neural correlates of audiovisual speech processing in a second language. Journal of Brain and Language

Briet, G., Collige, V., & Rassart, E. (2014). La prononciation en classe. Presses universitaires de Grenoble

Brognaux, S., & Drugman, T. (2016). HMM-based Speech Segmentation: Improvements of Fully Automatic Approaches. IEEE/ACM Trans. Audio Speech Lang. Process., 24(1)

Cugelman, B. (2013). Gamification: what it is and why it matters to digital health behavior change developers. JMIR Serious Games, 1 (1)

Fette, I., & Melnikov, A. (2011). The Websocket Protocol, IETF, RFC 6455.

Hamari, J., Koivisto, J., & Sarsa H. (2014). Does gamification work? A literature review of empirical studies on gamification. Proceedings of 47th Hawaii International Conference on System Sciences (HICSS)

Kalogeras, S. (2013). Media-education Convergence: Applying Transmedia Storytelling Edutainment in E-Learning Environments. International Journal of Information and Communication Technology Education 9(2).

Miller, A. S., Cafazzo, J. A., & Seto, E. (2014). A game plan: Gamification design principles in mHealth applications for chronic disease management. Health informatics journal, 22(2), 184-193

Møller, H. (1992). Fundamentals of Binaural Technology. Applied Acoustics, 36, 171-218.

Müller, M. (2007). Information Retrieval for Music and Motion, chapter Dynamic Time Warping, 69-84, Springer, Berlin, Heidelberg

Soens, P., & Verhelst, W. (2012). On split Dynamic Time Warping for robust Automatic Dialogue Replacement. Signal Processing, 92, 439-454

Soens, P., & Verhelst, W. (2012b). An iterative bilinear frequency warping approach to robust

speaker-independent time synchronization. Proceedings of 20th European Signal Processing Conference (EUSIPCO)

Stadniczuk, D., Bauckmann, G., & Suendermann-Oeft, D. (2013). An Open-Source Octave Toolbox for VTLN-Based Voice Conversion. Proceedings of International Conference of the German Society for Computational Linguistics and Language Technology

Turetsky, R., & Ellis, D. (2003). Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses. Proceedings of 4th International Symposium on Music Information Retrieval (ISMIR)

Verhelst, W., & Roelands, M. (1993). An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Verhelst, W. (1997). Automatic post-synchronization of speech utterances. Proceedings of 5th European Conference on Speech Communication and Technology

Verhelst, W., & Brouckxon, H. (2003). Rejection phenomena in inter-signal voice transplantations. Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

Zhao, Y. (1997). The Effects of Listener' Control of Speech Rate on Second Language Comprehension. Applied Linguistics, 18(1), 49-68

BIOGRAPHICAL INFORMATION

Georgios Athanasopoulos combines academic and industry experience in the complete life-cycle of various projects. He received a diploma in Electrical Engineering from the University of Patras in 2004 and a post-graduate degree in Signal Processing from the National University of Athens in 2006. In 2016, he obtained his PhD from Vrije Universiteit Brussel, where he worked on multichannel audition for robots. He is currently a postdoctoral researcher at the Pixels and Interactions Lab, Université catholique de Louvain. His main research activities are in the area of audio & speech innovations, acoustic interfaces, rapid prototyping and integration.

Céline Lucas is a speech therapist for adult neurology (Master's degree from Medicine Faculty of Lille University). She has 16 years of practice in Paris university hospitals with brain-damaged patients (stroke, head trauma, dementia diagnosis). She joined Université catholique de Louvain in 2014 as researcher in e-health and patient empowerment for chronic diseases. In 2016, she received a First Spin-Off grant to launch the GRAAL project. GRAAL aims at developing a mobile interactive environment to improve foreign language pronunciation.

Alessandro Cierro and Robin Guérit work at Université catholique de Louvain as interaction designers. They are pursuing a Master's degree in Information and Communication Science and Technology at Université catholique de Louvain. They develop digital media with human-centered design as a key focus. Their interests are on different topics such as creative technologies, serious games and e-learning. Their projects tend to emphasize the pedagogical value of fun in connection with users' needs. Their Master's thesis aims to design and analyze digital learning environments in real context.

Kaori Hagihara obtained her Master's degree in Information Science from Nara Institute of Science and Technology in Japan. She stayed in the Massachusetts Institute of Technology in 2005 as an intern for 10 months. Thereafter, she joined Université catholique de Louvain as a researcher in the PiLAB team. She worked on multiple camera video surveillance systems for train security. She was also involved in the development of the OpenJPIP software: an implementation of JPEG 2000 Part9 (JPIP). She is today involved in the ParkAR project co-funded by the Walloon region and Alterface company. She is developing a multi-agent based multi-camera system for group interactions in augmented distributed spaces.

Julie Chatelain obtained a Master's degree in Computer Science from the Université catholique de Louvain in 2017. She completed her Master's Thesis on e-health applications for the empowerment of diabetic patients. She is now involved as a researcher in the UserMedia project funded by the Walloon region. She is developing web mobile architecture (Mean stack and Unity) for interactive e-learning applications.

Sébastien Lugan graduated his Master's degree at ESIEE (École Supérieure d'Ingénieurs en Électronique et Électrotechnique), his DEA (Diplôme

d'Études Approfondies) at Gaspard-Monge institute of electronics and computer science of Université Paris-Est Marne-la-Vallée (UPEM), and his PhD in engineering sciences from École Polytechnique de Louvain, Université catholique de Louvain. Since 2003, he is a member of the Pixels and Interactions Lab of the ICTEAM (Institute of Information and Communication Technologies, Electronics and Applied Mathematics) at Université catholique de Louvain.

Benoît Macq is currently Professor at Université catholique de Louvain (UCL), in the Telecommunication Laboratory and Pro-Rector of the University. He did his doctoral thesis on perceptual coding for digital TV under the supervision of Prof. Paul Delogne at UCL. He was researcher at Philips Research in 1990 and 1991. He has been senior researcher of the Belgian NSF. Benoît Macq has been visiting scientist at Ecole Polytechnique Fédérale de Lausanne and at the Massachusetts Institute of Technology, Boston. Benoît Macq is teaching and doing his research work in image processing for visual communications. His main research interests are image compression, image watermarking, image analysis for medical and immersive communications.