Discovering the Italian literature: interactive access to audio indexed text resources

Vincenzo Galatà, Alberto Benin, Piero Cosi, Giuseppe Riccardo Leone, Giulio Paci, Giacomo Sommavilla, Fabio Tesser
Institute of Cognitive Sciences and Technologies, National Research Council (ISTC-CNR)
Via Martiri della Libertà, 2 - 35132 Padova, Italy
{vincenzo.galata, alberto.benin, piero.cosi, riccardo.leone, giulio.paci, giacomo.sommavilla, fabio.tesser}@pd.istc.cnr.it

Abstract

In this paper we present a web interface to study Italian through the access to read Italian literature. The system allows to browse the content, search for specific words and listen to the correct pronunciation produced by native speakers in a given context. This work aims at providing people who are interested in learning Italian with a new way of exploring the Italian culture and literature through a web interface with a search module. By submitting a query, users may browse and listen to the results through several modalities including: a) the voice of a native speaker: if an indexed audio track is available, the user can listen either to the query terms or to the whole context in which they appear (sentence, paragraph, verse); b) a synthetic voice: the user can listen to the results read by a text-to-speech system; c) an avatar: the user can listen to and look at a talking head reading the paragraph and visually reproducing real speech articulatory movements. In its up to date version, different speech technologies currently being developed at ISTC-CNR are implemented into a single framework. The system will be described in detail and hints for future work are discussed.

Keywords: Italian, Audio Resource, Web Interface

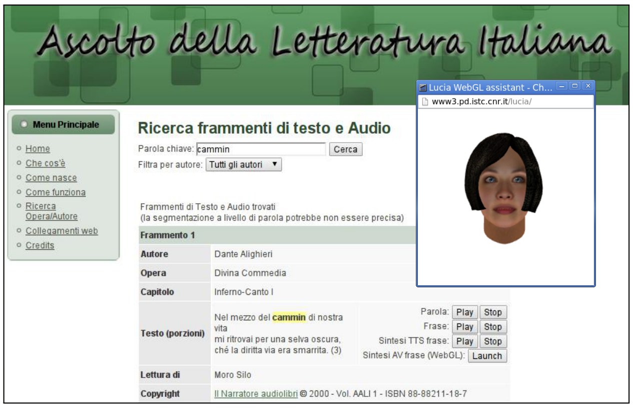
1. Introduction

Rapid technological advances, allowing the storage of huge amounts of data as well as their transmission through the internet, make it possible to develop new and interactive communication tools and interfaces that provide any potential user with an access to resources previously not available. As stated by (Calzolari et al., 2012, p. 41), Italian language counts more than 60 million native speakers around the world and represents the 20th most spoken language worldwide, thanks also to 125 million Italian second language speakers and to large emigrant communities still speaking Italian. The main objective of this project is to give the broad population of Italian users the chance to access Italian literature resources, providing people learning Italian with examples of how words are uttered by native speakers.

The proposed framework, shown in Figure 1, is based on a previous prototype by (Drioli and Cosi, 2008) which has now been developed as part of a broader project called "Wikimemo.it, Il Portale della Lingua e della Cultura Italiana"[1], founded by the Italian Ministry of Education, University and Research (MIUR) and aiming at promoting Italian culture through reading and listening. This is achieved with a web interface that allows users to browse the indexed content by searching for specific words and by listening to how they are pronounced by native speakers within a given context.

The proposed tool integrates several speech technologies currently being developed at ISTC-CNR that add the possibility for the user to listen to synthetic speech as well as look at a human-like avatar reading aloud Italian literary texts.

This represents a first step for the development of new technological and advanced Italian language teaching resources: an interactive talking head could be effective, for example, in helping young kids to enjoy alphabetization exercises, or deaf users to "lip read" the pronunciation of words pronounced by an avatar.



Ricerca frammenti di testo e Audio

Parola chiave: cammin

Filtra per autore: Tutti gli autori

Frammenti di Testo e Audio trovati
(la segmentazione a livello di parola potrebbe non essere precisa)

Frammento 1		
Autore	Dante Alighieri	
Opera	Divina Commedia	
Capitolo	Inferno-Canto I	
Testo (porzioni)	Nel mezzo del cammin di nostra vita mi ritrovai per una selva oscura, ché la diritta via era smarrita. (3)	Parola: Play Stop Frase: Play Stop Sintesi TTS frase: Play Stop Sintesi AV frase (WebGL): Launch
Lettura di	Moro Silo	
Copyright	Il Narratore audiolibri © 2000 - Vol. AALI 1 - ISBN 88-88211-18-7	

Figure 1: A screen-shot of the interface showing the search result for the keyword "cammin" (highlighted in yellow): all the available information for the given search results is given in a table format. The buttons to play the results are visible on the right side of the table.

In the framework here proposed the avatar's task is limited to the reading of Italian literature resources, whereas in the Wikimemo project its role is to guide the users while surfing on a web site by means of a speech based interface: traditional interfaces (keyboard and mouse) represent always an initial barrier for those approaching the high-tech world, especially for elders.

The development and study of more friendly human-machine interfaces has been a very active area for a long time. In particular, the anthropomorphism of machines through the use of avatars communicating with natural language reduces and eliminates the need of a long training

[1] "Wikimemo.it, The portal of the Italian language and culture".

phase which often causes "aversion" towards modern technologies. Numerous research projects followed this trend: the robo-receptionist VALERY[2], the European project SOPRANO (Sixsmith et al., 2009), the domotic environment ROBOCARE (Cesta et al., 2009) developed at ISTC-CNR and the recent English project RITA (Simmons et al., 2003), just to name a few.

The next sections present and briefly describe the technologies that allow the user to browse and listen to the results through the voice of a native speaker, a synthetic voice and an avatar.

In particular, Section 2. describes the audio indexing procedure by means of Forced Alignment (FA). Section 3. and Section 4. describe the synthetic voice and the avatar, respectively, that are based on audio and audio/visual (AV) speech synthesis. These two modalities are mainly used to read resources when no indexed audio is available. Section 5. provides technical details about the environment that runs the interface. In Section 6. figures of speech in the Italian literature are presented as an example of potential use of the framework here outlined.

Finally, some conclusions and hints for future directions are briefly addressed.

2. Forced Alignment and audio indexing

Each literary resource included in this project is divided into chapters, and each chapter may also have a corresponding audio read by a native speaker, so the possibility to listen to it has been provided. Forced Alignment (FA) has been performed both at word and sentence level for the orthographical transcriptions and their corresponding speech audio data. The resulting information has been indexed so that users may query the system and retrieve specific portions of the audio track: a complete paragraph or just a few words.

Because FA technology is not designed to process very long audio files, the audio files has been splitted in sentences by means of a sentence-level FA carried out through EasyAlign (Goldman, 2011). The resulting segmentation has been manually revised before performing word-level FA. Word-level FA has been carried out off-line by means of Sonic (Pellom and Hacioglu, 2001) with a specific technology for Italian described in (Paci et al., 2013).

FA information is stored in a segmentation file that contains, for each word, beginning and end time positions as well as beginning and end time positions for each sentence, paragraph or verse. The segmentation file is used by a server-side application that splits the stored audio file to match the query results.

3. Italian MaryTTS

Whenever audio tracks are not available for a keyword, or if the user wants to compare real with synthetic speech, the system provides text-to-speech (TTS) synthesis capabilities. Moreover the TTS is able to generate phoneme duration information necessary to drive the synchronization between the speech and the articulatory movements of the talking agent (avatar).

The TTS facility has been implemented by means of the open-source speech synthesis engine MaryTTS (Modular Architecture for Research on speech sYnthesis) system (Schröder and Trouvain, 2003). Recently (Tesser et al., 2013) released language modules and voices for Italian. The voice used in this work is the result of an automatic corpus driven building procedure. Two voices are used in the system: a male voice, trained on the audio version of a weekly magazine that publishes the Italian translation of international articles[3]; and a female one, trained on a corpus of selected sentences uttered by a young native Italian non-professional female speaker (*Lucia*).

Both voices have been built using semi-automatic paragraph segmentation, followed by the HTK-based training procedure for forced alignment and HTS (Zen et al., 2007) to build the HMM-based voice.

4. WebGL avatar

Users interacting with the web interface are also offered the opportunity to listen and look at a female talking head reading the paragraph which contains the keyword. This task has been achieved through LUCIA-WebGL (Leone and Cosi, 2011; Cosi et al., 2011), an MPEG-4 Facial Animation Parameters (FAP) driven talking head that implements a decoder compatible with the "Predictable Facial Animation Object Profile" (Pandzic and Forchheimer, 2003). LUCIA-WebGL is totally based on real human data collected by means of ELITE, a fully automatic movement analyzer for 3D kinematics data acquisition by (Ferrigno and Pedotti, 1985). This information is used to create the lip articulatory model and to drive directly the talking face in order to visually generate and reproduce real speech articulatory movements and human-like facial movements. LUCIA-WebGL relies on a high quality 3D model and a fine co-articulation model, that is capable of reproducing quite precisely the real cinematic articulatory movements (Cosi et al., 2003). The WebGL client-server architecture separates the visualization/interaction process (on the WebGL client) from the generation of the audio/visual streams necessary for the animation of the talking head (on the server side). These streams have a very low bit-rate and can perform very well also with slow connections and mobile devices.

LUCIA-WebGL currently speaks with the Italian version of MaryTTS (see section 3.).

WebGL is currently available by default on desktop browsers such as Google Chrome and Mozilla Firefox as well as being supported by others. WebGL applications are also supported by by most of the mobile browsers with a few exceptions.[4]. Unfortunately, due to particular commercial policies, Apple does not allow WebGL applications to run natively on iOS: a possible solution is presented in (Benin et al., 2012) and is left for future work.

5. Technical details

As described above, the interface allows users to choose among different kinds of modalities to listen to audio in-

[2] http://www.kent.ac.uk/sspssr/ccp/rita/index.html

[3] "Internazionale", http://www.internazionale.it

[4] See the continuously updated Wikipedia page dedicated to WebGL, http://it.wikipedia.org/wiki/WebGL

dexed Italian literature resources.

The web interface is currently implemented in Joomla!1.5 running on a LAMP (Linux, Apache, MySql, Php) server. The search module and the queries are written in Php and Html.

A MySQL database has been populated with textual resources for the accessible works. Copyright information and details about the stored literary text (author, title of the work, etc.) are also included.

The information related to time alignment for audio retrieval is contained in text files, one for each literature resource stored in the database.

After querying the system with a desired keyword, the results are printed on an Html page in a user-friendly table format.

To avoid speeding down the system during the output of the results on the Html page, the media splitting process is carried out in background. The output audio files and the commands for playback operations are embedded into a completely free and open source cross-browser JavaScript media library, developed as a jQuery Plugin[5]. All the extracted portions are stored as temporary files allowing also multiuser access. A cronjob bash script periodically deletes the temporary files so to avoid space consumption on the webserver.

The audio outputs are generated in different ways according to the consultation modality including: 1) the voice of a native speaker, 2) the synthetic voice and 3) the avatar.

A short description of the three modalities' working principles is hereafter provided.

5.1. The voice of a native speaker

When an indexed audio track is available in the database, the user is given the chance to listen to the query term or to the whole context in which it appears (sentence, paragraph or verse) using a real voice of a professional speaker.

In this particular browsing modality, some limits are encountered. Since the FA has been carried out at word level, it is currently possible to search only for whole-word keywords of at least three characters. The reasons for this limit are the following: a) the FA has been manually controlled only at sentence level; b) smaller words (articles, pronouns etc.) are more prone to be mis-aligned than longer words; c) searching for text portions smaller than three chars could produce many unwanted results.

With reference to the query results for the voice of the native speaker (if present in the database), the generation of the media files is demanded to a server-side software application[6].

Mp3Splt splits the stored audio file of the whole work into smaller parts. Using the audio indexing information obtained as in Section 2., only the audio of the searched keyword is extracted. The same is done for the sentences, paragraphs or verses in which the keyword appears.

5.2. The synthetic voice

The database may also contain literature resources for which no audio is available. In those cases the user is offered the possibility to listen to the results by means of a text-to-speech system (see Section 3.).

In this case, by clicking on the play button, the user sends a link containing the query result from the database directly to the server running MaryTTS.

5.3. The avatar

A further possibility offered to the user is to listen to and look at a talking head reading the paragraph and visually reproducing real speech articulatory movements (see Section 4.).

For this purpose, an embedded web-server application controlling the WebGL animation for the avatar has been developed through the microhttpd library[7].

Clicking on the button that launches the AV animation, Apache 2 mod-proxy forwards the request to the microhttpd web-server which synthesizes the selected text by means of the AV engine module: the synchronized audio and FAP streams are used to play the aligned AV portion in the WebGL client window. The AV streaming is played in a separate window that automatically closes when the AV streaming is completed.

6. An example of usage: discovering the figures of speech in the Italian literature

The tool we present here is about audio-indexing of Italian literature resources and the possibility to listen to them by searching and browsing through specific keywords. In this paragraph we present an example of how the proposed framework may be used to explore and to listen, for instance, to particular figures of speech that are present in the Italian literature.

The figures of speech including repetition (e.g. alliteration, onomatopoeia and pun) are here particularly interesting because they involve the sounds of speech. Exploring those figures by listening to a real (or synthetic) voice offers the user the opportunity to grasp and to appreciate more directly and realistically the effect the writer/author intends wants to convey to the reader. Some examples for each of the three figures of speech are presented in the next subsections. The effect that those figures produce with the sounds of speech is easy to appreciate if spoken aloud from the voice of a professional actor. In the following paragraphs we provide a few examples that may be found in the database currently available.

6.1. Alliteration

Alliteration modifies and intensifies the phonetic relationship between words. It involves the repetition of "homophonous accented, syllable initial phonemes, as in house and home, cash and carry, tea for two, usually for stylistic or poetic effect." (Bussmann et al., 1996, p. 42).

In D. Alighieri's *Tanto gentile e tanto onesta pare*, a sonnet taken from la *Vita Nova*, different alliterations can be found to soften the reading as in the pairing of nasal sounds "n, m,

[5] See `http://jplayer.org/` for more technical details

[6] Mp3Spl-project is a utility to split mp3, ogg vorbis and native FLAC files selecting a begin and an end time position, without decoding. See `http://mp3splt.sourceforge.net/mp3splt_page/home.php`

[7] `https://gnu.org/software/libmicrohttpd/`

gn" with dental sounds "t, d" and the repetition of vowels and consonants.

Ta**nt**o ge**nt**ile e ta**nt**o onesta pare
l**a** donn**a** mi**a** qu**a**nd 'ell**a** **a**ltrui s**a**lut**a**,
ch'o**gn**e li**ng**ua deve**n** **t**rema**nd**o muta,
e li occhi no l'ardiscon di guardare.

Another example may be found in D.Alighieri's *Divina Commedia*, "Inferno, Canto V", where there is a reiteration of vowels and consonants:

Amor, ch'**a** nullo **a**m**a**to **a**m**a**r perdon**a**, (v. 103)
[...]
E **c**addi **c**ome **c**orpo morto **c**ade. (v. 142)

6.2. Onomatopoeia

An onomatopoeia is a word that phonetically imitates or suggests the source of the sound that it describes. In other words (Bussmann et al., 1996, p. 532), "an onomatopoetic (onomatopoeia) representation of a concept, often consisting of reduplicated syllables and not adhering to the phonotactic structure of the given language."

An example of onomatopoeia can be found in D. Alighieri's "Paradiso, Canto X":

[...] che l'una parte e l'altra tira e urge,
tin tin sonando con sì dolce nota,
che 'l ben disposto spirto d'amor turge; (v. 144)

6.3. The pun

The pun (or paronomasia) is another figure of speech involving repetition. According to (Bussmann et al., 1996, p. 968) it is considered: "A play on words through the coupling of words that sound similar but which are very different semantically and etymologically [...]."

The pun can be used to establish a peremptory association between two concepts, in order to enhance the musicality of a verse or for humorous scopes.

Some examples are, again, present in D. Alighieri's *Divina Commedia*. The first one is taken from "Inferno, Canto I":

[...] e non mi si partia dinanzi al volto
anzi 'mpediva tanto il mio cammino
ch'i' fui per ritornar più **volte volto**. (v. 36)

The second one is instead taken from "Paradiso, Canto III":

E questa sorte, che par giù cotanto,
però n'è data, perché fuor negletti
li nostri **vòti**, e **voti** in alcun canto. (v. 57)

7. Conclusions and future works

A system merging together different speech technologies has been presented (accessible on-line at `http://www3.pd.istc.cnr.it/ali/`). In its current state, the system is already a useful tool for learners of Italian getting them closer to the practice of reading and allowing them to interact with indexed text resources. The database is currently being expanded with new entries and more sophisticated analysis tools will be developed to simplify the database population: one of the problematic issues of this current system is, in fact, related to database population that is currently being carried out manually for each entry.

Further investigation will be carried out to improve the search module hence allowing users to interrogate the system with more sophisticated queries. The introduction of other high-level indexing information, such as Part-Of-Speech (POS) tagging, syntactic dependency trees and prosodic labeling, may improve the usefulness of the system for researchers interested in other aspects of spoken and written language. These aspects are left for future work.

The web-interface has been tested with different browsers to ensure high robustness and, so far, a good support is available for both Desktop and Mobile browsers, with only a few exceptions. On Apple iOS systems, the use of WebGL is not permitted: this issue will be addressed in future work as proposed in (Benin et al., 2012).

Finally, thanks to the integration of different technologies and besides the few limitations highlighted in the previous sections, the proposed framework represents a robust resource for Italian language learners and an interesting challenge for the creation of more sophisticated and appealing educational tools.

8. Acknowledgements

The authors would like to thank "Il Narratore audiolibri" which kindly allowed the use of their audio-books containing the voice of professional speakers reading aloud some works of the Italian literature featured in the web interface. Vincenzo Galatà's work has been supported by WIKIMEMO.IT (The Portal of Italian Language and Culture, FIRB Project, RBNE078K93, Italian Ministry of Education, University and Research). This research was partly funded by EU-FP7 project ALIZ-E (ICT-248116).

9. References

Benin, A., Leone, G. R., and Cosi, P. (2012). A 3D Talking Head for mobile devices based on unofficial iOS WebGL support. In Mouton, C., Posada, J., Jung, Y., and Cabral, M., editors, *Web3D*, pages 117–120. ACM.

Bussmann, H., Trauth, G., and Kazzazi, K. (1996). *Routledge dictionary of language and linguistics*. Taylor & Francis, London, New York.

Calzolari, N., Magnini, B., Soria, C., and Speranza, M. (2012). *La Lingua Italiana nell'Era Digitale – The Italian Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer.

Cesta, A., Iocchi, L., Leone, G. R., Nardi, D., Pecora, F., and Rasconi, R. (2009). Robotic, sensory and problem-solving ingredients for the future home. In Monekosso, D., Remagnino, P., and Kuno, Y., editors, *Intelligent Environments*, Advanced Information and Knowledge Processing, pages 67–87. Springer London.

Cosi, P., Fusaro, A., and Tisato, G. (2003). LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. In *Proceedings of Eurospeech 2003*, pages 2269–2272, Geneva, Switzerland. Eurospeech.

Cosi, P., Leone, G. R., and Paci, G. (2011). LUCIA: An open source 3D expressive avatar for multimodal H.M.I. In Antonio Camurri, Cristina Costa, G. V., editor, *INTETAIN 2011, Intelligent Technologies, for Interactive Environments*, pages 1 – 10.

Drioli, C. and Cosi, P. (2008). Audio indexing for an interactive Italian literature management system. In *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, page 2170.

Ferrigno, G. and Pedotti, A. (1985). Elite: A digital dedicated hardware system for movement analysis via real-time tv signal processing. *IEEE Transactions on Biomedical Engineering*, pages 943–950.

Goldman, J. P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *INTERSPEECH 2011*, Firenze, Italy.

Leone, G. R. and Cosi, P. (2011). LUCIA-WebGL: A Web Based Italian MPEG-4 Talking Head. In Salvi, G., Beskow, J., Engwall, O., and Al Moubayed, S., editors, *Proceedings of the International Conference on Audio-Visual Speech Processing 2011*, pages 123 – 126.

Paci, G., Sommavilla, G., and Cosi, P. (2013). SAD-Based Italian Forced Alignment Strategies. In Magnini, B., Cutugno, F., Falcone, M., and Pianta, E., editors, *Evaluation of Natural Language and Speech Tools for Italian*, volume 7689 of *Lecture Notes in Computer Science*, pages 322–329. Springer Berlin Heidelberg.

Pandzic, I. S. and Forchheimer, R., editors. (2003). *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. John Wiley & Sons, Inc., New York, NY, USA.

Pellom, B. L. and Hacioglu, K. (2001). SONIC: The University of Colorado continuous speech recognizer TR-CSLR-2001-01. Technical report, University of Colorado, Boulder, Colorado, March.

Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

Simmons, R., Goldberg, D., Goode, A., Montemerlo, M., Roy, N., Sellner, B., Urmson, C., Bugajska, M., Coblenz, M., Macmahon, M., Perzanowski, D., Horswill, I., Zubek, R., Kortenkamp, D., Wolfe, B., Milam, T., Inc, M., and Maxwell, B. (2003). Grace: An autonomous robot for the aaai robot challenge. *AI Magazine*, 24:51–72.

Sixsmith, A., Meuller, S., Lull, F., Klein, M., Bierhoff, I., Delaney, S., and Savage, R. (2009). Soprano an ambient assisted living system for supporting older people at home. In Mokhtari, M., Khalil, I., Bauchet, J., Zhang, D., and Nugent, C., editors, *Ambient Assistive Health and Wellness Management in the Heart of the City*, volume 5597 of *Lecture Notes in Computer Science*, pages 233–236. Springer Berlin Heidelberg.

Tesser, F., Paci, G., Sommavilla, G., and Cosi, P. (2013). A new language and a new voice for MARY-TTS. In Galatà, V., editor, *9th national congress, Italian Association for Speech Sciences*, pages 435–443, Venice, Italy.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *The 6th International Workshop on Speech Synthesis*, pages 294–299, Bonn, Germany.