Acoustic Data Analysis from Multi-Sensor Capture in Rare Singing: Cantu in Paghjella Case Study

by
Lise Crevier-Buchman, Angélique Amelot, Samer K. Al Kork, Martine Adda-Decker, Nicolas Audibert, Patrick Chawah, Bruce Denby, Thibaut Fux, Aurore Jaumard-Hakoun, Pierre Roussel, Maureen Stone, Jacqueline Vaissiere, Kele Xu, Claire Pillot-Loiseau

Reprinted from

International Journal of
Heritage in the Digital Era

volume 4 number 1 2015

Acoustic Data Analysis from Multi-Sensor Capture in Rare Singing: Cantu in Paghjella Case Study

Lise Crevier-Buchman[1], Angélique Amelot[1], Samer K. Al Kork[2,3], Martine Adda-Decker[1], Nicolas Audibert[1], Patrick Chawah[1], Bruce Denby[2,3], Thibaut Fux[1], Aurore Jaumard-Hakoun[2,3], Pierre Roussel[3], Maureen Stone[4], Jacqueline Vaissiere[1], Kele Xu[2,3], Claire Pillot-Loiseau[1]

[1]Phonetics and Phonology Laboratory, LPP-CNRS, UMR7018, Univ. Paris3 Sorbonne Nouvelle

[2]Université Pierre Marie Curie, Paris, France

[3]Signal Processing and Machine Learning Lab, ESPCI Paris-Tech, Paris, France

[4]Vocal Tract Visualization Lab, Univ of Maryland Dental School, Baltimore, USA,

lise.buchman@numericable.fr

[Received date; Accepted date] – to be inserted later

Acoustic Data Analysis from Multi-Sensor Capture in Rare Singing: Cantu in Paghjella Case Study

Lise Crevier-Buchman, Angélique Amelot, Samer K. Al Kork, Martine Adda-Decker, Nicolas Audibert, Patrick Chawah, Bruce Denby, Thibaut Fux, Aurore Jaumard-Hakoun, Pierre Roussel, Maureen Stone, Jacqueline Vaissiere, Kele Xu, Claire Pillot-Loiseau

Abstract:

This paper deals with new capturing technologies to safeguard and transmit endangered intangible cultural heritage including Corsican multipart singing technique. The described work, part of the European FP7 *i-Treasures* project, aims at increasing our knowledge on rare singing techniques. This paper includes (i) a presentation of our light hyper-helmet with 5 non-invasive sensors (microphone, camera, ultrasound sensor, piezoelectric sensor, electroglottograph), (ii) the data acquisition process and software modules for visualization and data analysis, (iii) a case study on acoustic analysis of voice quality for the UNESCO labelled traditional *Cantu in Paghjella*. We have identified specific features for this singing style, such as changes in vocal quality, especially concerning the energy in the speaking and singing formant frequency region, a nasal vibration that seems to occur during singing, as well as laryngeal mechanism characteristics. These capturing and analysis technologies will contribute to define relevant features for a future educational platform.

1. INTRODUCTION

The main objective of i-Treasures project "Intangible treasures - capturing the intangible cultural heritage and learning the rare know-how of living human treasures" [1], is to develop an open and extendable platform to provide access to Intangible Cultural Heritage (ICH) resources, and to contribute to the transmission of rare know-how from Living Human Treasures to apprentices. In order to facilitate the transmission of such learning information, we are working on an educational platform that makes the link between the master and the apprentice by means of a variety of sensors and developed software [2].

Manifestations of human intelligence and creativeness constitute our ICH,,some of them being in need of urgent safeguarding. Therefore, the i-Treasures project deals with a number of traditional European ICH, amongst others, the singing techniques of the UNESCO (2012) inventory of ICH [3]. The aim of this paper is to present new methodology to capture rare singing, with multiple sensors, to better understanding their acoustic specificities, and to contribute in the elaboration of training program and pedagogical tools.

To explore the complex and mainly hidden human vocal tract, non-invasive sensing techniques have been used including modelling and recognition of vocal tract operation, voice articulations, acoustic speech and music sounds. Our system, based on vocal tract sensing methods developed for speech production and recognition [4], consists of a prototype lightweight "hyper-helmet" (Fig. 1). Multi-sensor data acquisition, visualisation and analysis protocols have also been designed to allow multi-media synchronous recording of singing voice [5].

The paper is structured as follows. Section 2 presents the recording protocol and the methodology to capture raw data and launch analysis: software designed for data recording and acquisition (i-THRec) and software designed as a MatLab tool for visualisation and analysis (i-THAn). In section 3, we will present a case study centred on voice quality and vowel articulation in Corsican *Cantu in Paghjella*

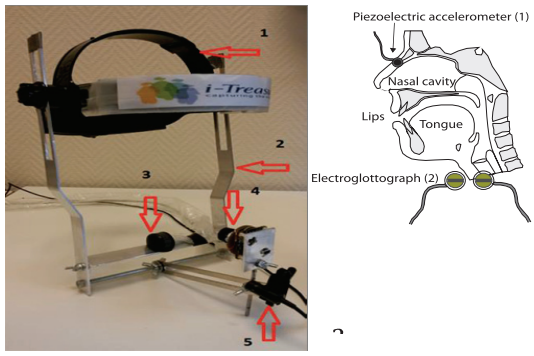


Figure 1. (a) Multi-sensor Hyper-Helmet: 1) Adjustable headband, 2) Probe height adjustment strut, 3) Adjustable US probe platform, 4) Lip camera with proximity and orientation adjustment, 5) Microphone. (b) Schematic of the placement of non-helmet sensors, including the (1) accelerometer piezoelectric, (2) electroglottograph (EGG)

from our *in situ* data collection. Finally, we will conclude on the usefulness of our multi-sensor acoustic data stream acquisition system to enhance knowledge of rare singing techniques for learning scenarios.

2. METHODS

To meet the requirements of the rare singing use case and to define relevant features [6], it is necessary to build a recording system that can follow the configurations of the vocal tract – including tongue, lips, vocal folds and soft palate – in real time, and with sufficient accuracy to link image features to actual, physiological elements of the vocal tract. Furthermore, the vocal tract acquisition system must be able to "synchronously" record multi-sensors data.

The following describes the sensors that were used, the dedicated software developed to manage and record sensors and a MATLAB tool allowing visualizing the recorded data.

2.1 Non-Invasive Sensors

To capture the complex and specific articulatory strategies of different types of singing, five sensors are used to identify vocal tract movements and define reliable features for educational scenarios.

The helmet allows simultaneous collection of vocal tract and audio signals. As shown in Fig. 1 (a), it includes an adjustable platform to hold a special custom designed 8MC4X Ultrasound (US) probe in contact with the skin beneath the chin. The probe used is a microconvex 128 elements model with handle removed to reduce its size and weight, which captures a 140° image allowing full visualization of tongue movement. The US machine chosen is the Terason T3000, a system which is lightweight and portable yet retains high image quality, and allows data to be directly exported to a PC via the Firewire port. A video camera (model DFM 22BUC03-ML, CMOS USB mono) is positioned facing the lips. Since differences in background lighting can affect computer recognition of lip motion, the camera is equipped with a visible-blocking optic filter and infrared LED ring, as is frequently done for lip image analysis. Finally, a commercial lapel microphone (model C520L, AKG) is also affixed to the helmet to record sound.

Two non-helmet sensors are directly attached to the body of the singer as indicated in Fig. 1(b). A piezoelectric accelerometer (model Twin spot from K&K sound) attached with double adhesive tape to the nasal bridge of the singer captures nasal bone vibration, which is indicative of nasal resonance during vocal production [7]. Nasal vibrations are important acoustic features in voice perception and has been the topic of numerous phonetic and speech processing studies. It is also implied in some singing techniques that use the nasal cavity as

a resonator in order to modify the timbre of the voice [7].

An ElecroGlottoGraph (EGG, Model EG2-PCX2, Glottal Enterprises Inc.) is placed on the singer's neck. This sensor's output is a signal that is proportional to the vocal fold contact area. By using the DEGG (Derivative ElecroGlottoGraph) signal, opening and closing instants can be identified which are useful to compute the open quotient. [8]. The DEGG is also very helpful for advanced analyses such as inverse filtering [9] aiming to predict the output signal from the glottis, which is essential in the speech production and perception process.

2.2 Data Acquisition: Capturing and Recording

Since configuring separate sensors and recording their outputs may be complicated if they are managed individually, a common module has been specifically designed. The proposed module, named i-THRec (i-Treasures Helmet Recording software), contains multiple Graphical User Interface (GUI) forms, each of them aimed at one of the following objectives: (i) creating directories to organize and store the newly acquired data into corresponding sub-folders (ii) writing.xml files that contain song lyrics to be performed (iii) calibrating the sensors and supervising their performances (iv) operating the recording session and replaying already saved data [10].

A snapshot of the recording windows is illustrated in Fig. 2. Nevertheless, i-THRec does not perform the actual interface with the sensors. The data acquisition from the sensors is handled by using the Real-Time Multi-sensor Advanced Prototyping software [11] (RTMaps®, Intempora inc.). The latter has the ability to acquire, display and record data, based on Synchronized Time stamped Data and could be sufficient by itself since this software included their GUI (i.e. RTMaps studio). However, we prefer to use RTMaps SDK as a toolkit serving in an i-THRec lower layer in a favor of user-friendly software. These data

Figure 2. Screen snapshot of the recording session software [10]. Top: display of Cantu in Paghjella lyrics. Below: 5 streams of the corresponding sensors, from left to right and top to bottom: lips from the camera, tongue contour from the US, time signals from EGG, microphone and piezoelectric sensor.

are henceforth ready to be post-processed using a developed MATLAB graphical user interface (GUI) named i-THAn (i-Treasures Helmet Analysis software).

2.3. Data visualization and analysis

The module referred to as i-THAn (i-Treasures Helmet Analyser) is a MATLAB multimedia tool that manages the data from the multi-sensor hyper-helmet streams captured by i-THRec through RTMaps (Fig. 3, *Left*). Each data stream is recorded in standard format (wav file for analogue signals and raw file for video streams) readable by lots of software. However, the file containing time information is a format specific to RTMaps. This file is essential to synchronously read the data. In order to overcome the limitation of viewing the data only on the computer where RTMaps is installed, a MATLAB GUI has been developed allowing viewing, checking and analysing the signals.

i-THAn software can also play back the audio and video data and extract part of the recording. The aim of this module is to validate the synchronicity of all data streams. In particular, we need to check for potential image data loss due to system overload during capturing, to display synchronized signals and images, to check for noise due to sensor movement, or thermal drift and to check for possible saturation of signals. It also provides a comprehensive set of capabilities to monitor the quality of acquired data regularly, to create measurement reports, figures, images and various documentations.

The current version of i-THAn includes tools dealing with the speech, the EGG and the piezoelectric signals. The pitch information, the open quotient and the spectrogram can be computed and viewed synchronously with the signals.

The operation of i-THAn is illustrated in the screenshot (Fig. 3, *Right*), which shows several types of analysis performed on the data of a *Corsican Paghjella* singer producing a sustained sung vowel /i/. The upper panel shows a narrow-band spectrogram of the vowel, where the harmonics are visible, and the vibrato of the voice with approx. 5

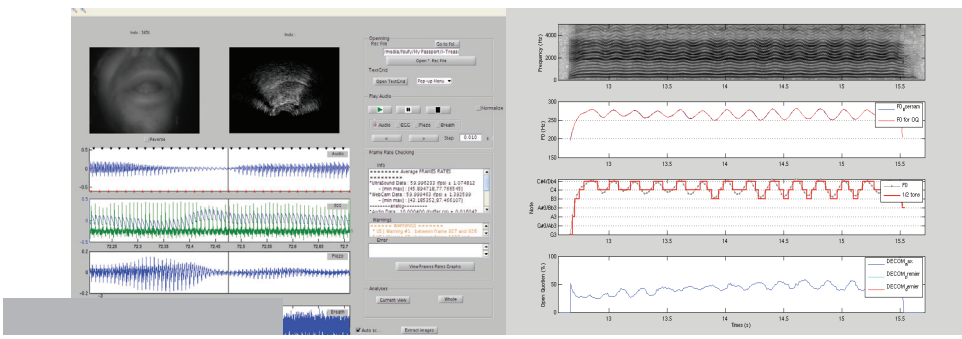


Figure 3. (Left) Screen shot from i-THAn for a Corsican *Paghjella* recording of the sustained singing /i/ vowel. The lip and tongue images; from top to bottom: the acoustic signal, the EGG waveform (blue) and it's derivate (green), the piezoelectric signal. (Right) Analyses figure showing from the top to the bottom: the narrow band spectrogram, the fundamental frequency (F0) of the speech directly on the EGG signal and the F0 used to compute the open quotient (Oq), the F0 in a musical note scale and the Oq.

cycles per second can also be identified. The lower panel shows deferent representations of the Oq.

3. CASE STUDY: THE POLYPHONIC CANTU IN PAGHJELLA

The secular and sacred *Cantu in Paghjella* polyphonic chant of Corsica, joined UNESCO's endangered list of intangible cultural heritage at the end of 2009. It designates the male chant interpreted a cappella by three voices (a *seconda*, a *bassu* and a *terza*) [12,13]. It is still transmitted orally, by intergenerational contact and endogen imitation. The traditional Corsican singing, including the *Cantu in Paghjella*, is often described as highly ornamented (melismatic), with vowel nazalisation and sometimes, glottal constriction with frequent use of reduced intervals (quarter-tones…) [14].

Even if some singers master the solfeggio, the members of this community must learn the skill orally, either by familial transmission, from master to disciple, through exposure to secular or sacred performances, or by the intermediary of audio or audio-visual documents [12].

Only few scientific researchers have studied the polyphonic Corsican singing tradition. Therefore, in the scope of our i-Treasures project we aimed to contribute to the development of a systemic methodology for the preservation, renewal and transmission of rare knowledge to future generations.

The objectives are to explore the voice quality, vowel articulation and tessitura of voice by analysing acoustic, EGG, and piezoelectric accelerometer signals.

3.1. Specific Spoken and Singing Voice Quality in Cantu in Paghjella

In order to study the different aspects of rare singing technique in *Cantu in Paghjella*, and to extract information and features for automatic classification and pedagogical activities and transmission, we collected material of different degrees of complexity: (i) isolated vowels in singing and spoken tasks (/i/, /u/, /e/, /o/, /a/), and (ii) sung vowels extracted from the whole chant. Spoken and sung isolated vowels are compared to vowels embedded in text to capture specific acoustic modifications when singing. With the acoustic signal, we studied the vocalic space through the vocalic triangle and compared spoken and sung situations. Furthermore we analysed the piezoelectric accelerometer signal to compare the use of nasal cavities in the singing situation. The laryngeal behaviour at the glottic level was analysed by calculating the open quotient from the EGG signal. These parameters were expected to contribute to a better understanding of specific singing situations.

Our case study was based on the recording of one expert Corsican *Paghjella* singer BS. He first produced spoken and sung voice with major Corsican vowels and consonants, and then performed two *Paghjella* songs (*Alto Mare* and *O Columba*) in his tessitura, the *secunda* voice.

3.2. Results and Discussion

We used the procedure described previously to record, capture and analyse the spoken and singing performance of our *Cantu in Paghjella* expert singer using the multi-sensor Hyper-Helmet.

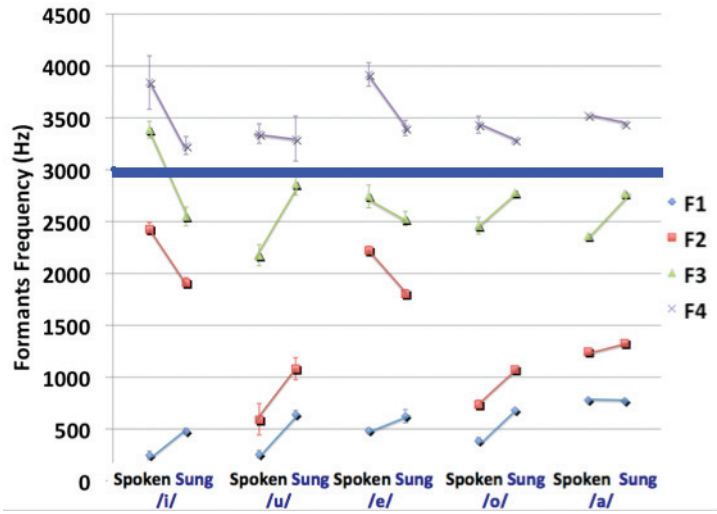
Vowel Pitch. The main 5 vowels [i, u, e, o, a] were produced in speaking and singing voice and repeated 6 times. The mean fundamental frequency (F0) was 128Hz (SD 34) and 259Hz (SD 17) for the spoken and sung vowels respectively.

Formant Frequency. We looked at the displacement of the formant frequencies from spoken to sung voice for the five vowels. The aim was to follow the energy reinforcement in singing and the articulatory adaptation. Formant frequency represents the energy that characterises the vocalic timber and the power of the voice.

The singer's formant is a prominent spectrum envelope peak near 3 kHz that appears in voiced sounds sung by professional singers to make the voice easier to hear. It can be explained as a clustering of formants [15].

Fig. 4 shows mean and standard deviation values of the formant frequencies (F1 to F4) for spoken and sung isolated vowels. The frequency was taken at the middle of each vowel in the spoken and singing mode. According to Sundberg [15], i) the second and third formant frequencies in the front sung vowels do not reach the high values they have in speech; ii) the fourth formant frequencies vary much less in singing than in speech. Sundberg (1987) described an "extra" formant corresponding to the clustering of the third and the fourth formants in the spoken vowels. According to this author, this "extra formant" exists also for spoken vowels but to a higher frequency than sung ones. In our data, there is a clustering of the F3 and F4 frequencies near 3000Hz from speech to singing especially for back vowels; iii) the F1 increase from speech to singing for each vowel is due to the F0 increase and probably due to mandible aperture; iv) the F2 frequency decreases from speech to singing only for anterior vowels /i/ and /e/ because of the « darkening » and « covering » of such vowels in singing [15]. It's not necessary for /u/ and /o/ which are already dark vowels. The rising F3 towards 3000 Hz can participate in higher acoustic energy.

Figure 4. Mean and standard deviations value of formant frequencies (Hz) F1 to F4 for spoken and sung isolated vowels. The bold line around 3000 Hz is situated between the F3 and F4, where the singing formant is expected.

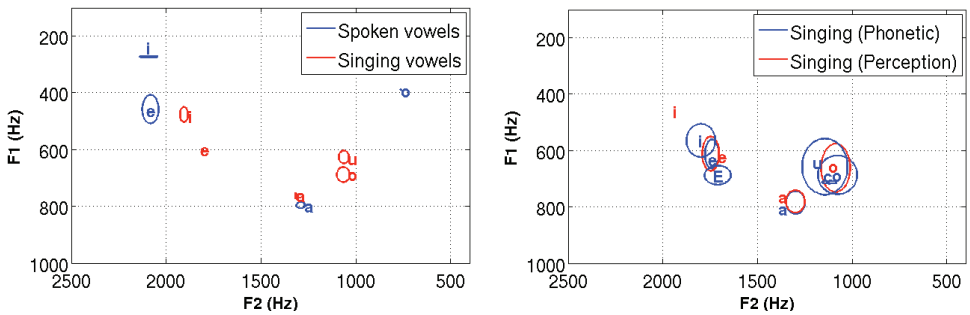


Vocalic Triangle. We measured the average value of the formant frequency F1 and F2 for the 5 vowels in various production contexts (isolated, spoken/singing, and singing). When considering the chant, we extracted the vowels from two different procedures; one perceptual annotation by listening to the song, and one phonological annotation by considering the expected vowel from the written text. The aim was to identify changes in the vocalic inventory when singing.

The results are presented in Fig. 5. When singing, we noticed a confusion between /i/ and /e/ and between /u/ and /o/ in both perceptual and phonologic singing vowels. The higher F1 is related to the production of a more open vowel (/i/ becomes /e/) and F2 is more centralized, corresponding to a less precise articulatory target or more centred vowel.

LTAS (Long Term Average Spectrum). We looked at the spectral distribution comparing spoken and singing mode for all the vowels separately. There is an increase in energy from 1500Hz to 3500Hz for the sung vowels. The peak observed at 3500Hz could be considered as

Figure 5. Left: F1/F2 for spoken (blue) and sung (red) isolated vowels. Right: F1/F2 for sung vowels extracted from the chant (red: vowel identified perceptually; blue: vowels identified phonologically).



intermediate between the speaking [16,17] and the singing formant. The results can be seen in Fig. 6. Interestingly, although the singer is in a singing mode, he has a tendency to use a spoken mechanism to project his voice. It can be seen by a larger peak at around 3000 Hz than expected for the singing formant.

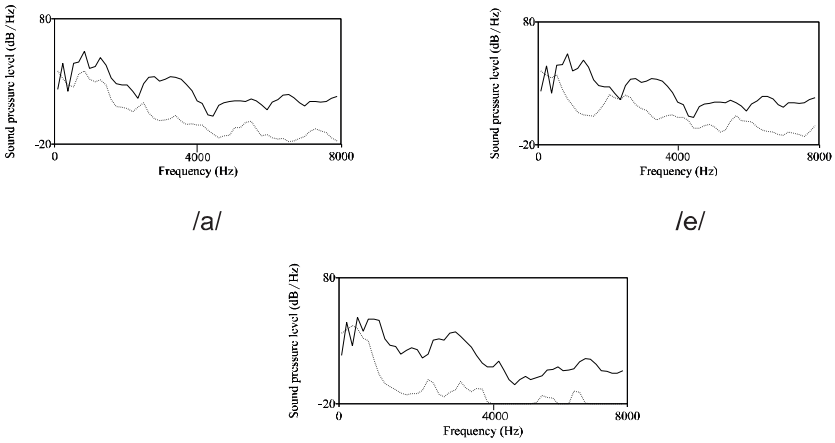


Figure 6. LTAS for 3 isolated vowels (/a/, /e/ and /o/) in singing task (solid line) and in spoken task (dotted line), bandwidth 150Hz.
å

Nasal Vibration. The aim of these measurements was to identify the nasal component of sound in the singing mode as a specificity of these chants. We calculated the root mean square for acoustic oral (from the signal of the microphone) and acoustic nasal signal (from the piezoelectric accelerometer signal).

During speech, changes in vocal intensity were relatively low and during nasalization the accelerometer signal grew significantly [7]. Our data in Fig. 7 show an important nasal vibration during oral vowel production in the singing task. The results showed the importance of the nasal cavity during *Paghjella* singing.

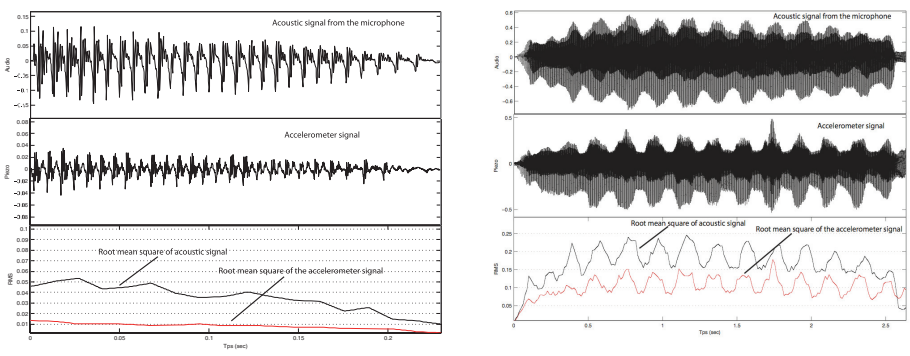


Figure 7. The two figures show the acoustic (top) and accelerometer (mid) signals for the same vowel /a/ in spoken (left) and singing (right) task. The black line in the RMS measurements (bottom) corresponds to the oral signal and the red line to the RMS of the accelerometer signal.

Laryngeal behaviour. The laryngeal mechanism was measured by calculating the Open Quotient (Oq) extracted from the EGG signal at the glottis level for each spoken and sung vowel. In our singer, the singing Oq is lower than in speech (F0: 263Hz, Oq: 0,4 and F0: 127Hz, Oq: 0,5 respectively), reflecting a strong laryngeal muscle contraction, like in "pressed" phonation. This behaviour participates in the acoustic enhancement.

4. CONCLUSIONS

We developed innovative methodologies for multimodal voice analysis and we used five sensors to record and identify vocal tract movements and define reliable features for educational scenarios. Our visible real-time acoustic specificities for singing sound, nasality and laryngeal involvement can be considered as valuable information for the apprentice. Additional novelty comes from the fact that the technology will be first applied to traditional songs. New technical problems and constraints may require further research, however a good basis will exist given that i-Treasures will provide modules that analyse the most important components of an artistic performance.

The applications developed within the project can be extended in the future for other types of cultural heritage, as well as for teaching and learning specific skills.

ACKNOWLEDGEMENTS

This work was partially funded by the European FP7 i-Treasures project (Intangible Treasures - Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures FP7-ICT-2011-9-600676-i-Treasures). It was also supported by the French Investissements d'Avenir -Labex EFL program (ANR-10-LABX-0083).

REFERENCES

[1] Intangible treasures - capturing the intangible cultural heritage and learning the rare know-how of living human treasures," http://i-treasures.eu/

[2] Dimitropoulos, K., Manitsaris, S., Tsalakanidou, F., Nikolopoulos, S., Denby, B., Kork, S.A., Crevier-Buchman, L., Pillot-Loiseau, C., Dupont, S., Tilmanne, J., Ott, M., Alivizatou, M., Yilmaz, E., Hadjileontiadis, L., Charisis, V., Deroo, O., Manitsaris, D., Kompatsiaris, I., and Grammalidis, N.: Capturing the intangible: An introduction to the i-treasures project, *Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Lisbon, Portugal (2014)*

[3] UNESCO: "Convention of the safeguarding of intangible cultural heritage of UNESCO," http://www.unesco.org/culture/ich/en/convention"

[4] Cai, J., Hueber, T., Denby, D., Benaroya, E.L., Chollet, G., Roussel, P., Dreyfus, G., and Crevier-Buchman, L.: A visual speech recognition system for an ultrasound-based silent speech interface. *Proceeding of International Congress Phonetics Sciences, Florence, Italy, 384-387 (2011)*

[5] Al Kork, S.K., Jaumard-Hakoun, A., Adda-Decker, M., Amelot, A., Crevier-Buchman, L., Chawah, P., Dreyfus, G., Fux, T., Pillot, C., Roussel, P., Stone, M., Xu, K., and Denby, B.: A Multi-Sensor Helmet to Capture Rare Singing, An Intangible Cultural Heritage Study*, Proceedings of 10th International Seminar on Speech Production, Cologne, Germany (2014).*

[6] Jaumard-Hakoun, A., Al Kork, S. K., Adda-Decker, M., Amelot, A., Crevier-Buchman, L., Fux, T., Pillot-Loiseau, C., Roussel, P., Stone, M., Dreyfus, G., and Denby B.: Capturing, analyzing, and transmitting intangible cultural heritage with the i-Treasures project", *Proceedings of Ultrafest VI, Edinburgh (2013)*.

[7] Stevens, K.N., Kalikow, D.N., and Willemain, T.R.: A miniature accelerometer for detecting glottal waveforms and nasalization, *Journal of Speech and Hearing Research*, 18, 594-599 (1975)

[8] Henrich, N., Roubeau, B., and Castellengo, M.: On the use of electroglottography for characterisation of the laryngeal mechanisms, *Proceedings of Stockholm Music Acoustics Conference, Stockholm, Sweden (2003)*.

[9] Henrich, N., d'Alessandro, C., Castellengo, M., and Doval, B.: On the use of the derivative of electroglottographic signals for characterization of non-pathological voice phonation, *Journal of the Acoustical Society of America*, 115 (3), 1321-1332 (2004)

[10] Chawah, P., Al Kork, S. K., Fux, T., Adda-Decker, M., Amelot, A., Audibert, N., Denby, B., Dreyfus, G., Jaumard-Hakoun, A., Pillot-Loiseau, C., Roussel, P., Stone, M., Xu, K., and Crevier-Buchman, L.: An educational platform to capture, visualize and analyze rare singing. *Proceedings of Interspeech, Singapour (2014)*

[11] RTMaps: http://www.intempora.com/rtmaps4/rtmaps-software/overview.html

[12] Bithell, C.: Transported by song – Corsican voices from oral tradition to world stage, *Bohlman & Stokes eds., The Scarecrow Press (2007)*

[13] Peres, M.: Le chant religieux corse. Etat, comparaison, perspectives (1996)

[14] Hergott C.: Patrimonialisation d'une pratique vocale: l'exemple du chant polyphonique en Corse, PhD Thesis, Université de Corse (2011)

[15] Sundberg J.: The Science of the Singing Voice, *DeKalb, Ill: Northern Illinois University Press (1987)*

[16] Leino, T.: Long-term average spectrum study on speaking voice quality in male voices. *SMAC93 Proceedings of the Stockholm Music Acoustics Conference, Stockholm, Sweden (1993)*

[17] Bele, I.V.: The speaker's formant, *Journal of Voice*, 20 (4), 555–578 (2006)