A Multi-Sensor Helmet to Capture Rare Singing, an Intangible Cultural Heritage Study

S. K. Al Kork[1,2], A. Jaumard-Hakoun[1,2], M. Adda-Decker[3], A. Amelot[3], L. Buchman[3], P. Chawah[3], G. Dreyfus[2], T. Fux[3], C. Pillot-Loiseau[3], P. Roussel[2], M. Stone[4], K. Xu[1,2], B. Denby[1,2]

[1]*Université Pierre Marie Curie, Paris, France,*

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

[3]*Phonetics and Phonology Laboratory, LPP-CNRS, UMR7018, University Paris3 Sorbonne Nouvelle*

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

samer_alkork@hotmail.com, aurore.hakoun@espci.fr, madda@univ-paris3.fr, angelique.amelot@univ-paris3.fr, lise.buchman@numericable.fr, patrick.chawah@gmail.com, gerard.dreyfus@espci.fr, thibaut.fux@univ-paris3.fr, claire.pillot@univ-paris3.fr, pierre.roussel@espci.fr, mstone@umaryland.edu, kelele.xu@gmail.com, denby@ieee.org

Abstract

A portable helmet based system has been developed to capture motor behavior during singing and other oral-motor functions in a non-laboratory experimental environment. The system, based on vocal tract sensing methods developed for speech production and recognition, consists of a lightweight "hyper-helmet" containing an ultrasonic (US) transducer to capture tongue movement, a video camera for the lips, and a microphone, coupled with a further sensor suite including an electroglottograph (EGG), nose-mounted accelerometer, and respiration sensor. The system has been tested on two rare, endangered singing musical styles, Corsican "Cantu in Paghjella", and Byzantine hymns from Mount Athos, Greece. The versatility of the approach is furthermore demonstrated by capturing a contemporary singing style known as "Human Beat Box."

Keywords: portable speech collection system, ultrasound, EGG, accelerometer, data acquisition, intangible cultural heritage, i-Treasures project.

1. Introduction

A major objective of the i-Treasures project is to provide students with innovative multi-media feedback to train specific articulatory strategies for different type of rare singing, considered an endangered Intangible Cultural Heritage. To this end, i-Treasures will carry out vocal tract capture during rare singing performances to enable study of production mechanisms, and to define reliable features for a subsequent animation of these articulatory movements for use in educational scenarios and automatic classification tasks. To accomplish this, it is necessary to build a system that can record the configuration of the vocal tract – including tongue lips, vocal folds and soft palate – in real time, and with sufficient accuracy to establish a link between image features and actual, physiological elements of the vocal tract.

Ultrasound, US, is a popular non-invasive technique for real time imaging of the vocal tract. Examples of portable devices that acquire US images of the tongue and video of the lips, for applications in speech synthesis and silent speech interfaces (Denby and Stone 2004) (Cai, et al. 2011), have been described in the literature. Ultrasound also requires no external magnetic field, and can thus also be readily complemented with other sensors. Here, we present a system based on a helmet containing a US probe, lip camera, and microphone, coupled with a suite of other sensors including electroglottograph (EGG), to measure and record vocal fold contact movement during speech; a piezoelectric accelerometer, for detecting the nasal resonance of speech sounds (Stevens, Kalikow and Willemain 1990); and a respiration sensor belt to determine breathing modalities (Tsui and Hsiao 2013).

The proposed system is advantageous in that 1) it is lighter and easier to wear for long periods than other solutions proposed in the literature (Wrench, Scobbie and Linden 2007); and 2) the combination with other sensors has the potential to greatly enhance our knowledge of rare singing techniques, and allow the extraction of sensorimotor features in order to drive a 3D avatar for learning scenarios.

2. Methods

Figure 1 presents a schematic overview of the modules contained in the capture system. Each sensor first requires specific gain tuning and/or zero calibration protocols, before the streams are simultaneously and synchronously recorded with the RTMaps toolkit, which will be described in section 2.2.

2.1. Helmet design and sensor setup

The helmet allows simultaneous collection of vocal tract and audio signals. As shown in Figure 2, it includes an adjustable platform to hold the US probe in contact with the skin beneath the chin. The probe used is a microconvex 128 element model with handle removed to reduce its size and weight, which captures a 140° image to allow 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 Firewire. A video camera (from The Imaging Source) is positioned facing the lips (Figure 2). Since differences in background lighting can affect computer recognition of lip motion, the camera is equipped with a visible-blocking filter and infrared LED ring, as is frequently done for lip image analysis. Finally, a commercial lapel microphone (Audio-Technica Pro 70) is also affixed to the helmet to record sound.

The three non-helmet sensors are directly attached to the body of the singer as indicated in Figure 3. An accelerometer attached with adhesive tape to the nasal bridge of the singer captures nasal bone vibration related to nasal tract airway resistance, which is indicative of nasal resonance during vocal production. Nasality is an important acoustic feature in voice timbre. An EGG (Model EG2-PCX2, Glottal Enterprises Inc.) is strapped to the singer's neck to record a time dependent signal whose peaks are reliable indicators of glottal opening and closing instances (Henrich, et al. 2004). Finally, on the singer's chest, a

respiration sensor or “breathing belt” is affixed to measure breathing modalities during singing.

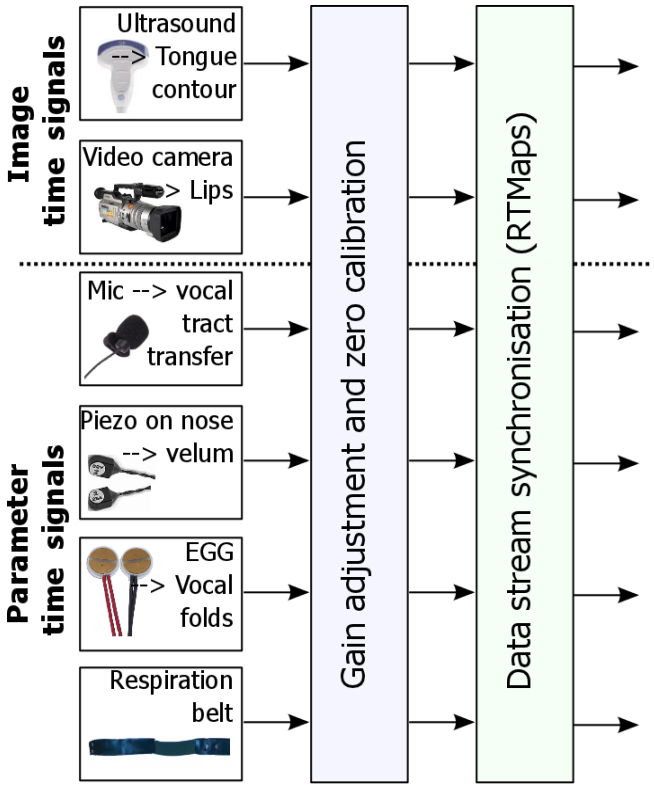

Figure 1 *Overview of the vocal tract capture system for the rare singing sub-use cases. Six sensors (left), capture motion of five physical events: (1) tongue, (2) lips, (3) the acoustic speech wave, (4) nasal resonance, (5) vocal folds, and (6) respiratory muscle. Each instrument is processed (middle) and digitally recorded (right).*

2.2. Data Acquisition and system design

2.2.1. System architecture

The data acquisition system must be able to synchronously record US and video data at sufficiently high frame rates to correctly characterize the movements of the tongue and lips, as well as the acoustic speech signals, the EGG, the accelerometer and the respiratory `waveforms. The acquisition platform was developed using the Real-Time, Multi-sensor, Advanced Prototyping Software (RTMaps®, Intempora Inc, Paris FR).

2.2.2. RTMaps real time user interface

The data acquisition platform has the ability both to record and display data in real time, and the acquired data can be stored locally or transferred over a network. Figure 4 displays a screen shot from the platform. Ultrasound and video images are streamed at a rate of 60 frames per second, then stored in either .bmp or jpeg format. Image size for US and camera are 320 by 240 pixels and 640 by 480 pixels respectively. The EGG, the microphone, the piezoelectric accelerometer and respiration belt are interfaced to a four-input USB sound card (AudiBox44VSL) whose output interfaces to the acquisition system. These four analog input signals are sampled at 44100 Hz with a 16 bit encoding. The sampled analog signals are saved to a .wav format.

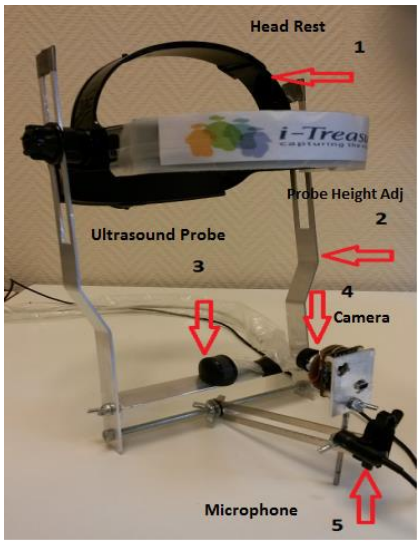

Figure 2 *Multi-sensor Hyper-Helmet: 1) Adjustable headband, 2) Probe height adjustment strut, 3) Adjustable US probe platform, 4) Lip camera with proximity and orientation adjustment, 5) Microphone.*

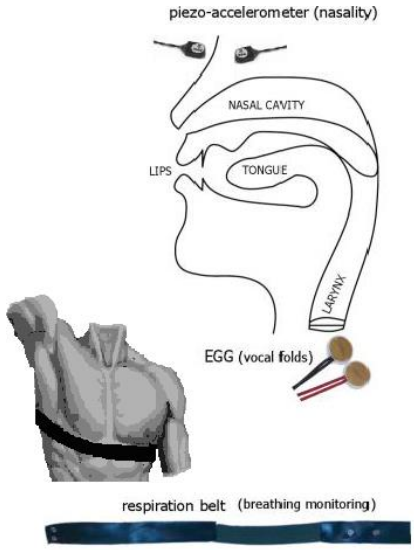

Figure 3 *Schematic of the placement of non-helmet sensors, including the (1) accelerometer, (2) EGG, and (3) respiration belt*

2.2.3. Definition of recording material

In order to study a variety of singing techniques, and to extract features for automatic classification and pedagogical scenarios, we will collect material of varying degrees of complexity: isolated vowels (/i/, /u/, /e/, /o/, /a/), CV syllables (/papapapapa/, /tatatatata/, /kakakakaka/...), sung phrases and entire pieces, where the material is to be produced both in spoken and singing modes. Byzantine chant, Corsican Paghjella, and the contemporary singing style known as “Human Beat Box”, HBB, have been chosen for study. For Byzantine chant, different styles (Mount Athos *vs* Ecumenical Patriarchate of Constantinople styles for example) have been selected. For Corsican Paghjella, we propose to study *versa* (melodies) from three different regions known for their traditional singing styles: Rusio, Sermanu and Tagliu-Isolacciu. The protocol to be used for Sardinian Canto a Tenore is still under discussion. For HBB,

basic material will be recorded as defined in (Proctor, et al. 2013), as well as short HBB phrases and longer performances in different styles, with details still to be defined.

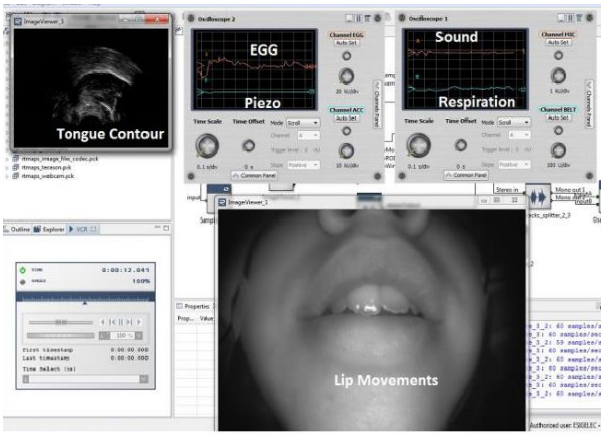

Figure 4 *RTMaps Data Acquisition User Interface, showing simultaneous recording and visualization of ultrasound tongue images, lip video, and the four analog sensors.*

3. Preliminary experimental results

3.1. Sensor synchronization

Sensor synchronization is crucial in our study since extracted sensor features are ultimately to be used to drive a 3D tongue avatar. To check synchrony, an event common to all sensor channels is created using the procedure illustrated in Figure 5. A syringe containing ultrasound gel is struck by a spring-loaded weight, ejecting a gel droplet that strikes the head of the ultrasound probe and shorts together the two sides of the EGG sensor. Vibration induced in the syringe body is detected in the nasality accelerometer, and the acoustic signal of the weight hitting the syringe is recorded by the microphone. Finally, the video camera normally used for the lips is positioned so that it can also capture the droplet as it is ejected. The resulting signals obtained from the 5 sensors are shown in Figure 6

We have calculated the timestamps at which the gel droplet occurred in all sensors. The time stamp at which the droplet was triggered for the microphone, piezo and EGG was 4.060s, 4.070s and 4.070s respectively. The droplet was captured by the camera and Ultrasound at image number 244 in sequence and at calculated time stamps of 4.066s based stream rate of 60 fps (Figure 6). The average latency among all sensors is thus of the order of 10ms.

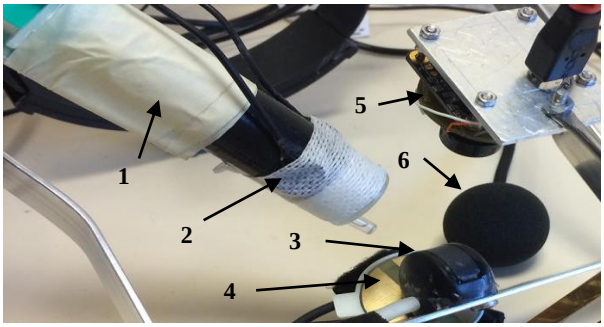

Figure 5 *Sensor synchronization experimental setup test. 1) Syringe, 2) Piezo, 3) EGG sensor, 4) US transducer, 5) Camera, 6) Microphone*

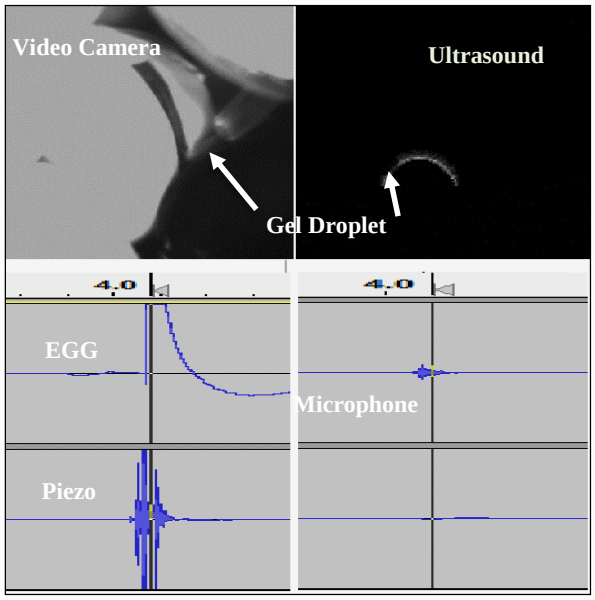

Figure 6 *Top left, arrow indicates gel droplet ejected from tip of syringe; top right, arrow indicates arrival point of droplet on ultrasound probe; bottom left and right, time waveforms of EGG, Piezo, and microphone, signals, see discussion in text*

3.2. Rare-singing data collection

Our data collection activities are focused on preparatory steps dealing with technical, operational and functional requirements. These steps are listed here below and some details are provided for the major steps.

3.2.1. Assessment phase of the hyper-helmet for the different singing types:

The voice capture system has been tested with one expert singer each for the Human Beat Box, HBB (Da Vox) (Figure 7a), the Paghjella (B. Sarocchi) Figure 7b, and the Byzantine (D. Manousis) (Figure 7c) musical styles. Each singer participated in a recording session to validate the helmet with respect to his or her style, and to assist us in specifying an appropriate data collection protocol. A recording session consists of three phases: 1) singer preparation (wearing of the helmet and body sensors), 2) sensor calibration and, 3) data collection proper. The three phases need to be optimized with respect to time delays in sensor setup and ease of use. Figure 6 illustrates the versatility of the helmet, which can be used with any head size and shape and does not impede the singing function.

The Byzantine expert singer (D. Manousis) produced vowels in both singing and speaking mode, before singing a dozen segments of Byzantine chants in both Mount Athos and Ecumenical Patriarchate of Constantinople styles. The Corsican Paghjella singer (B. Sarocchi) first produced spoken and singing voice using isolated vowels and connected CV syllables with major Corsican vowels and consonants, and then performed three Paghjella songs. For the HBB case, we undertook several testing and recording sessions with our expert, (Da Vox). A specific problem for HBB is the difficulty of stabilizing of the ultrasound probe in view of the large range of motion of the jaw in this singing style, as compared to the

other styles. Each of the three singing styles produced about 30 minutes of singing material, which are being used to develop and assess the next steps to be undertaken in the continuing development of our synchronous data collection platform, as well as our data calibration, data display and analysis modules.

Figure 7 *a) HBB expert singer (Da Vox), b) Paghjella expert singer (B. Sarocchi), and c) Byzantine expert singer (D. Manousis)*

3.3. Pilot data

In this section, some samples of the pilot data are presented. Figure 8 shows similar vocalic /o/-like sounds by three singers specialized in different singing styles: Byzantine chant (left), Cantu in Paghjella singer (secunda voice, middle) and HBB singer (right). Initial data of ultrasound tongue images and video camera lips image are displayed for all the above singers as shown Figure 9, Figure 10 and Figure 11.

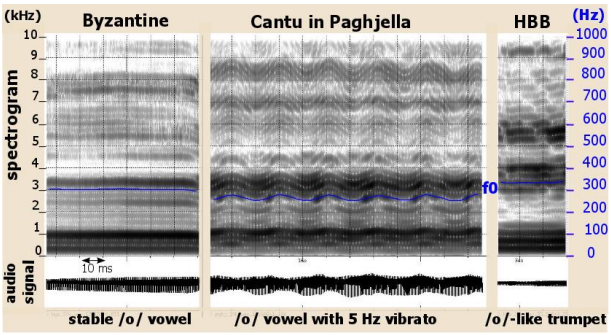

Figure 8 *Three vocalic /o/ samples of different style singing voice (Byzantine,Cantu in Paghjella, HBB). Spectrograms (10kHz band in black on the left) and f0 curves (in blue on the right) are shown in the upper panel, corresponding acoustic waveforms in the lower one.*

Figure *9 Byzantine Singing Case: Left) Ultrasound Tongue image. Right) Video camera lip image*

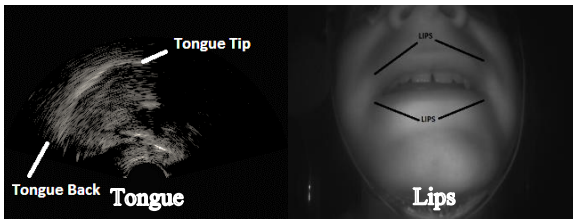

Figure 10 *HBB singing case: Left) Ultrasound Tongue image. Right) Video camera lip image*

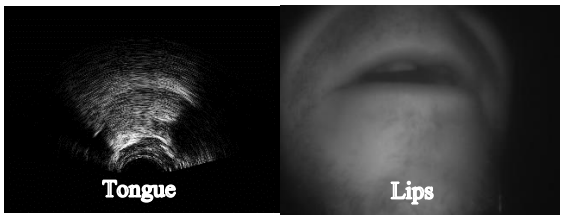

Figure 11 *Corsican singing case: Left) Ultrasound Tongue image. Right) Video camera lip image*

4. Discussion and conclusion

The vocal tract acquisition system for the preservation of rare singing techniques has evaluated on initial data taken on all three singing cases. Results are promising with respect to system reliability, portability and quality of the recorded data. The system still needs to be tested on a larger number of users, and the helmet better adapted to the HBB case due to rapid jaw movements encountered in this singing style. The next steps will involve integrating a feature extraction processing block for all the sensors into the RTMaps platform, as well as provide an easy user friendly interface to drive and control a 3D vocal tract avatar to be developed in future.

5. Acknowledgements

This work is funded by the European Commission via the i-Treasures (FP7-ICT-2011-9-600676-i-Treasures).

6. References

Cai, Jun, Thomas Hueber, Bruce Denby, Elie-Laurent Benaroya, Gérard Chollet, Pierre Roussel, Gérard Dreyfus, and Lise Crevier-Buchman. 2011. "A visual speech recognition system for an ultrasound-based silent speech interface." *In Proc. of ICPhS.* pp. 384-387.

Denby, B., and M. Stone. 2004. "Speech synthesis from real time ultrasound images of the tongue." *In Acoustics, Speech and Signal Processing; ICASSP.* pp 685-I688.

Henrich, N., C. d'Alessandro, M. Castellengo, and B. Doval. 2004. "On the use of the derivative of electroglottographic signals for characterization of nonpathological voice phonation." *Journal of the Acoustical Society of America* pp. 1321-1332.

Proctor, M., E. Bresch, D. Byrd, K. Nayak, and S. Narayanan. 2013. "Paralinguistic mechanisms of production in human "beatboxing": A real-time magnetic resonance imaging study." *Journal of the Acoustical Society of America (JASA)* 133 (2): pp. 1043-1054.

Stevens, K.N., D.N. Kalikow, and T.R. Willemain. 1990. "A miniature accelerometer for detecting glottal waveforms and nasalization." *Journal of Speech and Hearing Research JSHR* pp. 594-599.

Tsui, W.H., and Tzu-Chien Hsiao. 2013. "Method and System on Detecting Absominals for singing." *Proc. IEEE EMBC.* pp.1-8.

Wrench, A., J. Scobbie, and M. Linden. 2007. "Evaluation of a helmet to hold ultrasound probe." *Ultrafest IV.*