Gesture Recognition & Chanting Assessment For Byzantine Music Learning

Kostas Kokkinidis[1], Theodoros Mastoras[2], Athanasia Stergiaki[3]Paraskevi Kritopoulou[4]

SMN Lab, Department of Applied Informatics, University of Macedonia, 156 Egnatia Street, GR-546 36 Thessaloniki, Greece

Abstract

Recent works related to digital self-instruction environments, present scarce efforts to provide combined instruction for gestural and vocal skills. Based upon a recently introduced learning and teaching method of vocal music, this research utilizes existing technologies to achieve the development of such a learning environment. The presented system administers the learning experience in order to improve the motion, sound and rhythm related skills of the student. Student performance is compared with a pre-recorded instructor performance in order to provide customized feedback that bespeaks the flaws of the former performance. Motion and sound-capturing technologies are combined, and related feature extraction algorithms are applied. The gestural and vocal features of the instructor performance are compared off-line to those of the student performance, in order to detect the differences, while the tempo is indicated through gestures. The system evaluates constantly the performances in order to provide visual feedback based on their differences. The aim is for the student to reproduce the instructor performance in an approximate manner. An assessment formula for the student performance is proposed and tested for its validity and accuracy. The selected musical genre on which this system was applied is Byzantine music, since its complexity and variety tests the existing sound recognition algorithms.

Keywords: sensory-motor learning; Byzantine music; multimodal interaction; gesture recognition; singing voice assessment

1. Introduction

Singing is augmented speech in terms of tonality and rythm, that is comperhented as a form of music. In most genres of music, vocalists accompany their sound performance with gestures in order to sustain the rythm during their performance. Yet, there are singing educational methods that attemp to impart tonallity through gestures (Kodály et al, 1974). A gesture is a form of non-vocal communication that requires the contribution of physical movements by various body parts (Kendom, 2004). What differentiates gestures from general human movements that are performed during activities such as walking, is the intention to interact and the fact that they are not performed intermittently.

The technological evolution has rendered the ability to create new learning environments that do not require teacher supervision. On the presented research, recent advances in the field

of capturing multimodal data focused on gestures and sounds, are combined with algorithms able to decipher the information included in the captured data. The purpose is to deliver an interactive learning system that uses a novel vocal learning method. This learning method utilizes gestures to impart tonality (Patronas, 2018). More specifically, the student enacts simultaneously sounds and gestures performed mainly with the palm, to reproduce a musical chart while s/he is being recorded. The recorded data is analyzed off-line, to extract the position of the palm throughout the performance duration. Each note performed by the student is segregated and compared with a reference performance of an expert (teacher). The feedback is the comparison result displayed as a traffic sign to the student together with a grading evaluation. The learning method was applied on the music genre of Byzantine music. The system grading method was evaluated through an in-practice experiment.

2. State of the Art

Traditionally, the microphone serves as the main sound capturing technology, while there are various motion capture technologies. Marker-based systems have been used for modeling musical performances (Rasamimanana & Bevilacqua, 2009), while marker-less systems such as depth cameras like Kinect (Microsoft Kinect, 2014) are used even commercially, for gesture recognition. Visual recognition of gestures is also a field with many applications on Human Computer Interaction (Rautaray & Agrawal, 2015).

Furthermore, machine learning algorithms are used successfully on data recognition, such as Hidden Markov Models (HMM) (Baum & Eagon, 1967), Particle Filter (PF) (Caramiaux et al, 2015) and Random Decision Forests (RDFs) (Barandiaran, 1998) (Shotton et al, 2013). HMM calculates output posibilities according to the already created data, when the new input data are to be recognised. Furtermore, it has been used to recognize sound data in folk music (Chai & Vercoe, 2001). Other systems have combined HMM with Dynamic Time Warping (DTW) (Fang, 2009) to achieve sound recognition in Byzantine Music (Kokkinidis et al, 2016). Shotton et al attemted a pixel-wise classification to achieve human body parts recognition, while on Human - Machine interaction field was developed a methodology that relies on the finger gesture data acquisition, hand segmentation, fingertips localization and identification and high-level feature extraction (Tsagaris & Trigkas, 2018).

2.1 Existing Multimodal Systems

By definition, multimodal systems aim on recognizing multiple types of input signal, such as speech, facial expressions, heartbeat, pressure and others. Multimodal signal recognition systems have been developed in an effort to record a variety of human activities such as emotions (Stathopoulou & Tsihrintzis, 2011). Other systems were developed to capture and recognize performances on fields such as craftsmanship (Ververidis et al, 2016), dancing (Camurri et al, 2016) and music (Chen et al, 2016).

In general, multimodal signal input combines a variety of sensors, each of which detects a different signal type. Therefore, complex preparatory tasks are necessary to attune such input, since synchronization is required to process and analyze the recorded activity. The advantages of receiving multimodal input lies on the complexity created on the resulting data structures

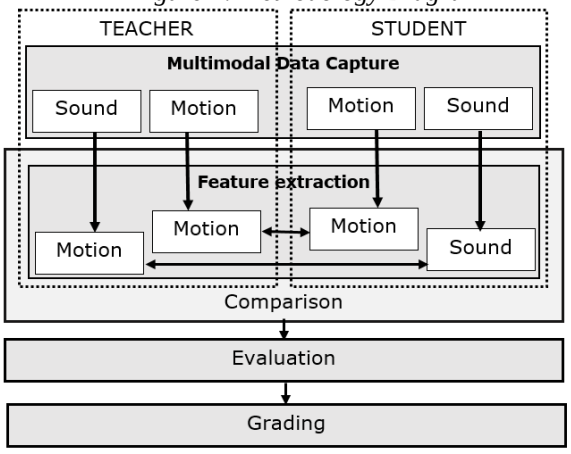
(Monaci, 2007) (Pitsikalis et al, 2015). The presented research focuses on recognizing dual-modal input signal, which originates in audiovisual recordings.

3. Methodology

The system extracts multimodal data from depth cameras. The time – space datasets are included in the frames and the sound of the recording. The selected, motion and sound capture sensor is Microsoft Kinect 2, which provides both the motion features and the sound reproduction. The gestural features are related to the position of the palm, in order to provide information on the rhythm of the student performance.

Consequently, these datasets are analyzed and segmented according to the tempo. In practice, the rhythm is used to segment the sound data and compare the student performance to the reference performance of the expert, for each note. The comparison is applied on the frequencies recorded that represent a note. Finally, through this comparison the student performance is evaluated in terms of the tonal and time divergence from the reference performance. The methodology diagram is displayed on Figure 1.

Figure 1: Methodology Diagram



3.1 Student Performance Evaluation Equation

As mentioned above, the gestures of the student are recognized, while in parallel the sound produced is evaluated for each musical note. For the performance evaluation step, the grading for pitch (equation 1) and rhythm (equation 2) are extracted separately.

For the pitch the aim is to return accurate grading results for both amateurs and excellent performances. Therefore, negative grading is used for the following cases:

- The performance is evaluated as excellent when the student performance-frequency divergence is within 10Hz radius from the reference.
- When the frequency divergence fluctuates from 10Hz to 60Hz – a distance slightly higher than a tone sound distance, the grading falls to 50%.
- Finally, the performance is not rewarded when the frequency divergence is greater than 140Hz– a four tones sound distance.

$$f(x) = \begin{cases} 100\%, & |x| \leq 10 \\ 100\% - \frac{(|x|-10)}{100}, & 10 < |x| \leq 60 \\ 50\% - \frac{(|x|-60)}{280}, & |x| > 60 \end{cases} \quad (1)$$

where f is the distance of the student performance frequency from the reference (expert's) performance and x is the frequency of the student performance

Similarly, to the above equation the rhythm-wise evaluation takes place within three student performance-time divergences and is based on a similar negative grading system.

- The performance is evaluated as excellent when the performance-time divergence is within 10% from the reference performance.
- When the performance-time divergence is within 10% to 80%, the grading falls to 50%.
- Finally, the performance is not rewarded when the performance-time divergence is greater than 80%

$$g(x) = \begin{cases} 100\%, & |x| \leq 10\% \\ 100\% - \frac{5\,(|x|-10\%)}{7}, & 10\% < |x| \leq 80\% \\ 50\%, & |x| > 80\% \end{cases} \quad (2)$$

where g is the percent difference on the pitch duration of the student from the reference. and x the percent difference on the pitch duration of the student from the reference.

The total grading displayed to the user is given by equation 3:

$$Total\ Grade = f(x) \cdot g(x) \quad (3)$$

.

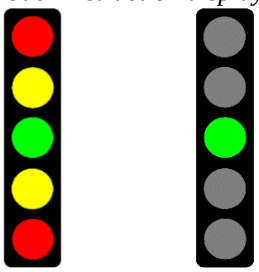
3.2 Visualisation concept to support Learning

Summative assessment is the most common process of student evaluation. The performance of the student is being translated as a numerical quantity which is presented as feedback at the end of the learning session (Race, 2014). Unfortunatelly, through feedback the student is not empowered to rectify the mistakes, although, feedback should aim at the modification of the student thinking and behaviour (Shute, 2008). Furthermore, shortness is an observed contemporary tedency on communication patterns among youth (Gold et al, 2010)(Hughes et al, 1998). Taking the above into consideration, we designed a simple and direct visualisation to display the feedback. The aim was to develop a brief and concise way to communicate the instruction. Thus, the traffic sign was selected as terse indication on what should be improved. Existing works have successfully used a similar type of instruction (Arnold & Pistilli, 2012). The benefits of this visualisation approach is primarily the high information transmition speed

since it is a familiar means of communication widely used on everyday life. Furthermore, the fact that the information is decoded rapidly by the student, results to a likewise rapid adjustment of the performance.

In detail, the traffic sign contains five lights of coded colour. The colour indicates the decline of the student performance, both quantitatively and qualitatively. The green light informs the student for a successfull performance, graded 90% or above. The yellow light and its position bespeaks whether the student declined from the reference frequency positively (high sign) or negatively (low sign). In this case the frequency reproduction is graded between 50% and 89%. Likewise, if the declination from the reference frequency is even higher, one of the red lights is displayed. On Figure 2 , the left traffic sign depicts the entirety of the possible lights to be displayed to the student, while the right sign displays the feedback as it is demonstrated to the student for the case of an excellent performance.

Figure 2: Feedback instruction displayed as traffic sign.



4. Case Study

This system was applied on the Byzantine music, a vocal music consisted among others, of chants and hymns. The distance between two notes (intervals) in Byzantine music differs from said intervals in western music system. Therefor in 1881 an Ecumenical Patriarchate appointed Committee proceeded to the establishment and definition of the Byzantine Music intervals resulting to new scales with specific frequencies for the notes (ek Madytwn, 1832) (Papadhmhtriou, 2005 A) (Papadhmhtriou, 2005 B). Thus, in the presented system, the absolute frequency value for each note is inessential, since the important is the intervals.

Similarities of non-equal-tempered intervals appear to the byzantine scales, called genre, such as Diatonic Scale, Soft Chromatic Scale, Hard Chromatic Scale, and Enharmonic Scale (Kypoyrgos, 1985). Byzantine genre possesses pitches that are micro-tonally distinctive from Western scale steps, so the ratio between two notes changes depending on the pitch and the genre. In more detail, while the western music system notation is set in absolute pitches, the Byzantine notation describes a melody as relative pitches and the intervals change depending on the genre. Moreover, in Byzantine music a chanting melody usually extends tonally until ± 1.7 intervals ("octaves") and performances at very low or high pitches, as occurs for example in opera, are rare. The captured chants were recorded for four variations. Said Byzantine modes, also called echos, were the "1st authentic", "1st plagal", "4th authentic" and "4th plagal" (Spanoudakis, 2013).

4.1 Recordings procedure

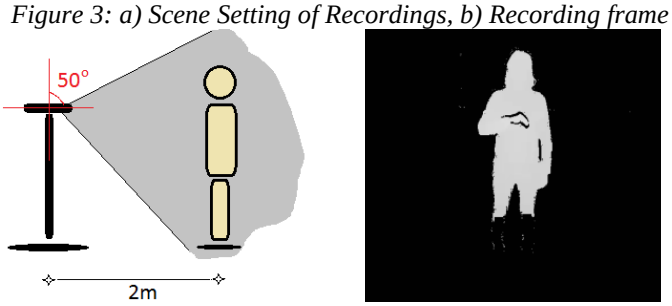
Three sets of recording sessions were conducted, for each mode (echos) mentioned above. The participants were a student and a teacher. The reference data used on the system were extracted from the teacher recordings. Later, the student performance was compared to it.

During the recordings session both performers were requested to stand at a certain distance from the depth sensor. Each recording took place separately, and no external source of instruction was provided during it. The raw data collected contained both depth images and sound data, which were synchronised.

4.1.1 Scene Setting

Figure 3 depicts the scene setting. During the recordings both performers stood at a distance of two meters from the depth sensor. The performer faced straight to the sensor, while the gradient of the direction of the sensor from the vertical axis was approximately 50 degrees.

Figure 3: a) Scene Setting of Recordings, b) Recording frame



The setting ensures that the palm of the performer is captured successfully. In general, the captured palm oscillates on the vertical axis (y axis), while limited activity is detected on the horizontal axis (x axis). The movement on the z axis (depth) is insignificant.

4.2 Data processing

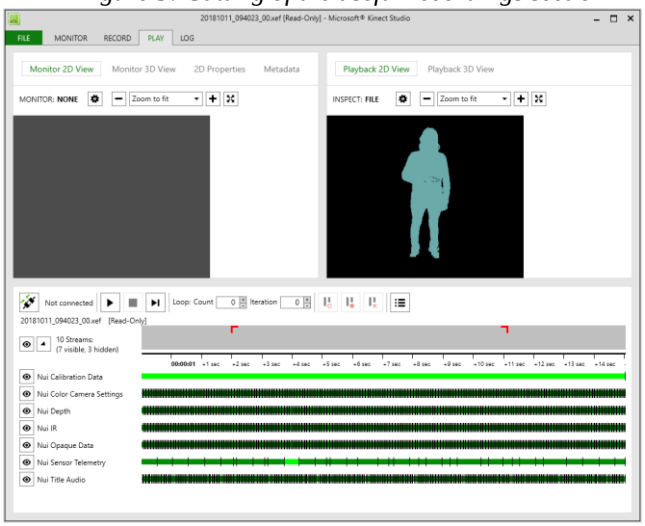
Besides the part of the performance that includes the chant and is useful to the system, a recording session contains a brief period of preparation that does not include gestures.

Figure 4: Recording segmentation to useful and preparatory

To extract the useful section, the beginning and ending of the performance were identified. Said timestamps of the recoding were marked on both the audio and the video file.

Consequently, a commercial sound recorder was used to subtract the useful part in a "replay-to-record" session (Figure 5).

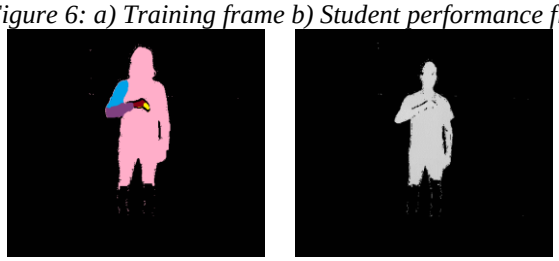
Figure 5: Cutting of the useful recordings section



4.2.1 Gesture's Feature Extraction

As mentioned above, the raw data provided from Kinect 2 is a depth video. The frames of this video were used to extract the features of the gesture using the ALGLIB SDK libraries (ALGLIB, 2017). In more detail, machine learning algorithms (Shotton et al, 2013) were applied on two datasets. The first dataset consisted of some recording frames selected from chosen teacher performances that came from various hymns and was used to train the algorithm. The second dataset was consisted of all the frames of a student performance. The features of the student gesture were extracted from it. On Figure 6 are displayed a training frame and a student performance frame.

Figure 6: a) Training frame b) Student performance frame

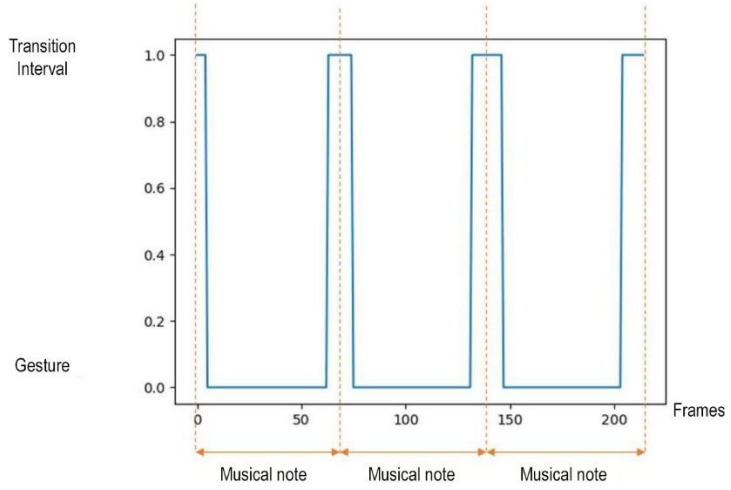


4.2.2 Note-wise Sound and Gesture Segmentation

To extract the sound features it is necessary to isolate each note from the audio recording and then recognize it. Although both gesture and audio recordings are continuous, the vocal note is more consistent compared to the ever-transiting gesture. Usually, a hymn consists of syllables that are sung in one or more genres. The movement of the performer's palm remains at a constant height for one or more syllables for a specific note. Thus, the palm movements are set in isotonic time intervals in large numbers, and they usually have a shorter duration than the written notes.

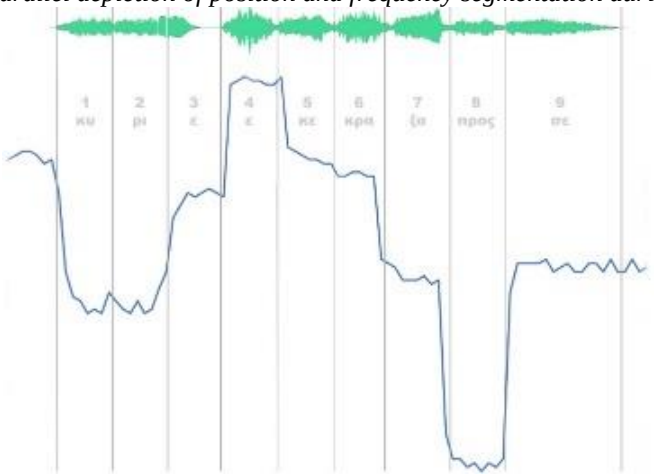
Moreover, in practice it was noticeable the fact that the movement of the hand precedes or follows the actual delivery of the corresponding hymn or note for a few milliseconds. Therefore, it is necessary to regard that there is a transitional context between gestures related to each note. On Figure 7 the segmentation of transitional and consistent contexts, is depicted in time.

Figure 7: Gesture segmentation on note depiction contexts and transitions



In more detail, there is the useful palm position, that refers to the note gesture-wise and the gesture context that leads to that position. The segmentation of the gesture on these contexts is beneficial to the feedback evaluation process. Furthermore, as mentioned above, the features used to recognize the gesture were the Cartesian coordinates in space (X, Y, Z) of the right-hand palm. The vertical oscillation of the palm represented mainly by the Y coordinate indicates its distance from the ground and proved to be the most characteristic trajectory of the gesture. On Figure 8 the gesture and sound extracted and segmented features are depicted in parallel.

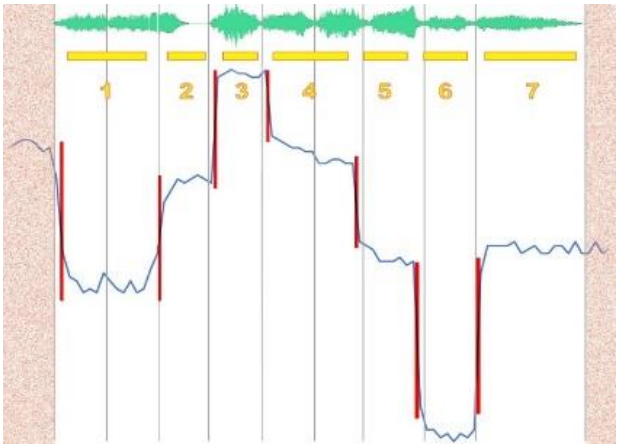
Figure 8: In parallel depiction of position and frequency segmentation during a performance



Despite the rejection of not useful recording parts mentioned during the data processing stage, it appears that a more accurate truncation of silent recording pieces at the beginning or end of the hymn is effective. Furthermore, by observing the data on Figure 9, it is easy to discern

horizontal "stability" segments that obviously appertain to the same tonal value. The steep slopes of the curve indicate the transitional frames between notes.

Figure 9: Red areas depict the additional truncation of the recording for hymn limits. The vertical lines segment the data according to the note reproduction



	Start	End
1	801 ms	2243 ms
2	2323 ms	2985 ms
3	3065 ms	3807 ms
4	3887 ms	5284 ms
5	5364 ms	6280 ms
6	6360 ms	7082 ms
7	7162 ms	8731 ms

Due to the hand movement tolerances mentioned above, the time intervals indicating separate notes are identified as the intervals between these steep slopes, using a time interval of ± 40ms before and after the detected transition. In the example of Figure 9, seven isotonic intervals are identified which are now treated as "notes" - based on the suggestion of the hymn gesture. These intervals are provided together with the audio recording to the evaluation stage.

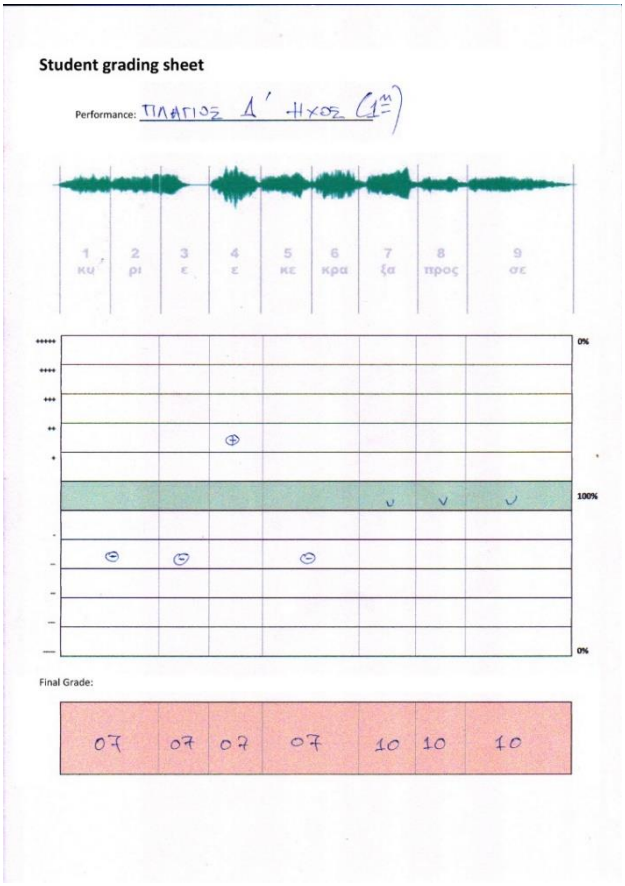
5. Experimental Student Evaluation and Feedback

To evaluate the system performance, eight student evaluations were contacted, two for each genre. The evaluation criteria were note-frequency related. The procedure followed was

- An expert teacher was requested to view the student recordings and grade the student's performance
- In parallel the student performance was graded by the presented system.
- The two acquired grades are compared, to estimate the efficiency of the system.

The divergence of the system grading to the teacher grading on mean square error (MSRE) was calculated MSRE = 0.0184. Taking into consideration that the maximum grade is 1 (corresponding to 100%), the calculated MSRE shows a mean deviation from teacher grading of about 1.8%. The divergence is due to the teacher's resilience and tendency to score "rounded" grades. Finally, the correlation coefficient *r* was used as the criterion of relevance between engine scores. Its value was calculated equal to r = 0.818, indicating an extremely high correlation, as is evidenced by the graphs.

Figure 10: Student grading sheet



6. Conclusion

This paper presented a system that implements a novel learning methodology of vocal music. The system compares the performance of a student with the reference performance of a teacher in order to provide off-line feedback. The performances are compared from the aspect of distance between two notes, the intervals. A skills evaluation methodology was developed on said system. The evaluation results of the system were compared to the evaluation results of the teacher. The comparison reveals that there is height accuracy on this grading methodology (r=0.818, MSRE=0.018) rendering this proposed approach promising. The vocal music genre on which the methodology was implemented on is Byzantine music. The pre-recorded performances executed by both student and teacher, were selected among four music scales called genres, out of total eight Byzantine music genres, to test the system.

Consequently, we intent to increase the number of student participants in order to conduct a more accurate evaluation of the system. Furthermore, the presented methodology is going to be implemented on more Byzantine music genres and hymns, testing the system's limits.

Acknowledgments

The presented research work was funded by the General Secretariat of Research and Technology of the Greek Ministry of Education, Research and Religious Affairs, for years 2016-2017, as a reward for participating to competitive EU programs, of the project Intangible Treasures - Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures-I-TREASURES, 600676/FP7-ICT-2011-9.

References

[1] ALGLIB Free Edition (2017). Retrieved February 2, 2017, from http://www.alglib.net/download.php

[2] Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 267-270). ACM.

[3] Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell*, *20*(8), 1-22.

[4] Baum, L. E., & Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, *73*(3), 360-363.

[5] Camurri, A., Volpe, G., Piana, S., Mancini, M., Niewiadomski, R., Ferrari, N., & Canepa, C. (2016, July). The dancer in the eye: towards a multi-layered computational framework of qualities in movement. In *Proceedings of the 3rd International Symposium on Movement and Computing* (p. 6). ACM.

[6] Caramiaux, B., Montecchio, N., Tanaka, A., & Bevilacqua, F. (2015). Adaptive gesture recognition with variation estimation for interactive systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *4*(4), 18.

[7] Chai, W., & Vercoe, B. (2001, June). Folk music classification using hidden Markov models. In *Proceedings of international conference on artificial intelligence* (Vol. 6, No. 6.4). sn.

[8] Chen, L., Gibet, S., Marteau, P. F., Marandola, F., & Wanderley, M. M. (2016, July). Quantitative evaluation of percussive gestures by ranking trainees versus teacher. In *Proceedings of the 3rd International Symposium on Movement and Computing* (p. 13). ACM.

[9] ek Madytwn., H., & Dyrrahioy, Ἀ. (1832). Theorhtikon. Mega tis moysikhs, htoi biblion didaktikon kai polytimon ths Moysikhs Episthmhs kai syggramma peri ths Byzantinhs Ekklhsiastikhs Moysikhs.

[10] Fang, C. (2009). From dynamic time warping (DTW) to hidden markov model (HMM). *University of Cincinnati*, *3*, 19.

[11] Gold, J., Lim, M. S., Hellard, M. E., Hocking, J. S., & Keogh, L. (2010). What's in a message? Delivering sexual health promotion to young people in Australia via text messaging. *BMC public health*, *10*(1), 792.

[12] Hughes, L. D., Done, J., & Young, A. (2011). Not 2 old 2 TXT: there is potential to use email and SMS text message healthcare reminders for rheumatology patients up to 65 years old. *Health informatics journal*, *17*(4), 266-276.

[13] Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

[14] Kodály, Z., Halápy, L., & Macnicol, F. (1974). *The Selected Writings of Zoltán Kodály. (Translated by Lili Halápy and Fred Macnicol.)*. Boosey & Hawkes Music Publishers.

[15] Kokkinidis, K., Stergiaki, A., & Tsagaris, A. (2016, July). Error prooving and sensorimotor feedback for singing voice. In *Proceedings of the 3rd International Symposium on Movement and Computing* (p. 32). ACM.

[16] Kypoyrgos, N. (1985). Merikes parathriseis panw sta basika diasthmata ths ellhnikhs kai anatolikhs moysikhs. *Moysikologia; 2*.

[17] Microsoft Kinect (2014), Retrieved May 18, 2014, from http://www.microsoft.com/en-us/kinectforwindows/

[18] Monaci, G. (2007). *On the modelling of Multi-modal data using redundant dictionaries* (No. THESIS). EPFL.

[19] Papadhmhtriou, P. D. (2005, July A). Oi klimakes ths Byzantinhs moysikhs kata thn moysiki epitroph toy 1881. Retrieved February 2, 2017, from http://portal.kithara.gr/modules.php?name=News&file=print&sid=729

[20] Papadhmhtriou, P. D. (2005, June B). Analogion. Methodos sygkerasmoy klimakwn – oi diatonikes klimakes toy Didymoy. Retrieved February 2, 2017, from from: http://byzantine-music.gr/Klimakes/diatonikh_sugkrash1881.html

[21] Patronas, G. (2018, November). Proseggish sth nea didaktikh methodo tis Byzantinhs Moysikhs. In *Proceedings of the 8th Presentation of the Greek Society for Music Education (G.S.M.E.), 37*(4), 339-351

[22] Pitsikalis, V., Katsamanis, A., Theodorakis, S., & Maragos, P. (2015). Multimodal gesture recognition via multiple hypotheses rescoring. *The Journal of Machine Learning Research*, *16*(1), 255-284.

[23] Race, P. (2014). *The lecturer's toolkit: a practical guide to assessment, learning and teaching*. Routledge.

[24] Rasamimanana, N., & Bevilacqua, F. (2008). Effort-based analysis of bowing movements: evidence of anticipation effects. *Journal of New Music Research*, *37*(4), 339-351.

[25] Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review*, *43*(1), 1-54.

[26] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... & Blake, A. (2011, June). Real-time human pose recognition in parts from single depth images. In *CVPR 2011* (pp. 1297-1304). Ieee.

[27] Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

[28] Spanoudakis, D. (2013, June). Shmeron krematai epi xyloy. Comparative musical analysis in fully developed Middle Byzantine and New Byzantine notation from manuscripts Dionysiou 564 (1445 AD) and Sancti Sepulcri 715 (19th cent.). *In Proceedings of the International Musicological Conference* (pp. 765-785).

[29] Stathopoulou, I. O., & Tsihrintzis, G. A. (2011). Emotion recognition from body movements and gestures. In *Intelligent Interactive Multimedia Systems and Services* (pp. 295-303). Springer, Berlin, Heidelberg.

[30] Tsagaris, A., & Trigkas, D. (2018, April). Mobile Gesture Recognition. In *Proceedings of the International Conference on Machine Vision and Applications* (pp. 13-17). ACM.

[31] Ververidis, D., Karavarsamis, S., Nikolopoulos, S., & Kompatsiaris, I. (2016, July). Pottery gestures style comparison by exploiting Myo sensor and forearm anatomy. In *Proceedings of the 3rd International Symposium on Movement and Computing* (p. 3). ACM.