# Machine learning via multimodal signal processing

K. Kokkinidis, A. Stergiaki, A. Tsagaris
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
{kostas.kokinidis, astrergiaki, tsagaris} @uom.edu.gr

***Abstract*— This paper proposes a methodology for recognition of vocal music (Byzantine music) via multi-modal signals processing. A sequence of multi-modal signals is captured from the expert's (teacher) and student's hymns performances, respectively. The machine learning system is trained using the values of particular features which are extracted from the captured multi-modal signals. After the system is being trained then it becomes able to recognize any hymn performance from the corpus. Training and recognition takes place in real time by utilizing machine learning techniques. The evaluation of the system was carried out with the cross - validation statistical method Jackknife, giving promising results.**

**Keywords—Human Computer Interaction; Singing Voice; Musical Gesture; Multimodal signal; Machine Learning; Jackknife implementation with Max/Msp**

## I. INTRODUCTION

The rapid development of ICT technology leads to the creation of new methods in software applications. Part of such methods are directly related to education and describe studies using teaching methods and insights in multiple fields. Many researches focus on human - computer interaction aiming to direct human contact with computer without requiring special human expertise's. Other research studies people's reaction to visual, acoustic and other signals, with artificial intelligence, computer vision, human-machine interaction (HCI) etc the most common.

Multi-modal signals are collections of various data arrays that demonstrate some common dependency, since they represent the same physical phenomenon. Different modalities are often captured by different sensors, and therefore they have different dimensionality and analysis, which often makes the definition of cross-modal correlations difficult and the joint analysis of this type of signals challenging. However, the synchronized presentation of different multi-modal signal allows discovering structures in the data revealing information that is unavailable when considering the modalities independently. [1][2].

Voice recognition is the ability of a machine to receive and interpret the dictation of a recording sound with the help of out coming devices such as microphone to understand and perform the vocal or singing commands. The perfect system of voice recognition is able to identify all the sounds (speech and singing) performed by humans. In practice, there are a great number of factors, such as noise, the user etc., that need to be analyzed for the proper identification of the audio signal, which makes this very difficult. This research is trying to study the future selection of multi modal signals (voice, movement etc), recognition and data extraction in order to train the multi modal identification system with the best solutions. The kind of music that is used is Byzantine Music (BM). The Byzantine music is ecclesiastical music of the Orthodox Church, which is interpreted by the human voice without accompaniment of musical instrument.

Any music excerpt has its own rhythm that an expert gives with the movement of his hands. The values of characteristics used to capture and model the performance are extracted from the performer's right hand movement. The above techniques make use of certain features of an object which can be detected by different kind of cameras. These cameras can be simple web cameras or cameras with depth map as the Kinect, the pmd and the LeapMotion.

The rest of this paper is organized as follows: In Section II, background details concerning machine learning techniques and validation methods are described. Related work is discussed in Section III. In Section IV the proposed methodology and an overview of the developed system is presented. Section V concludes the paper.

## II. BACKGROUND

### *A. Machine learning for Voice recognition*

The machine learning and recognition features accomplished through the MFCC [3] and the F0 [4] distance, as far as concerning the voice recognition. The Mel-Frequency Cepstral Coefficients or MFCC coefficients describe the shape of the spectrum and they are the most commonly used features in the field of voice recognition. To define the MFCC calculating process a sequence of actions is needed to be done. The formula which converts frequencies in Mel scale is given below:

$$M(f) = 1125 \ln(1 + f / 700) \quad (1)$$

The fundamental frequency $F_0 = 1 / T_0$. The fundamental period $T_0 = t_c - t_0$ is the time between two sequential glottal pulses. The fundamental frequency is the main method for examining the voice prosody.

The evaluation without weight of the frequency and its bandwidth is defined as the mean and the standard deviation

of the instant frequency f(t) which is computed by the following type:

$$F_u = \frac{1}{T}\int_{t_0}^{t_0+T} f(t)dt \qquad (2)$$

where $t_0$ and T defines the analysis frame of T duration from the $t_0$ time stamp.

*B. Machine learning for Motion recognition*

The motion characteristics of the performer's right hand movement are the geometrical characteristic of the coordinates of the hand in every frame of the video (Mass Centroid). The constraint of this method is to find the coordinates of the center of the hand mass, to be connected with the distance of the performer from the camera's depth and to calculate the size of the hand. For example, there is a difference in the movement between the hand of an adult and a toddler. In order to overcome these limitations and have better results, a group of adults with experience in depth camera participated, keeping a distance of 50-70 cm between the performer and the depth camera.

The description of this movement can be quantified using the movement characteristics depending on its use. One of the characteristics that describes a musical gesture and is being used is the center of mass (Centroid) of the hand. The coordinates of the center of mass $C_y$ $C_x$ on a system of axes (x, y) are defined by the following formulas:

$$C_x = \frac{\sum_n A_n C_{x_n}}{\sum_n A_n}, \quad C_y = \frac{\sum_n A_n C_{y_n}}{\sum_n A_n} \qquad (3)$$

Where $C_x$ is the distance of the Centre of mass C of y axis while $C_y$ is the mass center distance C from the x axis. The coordinates of the center of mass is $(C_x, C_y)$.

*C. Machine learning teqniques*

The classification of the Byzantine hymns took place with the use of HMMs (Fig.1) [5]. Hymns can be recognized by evaluating the trained HMMs with the help of the highest likelihood criterion. We expect the system to learn and recognize the hymns since the HMM is more flexible than the Markov model. [6]

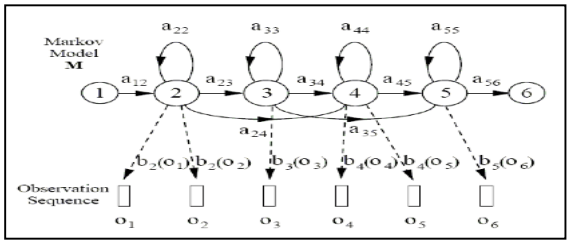


Fig. 1. Hidden Markov Models (HMM)

With the help of HMM the hymn will be identified before the performance ends. So, it is not necessary to complete the hymn. Finally, in order to synchronize the two time series (training and recognition), due to the different duration of each performance of hymn, we used the Dynamic Time Wrapping (DTW) algorithm (Fig.2).

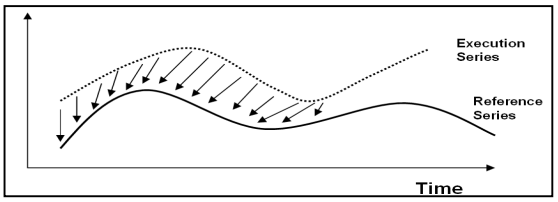


Fig. 2. Dynamic Time Wrapping (DTW) algorithm

*D. Model Validation method*

'Jackknife' or 'leave one out' is a cross-validation technique which was developed by Quenouille [7] [8]. This validation method is one of the most commonly used because it overcame the over-fitting problem. The parametrized method that is being used for the balance between noise and accuracy is the Recall & Precision metrics. This method divides the total number of hymns (i.e. recognized and not recognized hymns). The valid recognition's event is precision and the valid ground truth data is recall.

Precision (P) is given by the equation:

$$\mathrm{Pr}ecision = \frac{True\_Positive}{True\_Positive + False\_Positive} \qquad (4)$$

Where Recall (R) is given by the following equation:

$$\mathrm{Re}call = \frac{True\_Positive}{True\_Positive + False\_Negative} \qquad (5)$$

Finally the system evaluation is being performed by the 'Jackknife' cross-validation method, based on Precision and Recall metrics.

## III. RELATED WORK

The first basic principals in vocal and acoustic signals research start at 1950's [9]. The investigation of acoustic perception begins in Bell Laboratories at 1952 when a computerized system was built for digit recognition from a single-speaker using separate letters by locating the formants [10]. Similarly, Olson and Belar conducted a research, in the RCA labs, and found that a specific speaker recognized 10 different syllables [11]. Next researchers developed a dynamic programming application in voice recognition to aligning data samples in time, which is known as Dynamic Time Warping – DTW [12]. Additionally, Itakura's research suggests a method that is known as Linear Predictive Codding – LPC [13]. In the decade of 1980's, Hidden Markov Models – HMM embodied to the stochastic modeling [5]. Even though the appearance and use of Neural Networks was contemporary to the HMM, the HMM application has prevailed. However, another acoustic feature type is the MFCC coefficients [3].

The most significant method that has been proposed for the features extraction from the acoustic signal spectrum is Perceptually weighed Linear Prediction [14], which is based in the linear prediction with weights and RASTA-PLP [15]. These methods are usually presented in noisy acoustic signals. Nevertheless, there are many proposed models that are trying to model human hearing, like the cochlea [16] and the acoustic model. Although HMM is being used firstly for speech recognition, the implementation of HMM in motion recognition brought to us excellent results. A representative survey is the recognition of American Sign Language in real time via HMM. This system can recognize phrase's from a vocabulary with a score grader than 92% [17]. Various researchers began to use HMM in their survey in the field of arts. So, in music Wilson and Bodick (1999) uses HMM to recognize simple musical gestures from a maestro who was trying to give the rhythm of a musical excerpt [18]. Another excellent survey was made from Kolesink and Wanderlay (2004) who developed an HMM based algorithm which could recognize expressive gestures from a maestro. The score of recognition was grader than 97% [19].

## IV. PROPOSED METHODOLOGY & SYSTEM OVERVIEW

The methodology pipeline is given at Fig. 3. Firstly some hymns and the musical gestures which accompanied these hymns is being recorded in order to build the corpus. In our experiment we use four different hymns.

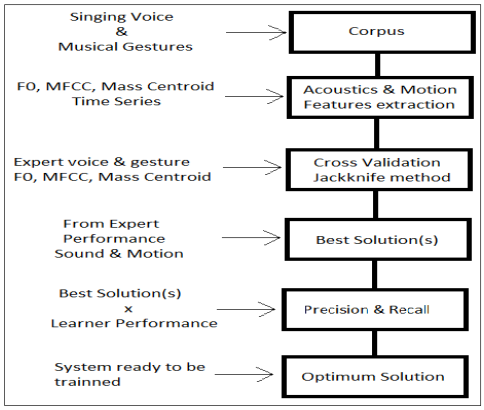


Fig. 3. Methodology pipeline

Each of these hymn (chanting and musical gestures) is being performed three times from the expert and the learner too. That happened due to performance changing every time that a chanter performs it. So we have to follow a methodology to find out the best solution(s) or the best performance(s). Once the corpus is being build then we are ready to apply our methodology. Efficient software was built to identify a hymn from a corpus. The acoustic features which have been extracted are the Fundamental frequency (F0) and the MFCC coefficients. In order the F0 and MFCC features to be extracted there has been used the ZSA Descriptors software which is been build at IRCAM, France [20]. The motion features which have been extracted are the coordinators (Cx, Cy) of mass Centroid of the right hand of performer. For the extraction of mass Centroid coordinators we implement some software in the main environment of MAX/MSP. Finally we construct a time series vector which includes all these 16 features. Using these vectors with machine learning techniques we recognize each hymn from the corpus. The software which implement the machine learning algorithms HMM and DTW is being developed in IRCAM, France, and it is called Gesture Follower (GF)[21]. It can be trained with one sample (hymn), 'one-shot' learning. These vectors after are being normalized they being used at all training and recognize phases from the system. Fig. 4 is a screen shot of our software application.

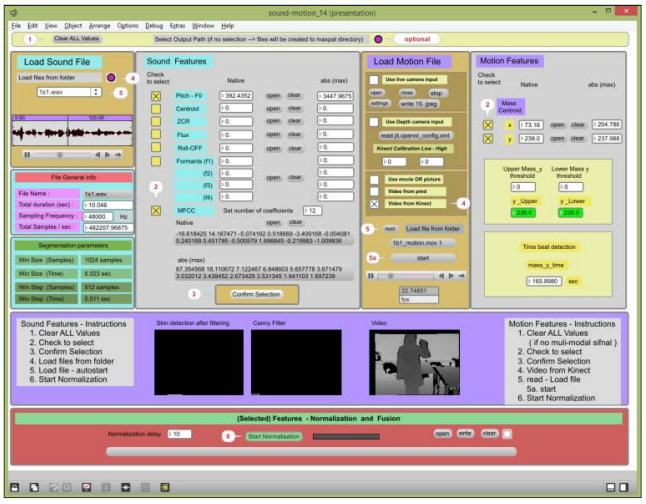


Fig. 4. Features (F0, MFCC, Mass Centroid) extraction

The evaluation of the system is based on the philosophy to find first of all which of the trainer execution is the best to train the system, in order to measure the effectiveness of the executions of the learner. For this reason the trainer is executing 3 times the 4 hymns.

Once the features vectors are being constructed from each hymn we may perform the first phase of hymns identification procedure. During this phase we read from a directory all the vectors - files – which concerns ONLY the expert performances. We perform the Jackknife Cross-Validation method in order to find out the best performance(s) of the experts executions of the hymns.

More analytically we evaluates all possible combinations of experts and students performances. In this experiment we have a total of 3^4=81 combinations. The system is being trained with each combination, which consists of four hymns, and performs recognition with the rest eight (8) performances. At the end of each fold, of a total of 81 folds, an array with values is being created. These values are the recognition percentages of the hymns due to training combination. After the end of whole phase (81 folds) we search to find the maximum value per column. Depending on which fold that cell belongs, we find the corresponding hymn from the training line (a combination

of four hymns by which we trained the system). These maximum values construct the best solution(s) (based on the execution of the expert) with which the system must be trained in order to achieve he best recognition percentages (results). In Figure 5, a screenshot of Jackknife method is presented. Furthermore, at phase 2, we training the system using each of the best solution(s), from Phase 1, in order to perform recognition with all LEARNER's performances. The learner performances are a total of (4 hymns x 3 performances each) 12 hymns. Using the results of hymns recognition we can evaluate the Precision & Recall metrics and by its score to find out the identification or not of any hymn of the corpus (the optimum training solution(s)). The evaluation method is in the early stage and has to be checked more for the efficiency.


Fig. 5. Jackknife Cross-Validation method

## V. CONCLUSION AND FUTURE WORK

In this paper, a methodology for recognition of vocal music performance based on multi-modal signal (singing voice and musical gestures) processing and its implementation in a corresponding system was proposed. The system was developed in Max/Msp environment. Machine learning techniques were applied via HMM and DTW based algorithms in order to model and train the system. Furthermore, the system was trained by the optimum solution which highlighted via the excellent rates of precision and recall metrics. In the advantages of the system one could include its ability to adapt easily to new hymns. In our future work, the methodology will be extended to support implementation of more complex musical gestures.

## REFERENCES


[1] G. Monaci, "On the modelling of multi-modal data using redundant dictionaries", Ph.D. dissertation, Dept. Elect. Eng., École polytechnique EPFL Univ., Lausanne, 2007.

[2] V. Pitsikalis, A. Katsamanis, St. Theodorakis, P. Maragos. (2015, Jan). Multimodal Gesture Recognition via Multiple Hypotheses Rescoring. *Journal of Machine Learning Research.* [Online]. 16(1). pp. 255-284.

[3] S.B. Davis, and P. Mermelstein. (1980, Aug.). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. [Online]. 28(4), pp. 357–366. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi =10.1.1.462 .5073&rep=rep1&type=pdf

[4] M. B. Siafarikas, "Speaker and speech recognition with the use of wavelets", Ph.D. dissertation, Dept. of Com. Eng. & Inf., Patra Univ., Patra, 2015.

[5] W. Chai and B. Vercoe, "Folk Music Classification Using Hidden Markov Models", in *Proc. of International Conference on Artificial Intelligence*, 2001.

[6] S. Manitsaris, A. Tsagaris, K. Dimitropoulos, A. Manitsaris, B. Denby. (2015). A visual perception of finger musical gestures in 3D space without any tangible instrument for performing arts. *The International Journal of Art and Technology*. 8(1).

[7] H. Abdi, L. J. Williams, "Jackknife", In Neil Salkind (Ed.), *"Encyclopedia of Research Design",* 2010.

[8] H.Abdi, and L.J Williams,. (2010). "Jackknife", in *N.J. Salkind, D.M., Dougherty, & B. Frey* (Eds.): *Encyclopedia of Research Design*. Thousand Oaks (CA): Sage., 2010 pp. 655-660

[9] L. R. Rabiner and B. Juang, *Fundamentals Of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[10] K. H. Davis, R. Biddulph and S. Balashek. (1952, Aug.). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*.[Online]. 24(6). pp. 637-642. Available: http://www.idemployee.id.tue.nl/g.w.m.rauterberg/presentations/HCI-history_files/..%5Cbell-labs.pdf.

[11] H. F. Olson and H. Belar. (1956). Phonetic typewriter. *The Journal of the Acoustical Society of America*. [Online]. 28(6). pp. 1072-1081. Available: http://asa.scitation.org/doi/abs/10.1121/1.1908561.

[12] T. K. Vintsyuk. (196, Jan.). Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*. [Online]. 4(1), pp. 52-57. Available: http://link.springer.com/article/ 10.1007% 2FBF01074755?LI=true.

[13] F. Itakura. (1975, Feb.). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*. [Online]. 23(1). pp. 67-72. Available: http://ieeexplore.ieee.org/document/1162641.

[14] H. Hermansky. (1990, Apr.). Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*. [Online]. 87(4). pp.1738-1752. Available: https://www.ncbi.nlm.nih.gov/pubmed/2341679

[15] H. Hermansky and N. Morgan. (1994, Oct.). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*. [Online]. 2(4). pp. 578-589. Available: http://ieeexplore.ieee.org/document/326616.

[16] R. F. Lyon and C. Mead. (1988, Jul.). An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing*. [Online]. 36(7), pp. 1119 – 1134. Available: http://ieeexplore.ieee.org/document/1639/

[17] T. Starner and A. Pentland, "Real-Time American Sign Language Recognition from Video Using Hidden Markov Models", in Motion-Based Recognition, M. Shah and R. Jain, Eds. Netherlands: Springer, 1997, pp. 227–243.

[18] A. D. Whilson & A. F. Bodick, "Real time Online Adaptive Gesture Recognition", in *Proc. of the International Conference on Pattern Recognition*, 1999.

[19] P. Kolesnik and M. Wanderlay, "Recognition, Analysis and Performance with Expressive Conducting Gestures", in *Proc. of the International Computer Music Conference* (ICMC 2004), Miami, USA, 2004.

[20] M. Malt, and E. Jourdan, "Zsa.Descriptors: a library for real time descriptors analysis", in *Proc. Of the 3rd Workshop on Learning the Semantics of Audio Signals,* 2015.

[21] F. Bevilacqua, F. Guédy, N. Sschnell, E.Fléty, N. Leroy, "Wireless sensor interface and gesture-follower for music pedagogy", in *Proc. of the International Conference of New interfaces for Musical Expression,* New York, USA, 2007.