Error prooving and sensorimotor feedback for singing voice

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

ABSTRACT
This paper presents a sensorimotor system for Byzantine Music. The main goal of this research is to detect some pre-defined errors in singing performance. After error-detection, the system uses a pre-defined error-dictionary in order to feedback. Through these feedbacks the potential chanter is being able to correct his performance. The system is being trained via experts MFCC features from a corpus of anthems. The recognition also takes place via MFCC but form student. The developed system is being able to evaluate in real time the pitch distance and furthermore the duration of two musician's performances, expert and student. The system may also evaluate the distance between two sequential musical gestures by which we may find the tempo of the hymn. After the pitch of these two hymns are being compared any identified errors will cause a feedback action to the student. This feedback corresponds to an error dictionary.

Author Keywords
machine learning; singing voice recognition; musical gestures recognition; error dictionary; sensorimotor learning

ACM Classification Keywords
H.5.5 [Information Interfaces and Presentation] Sound and Music Computing; I.2.10 [ARTIFICIAL INTELLIGENCE] Vision and Scene Understanding.

INTRODUCTION
In the evolution of technology, there has been noticed an increasing development of new applications in order to help the learning process. Easier ways between human and machines interactive (HCI) have been examined, so that ordinary people without special knowledge will be able to communicate with each other more effectively. For this reason, the collaboration of many scientific fields is essential, such as artificial intelligence (AI), human-machine interaction (HCI), computer vision, biology, psychology, etc. The technology that captures sounds and/or voices coming from a human with the help of technological equipment and recognizes them using a familiar procedure is, generally, defined as “Voice Recognition”. In the same way, a similar process referring to human gestures is called “Gesture Recognition”. This survey deals with the issue of feature selection-extraction, system training, error tracking and finally sensorimotor feedback. The features which have been selected for this survey are based on literature review; these are Pitch[1] and MFCC[2][3] for singing voice and Mass Centroid for musical gestures.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
MOCO'16, July 05-06, 2016, Thessaloniki, GA, Greece
© 2016 ACM. ISBN 978-1-4503-4307-7/16/07$15.00
DOI: http://dx.doi.org/10.1145/2948910.2948952

Modeling, Recognition & Comparison tone and tempo
The Hidden Markov Models (HMM) and more specifically their parameters have been trained from a data set which has been used for this purpose. In order to model our hymns we have used Forward HMM’s. Taking into account the criterion of the most likely performance, we are able to recognize the hymns by evaluating the trained HMMs. The procedure of the hymns’recognition is accomplished by using the HMM classifier. For each hymn we created a HMM, which parameters –according to the training data set- has been estimated (computed) by using the Baum-Welch algorithm. During the HMM de-coding phase, it is possible to reproduce a sequence of observations by calculating the possibilities of transition to a state from a previous one. These possibilities constitute a recognition criterion. The sequence of observations which institute the gesture is the one that HMM gives the greatest possibilities. If the possibilities (regarding to a defined threshold) of the sequence are too low for all HMM models, then the sequence is not considered to belong in any of the hymns which can be recognized by our system [4]. Additionally, due to different time periods of the predicted hymn’performance each time, it needs to be synchronized. More specifically, there must be a synchronization of the two data series, which is the series that came from the training (Hymn A) and the series which came from the recognition phase (Hymn B). In order to do that, before the beginning of the recognition process, takes place a time alignment procedure for the two series by using the Dynamic Time Wrapping (DTW) algorithm. Each hymn performance consists of two states. One state for sound (singing voice) and one for motion (musical gesture). Suitable vectors, created by pitch constitutes the HMM’s input which are modeling each hymn. Musical Gesture & mass centroid features are been used to define the time-distance between two hymns in order to compare the tonality of them.

Sensorimotor Learning

We can claim that the source of sensorimotor learning is the theory of Jean Piaget [5] which concerns the human period from the moment of his/her birth until the age of two (2) years that is known as the sensorimotor learning period. Jean Piaget himself has introduced the idea of Constructivism in which the acquisition of knowledge is accomplished through a continuous procedure of adaptation – interaction with the environment. By using Jean Piaget's theories, the computer science is making an attempt of approaching in different way system trainings since the algorithms of machine learning and artificial intelligence cannot contribute to the training procedure any time. This is due to the fact of the deficiency of the samples (history) that could be used as a base for the required training or even in the case where the training which takes place operates too slowly. Consequently, it is required a more intuitive approach of the matter to be accomplished.

STATE OF THE ART

Byzantine music

Pythagoras is mentioned as the establisher of the musical style, which Byzantine music afterwards was based on (microtonal music)[6]. Byzantine music has eight modes fourth tone and forth plagal tone. Initially, chanters were taught at Agia Sofia in Istanbul. The number of the chanters was determined and the teaching of music was made with specific gestures. With those gestures the teacher gave the beginning of the hymn, the rests of the psalmody, the signature time and the pace of the piece [6]. After the development of Byzantine notation, gestures in chanting were limited. In our project we tried to make a system that the student learns the tune, the interval, the rhythm and the pace of the Byzantine music. In order to collect more samples from the students, we used only one of the recorded hymns called "The Stone before Your Grave".

Voice recognition methods

The voice recognition which is based on computer usage has its roots in the '50s. In that period several researchers for the first time have started research based on acoustic and vocal data [7] [8]. In next decade researches concerning the voice recognition enriched by several techniques and methods such as -Dynamic Time Warping- (DTW) algorithm [9] and Linear Predictive Coding –LPC- method, that Itakuta refers to in his research [10]. In the '80s stochastic models appear through the presentation of Hidden Markov Models – HMM [11]. The appearance and usage of the Neural Networks is contemporary with HMM. nevertheless the implementation of HMM has prevailed. The use of MFCC features [2] from acoustic signals is also recommended by the bibliography for voice and/or songs recognition matters. A very powerful tool for gestures classification which combines the above techniques is named Gesture Follower (GF) and is being developed at Ircam institute at France [12].

Relevant research

Nowadays, research in computing music has gained popularity. Most applied research has been made to European music (Tonal System). Everyday new software programs are developed for the understanding of monophonic or polyphonic signals. For example software programs: "Ear training pro"[13], "sing and see"[14] and "smart music"[15]. In Greece the last decades various program researches have made a huge effort to rescue the country's music cultural tradition. Some of these programs are: "Polymnia", in which an automatic system of Optical recognition of Byzantine music writing has been created, "Damaskinos" in which a corpus of Byzantine Hymns has been made with recording of electro-glottal graph (EGG) [16], glottal flow and pressure of air, "VEMUS" (Virtual European MUsic School) [17], [18], in which, a virtual environment has been made for wind instruments such us Flute, Recorder, Clarinet etc. This program aim is to train students under the guidance of a virtual music professor[19].

Music beat tracking methods

Each sound has four features which makes it unique. These features are loudness, pitch, duration and timbre. Tempo is the music feature which is being related to the duration of an excerpt [20]. The tempo is assigned at the begging of the music excerpt in order to specify how fast or slowly a music piece must be performed. The common designation of the speed is the beats per minute (BPM). By comparing two similar audio music signals of two separated performances we focus on the detection of errors (Expert-student) as versus to time (Tempo) and to pitch (Tone). Firstly, we must do the segmentation of the music piece on every beat tracking. The time signature defines how many beats (pulses) are to be contained in each bar (measure), which note has one beat and how many notes the beat can have. In this way, we can focus on the exact time of error detection. Each beat contains specific numbers of notes performed in the correct pitch. In this way we may assume that beat tracking can be used as a meaningful temporal segmentation for higher level Music Information Retrieval (MIR) tasks such as structural segmentation of audio [21] and music similarity [22]. Beat tracking evaluation might be considered analogous to evaluation methods for onset detection [23]. Onset locations are obtained through an iterative process of hand-labeling time instants and listening back to the result [24]. We mention that some of beat tracking methods are the F-measure [25], Cemgil et al [26], PScore [27] etc.

METHODOLOGY & SYSTEM OVERVIEW

Error prooving and error dictionary

Our system (Fig.1) can firstly recognize as referred at Konstantinos-Hercules Kokkinidis [28] research and secondly compare a hymn, from the corpus. The system, furthermore, could provide feedback for the sensorimotor learning for the Byzantine music. The proposed system (Fig. 1) is constituted by three sub-systems. The sub-system of hymns'recording from the performances of the chants establishes the creation of our corpus. The hymns' processing sub-system, in which all the calculations take place, and finally the sub-system of sensorimotor learning. Initially, through our system we create a corpus of hymns by recording them (voice & motion). These hymns are been performed by chanters. Afterwards, we

proceed in the extraction of acoustic features phase. These features are the fundamental frequency (pitch or F0) and the first twelve (12) MFCC coefficients from voice and mass centroid form motion.

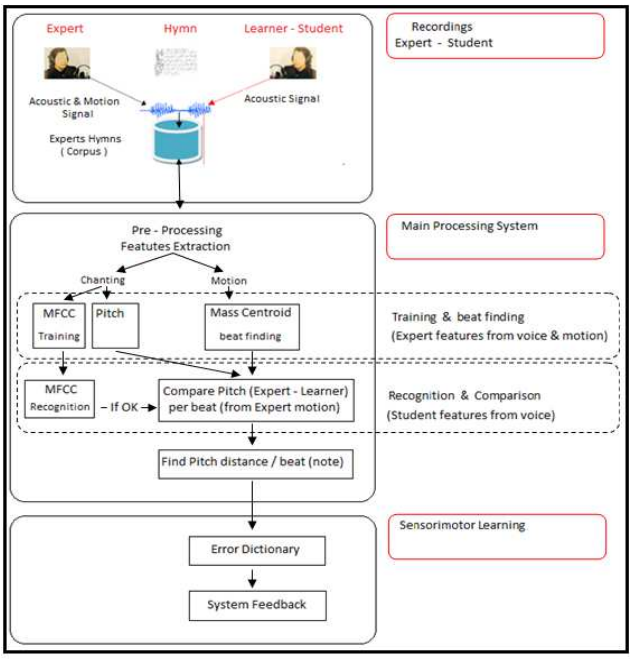

Figure 1. Methodology pipeline

The training of the system is being accomplished through machine learning techniques (HMM & DTW) [12] by sequences of training data which comprise MFCC. After the training, the system can recognize any hymn from our collection successfully [28]. The motion recognition via mass centroid features used for evaluating the tempo of the hymns. The technique which is used is implemented in GF software. A calculation of the acoustic feature named Pitch is being performed for each hymn in order to be evaluated the tone distance between the two hymns. A calculation of mass Centroid is also being performed from expert musical gesture, in order to find the time-distance between two gestures and furthermore to evaluate the Tempo of the hymn.

Error type	Comparison of errors	Error message
1	Arctic note (pitch)	"Your Arctic note (pitch) is not correct".
2	Ending note (pitch)	"Your ending note (pitch) is not correct".
3	Intermediate notes comparison.	"Your note(s) (pitch) is (are) not correct".
4	Duration of each note	"The duration of this note is not correct".
5	Duration of excerpt	"Your tempo is wrong".

Table I. Error dictionary

A threshold for each of the above calculations is being defined in order to set up the error tolerance. A detailed annotation for beat segmentation of music excerpt takes place by the expert musical gesture in order to find the time stump of notes. In this research we work out musical extracts in which every beat, of them, contains only one note.

We check the tonality and the duration of the note in the specific interval. Once any error is being identified, the system is able to feedbacks due to a pre-defined error dictionary for sensorimotor learning. The error dictionary and the sensorimotor feedback are described at table I.

Sensorimotor Feadback

The sensorimotor feedback message(s) which the system provides to student depends on the type of the error(s) identification. E.g. for error type 3 Fig. 2 shows the Error message.

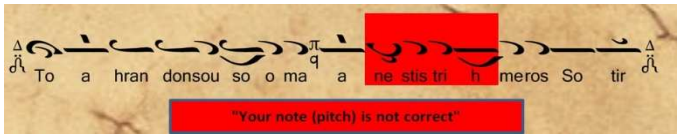

Figure 2. Sensorimotor feedback – Error type 3

Respectively, for error type 4 Fig. 3 shows the Error message.

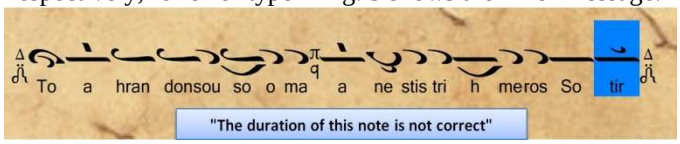

Figure 3. Sensorimotor feedback – Error type 4

EXPERIMENTAL RESULTS & SYSTEM EVALUATION

Chants are recorded in mono channel as 44100Hz sampling rate (samples per second) and 16 bits depth. The total duration of each command is approximately 10-30 sec with a number of total samples close to 450.000 per/hymn. Segmentation Window size is being defined at 512 samples as is the best window size (hamming) for this project.

For the evaluation of the system, the cross-validation method is used [29]. We create a population of performances concerning one hymn performed from student. Each performance may contain any error type but only once. We make this assumption in order to restrict the amount of combinations of errors (performances repetitions). Each student performance compared with the experts ones corresponding hymn. The error recognition efficiency is recorded. This will continue till the end of the combinations. The recognition efficiency is shown in Table II. The total combination evaluated by the statistical type (1).

$$8(n-2)(n-3) \qquad (1)$$

$$\text{Where n} = \text{Number of notes}$$

We evaluate 100 of 1920 total performances of student for each hymn for computations reasons. The hymn consist of 18 notes, so n=18. The following table regards the detection of the mistakes which arise from the comparison of beats (notes) of the two hymns (expert-learner) in specified intervals. The comparison is based on the usage of the Pitch feature. For each beat of the hymn, which defines the interval of comparison, the Pitch average is estimated. The percentages of the error detection between expert-learner are high, because of

the fact that the error tolerance (in case of error) was defined by us in a less than a semitone (<26Hz). At this point we must mention that student chanters are not qualified in entry level. Students - chanters may be characterized as senior chanters with advanced level of chanting experience and performance.

	Err-1	Err-2	Err-3	Err-4	Err-5	Recall
Err-1	100	0	0	0	0	100%
Err-2	0	100	0	0	0	100%
Err-3	0	0	85	15	0	85%
Err-4	0	0	19	81	0	81%
Err-5	0	0	0	0	100	100%
Precision	100%	100%	81%	82%	100%	

Table II. Error Recognition Efficiency

CONCLUSION AND FUTURE WORK

There has been presented an automatic recognition and comparison (of hymns) system. The system is able to recognize and compare hymns from a corpus. Based on the result of this comparison the system is able to provide feedback. The sensorimotor learning, which in fact is the system's feedback, is provided to the student based on an error-dictionary. In the proposed system the human-machine interaction (HMI) is simple. It is accomplished without a special or expensive equipment. The system's architecture can be enriched easily, with new hymns without the need of any difficult process.

In future research the methodology can be used to implement a hybrid system which may identify a variety of errors with automatic beat tracking process. Another open aspect is the evaluation of the system with no error combinations limits.

REFERENCES

[1] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. Journal of Acoustical Society of America, 99:3795–3806, June 1996.

[2] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4):357‑366, August 1980.

[3] M. B. Siafarikas, Speaker and speech recognition with the use of wavelets. Phd thesis, June 2015, *nemertes.lis.upatras.gr*

[4] Manitsaris S., Tsagaris A., Dimitropoulos K., Manitsaris A., Denby B., (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, Vol:8, No:01

[5] Piaget, J. (1952). The origins of intelligence in children. New York: International Universities Press.

[6] Panagiotopoulos, D.G. (1997). Theory and Praxis of Byzantine Ecclesiastical Music, 6th edition, "SOTIR" (first edition 1947, "ZOH"), Athens. (In Greek)

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

[8] K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits.The Journal of the Acoustical Society of America, 24(6):637‑642, 1952.

[9] T. K. Vintsyuk. Speech discrimination by dynamic programming. Cybernetics and Systems Analysis, 4(1):52‑57, 1968.

[10] F. Itakura. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 23(1):67‑72, February 1975.

[11] Chai, Wei and Barry Vercoe. 'Folk Music Classification Using Hidden Markov Models', Proceedings of International Conference on Artificial Intelligence, June 2001.

[12] Bevilacqua, F., Guédy, F., Sschnell, N., Fléty E. Leroy N., 'Wireless sensor interface and gesture-follower for music pedagogy'. In Proceedings of the International Conference of New interfaces for Musical Expression, New York, USA, pp 124-129, 2007.

[13] http://www.earmaster.com/

[14] http://www.singandsee.com/

[15] http://www.smartmusic.com/

[16] Kouroupetroglou, G., Delvinioti, D., & Chrisochoïdis, G. (2006). DAMASKINOS: Standard annotated collection of voices chanting Byzantine Ecclesiastical Music, National University of Athens, Department of Informatics and Telecommunications. (in Greek "ΔΑΜΑΣΚΗΝΟΣ")

[17] Vaia X., Ou.-Ei. (2007). Applications of music technology in music education, (unpublished master thesis), National Technical University. Athenes. (in Greek)

[18] Karvouni, Th. (2008). Music Composer, semiautomated synthesis system of Byzantine music. (Unpublished master thesis), National Technical University of Athens.(in Greek)

[19] http://www.tehne.ro/projects/vemus_virtual_ music_ school.html

[20] Chang, Y.-Y., Lin, Y.-C.: Music Tempo (Speed) Classification, CS229 Autumn 2005

[21] M. Levy and M. Sandler, "Structural segmenation of musical audio by constrained clustering," IEEE Transactions on Audio, Speech and Language Processing, vol. 156, no. 2, pp. 318–326, 2008.

[22] D. P.W. Ellis, C. Cotton, and M. Mandel, "Cross-correlation of beat-synchronous representations for music similarity," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, USA, April 2008, pp. 57–60.

[23] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, part 2, pp. 1035–1047, 2005.

[24] P. Leveau, L. Daudet, and G. Richard, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in Proceedings of 5th International Conference on Music Information Retrieval, Barcelona, Spain, 2004, pp. 72–75.

[25] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," Journal of New Music Research, vol. 30, pp. 39–58, 2001.

[26] A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempo gramre presentation and Kalman filtering," Journal Of New Music Research, vol. 28, no. 4, pp. 259–273, 2001.

[27] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," Journal of New Music Research, vol. 36, no. 1, pp. 1–16, 2007.

[28] Konstantinos-Hercules Kokkinidis and Athanasios Manitsaris, "Intelligent Sensorimotor Learning for Byzantine music", 4th International Conference on Modern Circuits and Systems Technologies (MO.C.A.S.T 2015).

[29] Abdi, H., Williams, L.J., 2010, « Jackknife », In Neil Salkind (Ed.), Encyclopedia of Research Design. Thousand Oaks, CA: Sage.