Intelligent Sensorimotor Learning for Byzantine music

(Acoustic signal proccessing using acoustic features)

Konstantinos-Hercules Kokkinidis
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
kkokk@uom.gr

Athanasios Manitsaris, Professor
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
amanitsaris@uom.edu.gr

***Abstract*—This paper presents a sensorimotor learning system for singing voice of Byzantine music. The goal of this research is to develop a human-machine interface, which will be able to guide and correct a potential chanter of Byzantine music. Using low features, extracted from a corpus of anthems, the system will be trained to recognize a specific chant from this corpus. After the training, intelligent techniques are used with a combination of Hidden Markov Models and Dynamic Time Wrapping algorithm for recognition. The system will be able to evaluate in real time the distance of the performance for two musicians, the teacher (expert) and the student (learner) and give guidelines to student. Custom software developed under max/msp development software in order to perform the sensorimotor learning section. Once the system is trained, it will then be tested with different chants. The distance between chanters performance is estimating in real-time and the experimental results shows that the interaction is efficient. The evaluation of the system took place by the cross-validation statistical method Jackknife. Precision and Recall metrics, estimated in order to validate the use of sensorimotor learning in Byzantine music.**

Keywords—Human machine Interaction; Singing Voice recognition; Sensorimotor Learning

I. INTRODUCTION

Nowadays, the evolution of technology leads to an increasing development of new applications, which tries to help the learning process of humans in various aspects. Scientists look for natural ways of interaction between human and machines (HCI), to be easier to communicate ordinary people without special knowledge. In this direction, many scientific fields are collaborating such as artificial intelligence, human-machine interaction (HCI), computer vision, biology and psychology.

A. Voice recognition and features

Voice recognition is the technology that captures words spoken by a human with a help of microphone and recognized them by speech recognizer. The process of speech recognition consists of different steps. An ideal situation in the process of speech recognition is that a speech recognition system recognizes all words performed by a human. This is practically difficult because the performance of a speech recognition system depends on a large number of factors. Features, users, noise, etc are some of the factors. This research deals with the issue of feature selection and extraction.

The features that are used in this research are selected based on the literature review and they are Pitch and Formants.

Pitch

The fundamental frequency $f_0 = 1 / T_0$, where $T_0 = t_c - t_0$ is the fundamental period, is the time between two sequential glottal pulses. The fundamental frequency is the main means for checking 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:

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

where t_0 and T defines the analysis frame of T duration from the t_0 time stamp [1].

Formants

The formants are altered due to the vocal tube geometry for the production of different sounds-phoneme. The overall length of vocal tube from the glottal to the lips defines the frequency band-width of formats. The typical length for men is 17cm resulting the mean values of frequency for the first three formants, to be 500Hz, 1500Hz, 2500Hz, while for the women the vocal tube is 14ch and the respectively format values are almost 600Hz, 1800Hz, 3000Hz.

The formats evaluation is computed by the tendency weighting of f(t), by dividing them with the squared a(t) , amplitude in order to result the weighting evaluations of the formats [1]. The formats evaluation is computed by the following formula :

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)[a(t)]^2 dt}{\int_{t_0}^{t_0+T} [a(t)]^2 dt} \quad (2)$$

B. *Hidden Markov Models (HMM) and Dynamic Time Wrapping (DTW)*

HMMs (Fig.1) are chosen to classify the Byzantine hymns, and their parameters are learned from the training data. With the help of the most likely performance criterion, the hymns can be recognized by evaluating the trained HMMs. Because the HMM is more feasible than the Markov model, we adopt the former to learn and recognize the continuous hand gestures to direct robots.

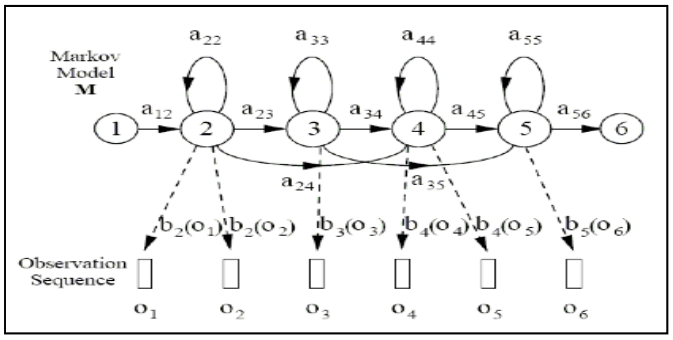

Fig. 1. Hidden Markov Models (HMM)

HMM classifier: The last phase is the classification of the hymn. The probability that each HMM has produced the input vector constitutes a recognition criterion. Each HMM models a hymn. Thus, the HMM which has the highest probability of having created the input sequence corresponds to the most probable hymn represented in this input vector.

An HMM can be determined by:

- {S} – Set of states, included initial state SI and final state SF·
- T – Array of transition probabilities, T = $\{t_{ij}\}$, t_{ij} is the transition probability from the state I to state j.
- Array of output probabilities, O = $\{o_j(x)\}$ for continuous HMM.

Forward HMM is used to model the dynamic features of each voice command. A model is created for each command and the parameters of models were appreciated by the sequences of training that were available for each command, using the Baum-Welch algorithm. During the decoding phase, each HMM is able to produce the sequence of observation. This probability constitutes a recognition criterion. The sequence is seen as belonging to the gesture of which HMM gives the highest probability. If a sequence gives too low probabilities (smaller than a threshold) in all models, then this sequence is not considered as belonging in any hymn which our system recognizes [3].

Thus, it allow us, with the proper training of the HMM model to predict the hymn during the speaking. It is not necessary to complete the hymn. With the help of hidden Markov Model the hymn will be predicted before the end of the speaking. Finally, the predicted hymn because of the different duration of each execution of the appropriate hymn at the same time, the Dynamic time Wrapping algorithm, is used to synchronize the two time series (training and recognition) (Fig.2).

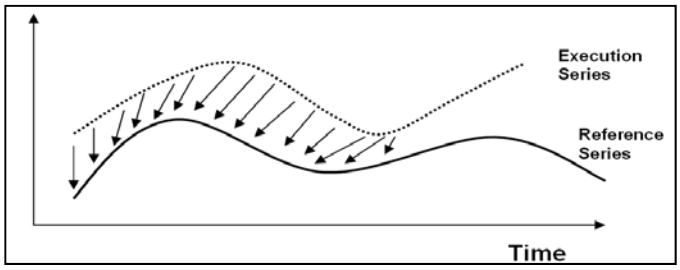

Fig. 2. Dynamic Time Wrapping (DTW) algorithm

C. *Sensorimotor Learning*

"From the moment where the system is trained for a set of human gestures, it can then be used in their pedagogic with sensorimotor criteria. (Jean Piaget)".

In that point there should be mentioned that the sensorimotor learning has its roots in Jean Piaget theory [4] concerning the period of the human thinking development from birth till the age of two years old, known also as the sensorimotor learning period. Jean Piaget himself, also has introduced the idea of Constructivism where the knowledge development is accomplished through a continual adaptation –interaction with the environment. IT science using the Jean Piaget theories attempts a different approach in the system training, where the machine learning algorithms and artificial intelligence cannot contribute to the training due to lack of history so that it can be used as a base for the requested training or even where the training takes place it functions too slowly. A brighter approach of the issue is needed.

II. STATE OF THE ART

Voice recognition

Voice recognition by using computer starts from the decade of 1950 when several researchers start researching for the very first time acoustic and vocal basic principles [5]. At Bell Laboratories, Davis, Biddulph and Balashek achieve to build a system which recognizes from a specific speaker individual letters by using formats [6]. That was the start of an intense research period which expanded in the next decades from USA to Japan and Soviet Union. The next step of this research are the Olson and Belar results in the RCA laboratories which recognized 10 different syllables from a unique speaker [7]. Various techniques and methods applied in the voice recognition. The dynamic programming applied in voice recognition for the aligning of data series in time. Today it is known as Dynamic Time Warping – DTW [8]. Another applied method is known as Linear Predictive Codding – LPC, which Itakura refers in his research [9]. In the 1980 decade we have the stochastic modeling appearance through the

implementation of Hidden Markov Models – HMM [10]. The appearance and use of Neural Networks is contemporary to the HMM, nevertheless the HMM application has prevailed. Another category of acoustic features which are used are MFCC coefficients [11]. For the features extraction from the acoustic signal spectrum, there have been proposed more efficient methods like Perceptually weighed Linear Prediction [12], which is based in the linear prediction with weights and also RASTA-PLP [13], which are used in noisy acoustic signals. There are also proposed other models which try to model the human hearing , like the cochlea model [14] and the acoustic model [15].

III. METHODOLOGY OVERVIEW

In this way, the aim of the methodology is first to identify a hymn from a corpus and second to provide useful feedback for the sensorimotor learning of the Byzantine music. At first, a recording phase takes place in order to build the corpus; then, we proceed to features extraction phase which are the fundamental frequency (pitch) and the first three formants (f1, f2, f3). Using these acoustic features the system is trained with machine learning techniques to recognize each hymn from the corpus (Fig 3). The technique which is used in both training and recognition is a combination of Hidden Markov Models (HMM) and Dynamic Time Wrapping (DTW) [16].

Fig. 3. Voice Recognition

Once the system is trained, any hymn form corpus can be recognized. When a hymn is performed by a trainee, it is modeled and classified according to the appropriate features. Finally a performed hymn can be graphically shown and compared with the one corresponding to corpus. A calculation of the acoustic feature named Pitch via the DTW algorithm gives necessary graphical feedback for the sensorimotor learning (Fig 4).

The evaluation of the system is performed by cross-validation technique by the computation of Precision and Recall metrics [17]. Cross-Validation is useful for overcoming the problem of over-fitting. Over-fitting is a term which refers when the model requires more information than the data can provide. It is one of the most commonly used model selection criteria. Based on a data splitting, part of the data is used for fitting each competing model (or procedure) and the rest of the data is used to measure the performance of the models.

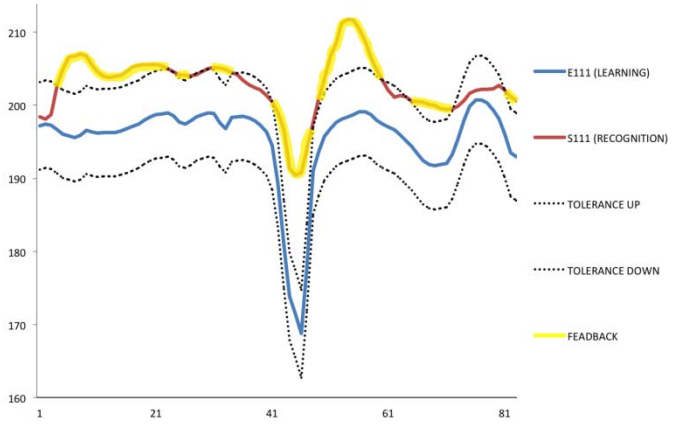

Fig. 4. Sensorimotor feadback

Precision-recall method is a parameterized method that is balanced between accuracy and noise. Recall is the number of correctly recognized hymns by the system divided by the total number of hymns (the hole corpus) (i.e. recognized and not recognized hymns). Precision is the ratio of correctly recognized hymns to the total number of hymns is recognized by the system (i.e. correctly recognized and wrongly recognized hymns). Recall measures how sensitive the recognized recognition module is and precision determines the accuracy of the system's module.

In these terms, precision is the probability that the recognition's event is valid, and recall is the probability that the ground truth data was recognized.

Equations (3) and (4) provide mathematical definitions of precision (p) and recall (r) for convenience.

Precision is the fraction of recognized hymns that are relevant to the search.

$$Precision = \frac{|Relevant\_hymns \cap Recognized\_hymns|}{Recognized\_hymns} \quad (3)$$

Recall in information retrieval is the fraction of the strokes that are relevant to the query that are successfully retrieved.

$$Recall = \frac{|Relevant\_hymns \cap Recognized\_hymns|}{Recognized\_hymns} \quad (4)$$

IV. SYSTEM OVERVIEW

The complete system consists of three subsystems, the voice recording system for the creation of corpus and the trainee performance of hymns, the features processing subsystem for training and recognition and the sensorimotor learning feedback subsystem. The voice recording and the training and recognition subsystem is implemented under Max/MSP programmable language. It is a programming tool that allows the user to create programs graphically and is concentrated in multimedia development, focusing primarily in the music field. The developed software works either in real-time by recording and

processing the hymns directly from the input, or with pre-recorded files.

The main patch gives to user the opportunity to record a hymn or to execute the features extraction in real time (Fig 3). The features will be extracted and the intelligent system will be trained (Fig 3) with the feature vector described in the methodology. For the training phase the features extraction and the training process are performed off-line.

Once the system is trained to recognize all the hymns of corpus, the trainee may perform a specific hymn of the corpus and the system after the recognition process sends it to the sensorimotor learning subsystem. The sensorimotor learning subsystem will provide intelligent feedback learning for the specific hymn [16].

V. Experimental results

Chants are recorded in mono channel as 48000 samples per second and 16 bits per sample. The total duration of each command is approximately 10-30 sec with a number of total samples close to 4.000 per/hymn. As voice is low bandwidth, these values are quite sufficient. As the window size, different values are tested and shown that 512 is the best window size (hamming) for this project.

There are 4 fixed hymns which can be used to provide sensorimotor learning feedback. Each chant performed three (3) times.

For the evaluation of the system it is used cross-validation (Leave one out). From the 3 repetition of 4 chants the first set is used to train the system and the other 3 sets are used for the recognition. The recognition efficiency is recorded. The second set of commands is used for training and the other 3 (1, 3-4) sets are used for the recognition process. This will continue till the end of the sets. Totally 81 executions are performed for each hymn. The recognition efficiency is shown in Table I.

	E11	E12	E13	E14	Recall
E11	81	0	0	0	100%
E12	0	81	0	0	100%
E13	0	15	66	0	81%
E14	0	0	0	81	100%
Precision	100%	84%	100%	100%	

TABLE I. Recognition efficiency

VI. CONCLUSION AND FUTURE WORK

An automatic singing voice recognition system is proposed, which is based on a combination of Hidden Markov models and dynamic time wrapping and enables to provide sensorimotor learning feedback for a corpus of hymns. The importance of hymn recognition lies in building efficient human–machine interaction. The algorithm is modeling the hymns via hidden Markov Model and dynamic time wrapping and the performed hymn can be predicted before the end of the chanting.

The compared of the Pitch feature between the two hymns (the one which the system be trained and the performed one by the trainee) will be transferred to the sensorimotor learning system to provide graphical sensorimotor learning feedback. The described architecture is simple, low computational costly, high reliability and high robustness to noise. The interaction is extremely simple and natural and does not require the trainee to use any other additional devices.

The efficiency of the system is very high with a precision rate between 84% to 100% and a recall rate between 81% to 100% for the 4 hymns, and average value 96% for precision and 95.25% for recall.

Although the proposed hymns for sensorimotor learning is from a restrict corpus, one of the advantages of the system is its easy adaptation to new hymns. In future research the methodology can be used to implement a hybrid system which can be combine singing voice and gestures.

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] 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

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

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

[5] 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.

[6] H. F. Olson and H. Belar. Phonetic typewriter. The Journal of the Acoustical Society of America, 28(6):1072‑1081, 1956.

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

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

[9] P. Birkholz and D. Jackel. Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In INTERSPEECH, 2004.

[10] 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.

[11] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America, 87(4):1738‑1752, 1990.

[12] H. Hermansky and N. Morgan. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, 2(4):578‑589, October 1994.

[13] R. F. Lyon and C. Mead. An analog electronic cochlea. IEEE Transactions on Acoustics, Speech and Signal Processing, 36(7), 1988.

[14] S. Seneff. A computational model for the peripheral auditory system: Application of speech recognition research. In ICASSP, volume 11, pages 1983‑1986, April 1986.

[15] 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.

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