# Optimizing Human Computer Interaction for Byzantine music learning: Comparing HMMs with RDFs

Paraskevi Kritopoulou*
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
pkritopoulou@uom.edu.gr

Athanasia Stergiaki*
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
astergiaki@uom.edu.gr

Konstantinos Kokkinidis
Department of Applied Informatics
University of Macedonia
Thessaloniki, Greece
kostas.kokkinidis @uom.edu.gr

***Abstract*—The current paper presents the results of an optimization study conducted on a self – training system of Byzantine chanting. The main aim of this work is the enhancement of the monophonic vocal music recognition step. Sound recognition technologies as implemented in practice, rely on classification algorithms. Various machine learning algorithms have already been applied for this purpose. In order to examine and select the most efficient one, we proceeded with a comparative evaluation of Hidden Markov Models and Random Decision Forests algorithms. Vocal performances of an expert-performer and an amateur-performer were captured to create a dataset for the training of the aforementioned algorithmic programs. Ensuing the training of the algorithms, Jackknife statistical method was applied to cross-validate the evaluation results of each algorithmic rule. The outcome of the evaluation highlights that Hidden Markov Models algorithm is more effective that Random Decision Forests.**

***Keywords*—Human Computer Interaction; Byzantine music; HMM; RDFs; Jackknife – Cross Validation**

## I. INTRODUCTION

Vocal music is perceived as augmented speech in relation to the tonal quality and rhythm. Currently, in the field of Human Computer Interaction (HCI) various music recognition techniques such as audio signal analysis methods and machine learning algorithms are applied to achieve music recognition.

Aiming to improve a previously developed interactive learning system that imparts vocal skills to the user, a comparative study of Hidden Markov Models (HMMs) and Random Decision Forests (RDFs) classification algorithms was conducted. Within the framework of this study, these machine learning algorithms are used for sound recognition. The case study on which this research was conducted is the genre of Byzantine music, a vocal music genre not accompanied by musical instruments that contains mainly chants and hymns. The audio signal was collected by microphones.

On the next session an existing work overview presents applications and evaluation results of the compared algorithms. Consequently, the methodology for training and operating the algorithms is exhibited and the training conducted by our team is displayed. On the last session are presented the results of the optimization study.

## II. EXISTING WORK

### A. Feature Extraction

In order to process the audio signal, it is necessary to extract sound features from the raw data. These features are used to train sound recognition algorithms. There have been various approaches towards extracting audio signal features. Algorithms such as Expectation-Maximization (EM) have been used for this purpose [21]. On other works, Grammatone Filter has been applied for sound feature extraction of the spectral energy of the audio data, returning positive and reasonable results during the recognition [27].

In general, the popular methods to extract sound features are divided in three large categories. Firstly, there are those who utilize time domain features, such as the Short-term energy, then those who employ frequency domain features, such as the Spectral Centroid, Spectral Flux and Fundamental Frequency, and lastly those who focus on Coefficient Domain Features, such as the Linear Prediction Coding and Mel-Frequency Cepstrum Coefficients (MFCCs) [23]. The latter coefficient features (MFCCs) are extracted by analyzing the spectrum [7]. Notably, the MFCC is widely preferred as criterion for speech signal feature extraction, since it presents high performance rates while being less complex [12]. Thus, it was selected on this study for feature extraction.

### B. Sound Recognition

Hidden Markov Models (HMMs) are stochastic models used for speech recognition [4]. Segments of the speech are represented with states, and probability-transition matrixes are used to transit from a state to the next [8]. HMMs have been repeatedly applied at sound for chord recognition [13] [18] [6], high level music recognition [19], or even music score recognition [3]. There are alternative models such as the Cochlea [14] that represent human hearing and Acoustic Model which is categorized as similar to HMM modeling. Yet, in qualitative evaluations HMMs seem to be more efficient [10]. Thus, it is usually selected for studies on musical art related projects.

Moreover, there have been variations of HMM in an effort to increase the recognition accuracy according to the case study. Multi-stream HMMs were created in an attempt to recognize frequency sub-bands [6]. HMMs trained with Expectation-Maximization algorithm were used for chord recognition [21]. For polyphonic music, merged-output HMMs were developed in order to address the problem of describing polyrhythmic scores [16]. Each approach mentioned above focuses on certain features of the music, trying to adjust the algorithm to these characteristics.

On the other hand, Random Decision Forests (RDFs) is an ensemble training method that aims on classification. For its training it relies on assembling multiple decision trees, while its output is the mean prediction of said trees, which concludes the class [2] [9]. RDFs have been widely used for feature

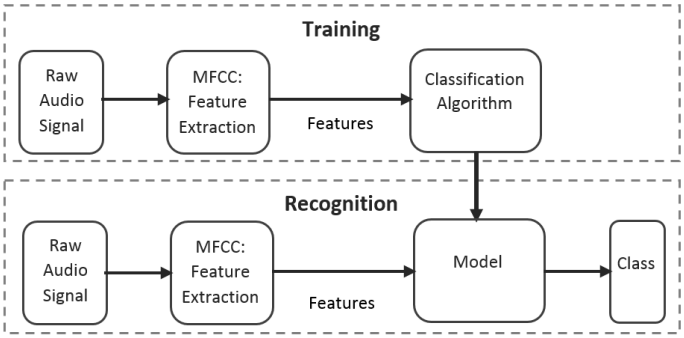


Fig. 1. Block Diagram of audio signal recognition

extraction on Image Processing [22]. Yet, this algorithm has been also used on various projects for sound recognition, and more precisely for music genre classification [25], background sound [20] and environmental sounds [26] recognition.

Although RDFs have been compared in respect to sound classification with algorithms such as Deep, Recurrent and Convolutional Neural Networks, or Support Vector Machine [5] [11], there have been few comparison studies with HMMs. It is noteworthy the fact that there have also been efforts to combine RDFs with Hierarchical HMMs on environmental sounds, to achieve higher recognition results [17]. To summarize, HMMs are considered as highly successful among pattern recognition techniques.

## III. METHODOLOGY

Both compared algorithms function for classification, thus they work on a similar way. Figure 1 displays the block diagram of their operation. In more detail, both HMMs and RDFs use supervised learning. Initially, the raw audio signal data are being processed so that MFCC sound features are extracted.

Mathematically, MFCC are coefficients that describe collectively the distribution of power on frequency. The computation process takes place by segmenting initially the audio signal into frames, to calculate the Discrete Fourier Transform (DFT) of each analysis frame. Consequently, DFT is multiplied by triangular band-pass filters whose central frequency and width are arranged according to Mel-scale [24]. After the conversion to Mel-scale, the acquired vectors are logarithmized by Discrete Cosine Transform (DCT) to remove inessential information.

The spectral energy is calculated by the equation:

$$E(i) = \frac{1}{\sum_{k=L_i}^{U_i}(V_i(k))^2} \sum_{k=L_i}^{U_i} \left(|X_k| V_i(k)\right)^2, \qquad (1)$$

where $L_i$ is the filter's lower bound,
$U_i$ is the filter's upper bound, and
$S_i$ is a normalization coefficient.

Both linear and non-linear properties of the raw audio signal are contained in the coefficients, rendering them capable to recognize dynamic sound features. Thus, MFCC are appropriate for audio signal feature extraction and are widely popular due to their efficient representation of speech data. The equation for frequency conversion in Mel-scale is:

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \qquad (2)$$

where f is the signal frequency.

The features computed from equation (2) form a set of numbers; on this work, the MFCC set has a size of twelve.

### A. Training

Consequently, the classification algorithm collects the sound features and after classifying each feature set, it labels it. This process signifies the completion of the training session that categorizes various feature sets into labels. The features processed during the training session are modeled to be used during the recognition session.

Especially for RDFs, during the training process multiple decision trees are generated to create a model that predicts the possibility for a certain feature, the MFCC set, to represent a class label. Each decision tree is created by defining randomly a subset of the initial MFCC set. It entails a specific number of set samples that is further divided to even smaller subsets, to the extent that the tree depth defines.

### B. Recognition

During recognition session, raw data that have not been used to train the algorithm are provided as input. In a similar process with the training session, MFCC of the respective audio signal are calculated. These features are consequently processed and compared by the model extracted during the training session, to find on which class it is most probable to belong to.

Respecting RDFs, the result of the majority of the tree classifiers define the prediction. Regarding HMMs, in the recognition stage the algorithm Dynamic Time Warping is additionally used. The audio signals are given as time series. For a successful recognition the input signal needs to be adapted in terms of its length, and to be synchronized with the time series that resulted from the training session. This process is graphically depicted on Figure 2.

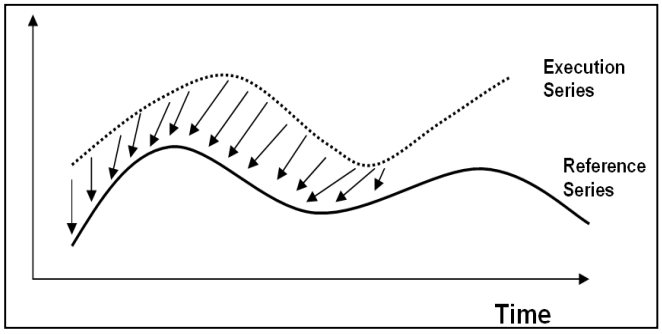


Fig 2 Dynamic Time Wrapping algorithm depiction

## IV. TRAINING

For the training of the algorithms were used three (3) iterations of four (4) hymns, performed by a professional chanter. Therefore, the classes are four. Moreover, the order in which the hymns are performed is essential. Permutations with Repetition are given by the equation below.

$$Permutations = n^r \quad (3)$$

In our case, (3) results to an evaluation combination of $3^4$=81 hymns. Out of these 81 combinations, the first one is applied on the training of the algorithms, while the rest are used for the recognition. This procedure is repeated for the total of the 81 combinations.

The performance was conducted as a monophonic rendition. For the recording a sample rate of 48ksps and a bit depth of 16 have been used. The training duration is approximately 1-30 sec for approximately 4.000 samples per hymn. These values are competent especially for the male voices that are placed at lower frequencies.

For each frame analysis the first twelve (12) coefficients were computed. The package of 12 MFC coefficients is regarded as a feature. MFCC were extracted by the ZSA Descriptors, a tool developed by IRCAM [15].

For HMMs, the window size was practically searched, and defined on about 512 msec for this training environment. For the training of the RDFs algorithm three decision trees of depth 4 were used. These values may seem low, but in regard to the amount of hymns used are appropriate to avoid over-fitting.

## V. RECOGNITION & VALIDATION

The recognition efficiency of each algorithm was tested by a set of four (4) recorded hymns, performed by a user of the interactive learning system. The user had no previous experience on Byzantine chanting. The features extracted by said hymns were used as input for each algorithm.

Following the completion of the recognition process, Jackknife cross-validation method was selected to evaluate the recognition results, since it is able to address over-fitting issues [1]. In more detail, an All-vs-One evaluation took place for every possible hymn combinations.

To access the recognition results, three metrics were used; Precision (Pre) and Recall (Rec) metrics, are selected to balance the noise and accuracy of recognition. Precision, is the event of the valid recognition, and is given by equation 4. Recall is the valid ground truth data given by equation 5. Finally, Accuracy (Acc) is calculated by equation 6:

$$Pre = \frac{TP}{(TP+FP)} \quad (4)$$

$$Rec = \frac{TP}{(TP+FN)} \quad (5)$$

$$Acc = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (6)$$

where:

- True Positive (TP) are correctly recognised hymns,
- True Negative (TN) are poorly performed hymns that were correctly failed to recognize,
- False Positive (FP) are poorly performed hymns that were incorrectly recognized as correct, and
- False Negative (FN) are poorly performed hymns that were recognized correctly

The recognition results by the HMMs algorithm, and the results of the Jackknife evaluation method are presented below. Table I is the confusion matrix for HMM. Precision and Recall metrics for each class (E1, E2, E3, E4) are extracted from said matrixes. The Accuracy is calculated as the sum of the recognition percentages of the correctly predicted classes, and is estimated at 0.95.

TABLE I. HMMS: JACKKNIFE CONFUSION MARIX

| Actual class | Predicted class | | | | |
|---|---|---|---|---|---|
| | **E1** | **E2** | **E3** | **E4** | **Total** |
| **E1** | **81** | 0 | 0 | 0 | 81 |
| **E2** | 0 | **81** | 0 | 0 | 81 |
| **E3** | 0 | 15 | **66** | 0 | 81 |
| **E4** | 0 | 0 | 0 | **81** | 81 |
| **Sum** | 81 | 96 | 66 | 81 | **324** |

Table II displays the calculated values for the rest of the metrics, for each class. Precision is the sum of all the recognition percentages of all the actual classes, for a predicted class (q.v. Table I). Recall is the sum of the recognition percentages of all the predicted classes for a certain class. The values of the metrics are quite high.

TABLE II. HMMS: PRECISION AND RECALL

| Actual class | Pre | Rec |
|---|---|---|
| **E1** | 1 | 1 |
| **E2** | 0.844 | 1 |
| **E3** | 1 | 0.815 |
| **E4** | 1 | 1 |

Similarly, Table III depicts the confusion matrix of the RDFs algorithm. The Acc was estimated at 0.885, a percentage of 6.5 below the HMMs respective one. It is evident that in comparison to RDFs, HMMs perform more efficiently sound recognition.

TABLE III. RDFS JACKKNIFE CONFUSION MARIX

| Actual class | Predicted class | | | | |
|---|---|---|---|---|---|
| | **E1** | **E2** | **E3** | **E4** | **Total** |
| **E1** | **79** | 2 | 0 | 0 | 81 |
| **E2** | 0 | **69** | 12 | 0 | 81 |
| **E3** | 0 | 20 | **58** | 3 | 81 |
| **E4** | 0 | 0 | 0 | **81** | 81 |
| **Sum** | 79 | 91 | 70 | 84 | **324** |

Likewise, Table IV displays the Precision and Recall metrics of the evaluation of the RDFs algorithm that are rather high.

TABLE IV. RDFS: PRECISION AND RECALL

| Actual class | Pre | Rec |
|---|---|---|
| **E1** | 1 | 0.975 |
| **E2** | 0.758 | 0.852 |
| **E3** | 0.829 | 0,0.716 |
| **E4** | 0.964 | 1 |

## CONCLUSIONS

An improvement research on a human computer interaction learning system for Byzantine music was conducted. The purpose was the optimization of sound recognition results. Hidden Markov Models and Random Decision Forests classification algorithms were compared for this reason. A combination of eighty-one (81) hymns were used to train and evaluate the algorithms using Jackknife cross-validation method. The metrics selected to evaluate the algorithms were Precision, Recall and Accuracy. The results highlight that regarding monophonic music recognition both algorithms are highly accurate presenting an accuracy of 95% for the HMMs, and 88.5% for the RDFs. Thus the HMMs algorithm presents a 6.5% higher accuracy than the RDFs algorithm.

Consequently, we plan on recording a bigger variety of hymns that correspond on a different scale, in order to perform a more in-depth evaluation.

## REFERENCES


[1] Abdi, H. & Williams, L. J. (2010). Jackknife. In Neil Salkind (Ed.). *Encyclopedia of Research Design*. Thousand Oaks (CA): Sage., pp. 655-660.

[2] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[3] Calvo-Zaragoza, J., Toselli, A. H., & Vidal, E. (2016, October). Early handwritten music recognition with hidden markov models. In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 319-324). IEEE.

[4] Chai, W. & Vercoe, B. (2001). Folk Music Classification Using Hidden Markov Models. In *Proc. of International Conference on Artificial Intelligence*.

[5] Chang, C., & Doran, B. (2016). Urban Sound Classification: With Random Forest, SVM, DNN, RNN, and CNN Classifiers. In *CSCI E-81 Machine Learning and Data Mining Final Project Fall 2016*. Harvard University Cambridge.

[6] Cho, T., & Bello, J. P. (2013). Mirex 2013: Large vocabulary chord recognition system using multi-band features and a multi-stream HMM. *Music Information Retrieval Evaluation eXchange (MIREX)*.

[7] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366.Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi =10.1. 1.462 .5073&rep=rep1&type=pdf

[8] Gaikwad, S. K., Gawali, B. W. & Yennawar, P. (2010). A Review on Speech Recognition Techniques, *International Journal of Computer Applications* Vol. 10, No. 3, pp. 16-24. http://dx.doi.org/10.5120/1462-1976

[9] Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

[10] Juang, B. H., & Rabiner, L. R. (2005). Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara*, *1*, 67.

[11] Kokkinidis, K., Mastoras, T., Tsagaris, A., & Fotaris, P. (2018, May). An empirical comparison of machine learning techniques for chant classification. In *2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST)* (pp. 1-4). IEEE.

[12] Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(12), 18006-18016.

[13] Lee, K., & Slaney, M. (2006, October). Automatic Chord Recognition from Audio Using a HMM with Supervised Learning. In *ISMIR* (pp. 133-137).

[14] Lyon, R. F. & Mead, C. (1988). 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/

[15] Malt, M. & Jourdan, E. (2015) Zsa.Descriptors: a library for real time descriptors analysis., In *Proc. Of the 3rd Workshop on Learning the Semantics of Audio Signals*.

[16] Nakamura, E., Yoshii, K., & Sagayama, S. (2017). Rhythm transcription of polyphonic piano music based on merged-output HMM for multiple voices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(4), 794-806.

[17] Niessen, M. E., Van Kasteren, T. L., & Merentitis, A. (2013, October). Hierarchical sound event detection. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[18] O'Hanlon, K., & Sandler, M. B. (2019, May). Comparing CQT and Reassignment Based Chroma Features for Template-based Automatic Chord Recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 860-864). IEEE.

[19] Qian, G. (2019, April). A Music Retrieval Approach Based on Hidden Markov Model. In *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)* (pp. 721-725). IEEE.

[20] Saki, F., & Kehtarnavaz, N. (2014, May). Background noise classification using random forest tree classifier for cochlear implant applications. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3591-3595). IEEE.

[21] Sheh, A., & Ellis, D. P. (2003). Chord segmentation and recognition using EM-trained hidden Markov models.

[22] Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., ... & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, *56*(1), 116-124.

[23] Thambi, S. V., Sreekumar, K. T., Kumar, C. S., & Raj, P. R. (2014, December). Random forest algorithm for improving the performance of speech/non-speech detection. In *2014 First International Conference on Computational Systems and Communications (ICCSC)* (pp. 28-32). IEEE.

[24] Volkmann, J., Stevens, S. S., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 208-208.

[25] Wang, L., Huang, S., Wang, S., Liang, J., & Xu, B. (2008, October). Music genre classification based on multiple classifier fusion. In *2008 Fourth International Conference on Natural Computation* (Vol. 5, pp. 580-583). IEEE.

[26] Wei, J. M., & Li, Y. (2013, December). Specific environmental sounds recognition using time-frequency texture features and random forest. In *2013 6th International Congress on Image and Signal Processing (CISP)* (Vol. 3, pp. 1277-1281). IEEE.

[27] Zhao, Y., Wang, H., & Cui, R. (2011). An approach to sound feature extraction method based on gammatone filter. In *Advances in Multimedia, Software Engineering and Computing Vol. 2* (pp. 371-376). Springer, Berlin, Heidelberg.