REAL TIME ADAPTIVE NOTE AND SWARAS RECOGNITION USING HMM AND NEURAL NETWORK

S.Hariram, Charu Sheila, S.Krishna, Arun Varadarajan & N.Venkateswaran
ABSTRACT

The paper describes a simulated prototype system for Swaras recognition using Hidden Markov Models (HMM) and Neural Networks. The goal of the paper is to develop a robust, autonomous system that would eventually enable a person to practice music for swaras set by direct communication with the computer. The system will provide immediate feedback on whether the pitch, scale, tune, ragas, match and are acceptable. This paper describes the development of underlying system recognition, a prototype simulated system for real time applications. Our procedure uses note identifiers, which has close conformity with Indian classical music concept of that for primary clustering of Ragas into groups. Our criterion is analyzed for the Arohi (ascending), Abrohi (descending) sequences. Our algorithm for the system was successfully tested on the portion of basic swaras of the E set (Re, Ne). From this basic algorithm one can develop a Lexicon through which Ragas can be identified and pitch extracted. The task becomes complex as the Indian classical singers do not follow a fixed scale concept and choose the base Sa at any pitch. This condition can be trained by using a Baum-Welch reestimation method for training a Viterbi procedure for recognition. The modeling of each segment of the system is carried out using HMM. In our paper we have also concentrated on the base dominant frequencies of swaras (Formants) for better understanding of the concept underlying Indian classical music.

KEYWORD

Formant Frequencies, Neural Networks, Pitch, Ragas, HMM.

1 INTRODUCTION

This paper basically is concerned with the application of hidden markov model (HMM)-based techniques for automatic music recognition where the system incorporating our algorithm can act as a tutor. The key pattern processing question, which is the basis of the work described in this paper, is whether acoustic modeling techniques, such as HMM, are sufficiently sensitive to distinguish between good and poor pronunciations of a known swara by an unknown person. These notions of “good”, and “poor”, pronunciation are largely subjective and therefore difficult to include formally and explicitly in the framework of a contemporary HMM-based swara recognition system. The purpose of the paper is a speaker-independent test of “good” pronunciation of ragas was defined by training HMM on data’s collected from persons who are experts in Indian classical music. In this section we will basically view traditional Indian classical music formats and divisions.
1.1 Swaras

The most basic unit of music is the swara (or note), which simply indicates the position in the audible spectrum occupied by a particular sound or the pitch of the sound. Actually, the spectral position is better described as swarasthana. Inherently, certain sounds ‘go together’ and certain others do not. This property was realized by man thousands of years ago and is indicated by the term harmony. There are seven swaras in Carnatic music, namely, Shadjam (Sa), Rishabam (Ri), Gandharam (Ga), Madhyamam (Ma), Panchamam (Pa), Dhaivatham (Da) and Nishadam (Ni). There is some theoretical basis for why there is an odd number (seven) of swaras and we will deal with this subsequently. For simplicity, let us fix the Sa at one kattai (I white key in the keyboard) and place the remaining swaras at the successive white keys in a keyboard. This provides us with a scale or a raga (in this case, containing all the seven swaras). As mentioned previously, ancient Vedic chants have but three swaras and somewhat later forms of music (Indian as well as other forms, eg. Chinese) use five swaras - e.g. the Sa, Ri, Ga, Pa and Da of the scale we just created. Our present system is based on seven swaras, and perhaps, a few thousand years from now, the human race will advance to a point of discriminating scales of more swaras. Carnatic music is based not on logarithmic division but on rational division. An octave is based on the ratio 1:2; Pa is located through the ratio 2:3; similar definitions exist for all the twelve swarasthanas. A few centuries ago, Western classical music too was based on rational division (the resulting scale was called as the natural scale), but this has given way to the equally tempered (also called chromatic) scale produced by logarithmic division. The difference is subtle, but quite important.

1.2 Ragas

Ragas are sometimes defined as melody types. The raga system is a method of organizing tunes based on certain natural principles. Tunes in the same raga use the same (nominal) swaras in various combinations and with practice; the listener can pick up the similarity. Each raga has a swaroopam (a musical form or image) that is defined by the swaras used, the gamakas given to these swaras, the sequence in which the swaras occur etc. This definition is termed as the raga lakshanam. Raga lakshanam usually contains the arohanam, avarohanam, details of the swaras which are chiefly responsible for the characteristic melody of the raga, characteristic swara phrases and general usage notes. It is intended more for the performer than for the listener. We shall first define arohanam and avarohanam. Arohanam is the sequence of swaras used in a raga in the ascending passages i.e. as the pitch goes up. Avarohanam is the sequence of swaras to be used in descent. The arohanam and avarohanam (or the scale) of a raga provide only a skeletal outline upon which the rest of the raga is formed. Ragas are not simply abstract collections of swaras that occur together to produce a tune. Each raga has a distinct image or swaroopam and it is this which defines a raga. Arbitrary selection of a set of swaras is unlikely to produce a distinct raga swaroopam and this is the reason for attributing the foundations of the raga system to nature. The ragas that we know of are the products of centuries of experimentation. Each ragam has associated with it a feeling that it induces in the listener and the performer.

1.3 Formant Analysis

The objective of formant analysis is to determine the complex natural frequencies of the vocal tract as they change during swara and musical production, which depends on vocal tract parameters. If the vocal-tract configurations were known, these natural frequencies could be computed. However, the musical speech signal is influenced both by the properties of the source and by the vocal tract. For example, if the source spectrum has a zero close to one of the natural frequencies of the vocal tract, it will be extremely difficult, if not impossible, to determine the frequency or the bandwidth of the particular formant. A side-branch element such as the nasal cavity creates a similar problem. We consider the analysis of formant frequencies because of the fact that certain musical tones, sounds,
notable vowels, may be identified and synthesized from that knowledge. These formants appear to be important information bearing elements of musical signals, speech waveforms. We can also apply the Markov models to formant frequency estimation, from which we can estimate the “spatial relations” and “locational information’s” among spectral lines, which are originally generated over frequency domain. These spectral lines can be taken and trained using Markov models. Switching from time domain to frequency domain drastically reduces the number of states on the Markov chain and the use of continuous parameters completely eliminates quantization error. The formant analysis of swaras becomes very important in knowledge-based extraction of relevant musical notes and statistical modeling of their distortion. This property of effective classification of swaras by formant methods helps in error analysis and also in hierarchically organizing the ragas. Two examples of formant plots obtained for /ne/ and /re/ belonging to seven swaras are illustrated in Figure 1 and 2 respectively.

![Formant plot for /ne/](image1.png) ![Formant plot for /re/](image2.png)

2. **STATISTICAL MODEL OF SWARAS**

Hidden Markov models have been introduced for modeling swara waveforms. The concepts of HMM, Viterbi Algorithm and Baum-Welch reestimation are discussed in APPENDIX A .

The swara signals $x_i(t)$ is generated by a discrete and finite sequence of actions

$$A = a_1(t_1) \ a_2(t_2) \ a_3(t_3) \ \ldots \ \ldots \ a_k(t_k)$$  \hspace{1cm} (1)

Where $a_k(t_k)$ denotes an action ending at time $t_k$; $a_1(t_1)$ represents the silence preceding the beginning of a sentence.

When a person sings a swara say /Sa/, a relation

$$R_i(Sa, A)$$  \hspace{1cm} (2)

Is applied which produces A. The relation $R_i$ depends on the musicians base alignment, his mood, state of health and history. As $R_i$ may produce several sounds $(A)$ for same $Sa$, probability distribution for all possible ‘A’ can be derived using a generative model. If the musician gives gap between swaras from a set of C concepts, then a third relation is applied:

$$R_3(C,S)$$  \hspace{1cm} (3)

$R_3$ may depend on the musician and his / her culture. Statistical models can be used for characterizing this relation.
3. DENSITY FUNCTIONS

The Swara signals are segmented into frames of equal length. For each frame a k-dimensional vector $\mathbf{x}$ of characteristic features are measured. The vectors are considered to be output symbols of N-state Markov chain. We need to model out the probability density function $b(y)$ assuming that to be spherically invariant. The $b(y)$ can be expressed by a certain function $f$ in the form:

$$b(y) = M^{-0.5} f ((y-m)^T M^{-1} (y-m))$$

with the mean $m = E(\mathbf{x})$, and $M$ is the covariance matrix.

At first the parameters of the word models based on the normal density function were calculated. From the training data of each word the corresponding maximum likelihood state sequences were evaluated and applied to divide the data into segments, each corresponding to the specific state of Markov functions. For a given set of parameters and a sequence of vectors the corresponding maximum likelihood state sequence and the conditional probability were computed simultaneously by means of Viterbi algorithm. Figure 3 illustrates the good approximation of the measured average values $f(s)$ for the density function, which we choose as ‘Laplacian density’ for its significant modeling and behavior. We also show the measured value plot for the $K_0$ density function in the same figure. The scattering of $K_0$ density leads to laplacian density analysis. With HMM using spherically invariant probability density functions recognition rates were determined for a set of test speakers. The recognition rates increase with the number of states and results for different density converge. For small number of states, the normal density yields sufficient results, but with increasing number of states, the recognition rates for the “Laplacian” density becomes superior.

4. NEURAL NETWORK DETAILS

We use a neural network a highly parallel, simple structure, having identical computational elements, for massive parallel computations. We use a Time delay neural network for our application as they have the ability to relate and compare current input to past history of events. For the network we adopt backpropogation learning algorithm, which is a gradient descent of the mean squared error as a function of weights. We now give a statistical plot for the learning effectiveness for both the methods and find that HMM based analysis gives better criterion of swara recognition than going for Neural network. Other neural networks like procedural networks gives still poor learning plots.

![Figure 3. Laplacian Density and $K_0$ Density plots for recognition rates](In the above figure “L” – shows Laplacian density plot)
5. EXPERIMENTAL RESULTS

Results if Figure 4 refers to set of experiments performed on a corpus of 100 words of a test vocabulary pronounced by a male speaker. The curve given in figure 2 gives the Laplacian density plots developed by straightforward dynamic programming using MATLAB-Ver 5.3. It is clear from Figure 4, which shows a comparison of the total errors from the recognition system on experiments from different musicians. The best performance was obtained from HMM (95%) and only (80%) from Time delay neural network. The experimental results show the capability of the statistical models and accuracy of HMM in understanding and recognition tasks.

![Figure 4. Results Of Statistical Estimation](image)

6. CONCLUSIONS AND FURTHER WORK

This paper has described the development of a prototype interactive, computerized music training, which uses HMM – based techniques for automatic musical recognition. The paper has covered some basic ideas about ragas, swaras, formant analysis and modeling for music recognition system. It has been our purpose to focus on basic concepts and logical ideas in developing musical recognition system, we have just given physical explanation of basic mathematics involved and have avoided long, drawn out proofs or derivations of the key result. We have also concentrated primarily on trying to interpret the meaning of HMM, and how it would be implemented in practice in real world systems. Some basic ideas about statistical modeling is also discussed in APPENDIX. The above paper shows the capability of statistical models and accuracy of HMM for swaras recognition, which leads to interesting applications in the field of continuous music recognition, and interactive system music tutor. Several other strategies, including most recent techniques like forward and backward chaining, syntactic models, hypothesized test models can be studied and implemented for more procedural models like our system designed above. Having a computer database to identify and recognize complex notes and tones can be developed for advancement of our system. These techniques are in current study and results will be developed after careful investigations. Moreover the presence of noise and echo can adversely affect the system design. This problem can be corrected by proper echo and noise canceller designs.
ACKNOWLEDGEMENTS

The authors wish to thank Mr. Madurai T.N. Seshagopalan, noted Indian musician for providing us with data’s on swaras and ragas along with waveforms. We would also like to thank Mr. Muthukaruppan (Principal-SVCE) and Dr. M.J.S. Rangachar, (HOD-ECE Dept) for their encouragements. We appreciate Dr. Narayanan Srinivasan, (Asst.Prof-ECE Dept) for his anonymous final review and help in correcting our mistakes. The authors also like to thank all our classmates for their support that was very important for the success of this paper.

APPENDIX A

Hidden Markov Models (HMMs) are a class of models for mimicking the probability density of a sequence of observed symbols. They are essentially stochastic finite state machines, which output a symbol each time they depart from a state. By specifying the state transition probabilities between states and the symbol generation model for each state, we can attempt to capture the underlying structure in a large set of symbol strings. In general, the operational paradigm is as follows: select a starting state according to some fixed probability distribution. At each time step, generate an output symbol by invoking the generative model of the current state, and then transition to a new state according to a static transition probability matrix.

A Hidden Markov Model is a Markov Chain with an associated output mechanism, which takes either states or transitions between states to either symbols or distributions on symbols. We refer to the Markov Chain as the underlying Markov Chain of the HMM. Hidden Markov Models appear in the literature in several forms, the most frequent being functions of a Markov Chain and State-output Hidden Markov Models. These forms are equivalent in the sense that for any HMM in one of these forms, there is an HMM in each of the other forms which defines the same process. The HMMs in this work will be Edge-output Hidden Markov Models, the elements of which are the set of states, the set of symbols, a stationary distribution on those states and, for each state, a joint distribution on symbols and next states.

Viterbi technique is an approximation technique. Viterbi training will attempt to optimize the parameters to maximize the likelihood of best path in the correct model. For this case, there is a consideration of posterior probabilities used in the estimation steps are assumed to either be zero or one. It can simply and without mathematical complexity stated as in following steps.

1. Assume an initial set of parameters for density estimators.
2. Determine the most likely state sequence
3. Update the parameters.
4. Access the solution and repeat the previous two steps as necessary.

Conceptually, segmentation of training data with a known model transcription is the same as recognition. Training can be done using different density functions like gaussian, Laplacian and other functions. In other words viterbi training is the estimation of best path for better convergence or learning of the system.

Baum-Welch training is a forward and backward training used to derive estimates for the probabilities of the hidden state and the transition variables, conditioned on the sequence of acoustic feature vectors. The training procedure is discussed in many classical books on probability as we concentrate mainly on the essence of the subject with respect to implementation than complex mathematics.
REFERENCES

