HIDDEN MARKOV MODEL APPROACH TOWARDS EMOTION DETECTION FROM SPEECH SIGNAL

K.Sathiyamurthy[1], T.Pavidhra[2], B.Monisha[3] and K.VishnuPriya[4]

Computer science and engineering, Pondicherry Engineering college, Puducherry, India

[1]sathiyamurthyk@pec.edu
[2]pavidhra0030@pec.edu
[3]monisha0026@pec.edu
[4]vishnupriya.pk94@gmail.com

ABSTRACT

Emotions carry the token indicating a human's mental state. Understanding the emotion exhibited becomes difficult for people suffering from autism and alexithymia. Assessment of emotions can also be beneficial in interactions involving a human and a machine. A system is developed to recognize the universally accepted emotions such as happy, anger, sad, disgust, fear and surprise. The gender of the speaker helps to obtain better clarity for identifying the emotion. Hidden Markov Model serves the purpose of gender identification.

KEYWORDS

HMM, Emotion recognition, Forward-Backward algorithm

1. INTRODUCTION

Need for natural way of communication between humans and machines arises as the computer plays a vital part in present world. This demands the computer to recognize its present situation and react differently which involves understanding a user's emotional state. Speech is the key input to an emotion recognition system. It is a potential way to communicate intentions and emotions. Speech features are extracted and classified using HMM and neural networks. Emotions help us to make decisions. The way men and women process their emotional memories also seem to differ. Women have better memories of their emotion than men. So they forget memories of the incidents happened before receiving emotionally charged information. This indicates that women are more pretentious to emotional content which means women and men process emotions in different parts of their brain. Evaluation of emotions and encoding of the memory is much more tightly integrated in women than in men. Hence gender plays a vital role in identifying the emotion. Hidden Markov Model is used for classifying the gender of the speaker. Markov model is a stochastic model for designing frequently changing systems in which the future states depend only on the present state and not on the sequence of events prior to the occurrence of the current state. A statistical Markov model with hidden states called hidden Markov model (HMM) is used in this paper. Next this paper illustrates the relationship related works.

Dhinaharan Nagamalai et al. (Eds) : CSITA, ARIN, ISPR - 2015
pp. 13–19, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.50802

2. RELATED WORKS

A work on A new architecture of intelligent audio emotion recognition done by Chien Shing Ooi, Kah Phooi Seng , Li-Minn Ang , Li Wern Chew [1]. It uses prosodic and spectral features. It has two main paths. Path 1 is for processing the prosodic features and path 2 for the spectral features. Bi-directional Principle Component Analysis (BDPCA) and Linear Discriminant Analysis (LDA) are done. The performance of the system is evaluated on eNTERFACE'05 and RML databases. Using audio emotion recognition the gender is not considered for detecting the emotion. Next work deals with gender recognition using Naive-bayes method.

A work on Gender Specific Emotion Recognition through Speech Signals [2] done by Vinay, Shilpi Gupta, Anu Mehra. The system has two modules one for Gender Recognition (GR) and the other for Emotion Recognition (ER). Features like pitch, energy and MFCC are extracted and then emotion recognition strategy is adapted. The result shows that a prior idea about the gender of speaker helps in increasing the performance of system. This approach has been implemented by using Naive Bayes method. But this system does not consider age which is also an important phenomenon for detecting the emotion. Next work deals with age and gender recognition.

A work on Automatic speaker age and gender recognition using acoustic and prosodic level information fusion [3] done by Ming Li ∗, Kyu J. Han, Shrikanth Narayanan. It has three sub systems. Gaussian mixture model using MFCC features, Support vector machine based on GMM mean super vectors and another SVM utilizing 450-dimensional utterance level features. Pitch, time domain energy, frequency domain harmonic structure energy and formant for each syllable are considered for analysis in subsystem. The subsystems have been used to achieve competitive results in classifying different age and gender groups. The database used to evaluate this approach is the aGender database. But this system does not use large database.

A work on **S**peaker state recognition using an HMM-based feature extraction method [4] done by R. Gajˇsek ∗, F. Miheliˇc, S. Dobriˇsek. This system uses acoustic features for recognizing various paralinguistic phenomena. The acoustic features are modeled with UBM by building a monophone-based Hidden Markov Model instead of representing UBM in the Gaussian Mixture Model (GMM). It is done by a two step process involving the transformation of the monophone-based segmented HMM–UBM to a GMM–UBM and then adapting the HMM–UBM directly. Both approaches supervise the emotion recognition task and the alcohol detection task. Here they used two corpuses, FAU-Aibo containing emotionally distinguished speech of children, and VINDAT for adult speech after alcohol consumption.

3. SYSTEM DESIGN

Fig 1 depicts the architecture of the proposed system. The proposed system consists of 4 stages. In the first stage random speech samples were collected from different speakers. This input is given to the signal processing stage where jAudio tool process the given input is processed to extract the required speech features. The features used are spectral centroid, spectral roll-off, spectral flux, spectral variability, root mean square, fraction of low energy window and zero crossing. Values for the above features are saved in an XML file. The values are classified into two ranges namely high and low based on all the values in the dataset as in table 1.

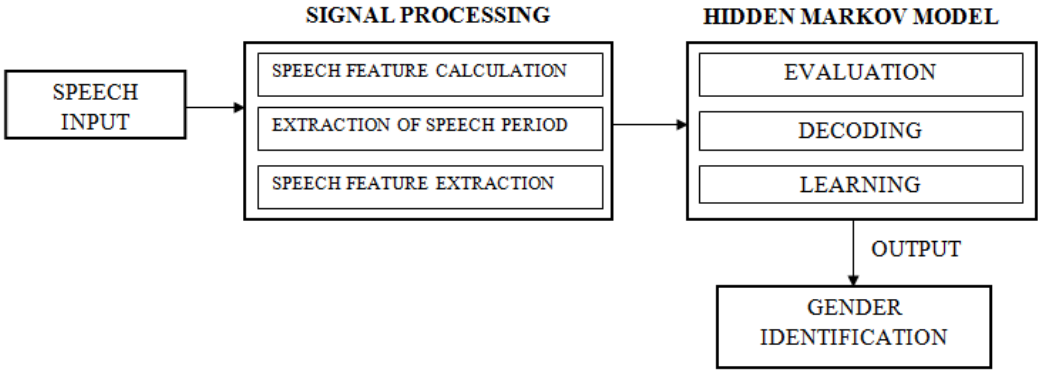

Figure 1. Architecture

Table 1. Data set

S. N	SPECTRAL CENTROID	ROLLOFF	FLUX	VARIABILITY	ROOT MEAN SQUARE	FRACTION FLOW ENERGY WINDOW	ZERO CROSSING	GENDER
1	HIGH	HIGH	HIGH	HIGH	LOW	HIGH	HIGH	MALE
2	HIGH	HIGH	LOW	HIGH	LOW	LOW	HIGH	MALE
3	HIGH	LOW	HIGH	LOW	HIGH	HIGH	LOW	MALE
4	LOW	HIGH	HIGH	HIGH	HIGH	LOW	HIGH	MALE
5	LOW	LOW	LOW	LOW	LOW	HIGH	LOW	FEMALE
6	LOW	HIGH	HIGH	HIGH	HIGH	HIGH	HIGH	FEMALE
7	HIGH	HIGH	HIGH	HIGH	HIGH	LOW	HIGH	FEMALE
8	HIGH	HIGH	LOW	HIGH	LOW	LOW	HIGH	FEMALE

4. IMPLEMENTATION OF PROPOSED WORK

HMM is used to recognize the gender of the speaker from the features in the XML file. HMM is modeled as in fig 2 with the gender {male, female} as the hidden states and the features {spectral centroid, roll-off, flux, variability, root mean square, fraction of low energy window, zero crossing} as the visible states.

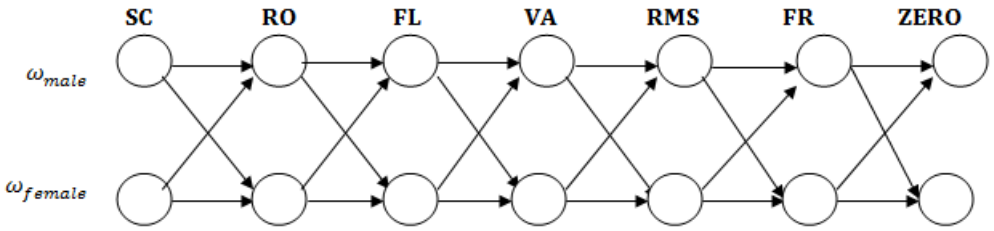

Figure 2. HMM MODEL

4.1. METHODOLOGY

We may have states at time t that are influenced directly by state at t -1.There are transition from one state to another (with certain probability) — Markov ModelThe states have a certain probability of generating various output symbols — the observations.Human can only see the

obervation, but not the underlying Markov Model. (Hence Hidden).The state at any time t is denoted $\omega(t)$. A sequence of states of length T is denoted by

$$\omega^T=\{\omega(1)\ ,\ \omega(2),\omega(3),\ldots,\omega(T)\}$$

Production of any sequence is described by the transition probability:

$$\mathrm{P}\ (\ \omega_j(\mathrm{t}+1)|\omega_i(\mathrm{t}))=a_{ij}$$

Transition prob. need not be symmetric, i.e.$a_{ij}\ \neq a_{ji}$ in general.We can observe some visible symbols v(t) at time t. However, the underlying state is unknown, i.e. hidden. In any state, we have a probability of emitting a particular visible state v_k(t), i.e. the same state may emit different symbols, and the same symbol may be emitted by different states.We denote this prob:

$$\mathrm{P}\ (v_k(\mathrm{t})|\omega_j(t))=\ b_{jk}$$

Because we can only observe the visible states, while the ω_i are unobservable, it is called Hidden Markov Model.

4.2. COMPUTATION

The underlying network is a finite state machine, and when associated with transition probabilities, they are called Markov networks. A final or absorbing state ω_0 is one which if entered, is never left. We require some transition must occur at each step (may be to the same state), and some symbol must be emitted. Thus, we have the normalized conditions.

$$\sum_j a_{ij}=1\ \forall_i$$
$$\sum_{\mathrm{k}} \mathrm{b_{jk}}=1\ \forall_{\mathrm{j}}$$

There are 3 central issues: (a). Evaluation process- Given a_{ij} and b_{jk}, Determine the prob. that a particular sequence of visible states V^T was generated by that model. Problem of the model produces a sequence V^T of visible state is which is mentioned in the equation below:

$$\mathrm{P}\ (\mathrm{V^T})=\sum_{\mathrm{r=1}}^{\mathrm{rmax}}\mathrm{P}\ (\mathrm{V^T}|\omega_{\mathrm{r}}^{\mathrm{T}})\ \mathrm{P}\ (\omega_{\mathrm{r}}^{\mathrm{T}})$$

where each r indexes a particular sequence $\omega_{\mathrm{t}}\ =\{\omega(1),\omega(2),\ldots..\omega(\mathrm{T})\}$ of T hidden states. In the general case of c hidden states, there will be rmax = cT possible terms. As we are working with a 1st order Markov process,

$$\mathrm{P}\ (\omega_{\mathrm{r}}^{\mathrm{T}})=\prod_{\mathrm{t=1}}^{\mathrm{T}}\mathrm{P}\big(\omega(\mathrm{t})\big)|\omega(\mathrm{t}-1)$$

i.e. a products of a_{ij}'s.The output symbol only depends on the hidden states, we can write

$$\mathrm{P}(\mathrm{V^T}|\omega_{\mathrm{r}}^{\mathrm{T}})=\prod_{\mathrm{t=1}}^{\mathrm{T}}\mathrm{P}(\mathrm{v}(\mathrm{t})|\omega(\mathrm{t}))$$

i.e. a product of b_{jk}'s. Hence,

$$\mathrm{P}\ (\mathrm{V^T})=\sum_{\mathrm{r=1}}^{\mathrm{rmax}}\prod_{\mathrm{t=1}}^{\mathrm{T}}\mathrm{P}(\mathrm{v}(\mathrm{t})|\omega(\mathrm{t}))|\mathrm{P}(\omega(\mathrm{t})|\omega(\mathrm{t}-1))$$

4.3. INTERPRETATION

Visible states v(t) is equal to the sum over all rmax possible sequence of hidden states of the conditional prob. that the system has made a particular transition, multiplied by the prob. that it then emitted the visible symbol in the target sequence.We can compute recursively define

$$a_j(t)=\begin{cases}0 & t=0 \text{ and } i \neq \text{initial state}\\ 1 & t=0 \text{ and } i=\text{initial state}\\ \left[\sum_i a_{i(t-1)a_{ij}}\right] b_{j(v(t))} & \text{otherwise}\end{cases}$$

α_j (t) denotes the probability of observing the sequence up to time t, and ending in state j.

$$\alpha_j(t) = P(v(1)\, v(2) \ldots\ldots v(t); I_t = i)$$

For the final state ω_0, we return α_0(T) for the final state. Computation Complexity O(c^2T). (b).decoding problem - Given the model and a set of observations V^T, determine the most likely sequence of hidden state $\omega_T$that led to those observation.We define backward variable:

$$\beta_j(t) = P(v(t+1); v(t+2); \ldots\ldots; v(T) \mid I_t = j)$$

$$\beta_i(t) = \begin{cases}0 & \omega_i(t) \neq sequence's final state\\ 1 & \omega_i(t) = sequence's final state\\ \sum_j \beta_j(t+1)a_{ij}b_j\big(v(t+1)\big) & otherwise\end{cases}$$

β_j(t) is the probability of starting from state j at time t, going through the observations and reach the final state.Time complexity: O(c2T). (c). learning problem - Given the coarse structure of the model (i.e. number of states, number of visible symbols), and a set of training observation of visible symbols, determine the parameters a_{ij} and b_{jk}.

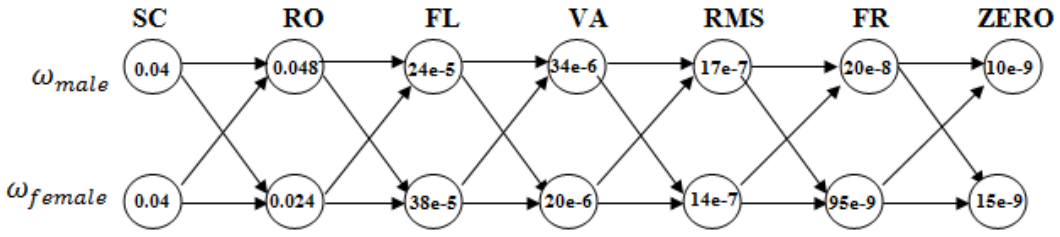

Figure 3. Forward Evaluation

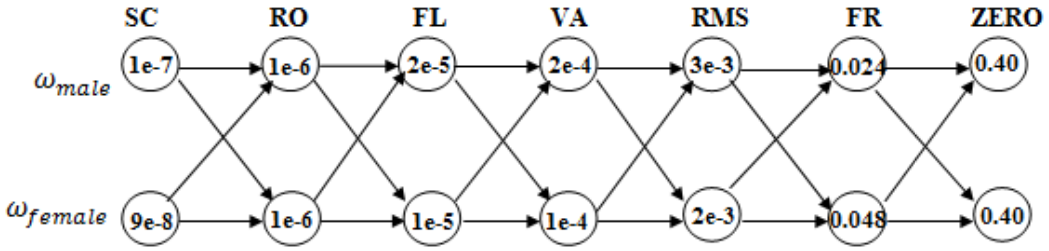

Figure 4. Backward Evaluation

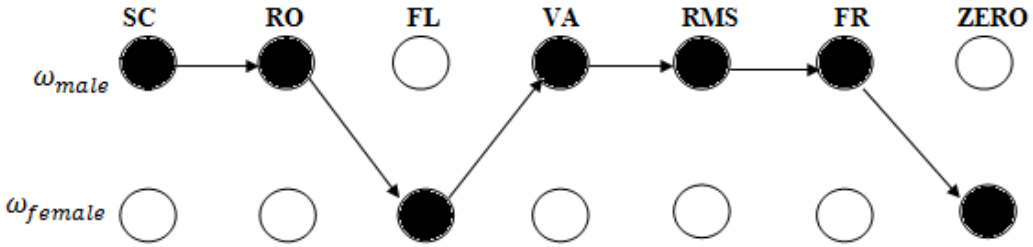

Figure 5. Pattern

5. CONCLUSION AND FUTURE WORKS

In this paper, the adoption of jahmm library for implementing HMM which has gender as the hidden state and speech features as the visible state. Speech features are extracted using jAudio tool. The future works include implementation of neural network for detecting the emotion with the help of speech features and the additional HMM output (Gender).

REFERENCES

[1] Chien Shing Ooi, Kah Phooi Seng, Li-Minn Ang & Li Wern Chew, (2014) " A new approach of audio emotion recognition", Expert System with Applications, Vol. 41, pp 5858–5869 .

[2] Vinay, Shilpi Gupta & Anu Mehra, (2014) "Gender Specific Emotion Recognition Through Speech Signals", International Conference on Signal Processing and Integrated Networks (SPIN), pp 727-733.

[3] Ming Li, Kyu J. Han, Shrikanth Narayanan, (2013) "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion", Computer Speech and Language, Vol. 27, pp 151–167.

[4] R. Gaj sek, F. Miheli c, S. Dobri sek, (2013) "Speaker state recognition using an HMM-based feature extraction Method", Computer Speech and Language, Vol. 27, pp 135–150.

[5] Ajmera, J., Burkhardt, F., (2008) "Age and gender classification using modulation cepstrum", In: Proc. Odyssey, p. 025.

[6] Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A. (2006b.) "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation", In: Proc. ICASSP, pp. 97–100.

[7] Ayadi, M. E., Kamel, M. S., & Karray, F. (2011) "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, 44(3), 572–587.

[8] Bhaykar, M., Yadav, J., & Rao, K. S. (2013) "Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM", Paper presented at the Communications (NCC), 2013 National Conference on (15–17 Feb).

AUTHORS

Dr. K. Sathiyamurthy is working as an Assistant Professor at Pondicherry Engineering College in the department of Computer Science And Engineering

T. Pavidhra is a student at Pondicherry Engineering College in the Department of Computer Science and Engineering

B. Monisha is a student at Pondicherry Engineering College in the Department of Computer Science and Engineering

K. Vishnu Priya is a student at Pondicherry Engineering College in the Department of Computer Science and Engineering