HMM Modeling of User Mood through Recognition of Vocal Emotions

Krishna Asawa and Raj Vardhan

Department of CSE & IT, JIIT-Noida, India
krishna.asawa@jiit.ac.in, me@rajvardhan.co.in

Abstract. This paper aims at defining a real-time probabilistic model for user's mood in its dialect with a software agent, which has a long-term goal of counseling the user in the domain of "coping with exam pressure". We propose a new approach based on Hidden Markov Models (HMMs) to describe the differences in the sequence of emotions expressed due to different moods experienced by users. During real time operation, each user move is passed on to a vocal affect recognizer. The decisions from the recognizer about the kind of emotion expressed are then mapped into code-words to generate a sequence of discrete symbols for HMM models of each mood. We train and test the system using corpora of the temporal sequences of tagged emotional utterances by six male and six female adult Indians in English and Hindi language. Our system achieved an average f-measure rating for all moods of approximately 78.33%.

Keywords: Mood detection, Hidden Markov models, affective computing.

1 Introduction

The examination of different software in various application areas like virtual training environments, portable personal guides, storytelling systems and interfaces of consumer electronics reveals that embodied conversational agents are widely used to provide users with a more human-like interface. But the ECA face the fundamental challenge of better understanding and integrating the affective cues used in communication by a user [8]. During an interaction with the agent, a user may be confident, ignorant, aggressive, cheerful, excited, bored, etc. The capability to recognize the current mood of the user could potentially assist the agent in better communication via selecting the appropriate words, phrases, dialog strategies etc. Here we attempt to integrate and model two major affective characteristics: emotions and moods [9] of an individual experienced and expressed in dynamic social interactions.

Emotion [4] is the complex psycho physiological experience of an individual's state of mind as interacting with biochemical (internal) and environmental (external) influences. A mood is a medium-term emotional state. Moods differ from emotions in that they are less specific, less intense, and less likely to be triggered by a particular stimulus or event. According to [3] conditions for mood-changes can be divided into (a) the onset of a mildly positive or negative event, (b) the offset of an emotion-inducing event, (c) the recollection or imagining of emotional experience, and (d) the

P.C. Vinh et al. (Eds.): ICCASA 2012, LNICST 109, pp. 229–238, 2013.
© Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2013

inhibition of emotional responding in the presence of an emotion-inducing event, (e) the personality traits such as optimism and neuroticism. Mood is an internal subjective state, but the emotions expressed by a user during an interaction can be monitored and analyzed to indicate the current mood. We classify an expressed fixed duration temporal pattern of emotions into one of the three classes of mood – Positive, Negative and Neutral.

In the first category of Positive mood (POS) [3], the user is in a desirable state of mind. Users seem to experience positive mood when they feel no sense of stress. Positive mood is usually considered a displaced state; users cannot pinpoint exactly why they are in a good mood. POS users are easier to interact with as compared to users in a negative or neutral mood. They have high probability of expressing emotions of happiness for longer durations during an interaction with the agent. Negative mood (NEG) [3], our second category is for the class of users that are in an undesirable state. Negative mood can be a consequence of chronic unresolved stress. We identify NEG users as those with high probability of expressing emotions of sadness for long durations. Finally, a third category of Neutral mood (NEUT) [3] is defined for users who display neutral emotions. Emotions expressed by NEUT users are generally neutral and the overall degree of happiness or sadness in an interaction is low.

Although the primary support of voice is to communicate, it can be also seen as an indicator of the psychological and physiological state of the speaker. Prosodic elements [2] transmit essential information with regard to the speaker's attitude, emotion, intention, context, gender, age and physical condition. The various prosodic features that characterize the emotions are variation in syllable length, loudness, pitch, formant frequencies of speech sound, accent, stress, rhythm, tone, and intonation. The discrete basic emotions considered in our study are happiness, surprise, sadness, anger and neutral.

In this paper, we investigate the issue of detecting the user's mood based on vocal emotions expressed by a user in due course of conversational speech with a software agent. We choose a three layered SVM classifier [5], having 85% recognition accuracy, to understand the kind of emotion carried by the utterance uttered by users during an interaction with the agent. Hidden Markov Models (HMMs) [1] are used as a suitable formalism to represent relationship of these temporal vocal emotion patterns with the user's mood. During real time operation, each user move is passed on to the vocal affect recognizer. The decisions from the recognizer about the kind of emotion expressed are then mapped into code-words to generate a sequence of discrete symbols for HMM models of each mood. The HMM model selected is the one that is more likely to produce the emotion sequence given as input.

2 Related Work

Embedding an affect-recognition component in an intelligent interactive system will enhance its ability to provide the necessary guidance, and make the chat sessions more interactive and thus effective. Affect, however is difficult to model because of its inherent complexity. Affect is a construct that subsumes heterogeneous group of

processes and there are many ways by which those emotions particular to a mood can be parsed at the human level [7]. One can argue that a computer cannot model emotions and feelings or general subjective impressions precisely because these are subjective entities.

The proven efficiency of identifying various differences in wide varieties of patterns encourages researchers to gauge the effectiveness of HMM modeling in the field of affect recognition from extracted low level feature classification and coded features classification to fusion of heterogeneous affective features. Other than several machine learning approaches like NNs (neural networks), k-NN (nearest neighbor) algorithm and SVM (support vector machines), HMM is also attempted to classify the given affect in the several emotions. In [13] speech data are parameterized with prosody related features and spectral features together with their first and second order derivatives. The temporal patterns of the emotion dependent prosody contours are modeled with the HMM structures. In an earlier work [17] author has proposed an expert-critic system based on HMMs to combine multiple modalities. The work reported by [16] has used HMM based head-nod and head-shake detection system, which provides the likelihoods of head-nods and head-shakes on the basis of tracked pupil positions. In the studies done in [18] Hidden Markov Models (HMM) are used to handle the temporal properties of the gesture(s). Levin et al. [10] proposed to use this formalism in dialogue pattern modeling: system's moves are represented in states while the user's moves are associated with arcs. Their goal is to solve the problem of defining the minimal cost dialogue strategy to adopt. Stolcke et al. [11] defines a discourse grammar using HMMs for Dialogue Act (DAs) prediction: they associate user moves with states in a HMM based dialogue structure in which transitions represent the likely sequencing of user moves. Evidence about DAs is expressed in terms of their lexical and prosodic manifestations. Novielli N. [12] studied on how the behavior of users changes according to their own goals and to their level of involvement in the advice-giving task using HMMs. In their models, states represent aggregates of either system's or user's moves, each with a probability to occur in that specific phase of the dialogue while the transitions represent the possible dialogue sequences.

For an agent to lively communicate with people in a natural way, it requires its understanding of human mood and personality. The research in the direction of detecting mood from perceived emotions by a software agent is very limited. In this study we have attempted to model the relationship of temporal expressed vocal emotions and its corresponding mood by wide range of personalities during interaction of a user with a software agent.

3 Architecture

During an interaction with the agent, each user move is classified into a discrete emotion value by the vocal affect recognizer, which is streamed to a decision pool. Figure 1 shows the overall architecture of our system for mood detection. Section IV and section V describe details of the emotion recognition phase and HMM modeling respectively.

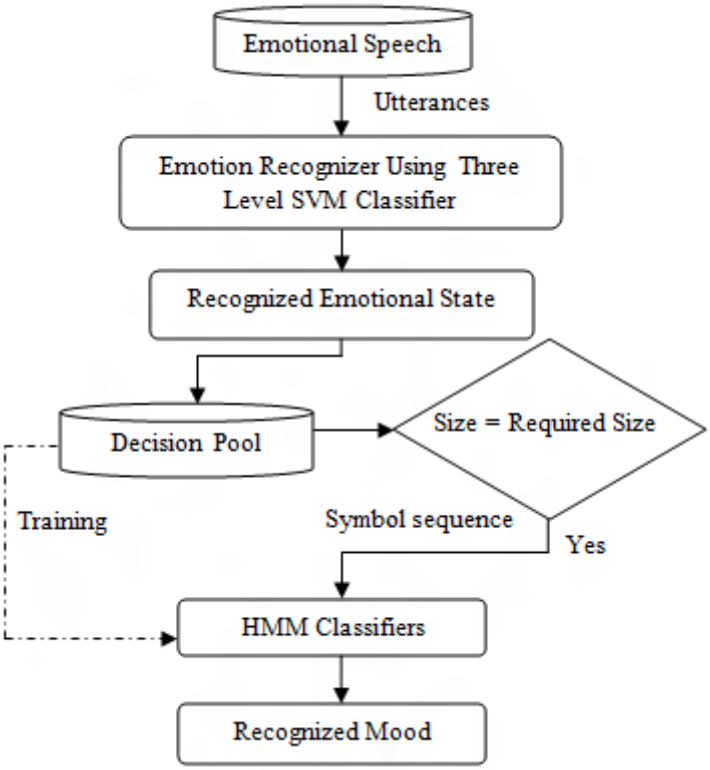

Fig. 1. Architecture of the mood detection system

4 Recognition of Vocal Emotions

The vocal emotion recognizer [5] proposed in our previous research recognizes discrete basic five emotions (anger, sadness, happiness, neutral, and fear) expressed in arbitrary duration utterances which may be language, cultural and subject independent. The various prosodic features used that characterize the emotions are variation in syllable length, loudness, pitch, formant frequencies of speech sound, accent, stress, rhythm, tone, and intonation. These features convey paralinguistic information such as emphasis, intent, attitude, and emotion of a speaker. Fusion of vocal segments with respect to classification and combination of diverse timing level are done at the pre-processing stage.

A three layered SVM classifier having RBF kernel is used that utilizes the dominant prosodic discriminant features at different layers. At the starting level, the features in consideration are intensity, shimmer, LPC and zero crossing rate so as to differentiate between high energy emotions (anger and happiness) with low energy emotions. At second level the various features considered are MFCC, pitch, jitter, harmonics. Along with MFCC and pitch, jitter distinguishes between anger and

happiness and shimmer distinguishes neutral from fear and sadness. The third level is used to differentiate sadness and fear since they are relatively similar and hence more specific features like fundamental frequency used to discriminate them.

The experiments are conducted on both the standard Berlin Database of Emotional Speech (EMO-DB) and self-recorded portrayed audio by Indians in the Hindi and English Language. Results obtained reveal that the system performed well with an average accuracy for all discrete basic five emotions of approximately 85%. A sample voice utterance which was tagged as 'Happy' by the affect recognizer can be seen in Figure 2.

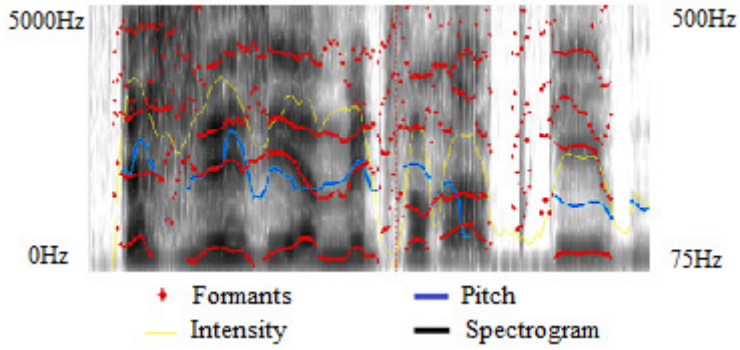

Fig. 2. Various prosodic features captured through Prat tool for a voice utterance corresponding to emotion 'happy'

5 HMM Modeling of User Mood

We decided to model the current mood of a user, by looking at differences in the emotion pattern obtained by analyzing their voice utterances. By using the formalism of HMMs, we are able to represent differences in the whole structure of these patterns among subjects with the kinds of mood we mentioned above.

5.1 User Mood Representation

Formally an HMM can be defined as a tuple: (Q, V, π, A, B), where:

- N = number of states in the model
- M = number of observation symbols
- Q = $\{q_1, q_2, \ldots, q_N\}$ is the set of states in the model
- V = $\{v_1, v_2 \ldots, v_M\}$ is the set of observations or output symbols
- $\pi = \{\pi_i\}$, $\pi_i = P_r(q_i \text{ at } t=1)$, $i \in S$, is the initial state distribution
- A = $\{a_{ij}\}$, aij = P_r(qj at t+1 | q_i at t), i, j ∈ S, is the state transition probability distribution
- B = $\{b_j(k)\}$, $b_j(k) = P_r$ (v_k at t|q_j at t), observation symbol probability distribution in state j

In our models states represent user moves, such that emission probability of the emotion *e* is probability of that emotion *e* being expressed in that state. The transitions represent the possible emotion sequences.

5.2 Learning the Model

The publicly available Berlin Database of Emotional Speech (EMO-DB) is used for the training of the affect recognizer model [5]. We used a database [19] of portrayed mood audio utterances from by six males and six females adult Indians in two languages viz. English and Hindi, to train and test the HMM models. This database was created with the help of a wizard of Oz experiment.

The database covers several different aspects including age, gender and background in computer science of the participating actors. During the building phase of the database in [19], 10 sessions were conducted, each on a different day with conditions suitable for recording. In every session, each actor had an interaction with the software agent for 15-20 minutes. To suite our requirements for testing, this database was further annotated. Each voice utterance was labeled with a discrete emotion value by 2 raters. Further the overall mood of users during a session interaction was annotated as per our definitions of positive, negative or neutral. The raters were asked to discuss the cases for which different annotations were given, and reach a common conclusion. These sequences of emotion labels are used to train HMM of the mood corresponding to the annotation of those interactions. For example, the set of sequences of emotion labels for interactions which were annotated with positive user mood will train the positive mood HMM, and so on. The corpus contains voice samples for 24 speech-based interactions with a total of 720 tagged voice utterances. 41% of this set of voice utterances was labeled positive, 27% as negative and 32% as neutral.

The Baum-Welch algorithm [6] is used to find the maximum-likelihood estimate of the parameters of a hidden Markov model given a set of observation sequences. The algorithm starts by assigning random parameters, which are iteratively adjusted according to the maximization function.

5.3 Model Description

Figures 3(a), (b) and (c) show respectively, the best 5-state HMMs for Positive, Negative and Neutral mood subjects. Tables 1, 2 and 3 show other parameters like emission and initial probabilities for the learnt HMMs. Abbreviated emotion labels {H-happy, Su-surprise, S-sad, A-anger, N-neutral} have been used in the tables. For instance, in Table 1 the value 0.149 at the intersection of the column *H* and the row *U1* gives the emission probability of emotion happiness at state U1. In Table 2, the value 0.65 at the intersection of the column P_i and the row *U2* indicates the initial probability of state U2 for the negative mood HMM.

In Figure 3(a) it can be seen that transition probabilities to U0, having the highest emission probability for emotion happiness, is relatively high. Clearly, a user in a positive mood will express emotions of joy for longer durations and if sad or angry, will have high probability of again starting to express neutral or happy emotions.

In Figure 3(b), states U2 and U3 have the highest emission probability of sadness and anger respectively. Transition probabilities to U2 and U3 being high explains why duration of emotions such as sadness and anger for a negative mood user are greater as compared to that of happiness, surprise or neutral. Similarly, in Figure 3(c), it can be seen that NEUT users show emotions of happiness or sadness for shorter durations as compared to neutral emotions.

Table 1. Emission and initial probabilities for positive mood HMM

State	H	Su	S	A	N	P_i
U0	.646	.150	.052	.047	.106	0.72
U1	.149	.609	.154	.073	.015	0.14
U2	.018	.085	.743	.057	.097	0.05
U3	.008	.016	.226	.734	.016	0.07
U4	.169	.029	.160	.027	.615	0.01

Table 2. Emission and initial probabilities for negative mood HMM

State	H	Su	S	A	N	P_i
U0	.700	.077	.036	.047	0.142	0.03
U1	.162	.549	.162	.101	.025	0.07
U2	.021	.087	.685	.091	.117	0.65
U3	.014	.040	.198	.738	.010	0.10
U4	.115	.018	.141	.028	.699	0.15

Table 3. Emission and initial probabilities for neutral mood HMM

State	H	Su	S	A	N	P_i
U0	.581	.164	.052	.062	.141	0.04
U1	.115	.621	.152	.095	0.016	0.03
U2	.015	.098	.682	.111	.095	0.03
U3	.006	.039	.201	.747	.007	0.08
U4	.148	.025	.149	.028	.650	0.83

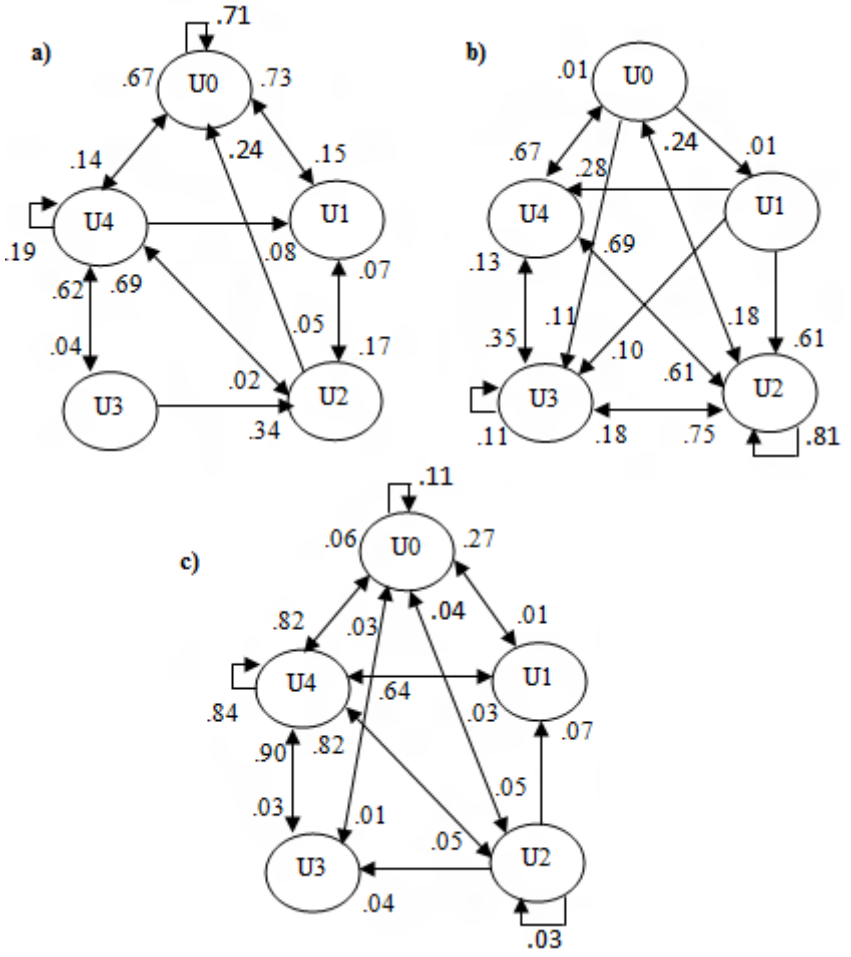

Fig. 3. HMM for *positive* (a), *negative* (b), *neutral* (c) mood

6 Testing

To evaluate the classification performance of the HMM models learnt, a 10-fold leave one out cross validation was performed on the annotated corpus. At every iteration *i*, the *i*-th instance of the data set is classified by choosing the model from all trained HMMs which maximizes the following:

$$\text{loglik} = \log P(i\text{-th case} \mid \text{HMM}_x), \text{ with } x \in \{\text{POS, NEG, NEUT}\}, \quad (1)$$

The HMM model selected is the one that is more likely to produce the emotion sequence given as input. For example, if we get that the positive HMM is most likely

to produce the sequence of emotions: {H, H, H, Su, S……, N, N, H}, it means that user is more likely to be in a positive mood. The probability is computed by using the forward algorithm [6]. All computations have been done using the open source JAHMM library [20] with programming done in java. Table 4 shows the results in terms of precision, recall and balanced f1 measure.

Table 4. Confusion matrix for positive, negative and neutral mood users

	Positive	Negative	Neutral	Precision	Recall	F-measure
Positive	.83	.05	.12	.83	.89	.86
Negative	.07	.74	.19	.74	.77	.75
Neutral	.09	.13	.78	.78	.71	.74

One of the reasons for robustness of the system being different for different classes of mood is the unequal distribution of training dataset. Some negative mood user interactions being confused for neutral mood can be attributed to certain similarities in behavior of negative and neutral mood users. Users of both these classes were found to have lower interest in the on-going interaction with the agent. While the negative mood users as a consequence of their mood tend to show more of emotions of sadness and anger, neutral users on the other hand are less engaged in the conversation and prefer to stay neutral.

We have discussed previously that in real-time mood detection, the voice inputs from a user first passes through a pre-processing phase where the emotion recognizer gives a resultant emotion label to it. The emotion recognizer used in our work has an accuracy of 85%, which further affects the accuracy of the overall system. Till now, we have limited our work to recognizing vocal emotions, however a multi-modal approach can be used to attain more efficiency as the attributes of HMM models are independent of the pre-processing done.

7 Conclusion

In this paper we model a problem of mood recognition faced by a software agent giving a session of dialect with a user. This study is based on temporal sequences of vocal emotions expressed by the user in different situations and context. The HMM model shows good recognition accuracy over wide varieties of emotional patterns corresponding to a mood. A pre-built 3 layer SVM based vocal emotion classifier which produces the input sequence to the HMM has been used to tag the emotion of the speech spoken by a user.

Acknowledgments. The work reported in this paper is supported by the grant received from All India Council for Technical Education; A Statutory body of the Govt. of India. vide f. no. 8023/BOR/RID/RPS-129/2008-09.

References

1. Charniak, E.: Statistical language learning. MIT Press, Cambridge (1993)
2. Fox, A.: Prosodic Features and Prosodic Structure. Oxford University Press (2000)
3. Morris, W.N.: Mood: The frame of mind. Springer, New York (1989)
4. Becker, P.: Structural and Relational Analyses of Emotion and Personality Traits. In: Zeitschrift für Differentielle und Diagnostische Psychologie (2001) (in German)
5. Asawa, K., Verma, V., Aggrawal, A.: Recognition of Vocal Emotions from Acoustic Profile. In: Proceedings of International Conference on Advances in Computing, Communications and Informatics, Chennai (2012)
6. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77(2), 257–286 (1989)
7. Davidson, R.: Honoring biology in the study of affective style. In: Ekman, P., Davidson, R. (eds.) The Nature of Emotion: Fundamental Questions, pp. 321–328 (1994)
8. Picard, R.W.: Affective computing: challenges. Int. J. Human- Comput. Stud 59(12), 55–64 (2003)
9. Batliner, A., Steidl, S., Hacker, C., Noth, E., Niemann, E.: Private emotions vs social interaction: towards new dimensions in research on emotions. In: Carberry, S., De Rosis, F. (eds.) Procs. of the Workshop on Adapting the Interaction Style to Affective Factors (2005)
10. Levin, E., Pieraccini, R., Eckert, W.: Using Markov decision process for learning dialogue strategies. In: Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing, vol. 1, pp. 201–204 (1998)
11. Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput. Linguist 26(3) (2000)
12. Novielli, N.: HMM modeling of user engagement in advice-giving dialogues. J. Multimodal User Interfaces 3, 131–140 (2010)
13. Schuller, B., Rigoll, G., Lang, M.: Hidden Markov Model-Based Speech Emotion Recognition. In: ICASSP, vol. 1, pp. 1–4 (2003)
14. Vlasenko, B., Wendemuth, A.: Tuning Hidden Markov Models for Speech Emotion Recognition. In: 33rd German Annual Conference on Acoustics, Stuttgart, Germany (2007)
15. Pantic, M., Bartlett, M.S.: Machine Analysis of Facial Expressions. In: Delac, K., Grgic, M. (eds.) Face Recognition, pp. 377–416. I-Tech Education and Publishing, Vienna
16. Kapoor Ashish Picard R.: Multimodal Affect Recognition in Learning Environments, Singapore (2005)
17. Kapoor, A., Picard, R.W., Ivanov, Y.: Probabilistic combination of multiple modalities to detect interest. In: ICPR (August 2004)
18. Elgammal, A., Shet, V., Yacoob, Y., Davis, L.S.: Learning dynamics for exemplar based gesture recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 571–578 (2003)
19. Vardhan, R., Asawa, K., Goel, S.: Emotion elicitation in a virtual dialog agent of an interactive counseling system. In: National Conference on Advances in Computer Sciences, Communication and Information Technologies, New Delhi (2012)
20. Jahmm – Hidden Markov Model (HMM): An Implementation of HMM in java, `http://www.run.montefiore.ulg.ac.be/~francois/software/jahmm`