ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

HMM-Based Analysis and Synthesis of Emotional Assamese Speech With Reference to its Prosody Features

Purnendu Acharjee[1], Jyotismita Talukdar[2]

Assistant professor, Dept. of CSc, AIMT, Guwahati, Assam India[1]

Assistant professor, Center of IT, UTM, Shillong, Meghalaya, India[2]

ABSTRACT :The present investigation focuses on how expressive content is apparent in the acoustic signal a speaker produces and also on listener reaction to the signal. In this paper the emotional features of Assamese language is presented. While the mono-thongs, diphthongs and trip thongs are considered with respect to different prosody we have seen that length of the Vowel not recognized as a distinguishing feature in Assamese . But in restrained deliberation, meaning-differentiating vowel length is a big issue towards the pronunciation of Assamese diphthongs and trip thongs. It is found that there is no change of meaning for vowel length written with symbols as short ('hrashya') or long ('dirgha'). Assamese word pronunciation does not differentiate between long and short vowels where as its orthography kept the prerequisite of short and long symbols. In the present study We are considering three basic emotional features of Assamese language : they are **normal(**Neutral**), angry and surprise**. It has been observed that the Assamese vowels /a//আ and /u//উ shows distinction up to frame 12th for **surprise** and **angry** emotions and from 17th frame onwards they seem to be similar in all the three emotions (prosody). But in case of the Assamese vowels /i//**ই**/ and /o//**ও** shows similarity up to 9th frame in Surprise and angry emotions and then shows dissimilarity up to 15th frame and then becomes flat at the end in all the emotions (prosody). For **monothongs** [say for Example: অ/a:/ আ/a/ ই/i/ এ/e/ ও/o/] it is observed that /a//আ/ and /u//উ/ shows clear cut distinction frame 9th-12th for Neutral, surprise and angry emotions and from 15th frame onwards shows similarity. On the other hand , in case of /i//**ই**/ and /o//**ও**/ **,** there is similarity up to 11th frame in **Neutral, Surprise and angry** emotions and but after that shows distinctions up to 17th frame. Similarly, for **diphthongs** ,like /a//আ/ with /i//**ই**/ and /u//উ/ shows distinction up to frame 13th for three emotions and rest frame onwards they show similarities in spectral behaviors. It is found that in the emptional speech , there are there are three **VOT patterns** found in Assamese language .(i) **(pre-voicing)Negative VOT**, where the vocal cords starts vibrating before the stop consonant release and an interval **from -125ms to -75ms**. (ii**) (simultaneous)Zero VOT**, where the vibration of the vocal cords starts vibrating more or less simultaneously to the release of plosive within an interval from **0 ms to +35ms,** and **(iii) (aspiration)Positive VOT**: where a delay pursues the plosive release and the vocal cords start vibrating after a **35ms to 100ms** interval. While Assamese native speaker utters sentences in emotions then it is seen that quite a few prosody factors affect this phonetic phonological characteristic such as place of articulation, syllable stress, rhythm, speech rate, number of syllables and vowel quality etc. Since values of VOT vary broadly depending on emotional status of a speaker so equally VOT length and sentence length need to be computed with the intention to establish what proportion of the sentence occupied by the VOT to obtain relative VOT (emotional states) in the target sentence.

KEYWORDS : emotion, VOT, HMM, ASR, GMM, synthesis,HPTS. cepstrum.

I. INTRODUCTION

The analysis of Speech emotion is highly associated with the speech production mechanism. The entire speech acoustics play an important role while interpreting the meaning of particular acoustic parameters . The air flow through the vocal tract, and thereafter, powered by respiration is the basis of all sound making with the human vocal apparatus. Interestingly,

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

the different types of sound produced by human depends on whether the air flow is set into vibration by rapid opening and closing of the glottis – phonation and thereby producing quasi periodic voiced sounds . In case of non periodic or unvoiced sounds, the air passes freely through the lower part of the vocal tract and is transformed into turbulent noise by friction at the mouth opening. Further, the acoustic filter characteristics of the vocal tract determines the quality of the sound . Thus, the entire mechanism of the sound production system is a complex process. In addition, this complexity further increases with the emotion of the speakers.

The most important advantage that lies in using **augmented feature** vectors is that the typical dynamic programming algorithms can be used to solve the HMM statistical problems, like the Viterbi and EM algorithms. In this case, the observation vectors are generally assumed to be statistically independent and in the training phase and also the correlations between them are not taken into account. As a result of this, it does not represent the temporal constraints of the training data and the constraints imposed on the generation of the speech features are from the output static features. This problem has been overcome by using the trajectory-HMM(Tokuda et al 2004, Zen et al 2007b). So in this trajectory model, the probability density function can be defined as a function of the static features and the combination of explicit relationships between the static and dynamic features are imposed through the normalization of the original likelihood P(O|q,λ). In this case, the EM and Viterbi algorithm is used for trajectory-HMM but the computations turns typically more complex. So static and dynamic features, known as (Static and Dynamic) SD-dimensional parameter vector o_t is as follows:

$$o_t = \left[c_t^\top, \Delta^{(1)} c_t^\top, \ldots, \Delta^{D-1} c_t^\top\right]^\top, \qquad \text{-----(1)}$$

Here $\mathbf{c_t}$ and $\Delta^{\mathbf{d}}\mathbf{c_t}$ represents the S-dimensional static and the ***d*-th** dynamic feature vectors. In HMM-based Emotional (Prosody) speech synthesis these vectors are calculated as follows:

$$c_t = [c_t(1), c_t(2), \ldots, c_t(M)]^\top \qquad \text{----(2)}$$

$$\Delta^{(d)} c_t = \sum_{\tau=-L_-^d}^{L_+^d} w^{(d)}(\tau) c_{t+\tau}, \qquad \text{----(3)}$$

Where $\mathbf{w^{(d)}(\tau)}$ is a window coefficient for calculating the ***d*-th** dynamic feature, $\mathbf{L^0 -= L^0 += 0}$ and $\mathbf{w^{(0)}}(0) = 1$. In this case number of dynamic feature vectors is often two (D = 3). It means the observation feature vector o_t defining the static coefficients, i.e., **delta(Δ)** and **delta-delta(Δ^2)** coefficients. The delta and delta-delta features are obtained by equations (4) & (5):

$$\Delta c_t = \frac{\sum_{\tau=-l}^{l} (c_{t+\tau} - c_t)}{\sum_{\tau=-l}^{l} \tau^2} \qquad \text{------(4)}$$

$$\Delta^2 c_t = \frac{1}{2} \frac{\sum_{\tau=-l}^{l} \tau^2 c_{t+\tau} - \frac{1}{L}\left(\sum_{\tau=-l}^{l} \tau^2\right)\left(\sum_{\tau=-l}^{l} c_{t+\tau}\right)}{\sum_{\tau=-l}^{l} \tau^4 - 1}, \qquad \text{----(5)}$$

Here *l* =no of frames per window and $\mathbf{L = 2l+1}$ is the width of the window that is used to calculate the dynamic features at frame t. Say for e.g., when a three-frame window is used in the HTS synthesizer, by the following formulas:

$$\Delta c_t = 0.5 c_{t-1} - 0.5 c_{t+1} \qquad \text{-----(6)}$$

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

$$\Delta^2 c_t = 0.25c_{t-1} - 0.5c_t + 0.25c_{t+1}$$

The duration model**[3],** which is used for speech recognition in conventional HMM, is not adequate for synthesis because it cannot capture the temporal structure of speech accurately. So an improved duration models is proposed and typically used in HMM-based Emotional (Prosody) speech synthesis.

II.Analysis

In the analysis, all types of linguistic, phonetic and prosodic parameters are calculated from the recorded speech corpus sentences using the text analysis tools. For this process each text and corresponding speech signal has to pass through a large set [4]of self designed question set as shown in Fig (1.1)**.**

Self Generated Question Set

```
QS "LL-Nasal"              {m^*,n^*,en^*,ng^*}
QS "LL-Fricative"          {ch^*,dh^*,f^*,hh^*,hv^*,s^*,sh^*,th^*,v^*,z^*,zh^*}
QS "LL-Liquid"             {el^*,hh^*,l^*,r^*,w^*,y^*}
QS "LL-Front"              {ae^*,b^*,eh^*,em^*,f^*,ih^*,ix^*,iy^*,m^*,p^*,v^*,w^*}
QS "LL-Central"            {ah^*,ao^*,axr^*,d^*,dh^*,dx^*,el^*,en^*,er^*,l^*,n^*,r^*,s^*,t^*,th^*,z^*,zh^*}
QS "LL-Back"               {aa^*,ax^*,ch^*,g^*,hh^*,jh^*,k^*,ng^*,ow^*,sh^*,uh^*,uw^*,y^*}
QS "LL-Front_Vowel"        {ae^*,eh^*,ey^*,ih^*,iy^*}
QS "LL-Central_Vowel"      {aa^*,ah^*,ao^*,axr^*,er^*}
QS "LL-Back_Vowel"         {ax^*,ow^*,uh^*,uw^*}
QS "LL-Long_Vowel"         {ao^*,aw^*,el^*,em^*,en^*,en^*,iy^*,ow^*,uw^*}
QS "LL-Short_Vowel"        {aa^*,ah^*,ax^*,ay^*,eh^*,ey^*,ih^*,ix^*,oy^*,uh^*}
QS "LL-Dipthong_Vowel"     {aw^*,axr^*,ay^*,el^*,em^*,en^*,er^*,ey^*,oy^*}
QS "LL-Front_Start_Vowel"  {aw^*,axr^*,er^*,ey^*}
```

Figure(1.1): A typical question set used in present study using Assamese words in HTK tool

This information is used by the HPTS (Hidden Markov Based Prosody Text to Speech) system in the form of labels which are further used for training all the context-dependent phone models (HMMs) as shown in fig(1.2). Most of the contexts are basically related to the counts, stretches from phone and to utterance level context positions and distances of stressed and accented syllables. Such contextual information used for my research in Assamese HPTS (Hidden Markov Based Prosody Text to Speech) is given below:

- Preceding, current, succeeding phones.
- Position of current phone in current syllable.
- Number of phones in preceding, current, succeeding syllable.
- Syllable Number of syllables in current utterance
- Position of current word in current phrase.
- Number of preceding, succeeding stressed syllables in current phrase.
- Accents of proceeding, current, and succeeding.

Both the (prosodic) Excitation and Spectral parameters calculated from the training speech corpus. The conventional method to estimate the envelope in HTS is Mel-Cepstral analysis.

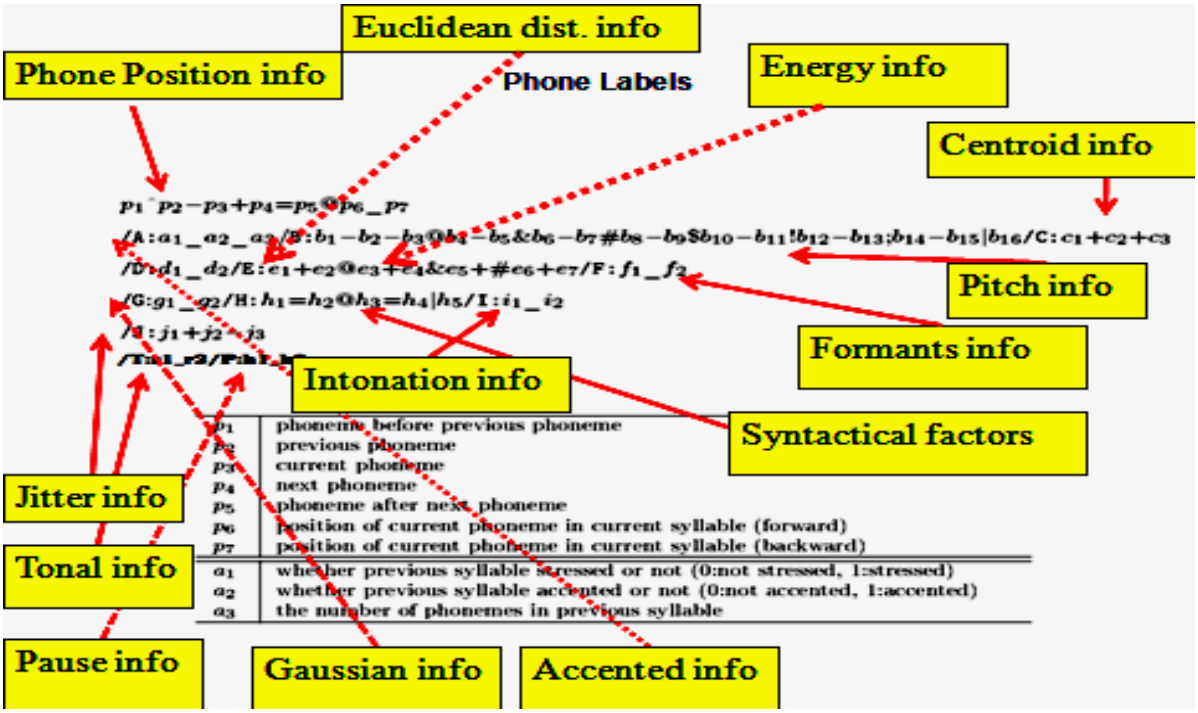
Figure (1.2): A typical e.g. of contextual information of an Assamese word **/deuta/** দেউতা

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015



Generating .lab files for each phoneme (Label the Signal):

To get the correct phone label each speech waveform we have to generate **.lab** files manually. When the speech region is marked for each signal we can get manually correct them. Here we have 3 successive regions: **start silence** (with label **sil**), the recorded phoneme (with label /**aa/** as **"**আ"), and **end silence** (with label **sil**). By rule, these 3 regions cannot overlap with each other but no matter if there is a little gap between them.

Remark:

The .lab file is a simple text file. It contains for each label a line of the type: **aa_S.lab**.

here '**S'** implies Surprised mood only

Start time	**end time**	**phone**	
4171250	**9229375**	**sil**	
9229375	**15043750**	**aa_S**	← "আ"
15043750	**20430625**	**sil**	

Figure (1.3): A typical e.g. of lab file used in present study for Assamese words in HTK tool

The numbers indicate the values of start and end sample time of phone label. Such type of file can be manually modified (for e.g. to adjust the start and end of a label).

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

III.STATISTICAL MODELING

In statistical modeling, HMM topology used in HTS with three or five-state left-to-right HMM. Each of the state output density function is modeled by a single Gaussian or Gaussian mixture distributions as shown in **figure (1.3).**
In this way the covariance matrix for each Gaussian mixture component generates as a diagonal covariance matrix. In terms of computational complexity, it is significantly more advantageous than a full covariance matrix. The Prosodic features and spectral parameters are included in the feature vector. The state duration densities of each word model for each emotional state have been modeled by Gaussian distribution[5]. The dimension of state duration density is found to be equal to the number of states in the HMM.
In the HTS, the process of re-estimation of the model parameters is performed by using the Hidden Markov Model Toolkit (HTK). In this training procedure, it uses the maximum likelihood estimation criterion. Finally, the prosodic and spectral parameters and state durations are clustered independently as per their own emotional state of speech because they have their own influential contextual factors.

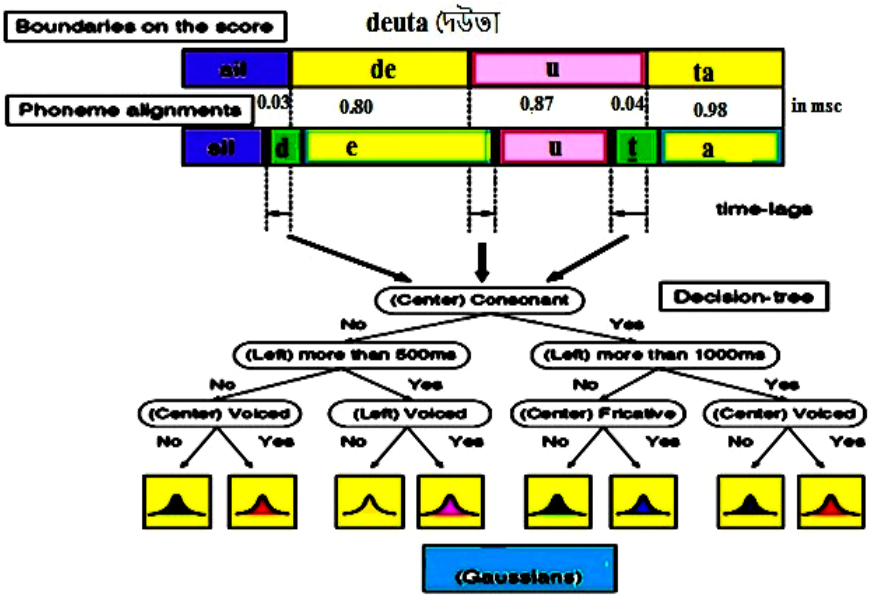


Figure (1.4): typical example of GMM distributions Assamese Male word /deuta/"দেউতা" (angry)

Configuration Parameters used in HTS :
Configuration file is a text file containing the information required for designing the HMM model for each phone. An e.g. configuration file is given below☹"#" can be used to introduce a comment)
Using a configuration file, an MFCC (Mel Frequency Cepstral Coefficient) analysis [6] is done.
For each signal frame:
- The first 12 MFCC coefficients [c1,…, c12] (where **NUMCEPS is taken as = 12**)
- The first "**null**" MFCC coefficient known as c0 defines the energy related to each signal since it is proportional to the total energy in the frame
- The 13 "**Delta coefficients**", estimates the first order derivative of [c0, c1… c12] as 1st prosody
- The 13 "**Acceleration coefficients**", estimates the second order derivative of [c0, c1… c12] as 2nd prosody.

Altogether, from each signal frame total number of 39 coefficient vector is extracted.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

Prosody Speech Feature Generation

There are some problems in generating the prosody speech parameter vector sequence O from HMM λ, for word transcription W. The problem is to maximize the output probability distribution with respect to O is as follows:

```
#
# Example of an acoustical analysis configuration file
#
SOURCEFORMAT = HTK              # Gives the format of the speech files
TARGETKIND = MFCC_0_D_A         # Identifier of the coefficients to use

# Unit = 0.1 micro-second :
WINDOWSIZE = 250000.0           # = 25 ms = length of a time frame
TARGETRATE = 100000.0           # = 10 ms = frame periodicity

NUMCEPS = 12                    # Number of MFCC coeffs (here from c1 to c12)
USEHAMMING = T                  # Use of Hamming function for windowing frames
PREEMCOEF = 0.97                # Pre-emphasis coefficient
NUMCHANS = 26                   # Number of filterbank channels
CEPLIFTER = 22                  # Length of cepstral liftering

# The End
```

Figure (1.5): A typical e.g. of HTK-configuration file

$$O^* = \arg\max_{O} P(O|W,\lambda,T) \qquad \text{------(6)}$$

To solve this problem we use the recursive method entirely based on the expectation maximization (EM) algorithm. Using EM algorithm, the HMGenS tool of the HTS system is used to generate prosody speech parameters. If the problem is still left then Viterbi-based method is applied to solve the optimization problem. The HTS system includes hts engine, a small run-time synthesis engine, which generates speech parameters based for developing HPTS (Hidden Markov Based Prosody Text to Speech) tool.

Creating monophone:

The first step is to define a prototype model in HMM training. The parameters of a model are very important; its main purpose is to define the entire model topology. For a phone-based HMM systems use 3-state left-right with no skips as following in figure (1.5):

Creating Monophone HMMs

```
<BeginHMM>
  <NumStates> 5
  <State> 2
      <Mean> 39
        0.0 0.0 0.0 ...
      <Variance> 39
        1.0 1.0 1.0 ...
  <State> 3
      <Mean> 39
        0.0 0.0 0.0 ...
      <Variance> 39
        1.0 1.0 1.0 ...
  <State> 4
      <Mean> 39
        0.0 0.0 0.0 ...
      <Variance> 39
        1.0 1.0 1.0 ...
  <TransP> 5
   0.0 1.0 0.0 0.0 0.0
   0.0 0.6 0.4 0.0 0.0
   0.0 0.0 0.6 0.4 0.0
   0.0 0.0 0.0 0.7 0.3
   0.0 0.0 0.0 0.0 0.0
```

Figure(1.6): Typical e.g. of a HMM 3-state left-right in Assamese with no skips

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

Creating HMM models for each phoneme with different emotions

The models consists of 4 **"active"** states namely {*S2*, *S3*, *S4*, *S5*}: where the first and the last states (here *S1* and *S6*), are known as **"non emitting"** states where no observation function is used. The observation functions b_i are known as single Gaussian distributions containing the diagonal matrices. The transition probabilities are in matrix ***aij*** as follows:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} \\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} \end{bmatrix}$$

$$\begin{bmatrix} 0.0 & 0.5 & 0.5 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.4 & 0.3 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.4 & 0.3 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 0.3 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

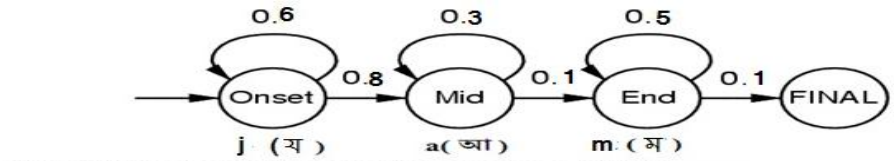
Figure (1.7): typical e.g. of transition probabilities are in matrix ***aij*** for the phoneme "/a/"

HMM Transition Diagram with HMM Transition Matrix for Assamese words:

For any real time systems, Standard Transition Diagram (STD) is the best way to describe the time dependent behavior. All real time systems follow the basic rule as :

- ✓ Each state referred in a system is an observable mode of behavior of the system.
- ✓ At any particular instant of time any Standard Transition Diagram (STD) can only be one unique state.
- ✓ Any real time system behavior can be described by one or more STD.

By definition "A Transition matrix is referred as a probability matrix which is used to describe the transition of a **Markav chain**"[7]. While using with HMM tool for Assamese words, say for e.g. /jam/ (যাম) in three different moods namely **Normal, Surprise** and **Angry**. The transition diagrams with the corresponding state transition matrix for the Assamese word **/jam/ (যাম)** is shown below:



Output probabilities for the phone HMM: a আ

Onset:	Mid:	End:
C1: 0.5	C3: 0.2	C4: 0.1
C2: 0.2	C4: 0.8	C6: 0.5
C3: 0.6	C5: 0.3	C7: 0.1

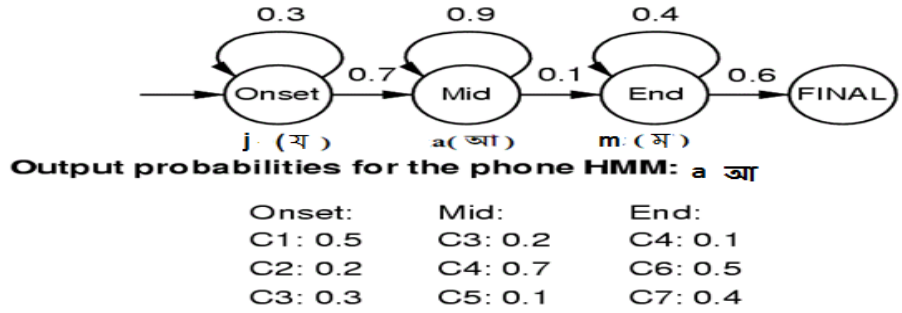
Figure(1.8): A typical e.g. of Transition diagram & matrix for Assamese word/jam/ (যাম) (Norml)

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

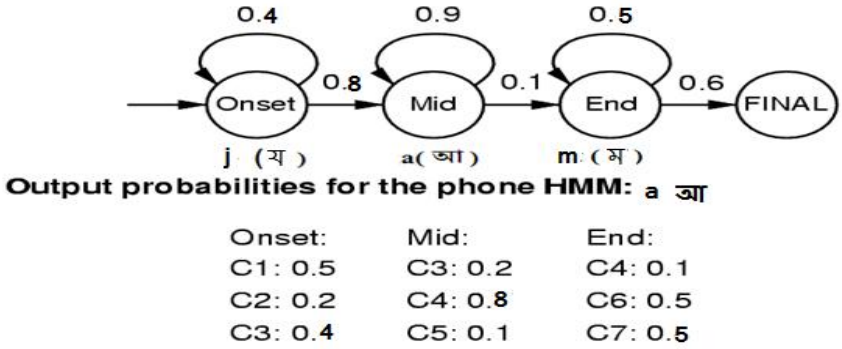
International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015



Figure(1.9):A typical e.g. of Transition diagram & matrix for Assamese word/jam/ (যাম) (Surprise)



Figure(1.10): A typical e.g. of Transition diagram & matrix for Assamese word/jam/ (যাম) (Angry)

IV. CO-VARIANCE ESTIMATION FOR ASSAMESE VOWELS WITH DIFFERENT EMOTIONS

The co-variance of HMM model for three emotional states namely Normal, Surprise and Angry are estimated by training through standard EM algorithm. The set all the state distributions are same. The formant mean is set to 500Hz and the bandwidth mean is set to 100Hz. The standard deviation is set to 500Hz and the bandwidth standard deviation is set to 100Hz. The mean of the delta-formant and delta-bandwidth set to 0Hz. The standard deviation for the delta-formant and delta-bandwidth set to 100Hz.

V. COVARIANCE VALUE COMPARISON FOR MALE AND FEMALE OF ASSAMESE VOWELS IN HMM

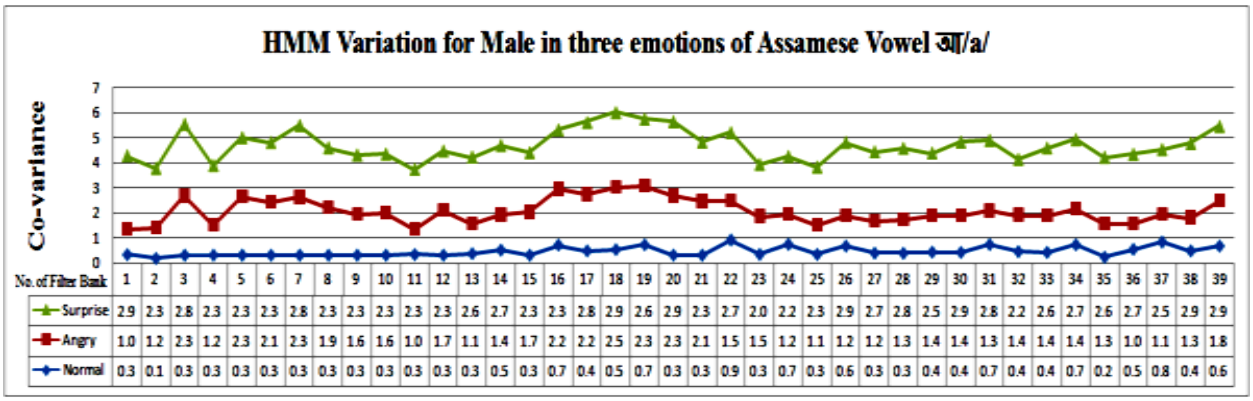
During analysis, it is found that the emotions are nicely visible in long vowels in Comparison to short vowels. So the comparison performed on long vowels such as /a/**আ**, /u/ **ঊ** of Assamese shows distinctive features in their utterances and also in the way of articulation. It is found that higher tone for Assamese /a/**আ** is uttered from a lower tone and went to sudden rise and after some time then sudden fall occurs. In case of vowel /u/**ঊ** for Assamese It is found that Assamese /u/**ঊ** has sudden rise in the beginning and also some nasalization in between and then slow fall down.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

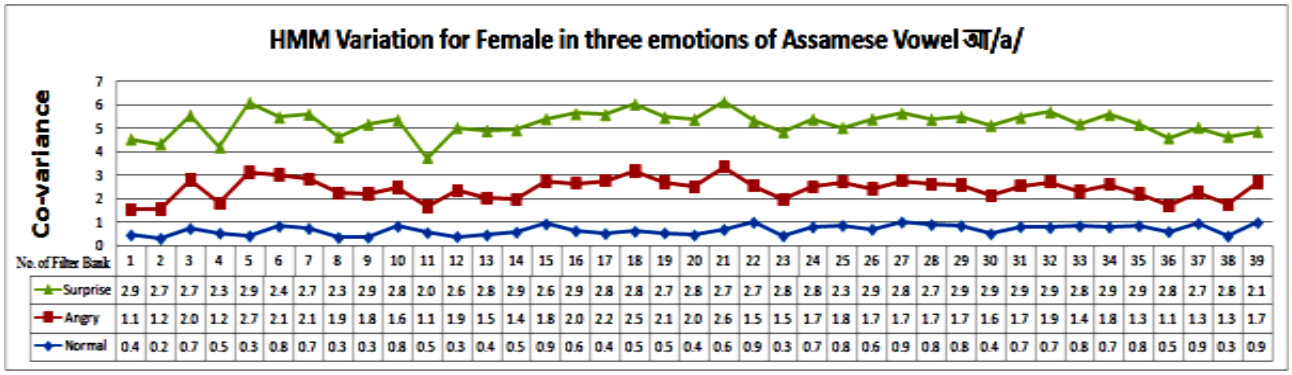
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015



(t)Time in ms

Figure (1.11):HMM co-variation for Male in three emotions of Assamese /a/**আ**



(t)Time in ms

Figure (1.12): HMM co-variation for Female in three emotions of Assamese /a/**আ**

VI. OBSERVATION AND CONCLUSION

In the present study it has been nicely observed that the Assamese vowels /a//আ/ and /u//উ/ shows distinction up to frame 12th for surprise and angry emotions and from 17th frame onwards they seem to be similar in all the three emotions (prosody). But in case of the Assamese vowels /i//**ই**/ and /o//**ও**/ shows similarity up to 9th frame in Surprise and angry emotions and then shows dissimilarity up to 15th frame and then becomes flat at the end in all the emotions (prosody).

The Assamese Stop consonants like (/p//প/, /b//ব/, /t//ট/, /d//দ/, /k//ক/, /g//গ/) are generated in three succeeding stages: a) the engaged articulators generate a complete obstruction of the air in the oral cavity (called closure); subsequent the closure, b) there is consonant release, which relate to the articulatory segment where the obstruction is undone (unwrap); lastly c) the vocal cords of the portion which adhere to the current consonant starts vibrating. Voice on Time is defined as the time period between the release of stop consonant and the start of vibration of the vocal cords of this consonant. There are three VOT patterns found in Assamese language.

When the stop consonants are investigated at different emotion it is observed as following:

a) The utterance of velar /k//ক/ (stop consonant) for Assamese word (leads by a silence), it is uttered by a native Assamese speaker results positive VOT 75-92 ms.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

b) The utterance of stop consonants /b//ব/, /d//দ/ and /g//গ/ results negative VOT -120ms to -80ms to - and the voiceless plosives /p//প/, /t//ট/ and /k//ক/are uttered results Zero VOT of mean values 16ms 21ms 45ms respectively.
c) The utterance of Voiceless plosive shows dynamic attitudes as in some cases it tend to be Zero VOT and also shows negative VOT.
d) The utterance of Voiceless stops /p^h//ফ/, /t^h//থ/ and /k^h//খ/ results positive VOT 57 ms, 73ms and 81 ms respectively.

HMM-based speech synthesis is one of the most recent application of the HMM models in speech technology. In my research I tried to design and develop a model on emotional speech synthesis using prosodic feature sets. It applies two types of technologies: 1) use of the generative model and EM algorithms for computing the HMM model parameters for building the emotional TTS and 2) to evaluate the likelihood $P(O|\lambda)$ in the synthesizer for producing the sound with proper emotions respectively. This HMM statistical emotional/prosody speech synthesizer aims to generate the most natural sounding human speech as much possible and also to model the speech with proper emotions which are characteristic of human speech, like speaker identity, expressiveness etc.

REFERENCES

[1] J. B. Allen and L. R. Rabiner. Proceedings of IEEE, 65, 1558 (1977).
[2] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai. Speech synthesis using hmm with dynamic features. IEEE, 389-392 (1996).
[3] Q. Yan and S. Vaseghi. Modeling and synthesis of english regional accents with pitch and duration correlates. Computer Speech and Language 24, 711-725 (2010).
[4] K. Kumar and R. K. Aggarwal. Hindi speech recognition system using htk. I. J. Comp. Busi. Res. 2, 2011.
[5] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17, 91 (1995).
[6] P. H. Talukdar et al. Cepstral measure of bodo vowels through lpc analysis. Journal of the CSI 34, 1 (2004).
[7] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. Ann. Math. Stat. 37, 1554 (1966).