Deterministic Annealing EM Algorithm for Developing TTS System in Gujarati

Nirmesh J. Shah, Hemant A. Patil, Maulik C. Madhavi, Hardik B. Sailor and Tanvina B. Patel

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IIICT), Gandhinagar-382007, India

{nirmesh_shah, hemant_patil, maulik_madhavi, hardik_sailor, tanvina_bpatel}@daiict.ac.in

Abstract

The generalized statistical framework of Hidden Markov Model (HMM) has been successfully applied from the field of speech recognition to speech synthesis. In this work, we have applied HMM-based Speech Synthesis System (HTS) method to Gujarati language. Adaptation and evaluation of HTS for Gujarati language has been done here. Evaluation of HTS system built using Gujarati data is done in terms of naturalness and speech intelligibility. Apart from this, a conventional EM algorithm-based HTS and recently proposed Deterministic Annealing Expectation-Maximization (DAEM) algorithm has been applied to Gujarati (a low resourced language) and it’s relative comparison has been done. It has been found that HTS in Gujarati has very high intelligibility. It was verified from AB-test that 70.5% times DAEM-based HTS has preferred over EM-based HTS developed for Gujarati language.

Index Terms: HMM, Expectation-Maximization, Deterministic Annealing Expectation-Maximization (DAEM).

1. Introduction

Text-to-Speech (TTS) means automatic generation of speech waveform from given input text [1]. There are mainly three methods available for TTS, viz., (1) model-based synthesis, (2) concatenative synthesis and (3) statistical parametric model-based synthesis. Model-based synthesis is rule-based synthesis which uses simplified model of speech production [2]. For example, in formant synthesis, fundamental frequency (i.e., $F_0$), voicing and noise levels are varied over time to synthesize waveform artificially [2]. Unit-Selection Synthesis (USS) method is an example of concatenative synthesis which comes under corpus-based methods [3]. In this method, at the time of synthesis, units were selected from pre-recorded database and concatenated in runtime. Since, speech sound units are selected from natural speech, this method provides high quality and naturalness. This method is very suitable for limited-domain synthesis but for unrestricted synthesis, it requires very huge amount of data whereas in HMM-based Speech Synthesis System (HTS), instead of storing entire studio recorded speech database, only model parameters are stored and used to synthesize waveforms [4]. HTS comes under Statistical Parametric Synthesis (SPS) method for TTS [5].

Gujarati is one of the official languages of India [6]. Gujarati is an Indo-Aryan language and evolved from Sanskrit [6]. Recently, USS method has been applied to Gujarati language [7], [8]. HTS has already been applied to other Indian languages [9], [10].

In this work, HTS have been applied to Gujarati language. Adaptation and evaluation of different HTS developed for Gujarati language have been presented. For training of the system, HTS uses Baum-Welch algorithm which is also called Expectation-Maximization (EM) algorithm for Hidden Markov Model (HMM). Deterministic annealing Expectation-Maximization (DAEM) algorithm was proposed to overcome local maxima problem of EM algorithm [11]. For English language, DAEM-based HTS has already been applied [12]. However, to the best of authors’ knowledge DAEM-based HTS for Gujarati language is not yet build. Recently, phone-based TTS has applied to Gujarati language [13]. In this work, we have done comparison of conventional EM algorithm and DAEM algorithm in the context of mixture density problem as well as in context of HTS for Gujarati language. Various subjective as well as objective evaluations methods have been applied to compare different HTS developed for Gujarati language.

2. HMM-Based Speech Synthesis System

HTS can be divided into two parts, viz., training part and synthesis part as discussed below.

Training part: Spectrum and excitation parameters are first extracted from speech database. Mel Generalized Cepstral Coefficients (MGCs) (with $\gamma=0$ and $\alpha=0.42$ for 16 kHz speech) and their dynamic features are generally taken as spectrum (i.e., vocal tract system) parameters and log $F_0$ and it’s dynamic features are taken as excitation (i.e., speech source) parameters. Logarithm of $F_0$ contour is taken due to original work of Fujisaki for representing log $F_0$ in terms of phrase and accent components in speech source [14]. Then, these features are modeled by context-dependent HMMs. These HMMs are trained using Baum-Welch algorithm. Here, spectrum, excitation and duration are going to be modeled in a unified framework [15].

Synthesis part: For given sentence (which has to be synthesized), its corresponding utterance is converted to context-dependent phoneme sequence. Then, according to the phoneme sequence, utterance HMM is constructed by concatenating context-dependent HMMs. Then, state duration
of HMMs is determined. After this, using speech parameter generation algorithm, spectrum and excitation parameters are generated [16]. Finally, using mel log-spectrum approximation (MLSA) filter, speech waveform is generated [17]. Figure 1 shows block diagram of HTS.

3. Deterministic Annealing Expectation-Maximization (DAEM) Algorithm

3.1. Derivation of DAEM algorithm

EM algorithm is a statistical technique which iteratively computes Maximum Likelihood (ML) estimate of model parameters in the presence of hidden or incomplete data [18], [19]. Though EM algorithm has many advantages, it suffers from local maxima problem. In particular, many times EM algorithm will converge to local maxima and not to global maxima. In order to solve this problem, Deterministic Annealing Expectation-Maximization (DAEM) algorithm was proposed which uses the principle of maximum entropy and statistical mechanics analogy [11]. In addition, in DAEM algorithm, the problem of maximization of log-likelihood is redefined as the minimization of thermodynamic free energy.

Let set X = {x₁, x₂, ..., xₙ} consists of observable data xₙ and unobservable data xₙ, where xₙ = {x₁, x₂, ..., xₙ} and xₙ = {x₁, x₂, ..., xₙ}. Here, xₙ is unobservable corresponds to xₙ. x = (x₁, xₙ) = (x₁, xₙ), k = 1, ..., N. Thus, x is complete data and xₙ is incomplete data. Now, let us assume that we know joint probability density of x and xₙ, which is given as p(x₁, xₙ | Θ), where Θ represents model parameters of the density to be estimated. The EM algorithm maximizes the complete data log-likelihood function by iteratively maximizing the expectation of complete data log-likelihood function which is given by,

\[ L_c(x) = \log p(x₁, xₙ | Θ) \]  

The EM algorithm works in following two steps, viz.,

E-Step: 
\[ Q(Θ | Θ') = E[L_c(Θ | : x) | x₁ : Θ'] \]

\[ = \int [\log p(x₁, xₙ | Θ') \times p(xₙ | x₀ : Θ')] dxₙ , \]

where Θ' is the model parameters at iteration t.

M-Step: 
\[ Θ^{t+1} = \arg \max_Θ Q(Θ | Θ') \]

where \( p(xₙ | x₀ : Θ') \) is the posterior density given by,

\[ p(xₙ | x₀ : Θ') = \frac{p(x₁, xₙ | Θ')}{\int p(x₁, xₙ | Θ') dxₙ} . \]

Eq. (2) is unreliable at an early stage of iteration. Hence, DAEM algorithm uses principle of maximum entropy to obtain another posterior function \( f(xₙ | x₀) \) with constrains \( E_f[L_c(Θ | : x) | x₀] = \) constant and \( \int f dxₙ = 1 \). By maximizing following functions using Lagrange method,

\[ J[f] = S + β(E_f[L_c(Θ | : x) | x₀] - \text{constant}) + λ(\int f dxₙ - 1) \]

where \( S = -\int f(xₙ | x₀) \log f(xₙ | x₀) dxₙ \) and \( β \) and \( λ \) are Lagrange multipliers, we get [17],

\[ f(xₙ | x₀) = \frac{p(x₁, xₙ | Θ')}{\int p(x₁, xₙ | Θ') dxₙ} . \]

where \( β \) is annealing control parameter. If we can make analogy of this to the annealing process, then \( (1/β) \) will corresponds to temperature. Taking logarithm on both side of eq. (4), we have,

\[ \frac{-1}{β} \int p(x₁, xₙ | Θ') dxₙ = -\log p(x₁, xₙ | Θ') + \frac{1}{β} \log f(xₙ | x₀ : Θ') \]

Furthermore, taking conditional expectation with respect to the distribution \( f(·) \), we get,

\[ \frac{-1}{β} \int p(x₁, xₙ | Θ') dxₙ = U_{θ}(Θ) - \frac{1}{β} S_{θ}(Θ) \]

where \( U_{θ}(Θ) = E_f[\log p(x₁, xₙ | Θ') | x₀] \), \( S_{θ}(Θ) = E_f[-\log f(xₙ | x₀ : Θ') | x₀] \) and \( F_{θ}(Θ) = E_f[\frac{-1}{β} \int p(x₁, xₙ | Θ') dxₙ] \),

\[ = \frac{-1}{β} \int p(x₁, xₙ | Θ') dxₙ \]

Hence,

\[ F_{θ}(Θ) = U_{θ}(Θ) - \frac{1}{β} S_{θ}(Θ) \]  

Consider, from statistical mechanics, we have \( F = U - TS \), where \( F \) is free energy, \( T \) is temperature, \( U \) is internal energy and \( S \) entropy. Eq. (8) has an analogy with it. Thus, by adding \( β \)-loop which is also known as annealing loop to EM algorithm, we can derive DAEM algorithm [11].

1) Set \( β ← β_{\text{min}} (0 < β_{\text{min}} ≤ 1) \)
2) Set \( Θ₀ \) and \( t ← 0 \)
3) Perform EM-steps until convergence
4) Set \( t ← t + 1 \)
5) Increase \( β \) (\( β = β \times 1.1 \))
6) If \( β < 1 \), repeat procedure from step-3 else stop.

3.2. Comparison in the context of mixture density estimation

Consider a one-dimensional (i.e., 1-D), two-component normal mixture distribution given by,

\[ p(x | Θ) = \frac{0.4}{\sqrt{2π}} \exp[-\frac{1}{2}(x-m₁)] + \frac{0.6}{\sqrt{2π}} \exp[-\frac{1}{2}(x-m₂)] . \]

100 sample points were generated from this mixture with \( m₁ = -2 \) and \( m₂ = 4 \). At the moment, we will forget \( m₁ \) and \( m₂ \), and assume that we have only 100 observation samples and we would like to estimate \( Θ = (m₁, m₂) \) using EM and DAEM algorithms. We know the observation points, however, we do not know as to which Gaussian pdf they have
come from (i.e., whether \( x^n \) is 1 or 2 is not known). Hence, it is incomplete data case. Free energy is defined as [11],

\[
F_p(\Theta) = \frac{1}{\beta} \log \sum_{x^n} p(x^n, x^n_0 : \Theta)^eta, \\
= \frac{1}{\beta} \sum_{k=1}^N \sum_{x^n} \frac{1}{\beta} \prod_{t=1}^T p(x_t, x^n_t : \Theta)^eta, \\
F_p(\Theta) = \frac{1}{\beta} \sum_{k=1}^N \log \sum_{x^n} p(x^n, x^n_0 : \Theta)^eta, \\
(10)
\]

where

\[
p(x^n_0, x^n_t : \Theta) = \begin{cases} 
\frac{0.4}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x - m_t)^2 \right\}, & \text{for } x^n = 1 \\
\frac{0.6}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(x - m_t)^2 \right\}, & \text{for } x^n = 2 
\end{cases}.
(11)
\]

Estimation for new mean at \( t+1 \) state is calculated from following formula which is derived by differentiating eq. (10) w.r.t. mean,

\[
m_{k+1} = \frac{\sum_{x^n} x^n P(w_t | x^n : \Theta)}{\sum_{x^n} P(w_t | x^n : \Theta)}, \\
(12)
\]

where \( P(w_t | x^n : \Theta) = \frac{p(x^n_0, x^n_t = i : \Theta)^eta}{\sum_{i=1}^K p(x^n_0, x^n_t = i : \Theta)^eta}. \\
(13)
\]

Negative value of free energy, i.e., log-likelihood of generated samples for different values of \((m_1, m_2)\) is shown in figure 2. In order to see a clear peak, upper portion of the figure 2 is zoomed by clipping the value less than (-300) of log-likelihood value. Thus, we can see two clear peaks in figure 3, viz., one is global maxima, at \((m_1, m_2) = (-2, 4)\) and local maxima at \((m_1, m_2) = (4, -2)\).

![Figure 2: Log-likelihood vs. \( \Theta = (m_1, m_2) \) with clipping.](image)

![Figure 3: Loglikelihod vs. \( \Theta = (m_1, m_2) \) with clipping.](image)

We have applied both EM and DAEM algorithms to the above problem and with different initial values which is different from experiment done in [11]. We got either global or local maxima for given initial value. Out of several initial values, results for some values are shown in Table 1. From this table, it is evident that DAEM is able to reach at global maxima as final value. However, EM algorithm stuck at local maxima for some cases. For some initial value (as shown in Table 1), both EM and DAEM algorithms are reaching at global maxima and similarly, for some initial values (as shown in Table 1) both EM and DAEM algorithms are reaching at local maxima. However, there is not a single case where EM algorithm is reaching at global maxima and DAEM algorithm is reaching at local maxima.

### Table 1: Simulation results of EM and DAEM algorithm applied to mixture density estimation problem

<table>
<thead>
<tr>
<th>No.</th>
<th>Initial value</th>
<th>Final value using EM algorithm</th>
<th>Final value using DAEM algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(-2,-2)</td>
<td>(3.8263,-1.8539)</td>
<td>(-1.8917,3.8070)</td>
</tr>
<tr>
<td>2</td>
<td>(6,6)</td>
<td>(3.8263,-1.8539)</td>
<td>(-1.8917,3.8070)</td>
</tr>
<tr>
<td>3</td>
<td>(0,0)</td>
<td>(3.8263,-1.8539)</td>
<td>(-1.8917,3.8070)</td>
</tr>
<tr>
<td>4</td>
<td>(-6,-4)</td>
<td>(-1.8917,3.8070)</td>
<td>(-1.8917,3.8070)</td>
</tr>
<tr>
<td>5</td>
<td>(1,-2)</td>
<td>(3.8263,-1.8539)</td>
<td>(3.8263,-1.8539)</td>
</tr>
</tbody>
</table>

Hence, from our observation given in Table 1, it is evident that DAEM algorithm is performing on an average better than EM algorithm. However, it is not guaranteed that it will come out of local maxima problem all the time.

### 3.3. Application of DAEM to Baum-Welch Algorithm

Baum-Welch algorithm is also an EM algorithm which is used as re-estimation algorithm in HMM. Here, observable data, corresponds to \( x_n \), is observation vector O and unobservable data, corresponds to \( x_n \), is state sequence q. The auxiliary function so called Q-function is given by [12], [20]

\[
Q(\Theta, \Theta') = \int \log p(O_q | \Theta) \times p(q | O_q, \Theta') dx_q. \\
(14)
\]

To apply, DAEM algorithm, we will consider \( U_{\beta}(\Theta, \Theta') \) which is negative of Q-function and \( p(q | O_q, \Theta') \) will be replaced by \( f(q | O_q, \Theta') \) and is given by [11],

\[
f(q | O_q, \Theta') = \frac{p(O_q | \Theta')^{\beta}}{\int p(O_q | \Theta')^{\beta} dq}. \\
(15)
\]

### 3.4. Adaptation of HTS to Gujarati

Due to unavailability of Gujarati standard corpus for speech synthesis applications, our first task is to develop speech database for Gujarati language. We have collected optimized sentences have been recorded by one female and one male (professional voice over artist) from professional studio [21]. Except, context-dependent modeling, every block of HTS is language-independent. However, contextual information is language-dependent. In this context, the HMM framework provides a general setup for sufficient context modeling. That can easily be adopted to other languages. For the phonemic representation of Gujarati language, a set of 49 phonemes that are broadly classified into silence (i.e., SIL), 33 consonants and 13 vowels, were taken. In context-dependent modeling, we require different groups of phonemes. For Gujarati, to do classification, we have used IPA chart of Gujarati for consonants and for vowels from [22].

Classifications of these phonemes were used for question set preparation.
4. Evaluation of HTS for Gujarati

4.1. Experimental Setup
In order to build HTS for Gujarati, we have taken total 3 hours of phoneme-level labeled speech database from a male and a female speaker. Phoneme-level speech database has been created using automatic forced alignment technique [23]. Different HTS systems have been built for Gujarati language. We have taken total 105-dimensional Mel generalized cepstral coefficients, 3-dimensional log F0 and penta-phone contextual factors per frame with frame period 10 ms and with Hamming window size of 20 ms durations. Since HTS does not store speech sound units rather it stores statistical parameters, size of HTS developed for Gujarati is within range of 3-3.77 MB.

4.2. Evaluations of several HTS

1) Subjective Evaluations
DMOS: Five synthesized utterances from each HTS system was taken and degraded MOS (i.e., DMOS) analysis have been done [24]. DMOS is used to analyze speech quality w.r.t. original natural speech, i.e., how much quality of TTS value is degraded w.r.t. quality of natural speech. Files from different HTS systems are played randomly and subjects were asked to give score in 1 to 5. Headphones were used by listener to listen the files. Then DMOS are calculated for each system using 18 native subjects.

WER(%): Intelligibility test was performed in which they have to transcribe synthesized utterances and based on this word error rate (WER) (%) was calculated by following formula [25],
\[ WER(\%) = \left( \frac{I + D + S}{T} \right) \times 100, \]  
where I is the total number of insertions, D is total number of deletions and S is the total number of substitutions that one needs to do in order to make transcription 100 % correct at word-level and T is the total number of words. From Table 2, it is clear that compared to EM-based HTS, DAEM-based HTS performs well in terms of MOS and WER for same hours of training data.

Table 2: DMOS and WER analysis of different HTS developed in Gujarati language.

<table>
<thead>
<tr>
<th>Voice over Artist</th>
<th>Duration (hrs)</th>
<th>Algorithm</th>
<th>EM</th>
<th>DAEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DMOS</td>
<td>WER (%)</td>
<td>DMOS</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>1.35</td>
<td>19.31</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.25</td>
<td>18.56</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.08</td>
<td>11.12</td>
<td>3.02</td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>2.39</td>
<td>10.05</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2.5</td>
<td>8.56</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.39</td>
<td>9.26</td>
<td>2.72</td>
</tr>
</tbody>
</table>

We can see that DAEM-based HTS on female 2 hours speech database works better than EM-based HTS on 3 hours training database. In addition, note that as amount of training data increases, the system performance also increases.

AB-Test: In this method, we played same utterances which are synthesized from two different approaches (here EM-based and DAEM-based HTS). 10 subjects have to decide which is better and accordingly 1 score is given to that approach. If subject is unable to make decision then 0.5 score is given to both approaches.

Table 3: Result of AB-Test

<table>
<thead>
<tr>
<th>Preference (%)</th>
<th>EM-based HTS</th>
<th>DAEM-based HTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.50</td>
<td>70.50</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 shows the result of AB-test. It is clear that 70.5 % times DAEM-based HTS has preferred over EM-based.

2) Objective Evaluation
Mel Cepstral Distance (MCD) measure is used for measuring the accuracy of the spectral envelope of the synthetic speech with respect to natural speech. The computational details of MCD are given in [26].

Table 4: MCD analysis of various HTS.

<table>
<thead>
<tr>
<th>Voice over Artist</th>
<th>Duration (hrs)</th>
<th>MCD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EM</td>
</tr>
<tr>
<td>Female</td>
<td>1</td>
<td>3.05</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2.76</td>
</tr>
<tr>
<td>Male</td>
<td>1</td>
<td>3.33</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.20</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3.14</td>
</tr>
</tbody>
</table>

From Table 4, we can see that for male voice, DAEM-based HTS has less MCD score than EM-based HTS. For female voice HTS, 1 hour and 2 hours of DAEM-based HTS has less MCD values than EM-based HTS. While for 3 hours HTS, DAEM-based HTS has slightly higher score of MCD. For female HTS also, on an average, DAEM-based HTS performs better than EM-based HTS system.

5. Summary and Conclusions
In this paper, adaption to statistical framework of HTS was done to develop HTS-based TTS voice in Gujarati. Different HTS systems for Gujarati language have been developed with male and female speech data. Our system has high speech intelligibility. It is found that as amount of training data increases during HTS system building, system performance is also found to increase in terms of both speech naturalness and speech intelligibility. HTS with two different HMM training approaches, viz., EM-based HTS and DAEM-based HTS have been built. From AB-test, it is found that subjects prefer DAEM-based HTS voice more compared to EM-based HTS voice. DAEM-based HTS requires more training time than EM-based HTS. However, synthesis time is same for both the approaches.

6. Acknowledgements
The authors would like to thank Department of Electronics and Information Technology (DeitY), Govt. of India for sponsoring the consortium project, viz., Development of Text-to-Speech Synthesis Systems in Indian Languages. They also thank the consortium leader Prof. Hema A. Murthy (IIT Madras) and the authorities of DA-IICT to support this research work. In addition, they thank all the participants who took part in subjective evaluation of HTS systems.
7. References


