Boltzmann–Dirichlet Process Mixture: A Mathematical Model for Speech Recognition

Rajesh Kumar T1, Vijendra Babu D2, Malarvezhi P3, Velu C M4, Haritha D5, and Karthikeyan C6

1,5,6Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India
2Department of Electronics and Communication Engineering, Aarupadai Veedu Institute of Technology, Vinayaka Mission’s Research Foundation, Chennai, Tamil Nadu, India
3Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India
4Department of Computer Science and Engineering, Saveetha School of Engineering, Chennai, Tamil Nadu, India
Email: *t.rajesh61074@gmail.com

Abstract. This article deliberates a mathematical model for the estimation of speech signals probability density function. Speech recognition is analyzed using an integration of Boltzmann equations with Dirichlet Process Mixture sequences. Usually, environmental noise, white noise, echo noise interferes with the speech signal. So, the speech identification rate decreases abruptly. By estimating the noise sequences in the speech signal, the speech identification rate increases. Rather than using a conventional Gaussian Mixture Model (GMM) procedure to recognize a pure speech, an integration of mathematical equations of Boltzmann and Dirichlet Process Mixture is proposed in this article. An uttered speech signal is identified using mean, variance, and standard deviation generated by Boltzmann-DPM. For an added white, particle, shaver percentage of noises, the speech signal to noise ratio is improved and proved experimentally using the Nil filter, GMM filters, and Extended Kalman filter.

Keywords: Speech Signal Features, Boltzmann Equations, Dirichlet Process Mixture, Mathematical Model, Kalman Filter, Speech Recognition.

1. Introduction
Speech signal identification plays a major role in this modern era, as mobile phone manufacturing and usage is incredible among human lives. Researchers in the scientific and engineering fields mostly relate their work with mathematical models as a product for signal processing such as image and speech using Boltzmann equations. To extract the significant coefficients to recognize the speech signal, the Dirichlet process mixture is integrated with Boltzmann equations, as it supports generating the coefficients technically. The probability density function estimation is deliberated in the tutorial and journal papers [1], using the Dirichlet process mixture [2]. The utilization of Boltzmann equations for aggregation in neural networks, also the usage of Deep Boltzmann machine is discussed in the manuscripts [3]. The functions of the probability density for the preparation of the speech papers [4], using (DPM) Dirichlet Process Mixture [5] without using GMM model is discussed in the journal papers [6]. Bayesian method of estimation of PDF [7] in speech is deliberated in papers [8]. The estimation of
pfd, without the estimation of noises in some enhanced algorithm and utilization of speech dataset in AURORA II and TIMIT, are deliberated [9]. The whispered voice and non-audible murmur identification using [10] inverse filters are discussed in [11] manuscripts where the TIMIT database is used as a benchmark [12].

The emotion of speech using regional languages [13], with the help of MFCC integrated Hidden Markov Model (HMM), are presented in the manuscript [14]. In this manuscript, a mathematical model is proposed to identify and estimate pfd by integrating Boltzmann equations with Dirichlet process mixture using the TIMIT corpus datasets [15]. This manuscript is structured as follows: after the introduction and review are deliberated in section 1, the projected methodology is described in section 2, section 3 evaluates the results on speech identification, and finally summarized in section 4.

2. Proposed Methodology

The speech signal is recognized using a proposed integration of the Boltzmann machine model with the Dirichlet Process Mixture. The traditional method of recognition of speech using the Gaussian Mixture Model is exempted. This technique follows Markov and Gaussian's procedure because along with features, noise is added, and a particle filter is applied for noise estimation. The mixing of BMM & DPM is illustrated in Figure 1, where the extracted features like spectral skewness, pitch chroma, and spectral centroid are induced with white noise or particle noise for our speech recognition analysis.

![Block diagram of the proposed methodology](image)

**Figure 1.** Block diagram of the proposed methodology

2.1 Extracted features from the input speech signal

*Spectral skewness:* "It specifies the coefficient of the skewness of the spectrum also it is the measure of the symmetry of the distribution." Typically, the skewness coefficient can be defined as the “ratio of skewness to the third power of standard deviation,” besides, the skewness coefficient is given in Eq. (1).

\[
SS = \sum_{i=1}^{g} \left( \nu - \mu \right)^3 \times \omega \sigma^{-3}
\] (1)

Whereas \( \omega \) denotes spectrum width of an input speech signal, \( \nu \) denotes individual spectrum, the standard deviation is \( \sigma \) \( \mu \) mean and \( SS \) indicates input signals skewness.

*Pitch Chroma:* It is mathematically defined for the input speech signal \( Y \) as per Eq. (2).

\[
C_f(\nu) = \sum_{i=0}^{g-1} S_{bf} (\nu + i \alpha)
\] (2)

Whereas \( \nu \) indicates chroma index or pitch class of integer value as \( \nu \in [0, g - 1] \), \( i \) provides an octave index in an integer value and \( i \in [0, g - 1] \), \( g \) indicates the count of an octave, \( S_{bf} \) specifies log-frequency spectrum and \( \alpha \) is some bins per octave. Each pitch is divided into two components: the
peak of tone, like a collection of pitches of 12 count, is called an octave. The information related to pitch with one coefficient is defined as chroma.

**Spectral Centroid:** It indicates the energy with a central frequency of the spectral centers is high, and it is formulated based on the spectral magnitude of the frequency bin in the input signal also the frequency at the specific bin”. It is mathematically defined in Eq. (3).

\[
SC(\kappa) = \frac{\sum_{\kappa} \kappa \cdot \mathcal{R}(\kappa)}{\sum_{\kappa} \mathcal{R}(\kappa)}
\]

Whereas, the overall count of the bins is \(\tau\). The center of gravity of \(Y\) (input signal) is the spectral centroid. Further, with the aid of the center of frequency \(\mathcal{R}(\kappa)\)’s value, the value of \(Y\) (input signal) is normalized. From the input signal of speech, the features like spectral skewness, pitch chroma, and spectral centroid are extracted as \(y_1, y_2, y_3\) and vectored as \(Y\)

\[
Y = \{y_1, y_2, y_3\}
\]

Then windowing and framing of the signal are done. To discriminate the amplitude and phase, a discrete Fourier transform is applied. The inverse discrete Fourier transformation uses a logarithmic value of amplitude.

### 2.2 Noise added in pure speech signal mathematically

When a noise gets added with a pure speech signal, the Gaussian probability distribution function is elaborated and adds noise signal \([7]\). Before, the geometric model is

\[
prob(x) = \sum_{i=1}^{N} P(v_k)N(x, \mu_{x,k}, \sigma_{x,k})
\]

Whereas, \(prob(x)\) is probability function, \(N(x, \mu, \sigma)\) - Gaussian distribution. Let the input pure speech signal be \(Y_T\), \(n_T\) be the additive noise signal, auxiliary constants are \(\nu_T, w_T\) appended to the signal, then output \(Z_T\) is:

\[
Z_T = Y_T + \log(1 + \exp(n_T - Y_T)) + \nu_T
\]

\[
n_T = n_{T-1} + w_{T-1}
\]

\[
\nu_T \equiv N(0, \sigma_y) \& w_T \equiv N(0, \sigma_w)
\]

### 2.3 Dirichlet Processes Mixture for PDF Estimation

Let probability density function be a sound sequence of statistical function, then \(y_1,y_2,...,y_n\) of different features of speech signal estimates the \(\mathcal{R}(y)\) as:

\[
\mathcal{R}(y) = \int_{\theta} k(y \mid \theta) dE(\theta)
\]

Whereas \(E(\theta)\) and \(k(y \mid \theta)\) are Bayesian frameworks with random pdf measure\([7]\) and a sound variable is \(\theta\). Let the framework of speech be ‘S’ as \(S_1, S_2,...,S_i\) with a probability of \(E_0\) on space and ‘\(\alpha\)’ be a positive number in integer then the probability distribution function ‘E’ for a Dirichlet process is \(E = DP(E_0, \alpha)\), whereas ‘D’ indicates a distribution of Dirichlet in standard mode.

\[
(E(S_1), E(S_2),...,E(S_i)) \equiv D(E_0(S_1), E_0(S_2),...,E_0(S_i), \alpha)
\]

Using Bayesian framework, the polynomial distribution function of Dirichlet process over added noise in pure speech is presented in 10th equation, which is integrated with 11th equation, where delta function is \(\delta_{\theta_i}\):
\[ \theta_{t+1} | \theta_t \approx \frac{1}{t + \alpha} \sum_{k=1}^{t} \delta_{\theta_k} + \frac{\alpha}{t + \alpha} E_0 \]  

(11)

Now the random probability quantifier with Dirichlet process mixture and input signal with pure speech is mathematically estimated as the function with density:

\[ E \equiv DP(E_0, \alpha); \theta_k \equiv E; Y_T \equiv f(s | \theta_k) \]  

(12)

Then the probability distribution function is:

\[ N(s) = \int f(s | \theta) dE(\theta) \text{ with } \theta \in \Theta. \]  

(13)

2.4 Boltzmann machine Learning

A network of Boltzmann machine shown in Figure 2, having binary units with harmonious neural networks from visible to visible layer and interlinked hidden layer, is laterally connected.

\[ E(x, \Phi, y) = \sum_{i,j} W_{ji} \Phi_i \Phi_j - \sum_{i} \sum_{m} J_{mj} \Phi_m \Phi_i - \sum_{j} y_j \sum_{k} b_k \Phi_k - \sum_{i} \sum_{j} a_i \Phi_i \]  

(14)

Where, as \( v \) and \( j \) are intermittent hidden variables. The hidden and visible unit’s probability distribution functions are given by:

\[ p(o|\Phi_j = 1 | v, \Phi_{-j}) = \sigma(\sum_j W_{ji} \nu_i + \sum_{m} J_{mj} \Phi_m - a_i) \]  

(15)

\[ p(o|\nu_i = 1 | \Phi, \nu_{-i}) = \sigma(\sum_j W_{ji} \Phi_i + \sum_{k} L_{ki} \nu_k - b_i) \]  

(16)

The log-likelihood derivative concerning the model parameters \( w \) and \( L \) are given by

\[ \frac{\delta \log p(o; \theta)}{\delta j} = E_{P_{Boltzmann}}[\Phi \Phi^T] + E_{DPM}[\Phi \Phi^T] \]  

(17)
A separate Markov chain runs to approximate the training data vector in the sense every feature vector leads to positive phase \(E_{\text{Boltzmann}}[\phi^+]\) raised to ”the states of the visible units.” In addition to this, in the "negative phase, an additional Dirichlet Processor mixer runs to approximate the \(E_{\text{DPM}}[\phi^-].\) It is a computationally demanding learning technique with appropriate iterations. To get stationary distribution, which is said to be learning, it is a time taking process by Boltzmann machine integrated with DPM.

2.5 Proposed Algorithm for Boltzmann Machine Learning-DPM

Step 1: Initialize randomly the \(\theta_0\) and the DPM vector samples \(\{\tilde{N}(s)^{0,1}, \tilde{G}^{0,1}\}, \ldots, \{\tilde{N}(s)^{0,M}, \tilde{G}^{0,M}\}\) based on the samples of the features \(y=y_1,y_2,y_3\) are selected.

Step 2: For \(i=0\) to \(T\) // Number of Iterations

- In “Positive Phase”
  - For each training samples \(v_n, n=1\) to \(N\)
    - \(\mu\) is randomly initialized and executes the mean-field updates until it converges:
      \[
      \mu_j \leftarrow \sum_{i} W_{ji} v_j^+ + \sum_i \Phi_{mi} \mu_m - a_j
      \]  \hspace{1cm}(18)
  - Set \(\mu^0=\mu\)
  - In “Negative Phase”
    - For each negative samples from DPM \(m=1\) to \(M\)
      - “By running DPM and its samples, a novel binary state \(\tilde{N}(s)^{1,1:m}, \tilde{G}^{1,1:m}\) is obtained,” which was initialized as \(\tilde{N}(s)^{1,1:m}, \tilde{G}^{1,1:m}\) visible, hidden from equations 13 and 17.

Step 3: Update the Parameters of weights, Interlinks of hidden and visible layers, \(W^{1,1}, \Phi^{1,1}, L^{1,1}\) for every iteration.

Step 4: At last \(\alpha_t\) value will be decreased in weights, visible and hidden layer. So convergence will be very fast. It facilitates learning quickly.

2.6 Speechless/speech frames identification using mathematical Model of Boltzmann Machine Learning-DPM

The above algorithm made of Boltzmann integration of Dirichlet leads to earning a mathematical model to identify speech signal from noise. The original learned copy distinguishes the most extreme (SNR) Signal to Noise ratio to separate the speech full and speechless portion in the frame of speech. The separation of speech indicates recognizable proof of signal with speech. The speech with Noise portion may be distributed as \(D_{n} D_{n} = (y_j - \hat{n}_j)^2\), and the Speech portion of the input signal's full-frame \(D_{s} = (y_j - (\hat{s}_j + \log(1 + \exp(\hat{n}_j - \hat{s}_j)))^2\) is given as \(D_{si}\), and the differences lead to a delta function \(\Delta = D_{si} - D_{n}\). Whereas \(y_j\) refers to the signal's extracted features, \(\hat{s}_j\) refers to clear speech, and \(\hat{n}_j\) refers to noise estimated. Finally, the modified Boltzmann-Dirichlet coefficient constant:

\[
\hat{n}_j = \hat{n}_j - \xi_n \sqrt{D_{n}}; \hat{s}_j = \hat{s}_j + \xi_s \sqrt{D_{s}};
\]  \hspace{1cm}(19)

The recognition of speech signal is performed using the coefficients obtained in \(\xi_n\) and \(\xi_s\) based on the Boltzmann machine learning and Dirichlet Process Mixture hybrid.

3. Experimental Results and Discussions

A worldwide language, English uttered words of 680 count, lexically and phonemically created corpus is available in TIMIT dataset are used as a corpus for analysis. The uttered speech from all speaker's transcription words is of 16-bit speech waveform with 16 kHz. An SN Ratio of range from
0 decibels to 8 decibels is synthetically created white noise, particle noise, and shaver noise added in the pure speech signal presented in the experimental setup. The word dataset used for the analysis is shown in Figure 3.

1 female utters 103 words, and 2 males are the English language with three different kinds of a human named as David, Hazel, and Zira. These male and female voices are generated in the word file named dot wave file (*.wav) used for speech recognition. The words such as cursor movement in computer systems of 'up,' 'down,' 'left,' 'right' is used.

Table 1. Comparison of filters for added white, particle, and shave Noise in Speech Recognition

| SNR in decibels | White | Particle | Shaver |
|-----------------|-------|----------|--------|
|                 | No Filter | GMM | EKF | No Filter | GMM | EKF | No Filter | GMM | EKF |
| 0db             | 3.4 | 2.8 | 21.5 | 5.2 | 11.3 | 27.5 | 8.1 | 8.2 | 25.3 |
| 2db             | 15.2 | 9.1 | 44.2 | 29.4 | 17.2 | 58.2 | 32.5 | 18 | 56.2 |
| 4db             | 45.7 | 31.2 | 69.4 | 67.2 | 36.6 | 70.1 | 64.5 | 37.7 | 68.7 |
| 8db             | 68.2 | 48.2 | 72.6 | 82.2 | 55.1 | 85.3 | 61.3 | 52.2 | 79.7 |

This paper uses NIL filter, Conventional GMM filter, and Extended Kalman filter for comparison purposes. The speech recognition rate in percentage based on adding the white, particle, shaver noises, and the filters’ filters using no filters, conventional Gaussian mixture model filters, and extended Kalman filters and tabulated in Table 1. The input speech signals are added with three different noises for the designed mathematical model based on Boltzmann machine learning integrated with Dirichlet Process Mixture. Two different filters are implemented and checked with speech decibels at the rate of 0db, 2db, 4db, and 8db are verified, and it is stipulated in the speech recognition rate are shown in figure 4 as bar chart.
Regarding the above tabular column Table 1, the Figure 4 bar chart has been charted. The white, particle and shaver noises are added and filtered using GMM filter and Extended Kalman filter. When we apply an SNR value of 8db of speech voice as input, the EKF filtering percentage is 72.6 in added white noise, 85.3 in added particle noise, and 79.7 in added shaver noise. With no filter and GMM filter, the maximum percentage obtained for 8db are 68.2, 48.2 in added white noise, 82.2, 55.1 in added particle noise, and 61.3, 52.2 in added shaver noise. Among these, the percentage of filtering using the Extended Kalman filter leads to a maximum recognition of input speech signal. When we compare all the added noise and applied filters, the proposed Boltzmann-DPM mathematical model proves that this Kalman filter provides maximum recognition rate for the given speech English words.

4. Conclusion
A mathematical model is designed and verified for the recognition of speech signal using an integration of Boltzmann machine learning with Dirichlet Processing Mixture by estimating probability density function, in which pure speech signal is added with white, particle, and shaver noises and filtered using traditional GMM filter, extended Kalman filter are compared in this manuscript. Features of speech such as spectral skewness, pitch chroma, spectral centroid are extracted and mixed with the Boltzmann machine model and integrated with DPM. So, this method estimates the noise accurately. The experimental analysis of speech recognition by evaluating the SNR’s results and comparison proves that the Kalman filter provides the improved output.

References
[1]. John Paisley, "A Tutorial on the Dirichlet Process for Engineers Technical Report," Department of Electrical & Computer Engineering Duke University, Durham, NC, pp:1-23, 2018
[2]. Yuelin Li, Elizabeth Schofield, Mithat Gonen, "A tutorial on Dirichlet process mixture is modeling," Journal of Mathematical Psychology, Elsevier, vol.91, 2019.

[3]. Ruslan Salakhutdinov, "Learning and Evaluating Boltzmann Machine," University of Toronto, Canada, 2008, http://learning.cs.toronto.edu

[4]. Ruslan Salakhutdinov, Hugo Larochelle, "Efficient Learning of Deep Boltzmann Machines," Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, Italy, Vol 9, Issue W&CP 9, 2010.

[5]. Michael D. Escobar and Mike West, “Bayesian Density Estimation and Inference using Mixtures,” Journal of the American Statistical Association, Vol. 90, No. 430, 1995.

[6]. Kenko Ota, Emmanuel Duflos, Philippe Vanheeghe, and Masuzo Yanagida, "Speech Recognition and Speech Density Estimation by the Dirichlet Process Mixture," Proceedings of ICASSP 2008, IEEE Xplorer, pp. 1553-1556, 2008.

[7]. Kottas A., “Dirichlet Process Mixtures of Beta Distributions, with Applications to Density and Intensity Estimation.”, Proceedings on Learning with Nonparametric Bayesian Methods, 23rd ICML, 2006.

[8]. Francois Caron Manuel Davy, Arnaud Doucet, Emmanuel Duflos, Philippe Vanheeghe, "Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures." Proc. of Fusion'06, Florence, Italia, July 10-13, 2006.

[9]. J. C. Segura, A. de la Torre, M. C. Benitez, A. M. Peinado, "Model-Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks," Proc. EuroSpeech '01, Vol. I, pp. 221-224, Sept. 2001.

[10]. TIMITacoustic-phonetic continuous speech corpus; 2018. https://catalog.ldc.upenn.edu/ldc93s1.

[11]. Rajesh Kumar T, Suresh GR, Kanaga Subaraja S, Karthikeyan C., "Taylor-AMS features and deep convolutional neural network for converting non-audible murmur to normal speech," Computational Intelligence, Wiley Publisher, Pp.1–24. 2020.

[12]. Kurada, P., Maruvada, S., & Sanagapalleka, K. R., “Speech bandwidth extension using transform-domain data hiding” International Journal of Speech Technology, 2019

[13]. Shariff, M. N., Saisambasivarao, B., Vishvak, T., & Rajesh Kumar T., "Biometric user identity verification using speech recognition based on ANN/HMM." Journal of Advanced Research in Dynamical and Control Systems, 9(12 Special issues), pp. 1739-1748. 2017.

[14]. T. Rajesh Kumar, G. R. Suresh, S. Kanaga Suba Raja, “Conversion of Non-Audible Murmur to Normal Speech based on Full-rank Gaussian Mixture Model," Journal of Computational and Theoretical Nanoscience, 1546-1955, Iss.1,vol-15, Pp:185-190, 2018.

[15]. Bommadevara, H. S. A., Sowmya, Y., Pradeepini, G., "Heart disease prediction using machine learning algorithms," International Journal of Innovative Technology and Exploring Engineering, 8(5), 270-272, 2019.