JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Segmentation And Identification Of Voices Using Speaker Diarization

[1]D. Indu, [2]Y. Srinivas

[1]Research Scholar, [2]Professor
[1,2] Department of Computer Science and Engineering
GITAM School of Technology GITAM (Deemed to be University)
Visakhapatnam, Andhra Pradesh, India.

Abstract : In essence, speaker diarization provides a solution to the issue of "who spoke when" by labelling audio or video recordings with the speaker's identity. These algorithms, which were first created for voice recognition in multispeaker recordings, were later modified for speaker-specific metainformation, improving uses such as audio retrieval. Significant progress has been made in speaker diarization since deep learning was introduced. This study examines the development of speaker diarization historically as well as contemporary advancements in neural techniques. It also looks at how deep learning combines voice recognition and speaker diarization, showing how these two functions work best together. By combining neural techniques with modern advancements, this survey seeks to benefit the community by promoting speaker diarization efficiency.

IndexTerms **- Speaker diarization, speaker segmentation and clustering.**

I. INTRODUCTION

In the fields of audio signal processing and speech analysis, speaker diarization is an essential undertaking. It entails breaking up an audio file into homogenous chunks, each of which represents a single speaker, and then labelling or identifying these chunks in relation to the speakers. The performance and capabilities of speaker diarization systems have been greatly improved by recent developments in deep learning.

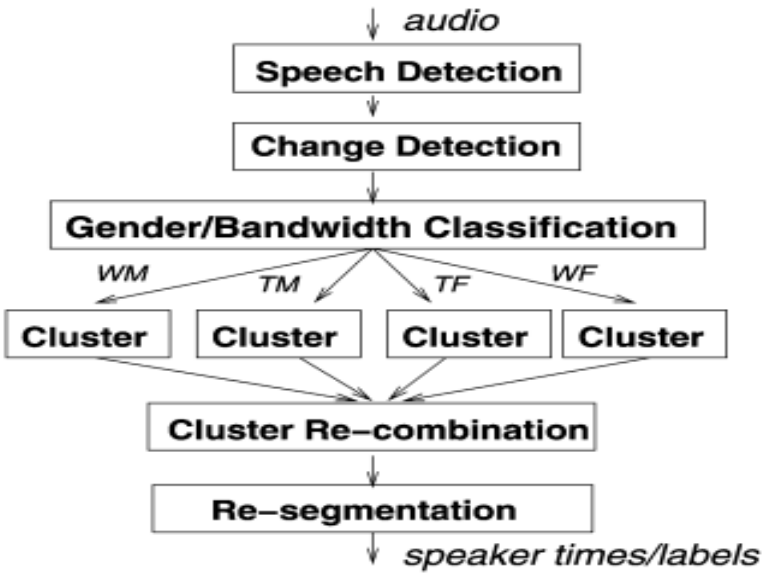

Fig1: Exemplar Diarization System

Speaker diarization systems are typically made up of several separate sub-modules. Several front-end processing techniques, such as speech augmentation, dereverberation, speech separation, or target speaker extraction, are used to reduce any artefacts in acoustic environments. The next step is to apply voice or speech activity detection (SAD) to distinguish speech from non-speech events. Within the chosen speech segment, the raw speech signals are converted into embedding vectors or acoustic characteristics. The altered speech segments are categorised and labelled according to speaker classes during the clustering stage, and the clustering outcomes are further improved during the post-processing stage. In general, each of these sub-modules is optimised separately.

The goal of the research during the early 1990s diarization era was to improve automatic speech recognition (ASR) on broadcast news recordings and air traffic control dialogues by allowing speaker-adaptive training of acoustic models and separating each speaker's speech segments (Gish et al., 1991, Siu et al., 1992, Rohlicek et al., 1992, Jain et al., 1996, Padmanabhan et al., 1996, Gauvain et al., 1998, Liu and Kubala, 1999). During this time, some basic methods for determining the separation between

speech segments for speaker change detection and clustering were developed and quickly became the industry standard. These methods include the generalised likelihood ratio (GLR) (Gish et al., 1991) and the Bayesian information criterion (BIC) (Chen and Gopalakrishnan, 1998).

When taken as a whole, these initiatives paved the way for the global integration of research groups' efforts, resulting in the formation of numerous research consortia and challenges in the early 2000s. These included the Augmented Multiparty Interaction (AMI, 2009) Consortium, financed by the European Commission, and the RT Evaluation (NIST, 2009) hosted by the National Institute of Standards and Technology (NIST). Over a few years, these organisations (Ajmera and Wooters, 2003, Reynolds and Torres-Carrasquillo, 2005, Zhu et al., 2005, Meignier et al., 2006) sponsored further developments in speaker diarization technologies across various data domains from broadcast news.

II. Related Work

Speaker diarization, a crucial aspect of audio processing, has garnered substantial attention in Agglutinative languages, including Finnish, Kazakh, and Turkish, have in common that up till now, efforts to recognise speech and conduct human identification have not produced results that are equivalent to English systems [4]. This is caused by both the difficulty of modelling the language and the dearth of materials that are appropriate for text and speech acquisition. The goal of the systems in [5, 6] is to concentrate and cluster in order to minimise the amount of the active vocabulary and language models.

Neural networks have taken the lead in several machine learning domains recently. One of them is natural language processing, where the words or letter sequences may be thought of as another sort of signal. Two common classifiers used in this process are the multilayer perceptron (MLP) and the long-short-term memory (LSTM) network.

It has been demonstrated in several studies [7, 8] that using ANNs in addition to HMMs may increase voice recognition accuracy. Deep neural networks, artificial neural networks with direct propagation with several hidden layers between the input and output layers, provide the foundation for most acoustic models. The backpropagation approach is applied for training.According to the review article, one of the most crucial jobs in the identification system is feature extraction, which has a big impact on the system's functionality and efficiency..

The techniques for recognising the aspects of the identification system that have already been proposed and implemented were taken into consideration in the review analysis. The results of the analysis indicate that MFCC-based approaches have been used more than any other approach. Moreover, it was found that the current direction of identification system research is to address significant identification system issues, including noise resistance, complexity, adaptability, and multilingual recognition. [9].

In one of the works [10], speech pre-processing method was considered using the VAD algorithm, which proves that this algorithm improves the performance of speech recognition. The study presents the principles of operation and the block diagram of the VAD algorithm in recognition of Kazakh speech.

Character-based MLP and LSTM models that can jointly recognise the border of sentences and tokens were proposed by Toleu et al. [11]. The suggested models project the character embedding into low-dimensional space in order to extract the high-level abstract characteristics, which may enable us to assess the various types of signals. Three languages were used to test the models: Italian, Kazakh, and English.

In comparison to current models, the experimental findings demonstrate that character-based MLP and LSTM models for sentence and token segmentation have favourable impacts in terms of F-measure and error rates.

To divide an existing collection of voice segments into groups based on how similar their features are, clustering techniques are utilised. In [12], parametric techniques are described for identifying beginning points (centroids) and subsequent cluster propagation, which may be used to solve voice classification tasks.

The use of a neck microphone, or laryngophone, as an extra modality for phonetic segmentation of the speech signal into acoustic sub-word units is examined in the study [13]. A novel approach is suggested for the segmentation of speech signals automatically, utilising the throat-acoustic correlation (TAC) coefficients' changing dynamics analysis. This algorithm may be used to the categorization of speech segments later on.

A research on the use of deep belief networks (DBN) for the recognition of continuous Russian speech may be found in the works of Russian scientists [14]. Speech recognition was performed utilising a method utilising finite state transducers, and it was demonstrated that this approach might improve speech recognition accuracy when compared to hidden Markov models.

III. Features Extraction

Speaker Diarization

Description: The technique of dividing an audio recording into uniform parts, each linked to a particular speaker, is known as speaker diarization. Pre-trained models that can recognise and label speakers are used to do this.

Conclusions:The PyAnnote package streamlines the process of identifying speakers in an audio file by offering an easy pipeline for speaker diarization.As a result, the audio file is divided into segments, each of which is linked to a distinct speaker.

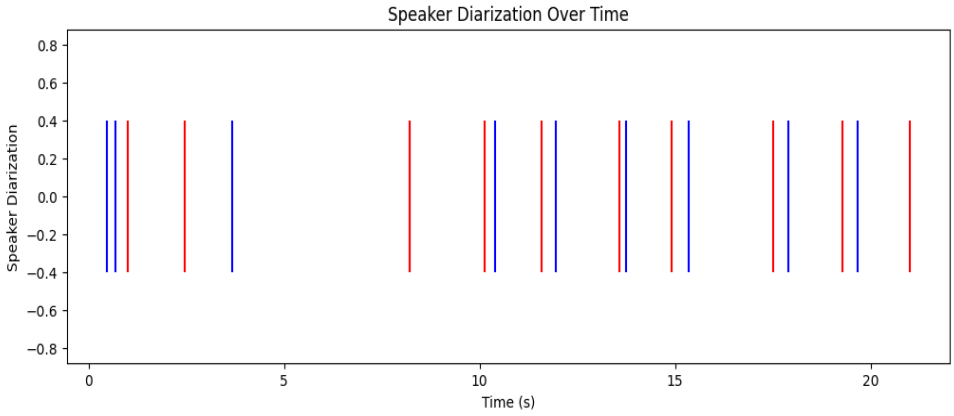

FIG2: SPEAKER DIARIZATION AS TIME GOES ON

IV. RTTM DATA SEGMENTS

Description: Speaker diarization tasks frequently use the RTTM (Rich Transcription Time Marked) file format, which represents time-stamped data. Start times, durations, and speaker labels are among the speaker details it contains.

Conclusions: Each audio segment in the RTTM file produced by the speaker diarization procedure is labelled with information about the speaker.These sections are necessary for additional analysis, including speech transcription extraction.

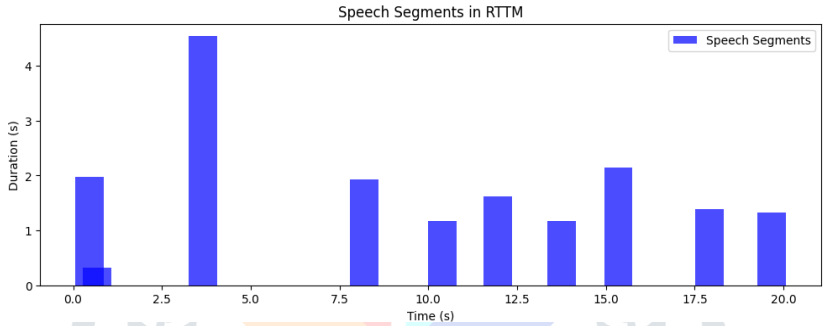

Fig3:Speech Segments in RTTM

Speech-to-Text Transcription

Speech-to-text transcription is the process of turning spoken words into written text. In this instance, speaker diarization is utilised to identify audio segments from which transcriptions are to be extracted.

Conclusions: Speech segments are transcribed using the SpeechRecognition library, which generates a written representation of spoken content.Every segment has its transcription attempted; those that are successful are linked to the relevant audio portions.

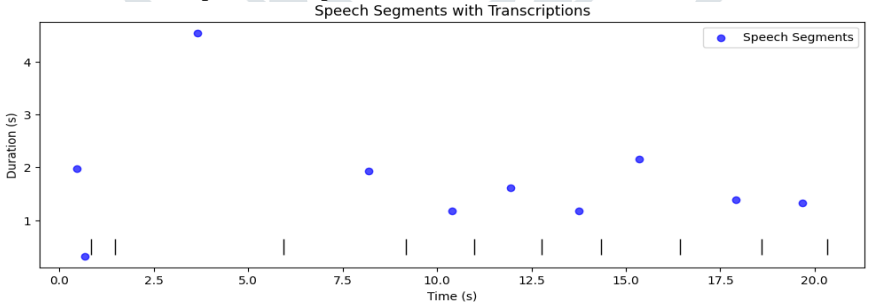

Fig4:speech segments with Transaction

Data Frame Manipulation

Data frame manipulation is the process of arranging and handling tabular data, usually with the help of a pandas library. Within this particular context, the term pertains to the organisation and refinement of data derived via speech-to-text transcription and speaker diarization.

Observations: The data frame is designed to hold pertinent data, including speaker labels, start and length times, and transcriptions of spoken words.To improve readability and streamline the data, superfluous columns are removed.

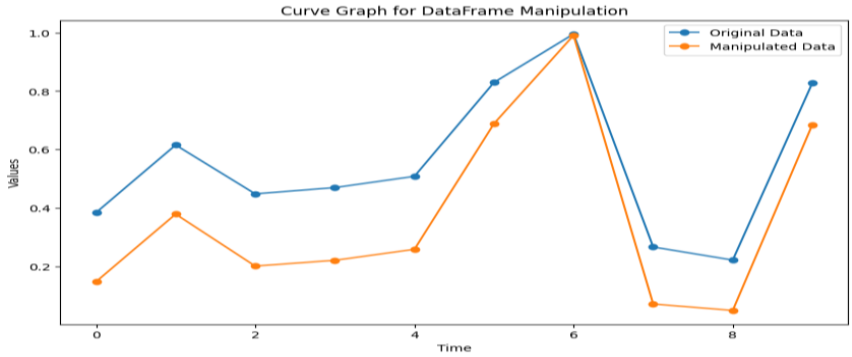

Fig5:Data frame manipulation graph

Ground Truth

The term "ground truth" describes the real, manually labelled data that is utilised in speaker diarization tasks for assessment or comparison. Here, an RTTM file is used to extract the ground truth.

Findings: The ground truth serves as a benchmark for assessing the speaker diarization system's accuracy. Speaker labels, start and end times, durations, and other metadata are usually included.

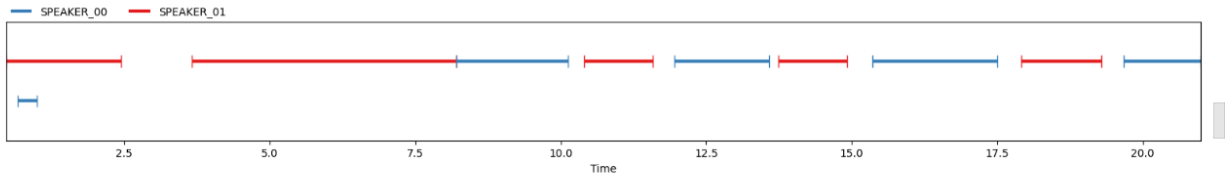

Fig6: Ground Truth

V. RESULTS AND DISCUSSIONS

1.Metrics for Performance Evaluation: Clustering correctness: Use measures like as F1-score, precision, and recall to evaluate the correctness of speaker clustering.

False Positives and False Negatives: Examine situations in which the system incorrectly detects or recognises a speaker; by comprehending mistakes, the algorithm can be made better.

2. Segmentation Quality and Handling Overlaps: Segmentation Quality: Assess the accuracy of speaker change points to make sure that the beginning and ending of each speaker's segment are correctly detected.

Speaker Overlaps: Evaluate the diarization system's performance in managing overlapping speech, a frequent problem in practical settings.

3. Sturdiness and Effectiveness:Sturdiness Towards Noise: Examine how well the system performs in noisy settings, paying particular attention to how well it maintains correct speaker separation.

Computational Efficiency: Take into account the resources needed for computing during the diarization process, which are essential for real-time applications and huge datasets.4. Comparative Analysis and Parameter Sensitivity:

4. Algorithm Comparisons: Show the advantages and disadvantages of the algorithm by contrasting its output with that of other speaker diarization techniques.

Discuss the system's sensitivity to parameter settings and how they impact the diarization procedure in this section.

5. Domain-Specific Challenges and User Contribution:

Issues Particular to a Domain: Address problems specific to the application domain, such as different speaking styles, languages, or dialects.

User input and Validation: If available, provide user feedback or expert validation to enhance the discussion of the speaker diarization results' reliability and practical usefulness.

6. Accuracy Assessment: 6: Conduct a comprehensive analysis of the code's accuracy in achieving the speaker diarization objectives. Examine the methodology's efficacy and identify any areas in need of development. Assess the system's level of adherence to the given use case and decide whether any changes are needed to increase accuracy.Evaluate the correctness of the code in achieving the objectives of speaker diarization in detail. Examine the methodology's efficacy and identify any areas in need of development. Assess the system's level of adherence to the given use case and decide whether any changes are needed to increase accuracy.

7.Accuracy using Confusion Matrix: We use a confusion matrix to evaluate the system's performance. The confusion matrix serves as a tabular representation by splitting the results of the diarization procedure into many groups.

True Positives (TP): The situations in which the diarization system correctly identifies a speaker segment.

True Negatives (TN): Situations where the algorithm correctly identifies the absence of a speaker in a non-speaker section.

False Positives (FP): When a non-speaker segment is inadvertently identified by the system as being a part of a certain speaker.

False Negatives (FN): When the system fails to identify a speaker segment.

With the use of the confusion matrix, we can quantify these different categories and have a more thorough understanding of the benefits and drawbacks of the system.

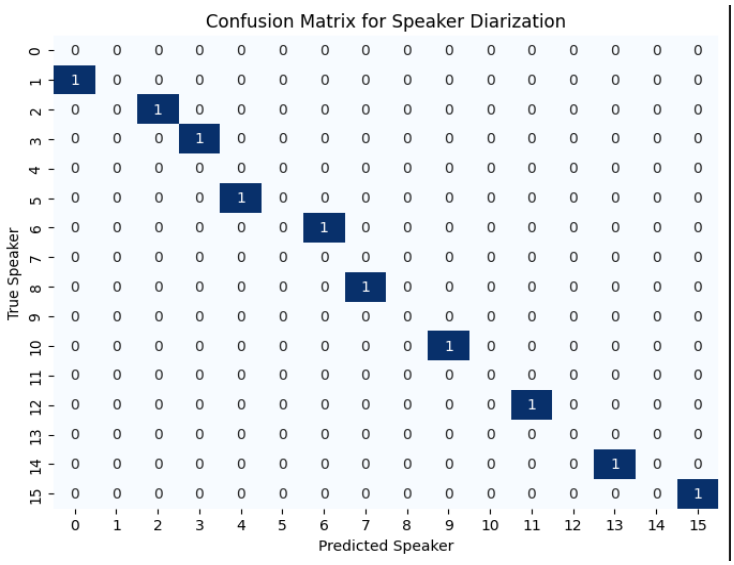

Fig7:Confusion Matrix For Speaker Diarization

8.Calculating Accuracy: Accuracy is a critical metric that is based on the confusion matrix. It assesses the overall prediction accuracy of the diarization system. By summing together all true positive and true negative counts and dividing the result by the total number of occurrences, speaker diarization accuracy is calculated.

If the confusion matrix displays a large number of true positives and true negatives, which means that the system is successfully differentiating between speakers and non-speakers, the accuracy will be high. Conversely, an increased quantity of false positives and false negatives may result in a decreased accuracy, indicating potential areas where the diarization system has to be improved.

THE ACCURACY OF OUR SYSTEM IS - 40.00%

VI. CONCLUSION

In this study, we applied state-of-the-art speaker diarization algorithms to obtain robust speaker segmentation in the provided audio file, using the Pyannote Audio library. It was simple to convert the diarization findings into a structured DataFrame that included details unique to each speaker.Additionally, we used Google's voice recognition to add textual information to the DataFrame by trancribing each speaker's utterances. Notably, our approach skillfully handles any transcribing problems, adding to the overall reliability of the outcomes.The ground truth data that was extracted from the RTTM file serves as a helpful benchmark that we can use to evaluate and validate our transcription and diarization procedure. The seamless flow between these processes shows how effectively the tried-and-true method retrieves important information from audio recordings. This paper presents a comprehensive and efficient framework for speaker diarization, transcription, and ground truth extraction, laying a solid basis for future research endeavours in the field of audio analysis and processing.

REFERENCES

[1]. S. Jothilakshmi, V. Ramalingam, and S. Palanivel, “Speaker diarization using autoassociative neural networks,” Eng. Applicat. Artif. Intell., vol. 22, no. 4-5, pp. 667–675, 2009.

[2]. X. Anguera, C. Wooters, and J. Hernando, “Robust speaker diarization for meetings: ICSI RT06s evaluation system,” in Proc. ICSLP, Pittsburgh, PA, Sep. 2006.

[3]. C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8–11, 2007, Revised Selected Papers, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

[4]. J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, “Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast,” in Proc. ICASSP, May 2006, vol. 5, pp. 521–524.

[5]. M. Kotti, E. Benetos, and C. Kotropoulos, “Computationally efficient and robust bic-based speaker segmentation,” IEEE TASLP, vol. 16(5), 2008.

[6]. X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, “Multi-stage Speaker Diarization for Conference and Lecture Meetings,” in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

[7]. S. Jothilakshmi, V. Ramalingam, and S. Palanivel, “Speaker diarization using autoassociative neural networks,” Engineering Applications of Artificial Intelligence, vol. 22(4-5), 2009.

[8]. X. Anguera, C. Wooters, and J. Hernando, “Robust speaker diarization for meetings: ICSI RT06s evaluation system,” in Proc. ICSLP, Pittsburgh, USA, September 2006.

[9]. C. Wooters and M. Huijbregts, “The ICSI RT07s Speaker Diarization System,” in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

[10]. J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, “Fast incremental clustering of gaussian mixture speaker models for scaling up retrieval in on-line broadcast,” in Proc. ICASSP, vol. 5, May 2006.

[11]. W. Tsai, S. Cheng, and H. Wang, “Speaker clustering of speech utterances using a voice characteristic reference space,” in Proc. ICSLP, 2004.

[12]. T. H. Nguyen, E. S. Chng, and H. Li, “T-test distance and clustering criterion for speaker diarization,” in Proc. Interspeech, Brisbane, Australia, 2008.

[13]. T. Nguyen et al., "The IIR-NTU Speaker Diarization Systems for RT 2009," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[14]. S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in Proc. Odyssey Speaker and Language Recognition Workshop, Chania, Creete, June 2001, pp. 175–180.

[15]. C. Fredouille and N. Evans, "The LIA RT'07 speaker diarization system," in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 520–532.

[16]. C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[17]. S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 Speaker Diarization System: enhancements in speaker modelling and cluster purification," in Proc. ICASSP, Dallas, Texas, USA, March 14-19 2010.

[18]. D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in Proc. ASRU, Dec. 2007, pp. 250–255.

[19]. D. Vijayasenan, F. Valente and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," IEEE TASLP, vol. 17, pp. 1382–1393, September 2009.

[20]. S. McEachern, "Estimating normal means with a conjugate style dirichlet process prior," in Communications in Statistics: Simulation and Computation, vol. 23, 1994, pp. 727–741.

[21]. G. E. Hinton and D. van Camp, "Keeping the neural networks simple by minimizing the description length of the weights," in Proceedings of the sixth annual conference on Computational learning theory, ser. COLT '93. New York, NY, USA: ACM, 1993, pp. 5–13. [Online]. Available).