Interspeech 2018
2-6 September 2018, Hyderabad

Information Bottleneck based Percussion Instrument Diarization System for Taniavartanam Segments of Carnatic Music Concerts

Nauman Dawalatabad, Jom Kuriakose, C. Chandra Sekhar, Hema A. Murthy

Indian Institute of Technology Madras, India
{nauman,jom,chandra,hema}@cse.iitm.ac.in

Abstract

An approach to diarize *taniavartanam* segments of a Carnatic music concert is proposed in this paper. Information bottleneck (IB) based approach used for speaker diarization is applied for this task. IB system initializes the segments to be clustered uniformly with fixed duration. The issue with diarization of percussion instruments in *taniavartanam* is that the stroke rate varies highly across the segments. It can double or even quadruple within a short duration, thus leading to variable information rate in different segments. To address this issue, the IB system is modified to use the stroke rate information to divide the audio into segments of varying durations. These varying duration segments are then clustered using the IB approach which is then followed by Kullback-Leibler hidden Markov model (KL-HMM) based realignment of the instrument boundaries. Performance of the conventional IB system and the proposed system is evaluated on standard Carnatic music dataset. The proposed technique shows a best case absolute improvement of 8.2% over the conventional IB based system in terms of diarization error rate.

Index Terms: Music Diarization, Information Bottleneck (IB), Carnatic Music, *Taniavartanam*, Percussion instruments

1. Introduction

Carnatic music is a classical music form popular in the southern part of India. Carnatic music attributes its origin to Samaveda [1] and is one of the oldest music forms in the world. This music form follows *gayaki* style where the emphasis is given to the vocal form of the music. A Carnatic music concert generally consists of a lead artist (typically a vocalist) who is accompanied by a violinist, and percussion instrument artists. The lead percussion instrument is usually the mridangam. A concert is made up of a sequence of items called compositions. The melody of these compositions are set to different rhythm patterns. Every concert is like a story that the artist wishes to convey, where audience continuously engages with the artist [2]. Every concert consists of a main item where the artist gives his/her best in terms of *manodharma* (improvisation) [3, 4]. This item also includes a solo performance by the percussion artists referred to as *taniavartanam*. Generally, a prime time concert will consist of at least two percussion artists. During the solo percussion performance, the percussion artists indulge in a *"call and response"* section where the lead percussion artist challenges the other artists with certain phrases, and the challenged artists give a fitting response.

The objective of this paper is to diarize the *taniavartanam* portion of the concert into the components corresponding to each of the instruments. The timbre of the instruments (for example, ghatam+mridangam or kanjira+mridangam or morsing+mridangam etc.) are quite different. This is similar to voice differences in speaker diarization. We therefore borrow ideas from the speaker diarization literature, to diarize percussion instruments in *taniavartanam*. Similar to speech production, where an utterance is made up of a sequence of syllables, here the segments are made up of a sequence of strokes. The length of the sequence is dependent upon the rhythm that is associated with the item. These segments can be as short as a few milliseconds. The stroke rate also varies quite significantly as one progresses from the beginning to the end of the *taniavartanam*. This makes the diarization of percussion solo a difficult task. Previous music information retrival (MIR) tasks on *taniavartanam* [5,6] have therefore only used mridangam solos, constraining the usage of available concert data for these tasks. Diarization of *taniavartanam* has a number of applications in music information retrival (MIR). Analysis of *taniavartanam* segments is crucial, as it can give insights into rhythm (or *tala*) that is used in the main item of the concert.

In the context of speaker diarization, a number of different efforts can be found in the literature [7, 8], where some focus on the model [9–11], the others focus on features [12–17]. In [18], authors use time delay of arrival as a secondary information that complements spectral features which helps to better discriminate between the speakers. These efforts do suggest the importance of domain information in diarization. Speaker diarization is a task of clustering and hence the segments to be clustered must have enough speaker-specific information for a better clustering solution. This is challenging in *taniavartanam* as the information is spread unevenly owing to the extemporaneousness of the performance.

There is a huge repository of Carnatic music concerts [19]. Authors in [20] showed that continuous recordings of Carnatic music concerts can be segmented into items for archival. They also showed that the main item of a concert can be identified by finding the segment corresponding to *taniavartanam*. Nevertheless, no attempts have been made to diarize the *taniavartanam* in the literature. This paper is an attempt to include additional meta information to the *taniavartanam* segment of a concert, where the audio is further diarized based on the instrument. Given a *taniavartanam* audio, the system finds the sections where each instrument is played. Information bottleneck (IB) [11] based method for speaker diarization is popular owing to its low runtime and error rates. In the first part of the paper, we directly use IB based speaker diarization system to diarize *taniavartanam*. The *taniavartanam* is divided uniformly into short segments and IB based clustering is performed. This is then followed by Kullback-Leibler hidden Markov model (KL-HMM) based realignment [21]. In the second part of the paper, we modify the IB based system for music diarization to make use of stroke rate as a secondary information to carefully initialize the segments to be clustered. This is important as the clustering solution depends on the information about the instrument present in the segments to be clustered. We refer to this system as *VarIB*. After *VarIB* clustering is done the instrument bound-

10.21437/Interspeech.2018-1203

aries are further refined using KL-HMM realignment. Both the systems are completely unsupervised.

The rest of the paper is organized as follows. Section 2 describes the IB based approach used to diarize percussion solos. In Section 3 we briefly describe the method used for onset detection. Section 4 presents the *VarIB* system that uses variable length segments for intialization. In Section 5 we show the performance of the proposed systems on Charsur datasets [22]. This section also discusses how the variable segmentation helps distribute the information uniformly among the segments. Finally, Section 6 concludes the paper.

2. Information bottleneck based diarization of *Taniavartanam*

Information bottleneck (IB) is an approach where the set of variables $\mathbf{X}$ are arranged into the set of clusters $\mathbf{C}$ such that the relevant information $\mathbf{Y}$ needed for clustering is preserved [11,23]. The IB approach is popular in diarizing human speech conversations to answer the question of *"who spoke when?"* In the context of the percussion instruments this idea translates to *"what instrument was played when?"*

In modeling Gaussian mixture model (GMM), the instrument-specific information is present in the components of the GMM. Hence, the GMM components become the relevant variable denoted as $\mathbf{Y}$. In agglomerative IB algorithm, a set of the fixed length audio segments $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ are clustered iteratively in a bottom-up manner. These segments are arranged into a set of clusters $\mathbf{C} = \{\mathbf{c_1}, \ldots, \mathbf{c_m}\}$ preserving the relevant information $\mathbf{Y} = \{\mathbf{y_1}, \ldots, \mathbf{y_k}\}$. The initial audio segments used in the IB algorithm come from dividing the whole *taniavartanam* into short equal sized segments $\mathbf{x_i}$. Here, the primary assumption is that each short audio segment $\mathbf{x_i}$ contains only one instrument.

The IB objective function given in Equation 1 is minimized such that it preserves the information about the relevance variable while producing a compact solution.

$$\mathcal{F} = \mathrm{I}(\mathbf{Y}, \mathbf{C}) - \frac{\mathbf{1}}{\beta}\mathrm{I}(\mathbf{C}, \mathbf{X}) \tag{1}$$

where $\mathrm{I}(\mathbf{A}, \mathbf{B})$ denotes the mutual information between the random variables $\mathbf{A}$ and $\mathbf{B}$, and β is a Lagrange multiplier.

After every iteration of bottom-up clustering, the change in the objective function can be calculated as,

$$\Delta\mathcal{F}(\mathbf{c_i}, \mathbf{c_j}) = (P(\mathbf{c_i}) + P(\mathbf{c_j})).d_{ij} \tag{2}$$

where $P(.)$ denotes the probability, and d_{ij} denotes the distance between the clusters $\mathbf{c_i}$ and $\mathbf{c_j}$ which is given by

$$d_{ij} = JS[P(\mathbf{y}|\mathbf{c_i}), P(\mathbf{y}|\mathbf{c_j})] - \frac{1}{\beta}JS[P(\mathbf{x}|\mathbf{c_i}), P(\mathbf{x}|\mathbf{c_j})] \tag{3}$$

The Jensen-Shannon divergence for $JS[P(\mathbf{y}|\mathbf{c_i}), P(\mathbf{y}|\mathbf{c_j})]$ is given by

$$\pi_i D_{KL}[P(\mathbf{y}|\mathbf{c_i})||q_Y(\mathbf{y})] + \pi_j D_{KL}[P(\mathbf{y}|\mathbf{c_j})||q_Y(\mathbf{y})] \tag{4}$$

where D_{KL} denotes the Kullback-Leibler divergence, $q_Y(\mathbf{y}) = \pi_i\ P(\mathbf{y}|\mathbf{c_i}) + \pi_j\ P(\mathbf{y}|\mathbf{c_j})$ and $\pi_i = P(\mathbf{c_i})/(P(\mathbf{c_i}) + P(\mathbf{c_j}))$.

After every iteration, two clusters with minimum $\Delta\mathcal{F}$ are considered for merging. The merged cluster $\mathbf{c_r}$ made from $\mathbf{c_i}$ and $\mathbf{c_j}$ is characterized as,

$$P(\mathbf{c_r}) = P(\mathbf{c_i}) + P(\mathbf{c_j}) \tag{5}$$

$$P(\mathbf{y}|\mathbf{c_r}) = \frac{P(\mathbf{y}|\mathbf{c_i})P(\mathbf{c_i}) + P(\mathbf{y}|\mathbf{c_j})P(\mathbf{c_j})}{P(\mathbf{c_r})} \tag{6}$$

The normalized mutual information after every iteration is calculate by $NMI = \frac{\mathrm{I}(\mathbf{Y},\mathbf{C})}{\mathrm{I}(\mathbf{Y},\mathbf{X})}$. The bottom-up clustering is repeated until the NMI based stopping criterion is reached. After the clustering process is completed, the boundaries are refined using KL-HMM based realignment [21].

3. Stroke onset detection

In this section, we briefly describe the algorithm for stroke onset detection [24]. The unsupervised group delay based onset detection algorithm is shown to give a performance at par with state-of-the-art supervised machine learning based onset detection algorithms on Carnatic percussion instruments [24]. As shown in Figure 1, minimum phase group delay processing is applied on the smoothed envelope of the derivative of the music signal to emphasise the peaks at onset locations.

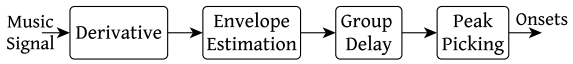

Figure 1: *Onset detection algorithm.*

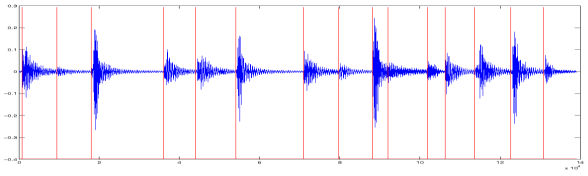

Figure 2: *Onset detection in Mridangam strokes. Onset points are marked with red boundary.*

The onsets detected by this algorithm as illustrated in Figure 2 give the stroke rate information in a concert. This information is used in the next section to initialize the segments for clustering.

4. Varying length segment for IB based diarization of taniavartanam

The difference between speaker diarization and percussion diarization is that the instrument switch can happen within a 20-30ms range, while it is of the order of a few seconds in speaker diarization. Further, the rate of the strokes can vary significantly across the concerts. In percussion instruments, the instrument-specific information is available only in strokes. Hence, it is important to capture this information properly to get a good clustering solution. Given the information on the stroke rate variation, we propose varying length segment initialization for IB based clustering.

In IB based diarization explained in section 2, the segments are uniformly divided in terms of duration. This can lead to different number strokes in different segments which affects the GMM modeling step in IB framework. This can be avoided by distributing the information evenly throughout the short segments. We initialize the segments in IB system using varying length segments. These segments are derived with the help of stroke detection algorithm as explained in section 3. Instead of fixed duration segments, we allow segments to take variable duration (length) within a meaningful range (this is defined

based on the domain information), while keeping the number of strokes constant in each segment.

Let $maxDur$ be the maximum allowed duration for any segment and $minDur$ be the minimum allowed duration for each segment. Let $minStrokes$ be the minimum allowed number of strokes in a segment. The segmentation is done such that the length of each segment is between $minDur$ and $maxDur$, and the number of strokes is equal to $minStrokes$. This ensures that each segment has enough number of strokes. The next segment is calculated from the time where the current segment has ended which avoids the overlapping segment problem.

The hyper-parameter $minStrokes$ controls the number of varying duration segments. The number of segments in the given concert is inversely proportional to $minStrokes$. However, the segment durations are bounded by $maxDur$ and $minDur$. This ensures that the number of segments is not too small or too large for a given concert. This is important for proper modeling of the GMM and also for estimating the posteriors in IB system. We then perform IB clustering on these varying duration segments. The clustering is stopped once the NMI threshold is reached.

Similar to the IB system, the clusters obtained from *VarIB* may have two kinds of errors. The first is that the initial segment may not be instrument homogeneous as it may have multiple instruments played in sequence. The second issue is that the initial segment itself is in a wrong cluster. Similar to speaker diarization, a KL-HMM realignment is performed after *VarIB* clustering. This helps to refine the boundaries. The boundaries obtained after realignment are the final output of the *VarIB* system.

5. Dataset, experimental setup and results

In this section, we describe the dataset used, experimental setup and the performance of both the systems.

5.1. Dataset

A subset of Charsur dataset is used in our experiments [22]. A subset of 50 concerts are chosen randomly from the whole dataset. The *taniavartanam* part from each of the concerts is obtained manually. The boundaries of the instruments for each of these *taniavartanam*s were also marked manually for evaluation. The duration of the *taniavartanam*s in all the 50 concerts is between 5 minutes to 25 minutes. These recordings were divided into 5 sets with 10 recordings in each set. One set (Tani-Dev) was used as development set to tune the hyperparameters, whereas the remaining four (Tani-1 to 4) were used for testing purpose.

5.2. Experimental setup

All the hyper-parameters were tuned to get the best performance on the development data. We used 19 dimensional MFCC features. The threshold on NMI and the value of β for both IB and *VarIB* systems were tuned to 0.4 and 10, respectively. The maximum number of clusters for IB and *VarIB* is fixed to be 3. The reason for this is that most concerts have two percussion instruments. The percussion instruments either play in tandem or together, leading to three different timbres. However, the NMI controls the clusters if the number of instruments is less than 3 and/or there is small/no overlapping segment in the *taniavartanam*. The initial segment length for IB was set to 2.0 seconds. However, for both the methods the minimum HMM duration constraint for realignment was kept to 1 second. This constraint was necessary to capture the fast switching of instruments that occurs in the last section of the *taniavartanam*. Shorter minimum HMM duration than 1 second introduced false boundaries. The $minDur$ and $maxDur$ for *VarIB* were fixed to 1 and 6 seconds, respectively. Hence, the *VarIB* is flexible to take segments between 1 and 6 seconds duration. We set the value of the threshold on the minimum number of strokes in a segment $minStrokes$ to 15 for *VarIB*. While implementing, we relax[1] the $minStrokes$ such that the segment can have more than $minStrokes$ strokes if and only if its duration is contained within $minDur$.

Table 1: *DER for all datasets are reported on forgiveness collar of 0.15 seconds. Performance of VarIB system is compared with that of the IB system for different initial segment duration.* Impr-I *denotes the absolute improvement in DER of VarIB system on each dataset with respect to the IB system with 2.0 seconds of fixed segment duration.* Impr-II *denotes the absolute improvement of VarIB system on each dataset with respect to the best DER obtained on IB system with any three of the fixed segment sizes. The corresponding relative improvements are given in the parentheses.*

Sys./Data.	Dev. Set	Test Set			
	Tani-Dev	Tani-1	Tani-2	Tani-3	Tani-4
IB (1.5)	17.9	20.4	23.0	21.6	13.9
IB (2.0)	17.6	19.0	19.4	22.3	13.1
IB (2.5)	18.0	18.5	17.9	19.9	15.2
VarIB	13.0	11.9	15.6	11.7	9.4
Impr-I (%)	4.6 (26.1)	7.1 (37.4)	3.8 (19.6)	**10.6** (47.5)	3.7 (28.2)
Impr-II (%)	4.6 (26.1)	6.6 (35.7)	2.3 (12.9)	**8.2** (41.2)	3.7 (28.2)

Since the *taniavartanam* concerts are continuous without considerable silence, voice activity detection (VAD) system is not required. However, one can use VAD for other types of concerts where there are considerably long pauses. Diarization error rate (DER) is used as the evaluation metric. This is calculated as the sum of the missed instrument (MI), the false alarms (FA) and the instrument error rate (IER). IER is the percentage of time the hypothesised segments are assigned to the wrong instrument. MI and FA are related to VAD errors. Since the concerts have negligible silence part, MI and FA are close to zero. Hence the values of DER and IER are similar. Nevertheless, we report the final performance in terms of DER.

We used the standard evaluation tool from NIST [25] to evaluate the performance in terms of DER of the proposed systems. A forgiveness collar of 0.15 seconds was used to exclude the errors at the boundaries within the 0.15 seconds range of the reference boundaries. This is important as it not only excludes the errors made by the system but also the errors made during the preparation of the reference boundaries. It should also be noted that the forgiveness collar of 0.15 seconds is stricter than the usual 0.25 seconds used for the task of speaker diarization [26].

5.3. Results

The results of the experiments are reported in Table 1 in terms of DER. We report the DER for IB system with 3 different initial segment durations since all of them showed similar performance on the development set. Since the DER for other initial segment durations for IB system were poor we did not report

[1]The relaxation is kept to avoid the corner case on last segment and also when VAD is used.

them. The duration (in seconds) of the fixed length initial segments for IB system are given in parentheses. Since the IB system with 2 seconds as the initial segment duration showed the best performance on the development dataset, we compare the performance of *VarIB* with this system. The Impr-I denotes the performance improvement of *VarIB* system over IB system with fixed segments of 2 seconds. The corresponding relative improvements are given in parentheses. We also compared the performance of *VarIB* with the best DER obtained among the three IB systems. This is mentioned in Impr-II. It can be seen from the Table 1 that the IB system showed reasonably good performance with fixed duration segments. The *VarIB* system showed significant improvement over all IB systems with different initial segment sizes. The best case relative improvement of 47.5% (absolute 10.6%) was observed for Tani-3 dataset with respect to the IB system with 2 seconds as the segment duration. The relative improvement of 41.2% (absolute 8.2%) was observed on Tani-3 dataset when compared with the best performing IB system (with 2.5 second) for that dataset. *VarIB* outperforms the IB systems for different segment durations on all the datasets. This shows the effectiveness of proper initialization of segments in *VarIB* system. If the *taniavartanam* contains overlapping segments (where both instruments are played simultaneously), a separate cluster for such segments is obtained in both the systems.

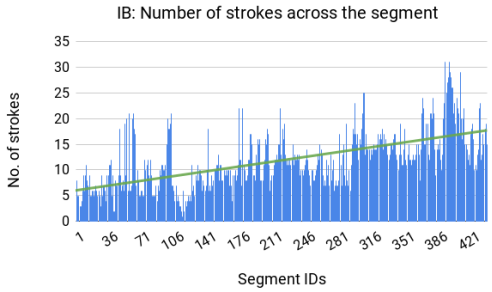

Figure 3: *Number of strokes in each of the fixed duration segments used in IB system. The information about the instrument is unevenly distributed. Its less in the starting part and more concentrated in the end part of the taniavartanam.*

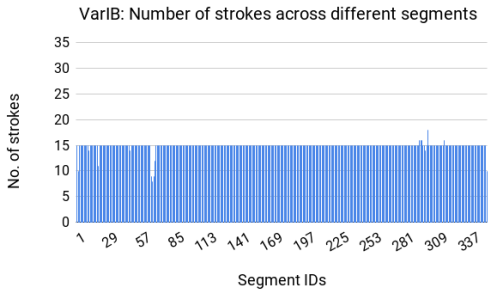

Figure 4: *Number of strokes in each of the varying duration segments used in VarIB system. VarIB equally distributed the information among different initial segments.*

In Figures 3, 4 and 5 the segment IDs shown on the X-axis are in the same order as the time. That is, the segment ID 1 denotes the segment from the start of the *taniavartanam* and the last segment ID is for the part where the *taniavartanam* ends.

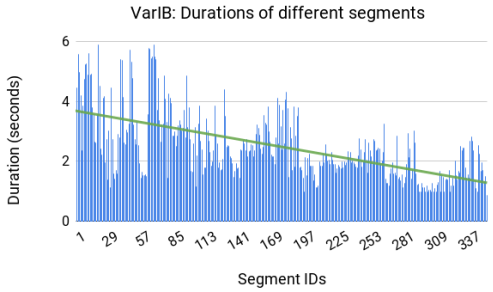

Figure 5: *Duration of each of the varying length segments used in VarIB system. The segment length reduces with the time. This is due to the high stroke rate towards the end of the concert.*

All the plots in all the figures were obtained from a sample concert[2] taken from the development set. Mridangam and ghatam are played in this concert.

Figure 3 shows the stroke distribution throughout the *taniavartanam* in IB system across the segments of fixed duration of 2 seconds. It can be seen that the strokes are unevenly distributed across the segments. Hence in the case of fixed duration segment in IB system, these segments have different levels of information. It should be noted that the stroke rate goes on increasing as the *taniavartanam* progresses. It is usually very high in the last section of the *taniavartanam*. And hence in the IB system uniformly divided segments have larger number of strokes in the last section which eventually leads to a poor balance of information across segments.

Figure 4 shows uniformly distributed strokes across varying duration segments in *VarIB* system. This is important as the initial segmentation affects the agglomerative clustering process. A good initialization leads to a better final solution. The dip near segment ID 75 shows the segments with few number of strokes. It showed around 8 strokes in 6 seconds ($maxDur$) duration. Also, in the latter part of the concert 18 strokes were found in short duration of 1 second segment as both the instruments were played fast and together. Figure 5 shows the durations of different segments in *VarIB*. It can be seen that as the *taniavartanam* progresses, the *VarIB* adjusts segment lengths such that the information in each segment remains similar. It can also be noted that the *VarIB* reduces the segment duration in the last section of the concert where the artists play fast strokes. The green line in Figures 3 and 5 shows the overall trend of the respective graphs.

6. Conclusion

In this work, the information bottleneck (IB) approach is used for the first time to diarize the *taniavartanam* part of a Carnatic music concert. We proposed a *VarIB* approach which initializes the segments in IB system with varying durations based on the stroke rate. Both the systems gave reasonable performance. However, *VarIB* showed significant absolute improvement of 8.2% over the IB system on the evaluation data.

The proposed technique opens up the possibility of research in music diarization from an information theoretic perspective. We plan to extend the technique further to instrument linking where similar instruments across the concerts can be linked.

[2]https://musicbrainz.org/release/5b24d8ca-c530-4332-b2c1-d42902169c99

7. References

[1] V. Raghavan, "Samaveda and music," *Journal of the Music Academy of Madras*, vol. 33, pp. 127–133, 1962.

[2] M.V.N. Murthy, "Applause and Aesthetic Experience, CompMusic," http://compmusic.upf.edu/node/151.

[3] A. Priyamvada, *Encyclopaedia of Indian music*. Anmol, 2007.

[4] B. Nettl, A. Arnold, R. M. Stone, J. Porter, and T. Rice, *The Garland Encyclopedia of World Music: South Asia: the Indian subcontinent*. Taylor & Francis, 1998, vol. 5.

[5] J. Kuriakose, J. C. Kumar, P. Sarala, H. A. Murthy, and U. K. Sivaraman, "Akshara transcription of mrudangam strokes in carnatic music," in *Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.

[6] K. Gogineni, J. Kuriakose, and H. A. Murthy, "Mridangam artist identification from taniavartanam audio," in *Twenty Fourth National Conference on Communications (NCC)*, Hyderabad, India, Feb. 2018.

[7] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[8] S. Tranter and D. Reynolds, "An Overview of Automatic Speaker Diarization Systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[9] J. Ajmera and C. Wooters, "A Robust Speaker Clustering Algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003, pp. 411–416.

[10] C. Wooters and M. Huijbregts, *Multimodal Technologies for Perception of Humans*. Springer Berlin Heidelberg, 2008, ch. The ICSI RT07s Speaker Diarization System, pp. 509–519.

[11] D. Vijayasenan, F. Valente, and H. Bourlard, "An Information Theoretic Approach to Speaker Diarization of Meeting Data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[12] S. Madikeri and H. Bourlard, "Filterbank Slope based Features for Speaker Diarization," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[13] S. Yella and H. Bourlard, "Information Bottleneck based Speaker Diarization of Meetings using Non-Speech as Side Information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 96–100.

[14] N. Dawalatabad, S. Madikeri, C. C. Sekhar, and H. A. Murthy, "Two-Pass IB based Speaker Diarization System Using Meeting-Specific ANN based Features," in *Proceedings of INTERSPEECH*, 2016, pp. 2199–2203.

[15] S. H. Yella, A. Stolcke, and M. Slaney, "Artificial Neural Network Features for Speaker Diarization," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 402–406.

[16] H. Bredin, "Tristounet: Triplet loss for speaker turn embedding," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5430–5434, 2017.

[17] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree, "Speaker diarization using deep neural network embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 4930–4934.

[18] D. Vijayasenan, F. Valente, and H. Bourlard, "An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 431–438, 2011.

[19] "Sangeethapriya," http://www.sangeethapriya.org/.

[20] P. Sarala and H. A. Murthy, *Segmentation of Continuous Audio Recordings of Carnatic Music Concerts into Items for Archival*. Sadhana (Springer), June 2017.

[21] D. Vijayasenan, F. Valente, and H. Bourlard, "KL Realignment for Speaker Diarization with Multiple Feature Streams," in *Proceedings of INTERSPEECH*, Sept 2009, pp. 1059–1062.

[22] "Charsur Dataset – MusicBrainz Database," https://musicbrainz.org/label/3e188240-9eb5-4842-b7b9-d6c2393211b7.

[23] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," *in NEC Research Institute TR*, 1998.

[24] P. M. Kumar, J. Sebastian, and H. A. Murthy, "Musical onset detection on carnatic percussion instruments," in *Twenty First National Conference on Communications (NCC)*. IEEE, 2015, pp. 1–6.

[25] "The NIST Rich Transcription," https://www.nist.gov/itl.

[26] "The NIST Rich Transcription 2006 Evaluation," https://www.nist.gov/speech/tests/rt/rt2006/spring/.