Speaker Diarization: A Review

Aishwary Joshi, Mohit Kumar, Pradip K. Das
Department of
Computer Science & Engineering
Indian Institute of Technology Guwahati
Guwahati, Assam 781039
Email: aishwary@iitg.ernet.in, mohit.2013@iitg.ernet.in, pkdas@iitg.ernet.in

***Abstract*—Speaker Diarization is the task of identifying start and end time of a speaker in an audio file, together with the identity of the speaker i.e. "who spoke when". Diarization has many applications in speaker indexing, retrieval, speech recognition with speaker identification, diarizing meeting and lectures. In this paper, we have reviewed state-of-art approaches involving telephony, TV shows, broadcasting and meeting data. Along with the state-of-art approaches, the major approaches that are commonly used in diarization are reviewed. Few possible future directions of this technology are also identified.**

I. Introduction

Speaker Diarization is the task of identifying speaker boundaries (start and end time) in an audio file along with the speaker identity. The term Speaker Diarization was not coined until 2003; rather it was earlier called Speaker Segmentation, which means from an audio file, segments belonging to the same speaker must be separated.

Speech Recognition in its early phase was just limited to distinguish speech and non-speech data from the audio and detecting the content of the speech part. Also, the audio file under consideration usually contained speech from a single speaker over a single channel. But presently, because of the advancement in speech technology, speech recognition tasks can be performed on audio files involving 2 speakers (Telephonic Domain) as well as more than 2 speakers (Broadcast Domain & Meeting Domain).

II. Early works involving telephony and news broadcast data

In 1997 Matthew A. Siegler came up with a very basic method for identifying speaker boundaries. In their work [19] this task was accomplished by dividing it into 3 subtasks: Segmentation, Classification and Clustering. Segmentation was performed by dividing initial speech data into small segments (window of 2 sec, mean and variance are calculated for that short period) such that each segment is expected to contain speech from the single speaker. Symmetric KL2 (Kullback Leibler) distance was calculated between two neighboring segments and if KL2 distance reached local maximum, a segment boundary is generated. KL2 distance reaching a maximum means that two segments are very much divergent and hence do not belong to the same speaker. By doing so multiple segments may contain speech from same speaker and hence they need to be merged. This problem was addressed in the Clustering step. On the other hand Classifications [14], step classifies the segments as half bandwidth vs full bandwidth according to an already trained Gaussian mixture model. All segments are finally clustered using agglomerative clustering, two segments with minimum distance metric are merged. For distance metric, both KL2 and Mahalanobis distances are used. However, KL2 distance outperforms Mahalanobis distance. A threshold is set on the distance metric to stop the Clustering Process.

The major problem with the KL2 distance is that threshold adjustments are required which is most likely to vary for a different audio file recorded in a different environment by a different set of speakers. So a clustering method is required that does not use this threshold adjustment and thus provides clustering that is independent of the channel and environment variation. In [6] BIC (Bayesian Information Criterion) based single point change detection method was introduced which can be extended to detect multiple change points in the audio file. The segmentation step remains the same as in [19]. Cepstral coefficients for each frame is obtained to detect the change points and it is detected in each segment (length 2 sec).

$$R(i) = N \log |\Sigma| - N_1 \log |\Sigma_1| - N_2 \log |\Sigma_2 \tag{1}$$

$$i = argmax_i R(i) \tag{2}$$

$$BIC(i) = R(i) - \lambda p \tag{3}$$

Where $\Sigma, \Sigma_1, \Sigma_2$ are covariance matrix from all data, from start to change point and from change point to the end respectively, p in (3) a is penalty term and λ is a penalty weight. A change is detected if BIC(i)>0

In [1], Robust speaker clustering algorithm was introduced which is based on [6] but with fully automatic stopping criterion. It applies clustering in an iterative manner. The algorithm runs iteratively and likelihood ratio along the best path increases until it reaches an optimal point and thereafter decreases. This optimal point defines the total number of speakers. The Algorithm is based on ergodic Hidden Markov model (HMM) where each state of HMM is a cluster and represents one speaker, each state consists of several sub-states. The probability density function for each state is a Gaussian Mixture Model (GMM) with M components and

978-1-5090-2684-5/16/$31.00 ©2016 IEEE

these components are shared across all sub-states. As in [19] and [6], similar method for speaker segmentation is followed where the input audio is divided into small segments i.e. over clustering of the data. After segmentation, HMM parameters are initialized and HMM is trained using EM (Expectation Maximization) algorithm. For training, the segment of data is obtained to maximize the likelihood of data, for given parameters of GMM. In each iteration, GMM parameters are updated based on segmentation. The last step is to merge or create clusters out of those segments, For this BIC based agglomerative clustering, discussed in [6], is followed. The advantage of BIC based clustering is that no tweaking of any parameter is required thereby providing a fully automatic stopping criterion. Results are tested with MFCC as well as LPCC but very little difference in terms of performance and diarization error rate was observed.

Since diarization is basically clustering the speech data from same speakers so initially, attempts were made to extend speaker recognition systems to diarization system. Not only speaker recognition, some of the hybrid system involving LIA and CLIPS system were introduced to improve the performance of diarization [17]. While both systems require a common pre-segmentation phase which classifies input speech as Male/Female, Speech/NonSpeech, and Wide/Narrow. LIA system [13] uses HMMs where each state of HMM represents one speaker and transition from one state to another represent speaker change. LIA system requires training and iteratively speakers are added by LIA system. CLIPS system [7] on the other hand are based on BIC detection based strategy. In CLIPS system Speaker change detection is applied individually on each class (Male/Female Wide/Narrow). In [17] two types of hybrid approaches are discussed; one is the piped and the other one is merging. In the piped system, CLIPS segmentation is followed by LIA re-segmentation which improves the purity of clusters obtained as final output. While in merging strategy output from both the systems are fused and LIA re-segmentation is applied on merged output. The merging strategy reduces the diarization error rate at the cost of not detecting some non-overlapping segments.

Iterative Clustering in [1] requires very large computation efforts because of the re-estimation of GMM parameters. In [5] multistage diarization system is proposed where instead of iterative BIC clustering, Agglomerative BIC clustering is applied once which saves on computation time and is then followed by Agglomerative SID (Speaker Identification) clustering. Multistage pipeline first employs speech activity detection (SAD) to extract speech content and the non-speech contents are discarded. After this step speech content is chopped into small segments and for each segment, GMM is trained and it is followed by Viterbi re-segmentation and agglomerative BIC clustering applied once. These steps are even followed in traditional diarization system but in [5] after these steps, classification is performed and then agglomerative SID clustering is applied on each class individually. The classification is done by Universal Background Model (UBM) with 128 diagonal Gaussians. Also, it may happen that SAD may not detect all the non-speech content present in the audio so these contents are removed by another step called SAD post filtering which removes small non-speech fragments.

III. Works involving meetings and TV shows broadcast data

The problems associated with telephony data, like presence of only limited number of microphones, can be tackled by various pre-processing steps that are used in approaches disccussed in the previous section. But in a meeting or TV show (recording of a live show) where the condition are not that much in control and so there will be overlapping between speech from one user to speech of other user and between noise/audio. Also, sometimes it will not be possible to have a dedicated microphone for each individual speaker of a meeting or a TV show so a common set of microphones are shared across all the speakers which result into the problem of session variability which is caused due to the varying distance of microphone from the speakers. Therefore, to resolve these problems of session variability different approaches are adapted for pre-processing of input audio file.

In [3] to resolve the problem of session variability in the presence of multiple distant microphones (MDM), Delay and Sum algorithm is applied on input signal obtained from microphones. Since all microphones are kept at different positions so the sound will be first recorded by the nearest microphone and then to the farthest one.This creates session variability between the microphones. This is modeled by Time Delay of Arrival (TDOA). Delay and Sum algorithm combine output from all channels into one channel and further processing then take place on the final output.The Algorithm requires a reference channel at each point which is selected by calculating the signal-to-noise ratio (SNR) at that point. Once the reference channel is identified for every other channel time delay of arrival (TDOA) is estimated by using GCC_PHAT (Generalized Cross-Correlation with Phase transform). Once this pre-processing step is done, the remaining process of diarization is same as the one followed in [26].

$$G_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|} \tag{4}$$

Given two signals $x_i(n)$ and $x_j(n)$,the GCC-PHAT can be found out by (4) where $X_i(f)$ and $X_j(f)$ are Fourier transforms of the two signal and []* denotes complex conjugate.

The major problem with meeting domain data is that it has a lot of overlapping speech which, if it is not segregated in pre-processing steps, will result into poor clustering which, in turn, will increase diarization error rate. So cluster purification becomes a necessary step to improve the performance of a diarization system. For purifying, cluster purity algorithms are suggested in [4]. There are two algorithms which perform cluster purity at segment and frame level. The first thing to understand the causes of cluster impurity. In meeting data domain primary reason for impurity is that initial cluster created from frame contains speech segments from different speakers. Also, initial clusters contain speech and non-speech

frames (for short silent segments). The first algorithm is a segment level purification which finds faulty segments and split them into new clusters. To do that, first a segment that best represents each model is found by estimating the normalized likelihood. Then, the delta BIC value is calculated between each segment and the best segment. If all the values of delta BIC are greater than a certain threshold, no splitting takes place. Otherwise, segments are split into new clusters. The second algorithm locates the individual frame that is causing cluster impurity and those frames are discarded once they are detected. This algorithm basically aims at finding non-speech frames that dont help in distinguishing speakers, these frames are discarded. To detect the non-speech frames, first silence modeling is carried out and then for each frame likelihood is calculated from silence model. All those frames having the high value of likelihood are discarded. The purified clusters are then clustered using agglomerative clustering.

Apart from state-of-art hierarchical diarization systems that are based on BIC, in the information bottleneck concept [23] is extended to diarization. Information Bottleneck (IB) principle is inspired from rate-distortion theory and aims at finding the most compact notation for input data. IB Clustering method is a bottom-up clustering method. The Information bottleneck states that data can be compressed till it doesnt lose necessary information describing its nature (relevance variables). IB has the segmentation method same as what has been followed in the state-of-art systems (dividing input audio into small chunks such that each chunk contains speech from one speaker). The difference is that once data is segmented then, lets say X denotes the number of data points which are to be clustered into C clusters, all these data points are expected to describe the nature of the data which they are representing.Segments contain speech from one and the only one speaker. So data points are expected to retain this information of the speaker which they are representing even after compression. If they fail to do so, clustering is stopped and the clusters at the point where this information can not be retained anymore represents the optimal number of clusters. So in information bottleneck principle, the information related to the nature of data are represented by a certain set of relevance variables. Let X denotes a set of speaker clusters that are to be clustered into C clusters, let Y denote the relevance variables that describes nature/information about X. So, as per the Information Bottleneck principle, the clustering C should preserve as much as information as possible about the actual data set X with respect to relevance variables Y .IB is based on information distortion theory which finds the compact representation of X such that it minimizes mutual information, $I(X, C)$, (mutual information between X and C) and preserves $I(C, Y)$ (mutual information between C and Y). Thus, the following objective function has to be minimized during IB clustering.

$$I(X, C) - \beta I(C, Y) \tag{5}$$

Where β is the trade-off between information to be preserved and compression of actual data. $I(X, C)$ and $I(C, Y)$ can be represented as

$$I(X, C) - \sum_{x \epsilon X, c \epsilon C} p(x)p(c|x) \log \frac{p(c|x)}{p(c)} \tag{6}$$

$$I(Y, C) - \sum_{y \epsilon Y, c \epsilon C} p(c)p(y|c) \log \frac{p(y|c)}{p(y)} \tag{7}$$

The diarization process starts with over-clustering which divides initial audio into many small chunks of equal size. Then by using Speech Activity Detection, non-speech chunks are discarded. A GMM is trained for each chunk with shared diagonal covariance matrix which sets the definition for set Y (Relevance Variables). Then agglomerative Information Bottleneck clustering [20] is applied which merges two clusters with minimum Jensen-Shannon divergence. These steps are repeated until the optimal number of clusters are obtained. To improve the segmentation, in the very last step, Viterbi segmentation is performed.

The performance of diarization system with meeting domain data is an issue because meeting data involves multiple distant microphones which cause the problem of time-delay of the arrival of sound. Then there are overlapping of speech with non-speech fragments and other noisy sounds. This greatly impacts the performance of diarization. Thus [10],[22],[24] deal with improving the performance of the system by adopting certain changes in the existing approaches to diarization.

In [10], bi-directional segmentation is performed on input data based on Generalized Likelihood Ratio and Bayesian information criterion. Bi-directional segmentation uses variable size window. Initial processing of segmentation is same that is the audio file is divided into various small chunks (2sec). Within each chunk, most probable change point is calculated by using GLR or BIC. So after this step, for each window we have a change point that corresponds to each chunk. Initially, fixed sized windows were used so there were high chances of missing out change points near the boundaries. Therefore, variable size window is used whose length varies according to chunk for which change point is to be calculated. For example, if we are calculating change point for 2nd chunk then from window size will be from the point where most probable change point for the 1st window is calculated and will extend up to the point where most probable change point for the 3rd window is calculated. The change detection method discussed above is then applied once from left to right and once from right to left. Union of the two are taken as the initial clusters for clustering step. Clustering step differs slightly from the one followed in the state-of-art system. Firstly, local clustering is done with 20 segments at a time and after that global clustering is done. After the Local BIC clustering on consecutive 20 segments is performed, Global BIC clustering is performed on the clusters obtained in through Local BIC clustering.

While in [22] focus is on segmentation, in [24] focus is on front-end pre-processing. Weiner Filtering is applied to improve the quality of speech activity detection. Once this is done, 36 MFCC (12 MFCC and its first and second order

derivatives) are calculated. Then SAD is applied within a window of certain size and shift. If after SAD, the segment is still greater than some pre-defined length, m, then in a window slightly more than m the frame with minimum energy is searched and in this way, segments have the size equal to m or less than m. The basic idea is that when one speaker stops and another starts, there is a region where energy will be low and thus that point is a change point. So by using energy as parameter change detection is carried out. Also, it will ensure cluster purity, each segment will have the speech from only one speaker. So the aim of clustering step that follows after this step is to merge clusters which belong to the same speaker. Further purification is carried out by training a GMM with whole data and diagonal covariance matrix and this GMM is called as GMM-root. Now for each initial cluster, a GMM is adapted from GMM-root using feature vector using MAP adaptation which results into the initial number of clusters. All these clusters are merged with the clusters which have high value of likelihood ratio. After this clusters are obtained again they are merged but this time by using BIC criterion. This way of cluster purification improves performance by reducing diarization error rate but requires heavy computation.

In presence of multiple microphones, delay and sum algorithm in [3] is most commonly used. In [24] the use of TDOA features are extended and along with the MFCC features, TDOA features are also used in diarization process. Unlike in most of the approaches where there is only 1 stream in input (MFCC), there are 2 streams of MFCC and TDOA in this approach. MFCC will be responsible for acoustic feature modeling and TDOA will model the time difference between multiple microphones. In baseline HMM/GMM system in presence of multiple streams, the linear combination of log likelihood ratio is used and most feature streams are modeled with different GMMs combined by weighting on their log likelihood.

$$P_{mfcc} \log b_{c_i}^{mfcc}(s_t^{mfcc}) + P_{tdoa} \log b_{c_i}^{tdoa}(s_t^{tdoa}) \quad (8)$$

P_{mfcc} ans P_{tdoa} are the weights of MFCC and TDOA features respectively and $b_{c_i}^{mfcc}(.)$ and $b_{c_i}^{tdoa}(.)$ are GMM models that are estimated separately from MFCC and TDOA feature. From equation (8), it is clear that the overall likelihood ratio is a weighted sum of MFCC and TDOA feature. The rest of the process of diarization follows the state-of-art method. For the calculation of TDOA feature vector, delay and sum (GCC-PHAT) [3] is used. If there are n channels, then the length of TDOA feature vector will be $n-1$ as at each point, one channel is selected as the reference channel. This way TDOA feature can be extended to HMM/GMM approach to diarization. However, these features can also be used in IB diarization system as well. In IB relevance variables are estimated by using Bayes rule. Stopping criterion for clustering is decided by Normalized Mutual Information (NMI). NMI is a fraction of original mutual information I(Y,X) preserved by clustering representation I(Y,C)

$$p(y|s_t^{mfcc}, s_t^{tdoa}) = p(y|s_t^{mfcc})P_{mfcc} + p(y|s_t^{tdoa}P_{tdoa}) \quad (9)$$

Most of the diarization system on meeting domain does not use training data as in meeting domain, it very much possible that training and testing data differ a lot. So training in such cases is a waste of time and it doesn not lead to improved performance but when broadcast and news data is concerned, training is useful. Subspace Gaussian Mixture Model (SGMM) speaker vector and KL-HMM approach to speaker diarization are combined in [16] which uses a priori data and extend IB diarization of [23]. In [23] experiment is conducted with UBM but no significant improvement in the performance could be seen. One of the reasons why UBM failed is that input audio is chopped into small segments in which speech information dominates over speaker information and hence those could not be captured by Universal Background Model. The main purpose of any model in speaker diarization is to capture speaker-related information along with speech. In [16] SGMM is used so that more speaker-related information can be captured. Each segment in segmentation phase is used to estimate one speaker vector from SGMM system and this SGMM system trained on the development set. SGMM is an acoustic modeling method where a common GMM structure is shared across all states. Each state is represented as a vector which maps to mean and weights of a states GMM. There are 2 major differences between GMM and SGMM. The first is that Gaussians are shared across all the states. The second difference is that the covariance matrix (Σ) is shared across all states in SGMM.

$$p(X|j) = \sum_{i=1}^{I} w_{ij} \mathcal{N}(x; \mu_{ji}, \Sigma_i) \quad (10)$$

$$\mu_{ji} = M_i v_j \quad (11)$$

$$w_{ji} = \exp w_i^T v_j / \sum_i^I \exp w_i^T v_j \quad (12)$$

Where I is number of Gaussians in the state and M_i and w_i are globally shared parameters

The aim of SGMM is to distinguish between two speakers. For this, speaker vector extraction process is applied. The mean vector can be split into 2 components: one component is speech specific and another component is speaker-specific.

Meeting domain is majorly affected by the presence of multiple channels which creates diarization for meetings a complex process. Broadcast domain, on the other hand, has dedicated high-quality microphones and thus does not require many pre-processing steps which are used for meeting domain. But in a broadcast domain, where diarization of TV shows have to be performed, the problems of diarization of large corpora comes into the picture. In the case of TV shows, though high-quality microphones are used, the recording of one show can be through a different microphone in a different environment. So the same speaker in one recording has recorded in a different environment or channel and on any other recording, the same speaker has recorded in another different environment or channel. Now in this case, if channel

and environment modeling is not done, then the same speaker in the different episode may be considered different which is not at all desirable. In [12] and [11], methods are described to deal with the diarization of large TV shows. The diarization system proposed has 2 steps, the very first step is diarization of the individual show and the second step is speaker linking where the same speaker across the various recording have to be clustered. Diarization step is usually same as the baseline system but speaker linking step may differ.

The diarization process in [12] is same as the one that was followed in [23],[24] i.e. information bottleneck clustering with one modification. The relevance variables were earlier defined with respect to GMM only but in [12], the relevance variables are defined on GMM-UBM. To model intersession variability between the recordings, Simplified JFA (Joint Factor Analysis) model is adapted. The speaker independent vector is formed by UBM typically trained on many speakers.

$$\hat{m} = m + Vy + Ux \tag{13}$$

Here $\hat{m}$ and m are the speaker adapted and speaker independent Gaussian mean supervectors of GMM, Vy is speaker-dependent low rank term that models speaker variation and Ux is session-dependent low rank term that models session variation.

The diarization is applied individually to each recording. The second phase is speaker linking which identifies speakers uniquely across all recordings. Each speaker cluster is modeled as a one Gaussian with full covariance which is speaker factor posterior distribution estimated by using JFA. Wards method [25] is used for merging the cluster. The two clusters are merged such that there is a minimum increase in within cluster variance after merging. Other methods for estimating cluster dissimilarity metric/score are cosine distance, symmetric KL divergence and 2-way Hotelling t-square. The method in [12] is just the addition of JFA to model session variability. The performance of the system can be improved if two diarization systems are merged/fused. This idea is implemented in [11]. The fusion has an advantage that it lowers diarization error rate at the expense of discarding certain portions of speech which are left out in the final output. The final output contains only those portions of speech where both approaches agree. The approaches used for fusion are baseline system that is, HMM/GMM using BIC criterion and IB diarization. Diarization of each recording is performed individually on HMM/GMM and IB diarization and final output is obtained by fusing them i.e. the parts where both approaches give the same result are taken. In output, there are clusters for all the speakers from all recordings, the next step is to merge clusters belonging to the same speaker. But before clustering, it is required to model the intersession variability which can be modeled by either JFA or approaches like single-factor eigenvector or even i-vectors can be used. The next step in diarization system dealing with large corpora is speaker linking. Agglomerative clustering is followed. Wards method is used to merge clusters. It selects 2 clusters for merging such that there is a minimum increase in the total within-cluster variance. The method discussed so far is able to deal with the portion where two approaches are agreeing on output. However, for parts where they do not agree will reduce the performance. For the left out parts, ergodic HMM is trained by agreed part which is already known as a result of the method discussed so far.

There is a trade-off between the diarization error rate and the computation time. In real time application computation time plays significant role, so diarization error rate within a certain limit is acceptable. In [2], [8], [9] diarization based on the binary key is used that performs diarization faster than the baseline system. In [2] each speaker is represented by a binary vector and this binary vector and as in any diarization system, speaker cluster (here vectors) are merged if they belong to the same speaker. In binary speaker diarization algorithm, there are two main tasks; acoustic block and binary block. In acoustic block binary keys are obtained. But to obtain them, first, features are extracted from the audio file, then a Binary Key background model (KBM), like UBM, is trained. In KBM initial clusters are trained for every 2 sec of input data and with certain overlapping. KBM also consists of anchor speakers which retain the full coverage of acoustic space. Anchor speakers are obtained by Selection algorithm. Each speaker cluster is represented by binary keys, so every speaker binary key is an N-dimensional binary vector where N is the number of anchor speakers. A matrix is constructed consisting of rows equal to the number of feature vectors and columns equal to number of Gaussian components indicating the Gaussian ID of the best matching Gaussian in KBM. The second task is binary key clustering. To cluster a certain metric has to be defined to compare to different vectors. Two different keys can be compared by assigning a similarity score between them.

Agglomerative bottom-up clustering merges the cluster or binary keys until the number of clusters becomes one. The optimal number of speakers is found out by applying t-test metric [18]. This implementation of the binary key is completely novel and is ten times faster than the baseline system and also has comparable performance.

In [8] and [9] method used in [11] are modified or improved. The most computational intensive process in binary key diarization is KBM training. In [11] for KBM training KL2 distance is used which requires high computation efforts as matrix operation like inverse, traces and determinant have to be calculated. Therefore, in [8] and [9], for calculating dissimilarity Gaussian mean vector cosine distance is used.

$$S_{cos}(a,b) = \frac{a.b}{|a|.|b|} \tag{14}$$

$$D_{cos}(a,b) = 1 - S_{cos}(a,b) \tag{15}$$

Here, $S_{cos}(a,b)$ is cosine similarity score.

In clustering, to find the optimal number of speakers instead of t-test, within-cluster sum of square method (WCSS) is used.

Given an initial clustering, WCSS finds how good a clustering solution is. A good cluster will have a low value of WCSS.

$$W(C_k) = \sum_{i,k} \sum_{x \epsilon c_i} ||x - \mu_i||^2 \quad (16)$$

Another problem in extending binary key diarization for cross-show diarization is intersession variability. Thus intersession and intra-speaker compensation are required. Channel compensation techniques like Nuisance Attribute Projection is used [21]. The entire cross-show diarization is done in 3 steps. The first step is to convert cluster obtained by diarization into the binary key which is then followed by inter-session and intra-speaker compensation and finally cross-show clustering is done. KBM for each recording is obtained separately and hence binary keys from two different recordings can not be directly merged. So Global KBM is trained on the whole dataset by using the process of [11].

IV. Conclusion

We looked at several seminal works during the course of this review on speaker diarization. From the analysis, it can be observed that HMM/GMM with BIC as stopping criterion, IB diarization and Binary key diarization are prominent ones that provide good performance. The Binary key diarization is fast, suited for real-time applications though can not be used in a system where very low diarization error rate is a requirement. Hybrid approaches tend to improve the performance of many diarization system as presented in [5], [15], [10], [11]. Investigation can be conducted to combine approaches like IB and binary key diarization. Also, speaker recognition approaches can also be incorporated. As of now, very few methods from speaker recognition domain are used in diarization as most of the approaches are clustering approaches only.

References

[1] Jitendra Ajmera and Chuck Wooters. A robust speaker clustering algorithm. In *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 411–416. IEEE, 2003.

[2] Xavier Anguera and Jean-François Bonastre. Fast speaker diarization based on binary keys. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4428–4431. IEEE, 2011.

[3] Xavier Anguera, C Woofers, and Javier Hernando. Speaker diarization for multi-party meetings using acoustic fusion. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 426–431. IEEE, 2005.

[4] Xavier Anguera, Chuck Wooters, and Javier Hernando. Purity algorithms for speaker diarization of meetings data. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

[5] Claude Barras, Xuan Zhu, Sylvain Meignier, and J-L Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1505–1512, 2006.

[6] Scott Chen and Ponani Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, volume 8, pages 127–132. Virginia, USA, 1998.

[7] Perrine Delacourt and Christian J Wellekens. Distbic: A speaker-based segmentation for audio data indexing. *Speech communication*, 32(1):111–126, 2000.

[8] Héctor Delgado, Xavier Anguera, Corinne Fredouille, and Javier Serrano. Fast single-and cross-show speaker diarization using binary key speaker modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(12):2286–2297, 2015.

[9] Héctor Delgado, Xavier Anguera, Corinne Fredouille, and Javier Serrano. Improved binary key speaker diarization system. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 2087–2091. IEEE, 2015.

[10] Elie El-Khoury, Christine Senac, and Julien Pinquier. Improved speaker diarization system for meetings. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4097–4100. IEEE, 2009.

[11] Marc Ferr, Srikanth Madikeri, Petr Motlicek, et al. System fusion and speaker linking for longitudinal diarization of tv shows. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5495–5499. IEEE, 2016.

[12] Marc Ferras and Hervé Bourlard. Speaker diarization and linking of large corpora. In *Proceedings of the IEEE Workshop on Spoken Language Technology*, pages 280–285, 2012.

[13] Corinne Fredouille, Jean-François Bonastre, and Teva Merlin. Amiral: a block-segmental multirecognizer architecture for automatic speaker recognition. *Digital Signal Processing*, 10(1):172–197, 2000.

[14] Herbert Gish and Michael Schmidt. Text-independent speaker identification. *IEEE signal processing magazine*, 11(4):18–32, 1994.

[15] Vishwa Gupta, Patrick Kenny, Pierre Ouellet, Gilles Boulianne, and Pierre Dumouchel. Combining gaussianized/non-gaussianized features to improve speaker diarization of telephone conversations. *IEEE Signal processing letters*, 14(12):1040–1043, 2007.

[16] Srikanth Madikeri, Petr Motlicek, and Hervé Bourlard. Combining sgmm speaker vectors and kl-hmm approach for speaker diarization. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4834–4838. IEEE, 2015.

[17] Daniel Moraru, Sylvain Meignier, Corinne Fredouille, Laurent Besacier, and Jean-François Bonastre. The elisa consortium approaches in broadcast news speaker segmentation during the nist 2003 rich transcription evaluation. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–373. IEEE, 2004.

[18] Trung Hieu Nguyen, Engsiong Chng, and Haizhou Li. T-test distance and clustering criterion for speaker diarization. In *INTERSPEECH*, pages 36–39. Citeseer, 2008.

[19] Matthew A Siegler, Uday Jain, Bhiksha Raj, and Richard M Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA speech recognition workshop*, volume 1997, 1997.

[20] Noam Slonim, Nir Friedman, and Naftali Tishby. Agglomerative multivariate information bottleneck. In *Advances in neural information processing systems*, pages 929–936, 2001.

[21] Alex Solomonoff, Carl Quillen, and William M Campbell. Channel compensation for svm speaker recognition. In *Odyssey*, volume 4, pages 219–226. Citeseer, 2004.

[22] Hanwu Sun, Bin Ma, Swe Zin Kalayar Khine, and Haizhou Li. Speaker diarization system for rt07 and rt09 meeting room audio. In *ICASSP*, pages 4982–4985, 2010.

[23] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard. Agglomerative information bottleneck for speaker diarization of meetings data. In *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, pages 250–255. IEEE, 2007.

[24] Deepu Vijayasenan, Fabio Valente, and Hervé Bourlard. An information theoretic combination of mfcc and tdoa features for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):431–438, 2011.

[25] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[26] Chuck Wooters, James Fung, Barbara Peskin, and Xavier Anguera. Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system. In *RT-04F Workshop*, volume 23, page 23, 2004.