

USC-SIPI REPORT #402

Robust Speaker Clustering Under Variation in Data Characteristics

by

Kyu Jeong Han

December 2009

**Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.**

ROBUST SPEAKER CLUSTERING
UNDER VARIATION IN DATA CHARACTERISTICS

by

Kyu Jeong Han

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(ELECTRICAL ENGINEERING)

December 2009

Dedication

This dissertation is dedicated with loving appreciation to my family for their endless love and support. Without them this work would never have been achieved.

Acknowledgements

First of all I would like to thank my advisor, Dr. Shrikanth S. Narayanan, for his inspiring and encouraging way to guide me to a deeper understanding of research work, and his invaluable comments during my graduate years. He taught me how to formulate problems and solve them with reasonable thinking, and how to express ideas and results efficiently. On top of all these, I really appreciate his patience and belief in me throughout my entire years as his students in Signal Analysis and Interpretation Laboratory (SAIL).

I also would like to thank the rest of my thesis committee: Dr. C.-C. Jay Kuo, Dr. Hong-Goo Kang, and Dr. Cyrus Shahabi, who reviewed my work and gave insightful comments. My special thanks go to Dr. Kang, who gave us a favor of accepting our invitation to this committee while he was working for Broadcom in Irvine, CA for his sabbatical year.

To my family, I appreciate and love you with all my heart. Without your love and unwavering belief in me it would have been impossible for me to complete Ph.D. work at USC. Specially I would like to thank my wife and daughter. I could not have overcome even a bit of huddle without them in the entire life in the United States. Seo Jung, my lovely wife offered by gracious God, you are the only one that deserves to share all my achievements during my USC years and even for the rest of my life. I LOVE YOU. My precious daughter, Jimin, I want you to know that you have motivated me as well as your mom to keep hard working on every side of our life since you were given to us. Please you never forget you are more than God's gift to us.

Finally, I would like to thank you, my Lord, for your guidance of my life and family through perseverance to know and enjoy you in the name of Jesus Christ.

Table of Contents

Dedication	ii
Acknowledgements	iii
List of Tables	vii
List of Figures	x
List of Algorithms	xiv
Abstract	xv
1. Introduction	1
1.1 Motivation	1
1.1.1 Pattern Classification and Clustering	2
1.1.2 Speaker Clustering	4
1.1.3 Focus Identification	4
1.2 Previous Work on Speaker Clustering	5
1.3 Problem Statement	7
1.4 Proposed Approaches	8
1.4.1 Perspective 1: Stopping Point Estimation	9
1.4.2 Perspective 2: Inter-Cluster Distance Measurement	10
1.4.2.1 Earlier Recursion Steps of AHSC	11
1.4.2.2 Later Recursion Steps of AHSC	11
1.4.2.3 Cluster Modeling	12
1.4.3 Application: Speaker Diarization	12
1.5 Contribution Summary	13
1.6 Dissertation Outline	14
2. Robust Stopping Point Estimation for AHSC	16
2.1 Introduction	16
2.2 Data and Experimental Setup	17
2.3 BIC-based Stopping Point Estimation for AHSC	18
2.3.1 Generalized Likelihood Ratio (GLR)	19
2.3.2 Bayesian Information Criterion (BIC)	24
2.3.3 BIC-based Stopping Point Estimation Method for AHSC	25

2.3.4	Tuning Parameter λ	28
2.3.5	Stopping Criterion under the Variation of Input Speech Data	29
2.4	ICR-based Stopping Point Estimation for AHSC	31
2.4.1	Information Change Rate (ICR)	31
2.4.2	Comparison of ICR with ICR-like Measures	33
2.4.3	ICR as a Homogeneity Decision Measure for Clusters	33
2.4.4	ICR-based Stopping Point Estimation Method for AHSC	34
2.5	Conclusions	38
3.	Robust Inter-Cluster Distance Measurement for AHSC	40
3.1	Introduction	40
3.2	GLR at Early AHSC Recursion Steps	41
3.3	Modification of AHSC	44
3.3.1	Constrained Cluster Selection for Merging	45
3.3.2	Pre-Classification of Short Speech Segments	47
3.3.3	Sequential Clustering prior to AHSC	48
3.4	Combination of GLR and ICR	49
3.4.1	(GLR+ICR)-based Inter-Cluster Distance Measurement	51
3.4.2	Proposed Measure in Modified AHSC Approaches	54
3.5	Selective AHSC	55
3.5.1	Modified AHSCs with Stopping Point Estimation	56
3.5.2	Selective AHSC	57
3.6	Conclusions	60
4.	Robust Cluster Modeling for Inter-Cluster Distance Measurement in AHSC	64
4.1	Introduction	64
4.2	Inter-Cluster Distance Measurement for AHSC	66
4.2.1	GLR-based statistical inter-cluster distance measurement	66
4.2.2	Conventional cluster modeling approaches	67
4.2.2.1	Single Gaussian cluster modeling	67
4.2.2.2	GMM cluster modeling	69
4.2.2.3	Experimental comparison	72
4.3	Incremental Gaussian Mixture Cluster Modeling	77
4.3.1	Proposed cluster modeling approach	77
4.3.2	Comparison and analysis	79
4.4	Conclusions	83
5.	Reliable Speaker Diarization based on Robust AHSC	84
5.1	Introduction	84
5.2	Speaker Diarization	85
5.3	SAIL Speaker Diarization System	87
5.3.1	Data Description and Experimental Setup	87
5.3.2	Speech/Non-Speech Detection	88
5.3.3	Speaker Change Detection	91
5.3.4	Speaker Clustering	91
5.3.4.1	IGMM Cluster Modeling	92

5.3.4.2	ICR-based Stopping Point Estimation	94
5.3.4.3	Comparison	96
5.3.5	Experimental Results	96
5.4	Refined Speaker Clustering	97
5.4.1	Selection of Representative Speech Segments	98
5.4.2	Participant Interaction Pattern Modeling	101
5.4.3	Experimental Results	103
5.5	Conclusions	105
6.	Conclusions	106
6.1	Contributions	106
6.2	Possible Future Research Topics	107
6.3	Final Remarks	109
	Bibliography	110

List of Tables

2.1	Development set of data sources. N_s : # of speaker identities (male:female) in each data source, T_s : total utterance time (sec.), N_t : # of speech segments, and T_a : average segment length (sec.). C , N , and I : data sources chosen from ICSI, NIST, and ISL meeting speech corpora respectively. . .	17
2.2	Evaluation set of data sources. The notation is same as that in Table 2.1.	18
2.3	Comparison of ICR with other measures utilizing the idea of normalizing GLR. C_x and C_y : two clusters consisting of M and N feature vectors respectively, α : parameter empirically determined, and n : dimension of feature vectors.	33
2.4	ICR-based stopping point estimation method vs. BIC-based stopping point estimation method. $c = \frac{1}{2} \{n + \frac{1}{2}n(n + 1)\}$, where n is the dimension of feature vectors. $n = 12$, $\eta = 0.18603$, and $\lambda = 12.0$	36
2.5	Global comparison (averaged speaker error time rate for the evaluation data set) of AHSC with the BIC-based stopping point estimation method and AHSC with the ICR-based stopping point estimation method	39
3.1	Distribution of three different merging types (M_{ss} , M_{sl} , and M_{ll}) at the first quarter of the entire merging recursions during AHSC for every data source in the development data set in Section 2.2. M_{ss} : merging between the speech segments shorter than 3 seconds, M_{sl} : merging between one speech segment shorter than 3 seconds and the other longer than or equal to 3 seconds, and M_{ll} : merging between the speech segments longer than or equal to 3seconds.	43
3.2	Accuracy of the GLR-based inter-cluster distance measure for AHSC depending on the merging types defined in Table 3.1. These accuracies were obtained only based on the first quarter of the entire merging recursions during AHSC for every data source in the development data set in Section 2.2.	43

3.3	Comparison of basic AHSC and its first modified version in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. Both of the clustering strategies use the GLR-based inter-cluster distance measure to select clusters for merging at every recursion step of AHSC, and perfect stopping point estimation is assumed. (For each result in the table, the corresponding standard deviation is presented as well.)	46
3.4	Comparison of basic AHSC and its second modified version in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. The same distance measure and assumption for stopping point estimation in AHSC as ones in Table 3.3 are applied.	47
3.5	Comparison of basic AHSC and its third modified version in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. The same distance measure and assumption for stopping point estimation in AHSC as ones in Table 3.3 are applied.	50
3.6	Comparison of AHSC with the GLR-based inter-cluster distance measure and that with our proposed method, in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. Perfect stopping point estimation for AHSC is assumed.	54
3.7	Average speaker error time rate for the evaluation data set in Section 2.2. This table compares AHSC and its three modified versions with both GLR-based and (GLR+ICR)-based inter-cluster distance measurement. Perfect estimation of the optimal stopping point for AHSC is assumed.	56
3.8	Average speaker error time rate for the evaluation data set in Section 2.2 when the ICR-based stopping point estimation method is applied. This table compares AHSC and its three modified versions with GLR-based and (GLR+ICR)-based inter-cluster distance measurement, as Table 3.7 does.	56
3.9	Average speaker error time rate for the evaluation data set in Section 2.2. This table compares selective AHSC with all the counterparts that have been dealt with in this dissertation.	63
4.1	Data source. N_s : number of speaker sources (male:female), T_s : total utterance time (sec.), N_t : number of speech segments, and T_a : average segment length (sec.).	73
4.2	Performance comparison of the two conventional cluster modeling approaches in terms of speaker error time rate (%). \mathcal{N} : single Gaussian cluster modeling and λ_x : GMM cluster modeling with x mixture components.	74
5.1	Training data set.	88
5.2	Testing data set.	88

- 5.3 Performance comparison of the proposed speech/non-speech detection process with and without updating the silence cluster, in terms of the two detection error rates for the training data set. 90
- 5.4 Comparison of 1) IGMM cluster modeling + ICR-based stopping point estimation, and 2) single Gaussian cluster modeling + BIC-based stopping point estimation, in terms of speaker-error-time rate for the testing data set. $\lambda = 25.0$ (for BIC-based stopping point estimation) and $\eta = 0.225$ (for ICR-based stopping point estimation), which are tuned based on the training data set. 95
- 5.5 Improved speaker diarization performance with the two approaches proposed in this section, i.e., representative speech segment selection and participant interaction pattern modeling. For the former approach we empirically set $N = 32$. Performance comparison is given in terms of average DER (%) across 10 data sources in the testing data set in Section 5.3.1. 103

List of Figures

1.1	Categorization of automatic pattern classification systems.	3
1.2	Application domains of speaker clustering.	5
1.3	Unreliable speaker clustering performance by AHSC across various input speech data. The entire data are 10 sets of segmented meeting conversations speech and each of them contains a number of speech segments in an utterance or sentence level. In each speech segment speaker-specific characteristics are homogeneous.	8
1.4	Overview illustration of dissertation contributions to robust AHSC and speaker diarization in terms of the variation of input speech data.	9
2.1	GLR for two clusters C_1 and C_2 along with the number of feature vectors in each cluster. The second order statistics of the corresponding cluster models are fixed at $\mu_1 = 0$, $\mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$	21
2.2	Comparison of speaker clustering performance (for the evaluation data set described in Section 2.2) with and without accurate stopping point estimation. For the BIC-based stopping point estimation method, we tuned λ to be 12.0. Average speaker error time rate degradation by incorrect estimation of the optimal stopping point is about 9.65% (absolute) per data source.	28
2.3	\ln GLR and $\ln(M + N)$ ($= \ln(N_1 + N_2)$ in this case) for the same clusters considered in Figure 2.1 along with the number of feature vectors in each cluster, with the fixed second order statistics of $\mu_1 = 0$, $\mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$	30
2.4	Distributions for correct and incorrect merging in terms of ICR. The threshold η is set so as to minimize classification error between the two distributions. The distributions were obtained based on our development data set, and feature vectors in every cluster considered corresponded to more than 30 seconds in amount of time.	35

2.5 \ln GLR, $\text{Th}_{\text{BIC}} = \lambda \cdot c \cdot \ln(M + N)$, and $\text{Th}_{\text{ICR}} = \eta \cdot (M + N)$ for C-6, where $\lambda = 12.0$ and $\eta = 0.18603$. The stopping point estimated by the ICR-based stopping point estimation method is identical to the optimal one in this case. 37

2.6 Comparison of speaker clustering performance for the evaluation data set with accurate stopping point estimation and with the ICR-based stopping point estimation method, for which $\eta = 0.18603$. Average speaker error time rate degradation by incorrect estimation of the optimal stopping point is less than 1% (absolute) per data source. 38

3.1 Figure 2.1 revisited. This figure displays GLR for two clusters C_1 and C_2 along with the number of feature vectors in each cluster. The second order statistics of the corresponding cluster models are fixed at $\mu_1 = 0$, $\mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$ 42

3.2 Segment length distributions for the development data set in Section 2.2. 44

3.3 Speaker error time rate by AHSC with perfect detection of the optimal stopping point for the development data set in Section 2.2. This figure compares performance for the entire speech segments given as an input to AHSC with that for the corresponding subset containing the segments longer than or equal to 3 seconds only. 45

3.4 Comparison of basic AHSC and its three modified versions proposed in terms of average speaker error time rate. 51

3.5 Soft ranking used in the proposed inter-cluster distance measurement method. If a certain pair of clusters have the normalized distance of 0.5, their soft ranking becomes 0.69 (grey line) in this system. 53

3.6 Extra performance improvement achieved if the proposed (GLR+ICR)-based inter-cluster distance measure were applied to the late recursion steps of the three modified versions of AHSC introduced in Section 3.3. The data sources used in this experiment are the development and evaluation data set in Section 2.2. Perfect estimation of the optimal stopping point for AHSC is assumed. 55

3.7 Figure 3.3 revisited, showing speaker error time rate by AHSC with perfect detection of the optimal stopping point for the development data set in Section 2.2. This figure compares performance for the entire speech segments given as an input to AHSC with that for the corresponding subset containing the segments longer than or equal to 3 seconds only. 58

3.8 Comparison of basic AHSC with the assumption of perfect estimation of the optimal stopping point and selective AHSC (including the BIC-based stopping point estimation method), in terms of speaker error time rate on the evaluation data set in Section 2.2. 60

3.9	Comparison of basic AHSC with the assumption of perfect estimation of the optimal stopping point and selective AHSC (including the ICR-based stopping point estimation method), in terms of speaker error time rate on the evaluation data set in Section 2.2.	61
4.1	Effectiveness of the EM procedures in the GMM cluster modeling approach for GLR-based inter-cluster distance measurement. Each subfigure compares distances between two pairs of clusters along with the number of iterations in the EM procedures for GMMs with 16 mixture components. One pair comes from the same speaker source (black curve) while the other is from different sources (grey curve).	71
4.2	Processing time comparison of the two conventional cluster modeling approaches. (For the GMM approach, four different mixture numbers are compared, i.e., 4, 8, 16, and 32.) (a) Full-shot version. (b) Zoomed-in version.	75
4.3	Clustering performance variation for Data 13 in the GMM cluster modeling approach. The circles denote speaker error time rates for the respective 10 sessions of the GMM approach with four different numbers of mixture components (i.e., 4, 8, 16, and 32), and the bold crosses are the corresponding mean values. The horizontal line presents the speaker error time rate obtained from the single Gaussian cluster modeling approach, which is 23.4%.	76
4.4	Performance comparison of the proposed and two conventional cluster modeling approaches in terms of speaker error time rate (%). For this comparison, the best performance of the GMM approach for each data source was chosen among the 4 candidates (4, 8, 16, and 32 mixture components).	79
4.5	Processing time comparison of the proposed and two conventional cluster modeling approaches. (For the GMM approach, four different mixture numbers are compared, i.e., 4, 8, 16, and 32.) (a) Full-shot version. (b) Zoomed-in version.	80
5.1	Speaker diarization: (a) Block diagram of a speaker diarization system. (b) Step-by-step graphical interpretation of how a given audio source is transcribed (in terms of “who spoke when”) by speaker diarization.	85
5.2	Performance of the proposed SAIL speaker diarization system on non-overlapped speech in the testing data set, in terms of DER.	97
5.3	IGMM cluster modeling. $\{C_i\}_{i=1}^5$ are initial clusters for AHSC, and a and b ($a + b = 1$) are weights for the respective constituent GMMs. The weights are determined by the cardinalities of $\{C_1, C_2, C_3\}$ and $\{C_4, C_5\}$, respectively. This figure illustrates how IGMMs grow through merging during AHSC.	98

5.4 Selection of representative speech segments for improved IGMM cluster modeling. In this case, C_2, C_4 , and C_5 are selected as representative speech segments to model $\{C_i\}_{i=1}^5$ 101

5.5 1st-order Markov chain model for participant interaction patterns when the estimated number of speakers is 4, where p_{ij} is the transition probability from the speaker S_i to the speaker S_j for $1 \leq i, j \leq m$ ($m = 4$ in this case). 101

5.6 Performance of the modified SAIL speaker diarization system on non-overlapped speech in the testing data set, in terms of DER. 104

List of Algorithms

1	Agglomerative Hierarchical Speaker Clustering (AHSC)	6
2	Modified Version 1 of AHSC	46
3	Modified Version 2 of AHSC	47
4	Modified Version 3 of AHSC	48
5	AHSC with combination of GLR and ICR as an inter-cluster distance measure	52
6	Selective AHSC	59
7	Leader-Follower Clustering (LFC)	89
8	Agglomerative Hierarchical Speaker Clustering (AHSC) revisited	92

Abstract

Speaker clustering refers to a process of classifying a set of input speech data (or speech segments) by a speaker identity in an unsupervised way, based on the similarity of speaker-specific characteristics between the data. The process identifies the speech segments of the same speaker source without any prior speaker-specific information of the given input data. This speaker-perspective, unsupervised classification of speech data can be applied as a pre-processing step to speech/speaker recognition or multimedia data segmentation/classification in various ways. Thus, speaker clustering has been recently attracting much attention in the research area of speech recognition and multimedia data processing.

One big, yet unsolved, issue in the research field of speaker clustering is *unreliable clustering performance under the variation of input speech data*. In this dissertation, we deal with this problem in the framework of agglomerative hierarchical speaker clustering (AHSC) in two perspectives: stopping point estimation and inter-cluster distance measurement. In order to improve the robustness of stopping point estimation for AHSC under the variation of input speech data, we propose a new statistical measure called *information change rate* (ICR), which can improve estimation of the optimal stopping point. The ICR-based stopping point estimation method is not only empirically but also theoretically verified to be more robust to the variation of input speech data than the conventional BIC-based method. In order to improve the robustness of inter-cluster distance measurement for AHSC under the variation of input speech data, we also propose *selective AHSC* and *incremental Gaussian mixture cluster modeling*.

These two approaches are proven to provide much more reliability for speaker clustering performance under the variation of input speech data.

Based on these results on robust speaker clustering under the variation of input speech data, we extend our interest to implementing a more robust *speaker diarization* system to the variation of input audio data. (Speaker diarization refers to an automated process that can annotate a given audio source in terms of “who spoke when”.) Focusing on speaker diarization of meeting conversations speech, we propose two refinement schemes to further improve the reliability of speaker clustering performance in the framework of speaker diarization under the variation of input audio data. One is *selection of representative speech segments* and the other is *interaction pattern modeling between meeting participants*, and both of them are experimentally verified to enhance the reliability of speaker clustering performance and hence improve the overall diarization accuracy under the variation of input audio data.

Chapter 1

Introduction

1.1 Motivation

Pattern classification refers to a process, by not only human beings but also machines, that categorizes information into pre-defined classes or classifies similar information among what is given to handle without any prior knowledge. This process is very common and we can count a number of examples inside/around us. A typical example can be found from the learning/cognitive systems of a human brain. With the help of various built-in pattern classification systems functioning in one's brain, he/she identifies who is the person to wake him/her up every morning, verifies if the car being towed across the street is his/hers or not, and recognizes that the music coming from a radio station is one of his/her favorites. We can also distinguish a mother and her son correctly out of unknown people based on their physical and behavioral resemblance detected by our brain systems, although we have never seen and known about them. For other instances, we can bring many state-of-the-art pattern recognition machines mimicking human brain functions, which are currently in a wide service in our daily life as a variety of forms, such as security-purposed biometrics, speech recognition solutions, and data mining applications. From these machine recognition systems, we obtain huge benefit in terms of both convenience and efficiency.

From an engineering point of view, pattern classification in general means a process by machines based on understanding of human pattern classification¹. In this regard, we, pattern classification engineers, have been trying to further expand the territory of relevant application domains beyond the currently deployed service areas including what has been mentioned above, e.g., biometrics, and there is still a significant amount of pattern classification research work actively going on around the world. Such a vast research effort can more enrich our future life as it has given us a lot of benefit thus far.

1.1.1 Pattern Classification and Clustering

Automatic pattern classification systems can be categorized into being supervised or unsupervised [18], [75]. In supervised classification, there are the respective class labels available for a part of the entire data set given to be classified, so we can utilize such a portion of labeled data to train a classifier. Then we can identify which class the rest of the data belong to, respectively, based on the trained classifier. On the other hand, unsupervised classification, also called *clustering*, is used when such a high-level information as class labels is not available for the given data set. In this case, the only way to classify the data is to measure their proximity, e.g., similarity or dissimilarity, and is to decide which data points or clusters belong to the same class based on the measured proximity. Since there is no prior class-dependent information available for the given data set, clustering is generally considered as a more challenging task than supervised classification. Clustering systems can be further categorized into being partitional and hierarchical. The former is used when it is known how many classes there exist in the given data set, while the latter is considered for the case that there is no such information at hand. The categorization of pattern classification systems is depicted in Figure 1.1.

Compared to supervised classification, clustering has attracted relatively more attention in recent years mainly due to information overload [46], [7]. As an enormous amount of new information are poured every moment out of various mass media and most of them

¹This is a research topic usually conducted by cognitive science rather than by engineering.

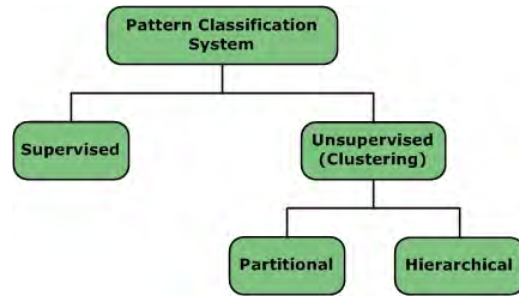


Figure 1.1: Categorization of automatic pattern classification systems.

are accessible because of the continuing growth of the Internet as well as the World Wide Web, there emerges a need for not only cost-effective but also time-efficient technologies that can handle the whole available information properly [41, 49, 56]. With most of the available information being stored as electric forms of data nowadays, data classification is necessary as the very first step for a proper information handling, and clustering offers such a required functionality. According to [18], there are a number of advantages in clustering in the era of information overload, some of which is listed as follows:

- It is almost impossible in practice to always guarantee an enough amount of labeled data for classification, because labeling a large amount of data might cost too much time/effort and be prohibitive in some applications.
- In the case that there existed the temporal dynamics of data patterns and we needed to consider them over time for a classifier, clustering could be applied to tracking such changes and possibly improve the overall classification performance based on it, which is hardly available in supervised classification.
- Clustering can provide a form of data-dependent “smart pre-processing” such as smart feature extraction. For this purpose, unsupervised data analysis by clustering could be utilized to gain some insight into the nature or structure of the given data set.

1.1.2 Speaker Clustering

As multimedia data, e.g, short environmental audio clips or a complex mixture of aural and visual sources such as movies and TV broadcast news, exponentially increase in number these days, how to properly classify and process a considerable amount of audio recordings, especially speech portions, becomes a critical topic. This brought data clustering concept into the research field of speech signal processing. There are various criteria that can be considered in terms of clustering speech data, such as gender, emotion, topic or genre, and so on. Such major criteria for speech data clustering also include speaker identity, based on which we can classify speech data by speaker-specific characteristics. This classification process is called *speaker clustering*.

Specifically, speaker clustering refers to the process of classifying a set of input speech data (or speech segments) by speaker identity in an unsupervised way, based on measuring the similarity of speaker-specific characteristics between the data. The process identifies which speech segments belong to the same speaker source without any prior speaker-specific information of the given input data. This speaker-perspective, unsupervised classification of speech data can be applied as a pre-processing step to speech/speaker recognition or multimedia data segmentation/classification in various ways. For instance, speaker clustering can provide speech recognition of a spontaneous conversation recording with unsupervised speaker adaptation capability, combined with speaker-specific segmentation of the recording. (Its possible application domains are further shown in Figure 1.2.) For this reason speaker clustering has been recently attracting much attention in the research area of speech recognition and multimedia data processing.

1.1.3 Focus Identification

Based on the aforementioned general benefit from pattern classification research and current importance of speaker clustering in speech signal and multimedia data processing, in this dissertation, we focus our research effort on speaker clustering and its relevant issues.

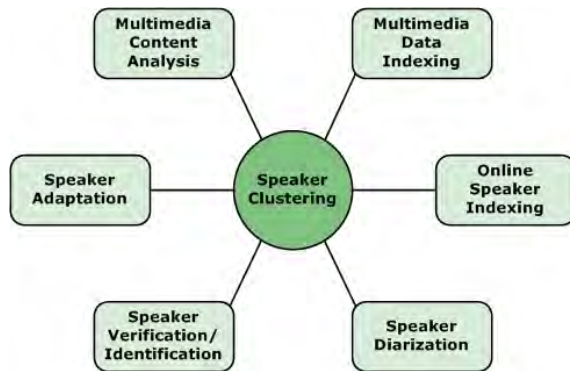


Figure 1.2: Application domains of speaker clustering.

1.2 Previous Work on Speaker Clustering

Since a simple speaker clustering framework based on a hierarchical approach² was introduced in early 1990s by Gish, *et al.* [26], there have been a lot of research work on speaker clustering thus far. Most of the work were initially relevant to broadcast news transcription systems [6, 9, 10, 12, 15, 19–25, 27, 28, 34–39, 43, 44, 47, 48, 57, 59, 61, 67–71, 74, 76, 77]. Speaker clustering was utilized and had been developed in those systems to improve the accuracy of speech recognition for transcription of broadcast news audio data, enabling to adapt original phoneme models (for speech recognition) based on unsupervised speaker-specific data classification results. As general interest in the research field of speech transcription and understanding moves from scripted/read speech data toward more challenging data domains like spontaneous meeting conversations, speaker clustering now gets more important. In addition, a number of current and potential systems for multimedia content analysis and data indexing requires speaker-specific information of multimedia data as a necessary prior knowledge to semantically understand the whole

²For your reminder, hierarchical clustering is utilized when there is no prior information of data at all including the number of data classes, as mentioned in Section 1.1.1. Since speaker clustering is mainly applied to applications where there is no available speaker-specific information of given speech data such as the number of speaker sources, hierarchical approaches are more natural to be considered in speaker clustering research than partitional ones.

Algorithm 1 Agglomerative Hierarchical Speaker Clustering (AHSC)

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: speech segments

$\hat{C}_i, i = 1, \dots, \hat{n}$: initial clusters

Ensure: $C_i, i = 1, \dots, n$: finally remaining clusters

1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$

2: **do**

3: $i, j \leftarrow \arg \min d(\hat{C}_k, \hat{C}_l), k, l = 1, \dots, \hat{n}, k \neq l$

4: merge \hat{C}_i and \hat{C}_j

5: $\hat{n} \leftarrow \hat{n} - 1$

6: **until** speaker clustering performance is estimated to have reached the lowest level

7: **return** $C_i, i = 1, \dots, n$

content of the data, so speaker clustering (being combined with speaker-specific segmentation) becomes into the spotlight more than ever. The Rich Transcription (RT) event that has been annually offered as one of mainstream benchmark tests since 2002 by the National Institute of Standards and Technology (NIST), therefore, includes speaker diarization³ system evaluation, considering it to be one of its main evaluation categories.

A typical strategy for speaker clustering is an agglomerative hierarchical approach [6, 9, 10, 12, 13, 15, 18, 19, 21–23, 25–27, 35, 43, 59, 61, 69, 74, 76, 77], which we usually call *agglomerative hierarchical speaker clustering* (AHSC). This strategy is considered as the best one for speaker clustering tasks, due to its simple processing structure and acceptable level of performance (although it is sub-optimal). Algorithm 1 shows how it works. It considers input speech data (or segments) as individual initial clusters and, at every recursion step, merges the closest pair of clusters in terms of speaker-specific characteristics among the entire candidate pairs. Such recursions continue until a certain stopping point where it is decided that an additional merging would not improve speaker clustering performance any further.

AHSC, since its prototypical introduction in [26], has evolved in terms of two main perspectives, as follows:

³Speaker diarization is an extended version of speaker clustering, referring to a process that divides and classifies speech data by speaker-specific characteristics and, as a result, can annotate the data in terms of “who spoke when” [63]. This process includes speaker-specific segmentation before speaker clustering, but the latter plays a much more critical role in the entire process than the former.

1. How to estimate when speaker clustering performance reaches the lowest level?
2. How to select the most homogeneous clusters (in terms of speaker-specific characteristics) for merging at every recursion step so as to achieve the minimum possible level of speaker clustering performance overall?

Toward addressing the first question, a stopping point estimation method based on *Bayesian information criterion* (BIC) [58] is now widely used as a standardized approach. It was introduced in 1998 by Chen and Gopalakrishnan [13], and since then most of speaker clustering applications have utilized it to estimate the optimal recursion stopping point in AHSC. In order to tackle the second question in the state of the art, on the other hand, *generalized likelihood ratio* (GLR) [26] has been popularly utilized. This statistical distance measure between speech data was empirically verified and is thus considered to be the best solution in terms of properly selecting the closest pair of clusters for merging during AHSC [63], among candidate measures including single/complete/average linkage, Euclidean/Mahalanobis distance or Kullback-Leibler divergence. A basic AHSC framework with these two schemes for stopping point estimation and inter-data (or inter-cluster) distance measurement is broadly adopted by many state-of-the-art speaker diarization systems [1, 3–5, 33, 42, 50, 54, 55, 65, 72, 73, 78].

1.3 Problem Statement

Despite its broad adoption in current state-of-the-art speaker clustering applications, AHSC has a big, yet unsolved, issue in its performance in terms of *robustness to the variation of input speech data* [63], [30]. We can clearly see the huge, negative effect of this issue on AHSC performance in Figure 1.3, which shows baseline experimental results⁴ for AHSC with BIC-based stopping point estimation and GLR-based inter-cluster

⁴AHSC performance was measured by speaker error time rate, which is one of official performance measures for speaker diarization, particularly for speaker clustering, in the NIST RT evaluation. The measurement tool is freely available at <http://www.nist.gov/speech/tests/rt/2006-spring>. We will discuss more in detail about this measure in subsequent chapters.

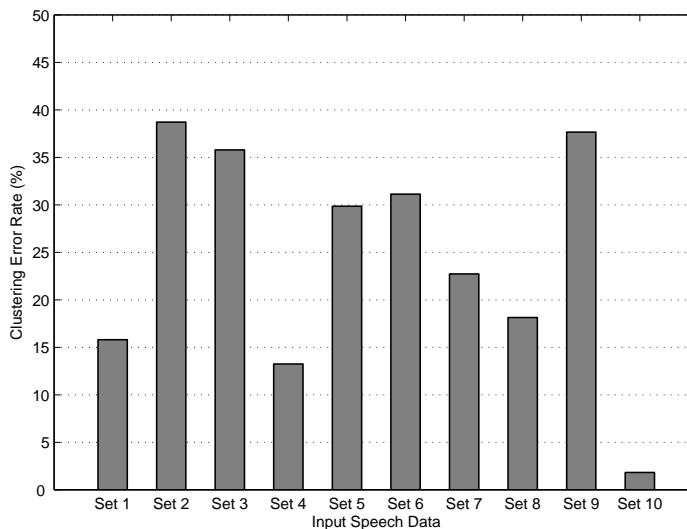


Figure 1.3: Unreliable speaker clustering performance by AHSC across various input speech data. The entire data are 10 sets of segmented meeting conversations speech and each of them contains a number of speech segments in an utterance or sentence level. In each speech segment speaker-specific characteristics are homogeneous.

distance measurement. While the performance for Set 10 is less than 5%, which is quite good, the performances for Sets 2, 3, and 9 are all more than 35%. The absolute difference is roughly over 30%, which is undesirable.

This unreliability problem in AHSC performance is caused because both of the BIC-based stopping point estimation method and the GLR-based inter-cluster distance measure are much influenced by the variation of input speech data. In this dissertation, we address this problem not only by analyzing its causes but also by proposing various algorithmic solutions to them. The next section is a brief list of our proposed approaches, which will be more explained later throughout the dissertation, respectively.

1.4 Proposed Approaches

An overview of our approaches to tackle the aforementioned unreliability problem in AHSC performance is shown in Figure 1.4. They can be categorized into two perspectives: stopping point estimation (the leftmost column of the upper figure) and inter-cluster

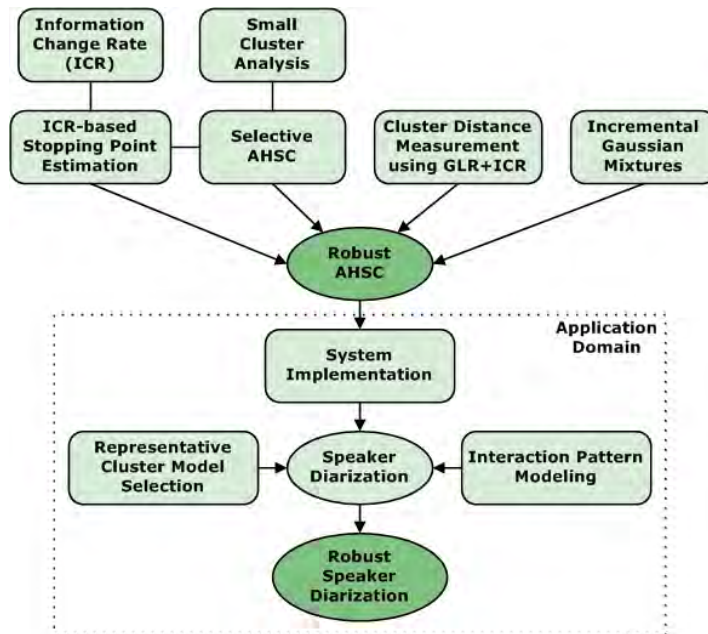


Figure 1.4: Overview illustration of dissertation contributions to robust AHSC and speaker diarization in terms of the variation of input speech data.

distance measurement (the rest of the upper figure). The proposed approaches are later applied to one of promising applications in the research field of speaker clustering, i.e., speaker diarization. In the framework of speaker diarization, we further propose a few refinement schemes for speaker clustering performance (the lower figure).

1.4.1 Perspective 1: Stopping Point Estimation

The conventional BIC-based stopping point estimation method for AHSC is not robust to the variation of input speech data, i.e., does not provide every input data set with reliable estimation of the optimal stopping point where clustering performance would not be improved any further with extra merging. A main reason for this robustness problem

in the method is that the stopping criterion used in the method is too sensitive to the variability of the following characteristics across input speech data:

- Total amount of time for the entire speech utterances in a given input data set
- Utterance time distribution over speaker sources, etc.

In order to improve the robustness of stopping point estimation for AHSC, we propose a new statistical measure, called *information change rate* (ICR), that can help better and more robustly estimating the optimal stopping point. The ICR-based stopping point estimation method is not only empirically but also theoretically verified to be more robust to the variation of input speech data than the conventional BIC-based one. We will take care of this subject in more detail in Chapter 2.

1.4.2 Perspective 2: Inter-Cluster Distance Measurement

As the BIC-based stopping point estimation method, the conventional GLR-based inter-cluster distance measure for AHSC does not provide reliable performance for every input speech data set. Specifically speaking, the GLR-based measure incorrectly selects a pair of heterogeneous clusters (in terms of speaker-specific characteristics) for merging at some recursion steps of AHSC, which causes the overall AHSC performance to degrade severely. A main reason for this problem is that the reliability of the GLR-based measure is affected by the variability of the following characteristics across input speech data:

- Utterance time distribution over speaker sources in a given input data set
- Total number of speech segments
- Average/individual time length per segment, etc.

In order to address this problem, we have three different viewpoints on it. The next three sub-sections show each viewpoint in the order.

1.4.2.1 Earlier Recursion Steps of AHSC

According to [29–32, 40, 66], a more specified reason for this unreliability problem in the GLR-based inter-cluster distance measurement is that GLR tends to get larger in proportion to the size of a cluster pair under consideration. As a result, the GLR-based measure has the following undesirable patterns:

- A pair of homogeneous clusters (in terms of speaker-specific characteristics) of small size might have a smaller GLR value and be regarded as mutually closer than those of large size.
- A pair of heterogeneous clusters of small size might have a smaller GLR value and be regarded as mutually closer than a pair of homogeneous clusters of large size.

These patterns cause merging between small size clusters to occur mostly in the early recursion steps of AHSC, for which it is highly likely that incorrect merging often happens in those steps due to insufficient data representing speaker-specific characteristics in small size clusters. Such an incorrect merging, especially in the early recursion steps of AHSC, could affect the subsequent merging process negatively because of the recursive structure of AHSC, and thus needs to be prevented as much as possible.

We propose several algorithmic approaches, e.g., *selective AHSC* in Figure 1.4, to figure out these undesirable patterns of the GLR-based inter-cluster distance measure by forcing merging between small size clusters to be kept from occurring in the early recursion steps of AHSC, which will be discussed more in Chapter 3.

1.4.2.2 Later Recursion Steps of AHSC

The aforementioned patterns of the GLR-based inter-cluster distance measure could also cause incorrect merging between heterogeneous clusters in the late recursion steps of AHSC, which might have much bigger impact on the overall clustering performance than incorrect merging in the earlier recursion steps of AHSC.

In order to prevent such an incorrect merging in the later recursion steps of AHSC, we propose an alternative distance measurement method to combine GLR and ICR for better resolution in selecting the closest pair of clusters. We will see more details later in Chapter 3 as well.

1.4.2.3 Cluster Modeling

Since inter-cluster distance is statistically measured in AHSC, selecting proper probability distribution functions (PDFs or pdfs) is required for individual clusters in order to obtain accurate distance between clusters. Ideal cluster modeling for cluster distance measurement within the framework of AHSC should account for variable cluster size, which grows when clusters are merged, and be dynamic enough to represent the statistical changes of data in clusters throughout the entire AHSC procedures. Since such changes in clusters during AHSC largely depend upon a number of input data characteristics, cluster modeling without dynamic representation capability would be much affected by the variation of input speech data, which is undesirable for reliable AHSC performance. Conventional cluster modeling approaches using either single Gaussian distributions or Gaussian mixture models (GMMs) are not ideal in this regard.

We introduce a novel cluster modeling approach with dynamic representation capability, called *incremental Gaussian mixture cluster modeling*. This new approach not only can better represent the statistical changes of data in clusters throughout AHSC than single Gaussian cluster modeling, but also provides slightly better clustering performance and has much lower computational complexity compared to GMM-based cluster modeling. We will handle this topic further in Chapter 4.

1.4.3 Application: Speaker Diarization

Based on our proposed approaches for robust AHSC to the variation of input speech data, we try to make speaker diarization (which is one of main speaker clustering applications and AHSC plays a critical role in it) further reliable across various input speech data

sets. For this purpose, we first implement our own speaker diarization system for analysis of spontaneous meeting conversations, called *SAIL*⁵ *speaker diarization system*, equipped with the ICR-based stopping point estimation method and the incremental Gaussian mixture cluster modeling strategy for GLR-based inter-cluster distance measurement. Then, we propose two schemes for clustering performance refinement in the framework of speaker diarization: *representative cluster model selection* and *interaction pattern modeling*. In Chapter 5, all of these will be taken care of in more detail.

1.5 Contribution Summary

This dissertation, as it has been mentioned thus far, handles how to make speaker clustering, particularly AHSC, more robust to the variation of input speech data in two main perspectives, i.e., stopping point estimation and inter-cluster distance measurement, and how to extend research results from work on robust speaker clustering toward an application domain. A summary for the contributions of the dissertation is as follows:

- Stopping Point Estimation Perspective
 - Proposes ICR, a new statistical distance measure between clusters, in order to avoid the negative effect of the variability of input data characteristics on stopping point estimation for AHSC.
 - Introduces a stopping point estimation method based on ICR for AHSC, which is more robust to the variation of input speech data than the conventional BIC-based one.

- Inter-Cluster Distance Measurement Perspective

⁵SAIL stands for Signal Interpretation and Analysis Laboratory, in which I have been a member since 2004 under Prof. Shri Narayanan, my advisor and committee chair for this dissertation.

- Proposes several modified versions of AHSC approaches so as to enhance the reliability of the GLR-based inter-cluster distance measure at the early recursion steps of AHSC.
 - Proposes a method to combine GLR and ICR so as to improve the reliability of the GLR-based inter-cluster distance measure at the late recursion steps of AHSC.
 - Proposes a dynamic cluster modeling method so as to account for variable cluster size throughout the entire AHSC procedures.
- Application
 - Proposes SAIL speaker diarization that can utilize our promising results from research work on robust speaker clustering.
 - Proposes two refinement schemes for clustering performance in the framework of speaker diarization.

1.6 Dissertation Outline

This dissertation is organized as follows. In Chapter 2, we address the robustness problem of the BIC-based stopping point estimation method for AHSC under the variation of input speech data. For this, we first take a short review of GLR and BIC, and then investigate a main reason for the problem considered. This investigation leads to understanding why a new statistical distance measure between clusters is needed for more robust stopping point estimation in AHSC under the variation of input speech data, which results in our proposal of ICR. In addition, we introduce a stopping point estimation method for AHSC based on ICR in this chapter. This stopping point estimation method is verified through experimental results to be more robust to the variation of input speech data than the conventional BIC-based one. In Chapter 3, we tackle the robustness problem of the GLR-based inter-cluster distance measure from both viewpoints of early and late

AHSC recursion steps. For this, we first examine why the reliability of the GLR-based inter-cluster distance measure severely varies across input data sources. Based on this examination, we propose several modified versions of AHSC approaches to improve the accuracy of the GLR-based inter-cluster distance measure, particularly at the early recursion steps of AHSC. Then we propose a supplement inter-cluster distance measure to utilize the advantages of GLR and ICR in order to tackle the robustness problem of the GLR-based inter-cluster distance measure at the late recursion steps of AHSC. All the methods proposed in this chapter are compared with original AHSC in terms of averaged performance across data sources, and are proven to provide benefit to the reliability of the GLR-based inter-cluster distance measure and thus the overall speaker clustering performance. In Chapter 4, we introduce incremental Gaussian mixture cluster modeling for inter-cluster distance measurement in AHSC. This dynamic cluster modeling approach not only provides AHSC with as comparable clustering performance as the conventional GMM-based one does, but also has a lot more feasibility in computational complexity. In Chapter 5, we apply our research results to speaker diarization. For this, we implement our own speaker diarization system and further modify it with two clustering performance refinement schemes. This dissertation is concluded in Chapter 6 with the final remarks on the work that has been dealt with thus far. We also mention our research's potential application domains other than speaker diarization in this final chapter.

Chapter 2

Robust Stopping Point Estimation for AHSC

2.1 Introduction

This chapter handles the robustness problem of the conventional BIC-based stopping point estimation method for AHSC under the variation of input speech data. This problem has a huge impact on speaker clustering performance because it results in incorrect estimation of the optimal stopping point for AHSC on some data sources¹, which might cause speaker clustering performance to be extremely worse than what it could be with exact estimation of the optimal stopping point during AHSC. In order to address the problem, we first propose *information change rate* (ICR), and then apply it to stopping point estimation for AHSC.

The chapter is organized as follows. In Section 2.2, we introduce the data sources used for experiments in this chapter including analysis and comparison. Experimental setup and relevant assumptions are also described here. In Section 2.3, the BIC-based stopping point estimation method is investigated. This section provides analysis on the cause of the sensitivity of the BIC-based stopping point estimation method to the variation of input speech data. In Section 2.4, based on the analysis in Section 2.3, we tackle the robustness problem of the BIC-based stopping point estimation method by proposing a novel alternative based on ICR. Through experiments on our evaluation data sources, the

¹This means that stopping point estimation in AHSC is perfectly done for some data sources while it is not for some others, which is why we call this issue a robustness problem to the variation of input speech data.

Table 2.1: Development set of data sources. N_s : # of speaker identities (male:female) in each data source, T_s : total utterance time (sec.), N_t : # of speech segments, and T_a : average segment length (sec.). C , N , and I : data sources chosen from ICSI, NIST, and ISL meeting speech corpora respectively.

	Development Set				
	C-1	C-2	C-3	N-1	I-1
N_s	7 (5:2)	7 (5:2)	6 (4:2)	4 (3:1)	4 (2:2)
T_s	1064.9	931.3	1148.5	835.7	477.7
N_t	417	278	243	178	118
T_a	2.5	3.3	4.7	4.7	4.0

proposed ICR-based stopping point estimation method is demonstrated to be more robust to the variation of input speech data than the BIC-based one. We conclude this chapter in Section 2.5 with comments on future work with regard to the ICR-based stopping point estimation method for AHSC.

2.2 Data and Experimental Setup

Tables 2.1 and 2.2 present the development and evaluation data sets used for the experiments reported in this chapter, obtained from 15 different meeting conversation excerpts (with the total length of approximately 3 hours and 45 minutes). The data sources are chosen from ICSI, NIST, and ISL meeting speech corpora². They are distinct from one another in terms of the number of speaker sources (N_s), gender distribution over speaker sources, total utterance time (T_s), number of speech segments (N_t), and average segment length (T_a). The development set will be used for tuning the parameters of the stopping point estimation methods (i.e., BIC- and ICR-based methods) that will be mentioned in this chapter, while the evaluation set will be used for performance calculation.

For the experiments presented in this chapter, we assume that there is no individual speech segment having more than two speaker sources or including overlapped utterances, in order to avoid any potential confusion in performance analysis. To enable this, we

²LDC2004S02, LDC2004S09, and LDC2004S05, respectively.

Table 2.2: Evaluation set of data sources. The notation is same as that in Table 2.1.

	Evaluation Set									
	C-4	C-5	C-6	C-7	C-8	C-9	N-2	N-3	I-2	I-3
N_s	5 (3:2)	9 (7:2)	7 (6:1)	6 (5:1)	4 (4:0)	9 (7:2)	4 (3:1)	6 (4:2)	8 (4:4)	3 (2:1)
T_s	674.5	423.2	2336.3	1664.9	1475.9	659.7	443.4	624.1	272.4	365.3
N_t	175	129	610	531	477	158	74	143	92	72
T_a	3.8	3.3	3.8	3.1	3.1	4.1	5.9	4.3	2.9	5.0

manually segmented each data source according to the reference transcription officially provided by the Linguistic Data Consortium (LDC) prior to the experiments.

Mel-frequency cepstral coefficients (MFCCs) are used as acoustic features. Through 23 mel-scaled filter banks, a 12-dimensional MFCC vector is generated for every 20ms-long frame of speech. Every frame is shifted with the fixed rate of 10ms so that there can be an overlap between two adjacent frames. In order to measure speaker clustering performance, the official scoring tool, i.e., md-eval-v21.pl³, distributed by NIST is used. This tool provides clustering performance as speaker error time rate.

2.3 BIC-based Stopping Point Estimation for AHSC

We begin this section by providing relevant background details on GLR and BIC. The former is, as mentioned in Section 1.2, a widely-used inter-cluster distance measure for selecting merging clusters at every recursion step of AHSC, and the latter is a well-known model selection criterion and is utilized for the stopping point estimation method considered in this section.

³This tool can be downloaded from <http://www.nist.gov/speech/tests/rt/2006-spring>, as mentioned in Section 1.3.

2.3.1 Generalized Likelihood Ratio (GLR)

Suppose that a pair of clusters C_x and C_y are given and they consist of n -dimensional acoustic feature vectors $x = \{x_1, x_2, \dots, x_M\}$ and $y = \{y_1, y_2, \dots, y_N\}$, respectively. Then, GLR for the given pair is computed as follows:

$$\text{GLR}(C_x, C_y) = \frac{P(x \cup y | H_1)}{P(x \cup y | H_2)}, \quad (2.1)$$

where

- H_1 (Unmerging Hypothesis): C_x and C_y are hypothesized to be left unmerged.
- H_2 (Merging Hypothesis): C_x and C_y are hypothesized to be merged so as to be a new cluster C_z , where $z = x \cup y$.

In order to mathematically calculate the two likelihoods in the right side of Eq. (2.1), the two hypotheses need to be modeled by probability mass or distribution functions (PMFs or PDFs) respectively. In this regard, single Gaussian modeling for each cluster considered (C_x and C_y for H_1 , and C_z for H_2) has been popularly utilized since [26]. In this chapter, we follow this approach as well because single Gaussian cluster modeling is much easier to be analyzed theoretically than other cluster modeling approaches such as one based on Gaussian mixture models (GMMs)⁴. Based on [26], C_x , C_y , and C_z are modeled by (multivariate) single Gaussian distributions f_X , f_Y , and f_Z with full covariance matrices respectively. Assuming that the PDFs represent random variables X , Y , and Z respectively, x , y , and z can be regarded (in the modeling framework of [26]) as the sequences of independently and identically distributed (i.i.d.) random variables drawn according to the PDFs f_X , f_Y , and f_Z of random variables X , Y , and Z respectively. The mean vectors and the covariance matrices of f_X , f_Y , and f_Z are

⁴We will discuss more in detail about cluster modeling for inter-cluster distance measurement in AHSC in Chapter 5.

determined by way of maximizing the likelihoods of x , y , and z for f_X , f_Y , and f_Z respectively. In other words,

$$\tilde{\theta}_x = (\mu_x, \Sigma_x) = (\mu_{f_X}, \Sigma_{f_X}) = \theta_{f_X}, \quad (2.2)$$

$$\tilde{\theta}_y = (\mu_y, \Sigma_y) = (\mu_{f_Y}, \Sigma_{f_Y}) = \theta_{f_Y}, \quad (2.3)$$

and

$$\tilde{\theta}_z = (\mu_z, \Sigma_z) = (\mu_{f_Z}, \Sigma_{f_Z}) = \theta_{f_Z}, \quad (2.4)$$

where μ_x , μ_y , and μ_z are the sample mean vectors, and Σ_x , Σ_y , and Σ_z are the sample covariance matrices obtained from x , y , and z respectively. μ_{f_X} , μ_{f_Y} , and μ_{f_Z} are the mean vectors, and Σ_{f_X} , Σ_{f_Y} , and Σ_{f_Z} are the covariance matrices of f_X , f_Y , and f_Z respectively. Under this framework, Eq. (2.1) can be re-written as

$$\text{GLR}(C_x, C_y) = \frac{p(x|f_X; \theta_{f_X}) \cdot p(y|f_Y; \theta_{f_Y})}{p(z|f_Z; \theta_{f_Z})} \quad (2.5)$$

$$= \frac{p(x|f_X; \tilde{\theta}_x)}{p(x|f_Z; \tilde{\theta}_z)} \cdot \frac{p(y|f_Y; \tilde{\theta}_y)}{p(y|f_Z; \tilde{\theta}_z)}. \quad (2.6)$$

Eq. (2.6) tells that GLR is always greater than or equal to 1 because both of the numerators in the equation are maximal out of the likelihoods of x and y respectively. In other words, $p(x|f_X; \tilde{\theta}_x) \geq p(x|f_Z; \tilde{\theta}_z)$ and $p(y|f_Y; \tilde{\theta}_y) \geq p(y|f_Z; \tilde{\theta}_z)$, where the equalities hold only if $C_x = C_y$ or $x = y$. This means that H_1 is always more likely than H_2 , and thus GLR is not adequate to indicate that one hypothesis is more likely than the other. Instead, GLR tells how much more likely H_1 is than H_2 . Therefore, the more likely H_1 is for a pair of clusters, the more distant the clusters are regarded in GLR-based inter-cluster distance measurement.

The drawback of GLR as an inter-cluster distance measure is, as mentioned in [29–32, 40, 66], that GLR tends to get larger as the total number of feature vectors within a pair of clusters under consideration increases. This can be clearly illustrated in Figure

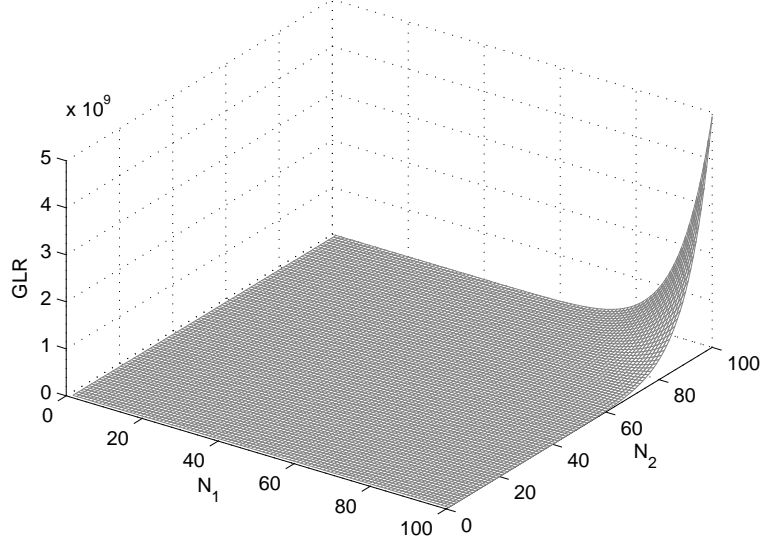


Figure 2.1: GLR for two clusters C_1 and C_2 along with the number of feature vectors in each cluster. The second order statistics of the corresponding cluster models are fixed at $\mu_1 = 0$, $\mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$.

2.1, which shows GLRs between two clusters C_1 and C_2 along with the corresponding numbers of feature vectors N_1 and N_2 . In order to observe the effect of the numbers of feature vectors in the clusters, we fixed the second order statistics of $\tilde{\theta}_1$ and $\tilde{\theta}_2$ arbitrarily. (In this case, $\mu_1 = 0$, $\mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$.) From this figure we can explicitly see the exponential rising-up of GLR as the numbers of feature vectors in the clusters increase. Consequently, in GLR-based inter-cluster distance measurement, a pair of homogeneous clusters (in terms of speaker-specific characteristics) of small size are likely to have a smaller GLR value and be regarded as mutually closer than those of large size. Besides, a pair of heterogeneous clusters of small size might have a smaller GLR value and be regarded as mutually closer than a pair of homogeneous clusters of large size, which is undesirable.

This undesirable tendency of GLR can be confirmed by analyzing GLR computation with a few basic concepts in the field of information theory. Let us begin this analysis

with Eq. (2.5). We can re-write the equation as below without loss of generality by applying logarithm to both sides:

$$\begin{aligned}
& \ln \text{GLR} (C_x, C_y) \\
&= \ln \frac{p(x|f_X; \theta_{f_X}) \cdot p(y|f_Y; \theta_{f_Y})}{p(z|f_Z; \theta_{f_Z})} \\
&= \ln f_X(x_1, x_2, \dots, x_M) + \ln f_Y(y_1, y_2, \dots, y_N) - \ln f_Z(x_1, \dots, x_M, y_1, \dots, y_N).
\end{aligned} \tag{2.7}$$

Considering that GLR computation intrinsically assumes the weak law of large numbers⁵ to be satisfied during its procedure, we can apply the asymptotic equipartition property⁶ (AEP) widely-known as the consequence of the weak law of large number² in the field of information theory to the right side term of Eq. (2.7). Then, the equation can be simplified to

$$\ln \text{GLR} (C_x, C_y) = -M \cdot h(X) - N \cdot h(Y) + (M + N) \cdot h(Z). \tag{2.9}$$

⁵The weak law of large numbers states that a sample mean and a sample variance converge in probability towards the expected value and the second central moment of a corresponding random variable respectively. In GLR computation, this law is inherent to Eqs. (2.2)-(2.4).

⁶This property can be explained as follows:

- Let x_1, x_2, \dots, x_M be the sequence of i.i.d. random variables drawn according to the PDF f_X of a random variable X . Then, according to [16], AEP states that

$$-\frac{1}{M} \ln f_X(x_1, x_2, \dots, x_M) = h(X) \text{ in probability,} \tag{2.8}$$

where h is entropy.

Since entropy for an n -dimensional multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ can be obtained (according to [16]) as a closed form of $\frac{1}{2} \ln(2\pi e)^n |\Sigma|$ where $|\cdot|$ is determinant, we can further simplify Eq. (2.9) to

$$\begin{aligned}
& \ln \text{GLR}(C_x, C_y) \\
&= -M \cdot \frac{1}{2} \ln(2\pi e)^n |\Sigma_x| - N \cdot \frac{1}{2} \ln(2\pi e)^n |\Sigma_y| + (M+N) \cdot \frac{1}{2} \ln(2\pi e)^n |\Sigma_z| \\
&= \frac{M+N}{2} \ln |\Sigma_z| - \frac{M}{2} \ln |\Sigma_x| - \frac{N}{2} \ln |\Sigma_y|, \tag{2.10}
\end{aligned}$$

where Σ_z has the following relation with Σ_x and Σ_y :

$$\begin{aligned}
& \Sigma_z \\
&= \frac{M \cdot \Sigma_x + N \cdot \Sigma_y}{M+N} + \frac{M \cdot \mu_x \mu_x^T + N \cdot \mu_y \mu_y^T}{M+N} - \frac{M \cdot \mu_x + N \cdot \mu_y}{M+N} \cdot \left(\frac{M \cdot \mu_x + N \cdot \mu_y}{M+N} \right)^T \tag{2.11}
\end{aligned}$$

because $z = x \cup y$.

Based on this, let us think of a simple instance. Suppose that we need to compute GLR between two clusters $C_{x'}$ and $C_{y'}$, where x' and y' are the sequences of i.i.d. random variables drawn according to the PDFs f_X and f_Y , and their cardinalities are $2M$ and $2N$ respectively. In other words, x' has the same second order statistics with x 's but twice the number of feature vectors within x . y' has such relation to y as well. Then, $\Sigma_{z'} = \Sigma_z$, and hence

$$\begin{aligned}
\ln \text{GLR}(C_{x'}, C_{y'}) &= (M+N) \ln |\Sigma_{z'}| - M \cdot \ln |\Sigma_{f_X}| - N \cdot \ln |\Sigma_{f_Y}| \\
&= (M+N) \ln |\Sigma_z| - M \cdot \ln |\Sigma_x| - N \cdot \ln |\Sigma_y| = 2 \cdot \ln \text{GLR}(C_x, C_y).
\end{aligned}$$

The above example indicates that $\ln \text{GLR}$ linearly increases (or GLR exponentially increases) with the fixed second order statistics as the numbers of feature vectors within

a pair of clusters under consideration get larger, which is consistent with what has been shown in Figure 2.1.

2.3.2 Bayesian Information Criterion (BIC)

BIC [58] was primarily intended for model (or PDF) selection, specifically for the problem of how to select the best model for given observations from candidate models. A basic model selection strategy based on BIC is as follows:

1. Compute BIC scores for all candidate models.

$$\begin{aligned} \text{BIC}(f) &= \ln p(x|f; \theta_f) - \mathbf{P}_f \\ &= \ln p(x|f; \theta_f) - \frac{1}{2} \#(\theta_f) \ln M, \end{aligned} \quad (2.12)$$

where $x = \{x_1, x_2, \dots, x_M\}$ represents given M observations, f is a model (or PDF), θ_f is a set of model parameters for f , and $\#(\theta_f)$ is the total number of model parameters for f .

2. Select the model whose BIC score is the highest as the best one to represent the observations.

The core of BIC is that the log-likelihood of given observations for a model is penalized by \mathbf{P}_f , which is determined by the total number of model parameters and the logarithm of the cardinality of the observations. This prevents the model having the most number of parameters from being chosen all the time as the best one, which is a well-known issue in model selection based on maximum likelihood without penalization.

2.3.3 BIC-based Stopping Point Estimation Method for AHSC

Keeping both GLR and BIC in mind, let us now investigate the BIC-based stopping point estimation method for AHSC. This conventional method to search for the optimal stopping point for AHSC (where speaker clustering performance would not be improved any further with extra merging) was originally introduced in [13] by Chen and Gopalakrishnan. It basically stops AHSC at the point where the closest pair among all pairs of remaining clusters are decided to be not homogeneous in terms of speaker-specific characteristics for the first time of the entire AHSC procedures, based on the reasoning that if the closest pair of clusters were heterogeneous then so would be any other pair of clusters, and thus there would be no more need for merging in AHSC. Decision of homogeneity for the closest pair of clusters at every recursion step of AHSC is done by comparing the BIC scores of the clusters for two hypotheses of ‘Unmerging’ and ‘Merging’. These two hypotheses are the same as those (H_1 and H_2) used in GLR computation in Section 2.3.1, and in this case H_2 supports homogeneity while H_1 supports heterogeneity. As in GLR computation, the two clusters considered are modeled by (multivariate) single Gaussian distributions with maximum likelihood parameter estimation. The details of how the BIC-based stopping point estimation method works for AHSC are as follows⁷:

⁷We used the same notation in Section 2.3.1 for single Gaussian modeling for clusters.

1. For the closest pair of clusters C_x and C_y consisting of feature vectors $x = \{x_1, x_2, \dots, x_M\}$ and $y = \{y_1, y_2, \dots, y_N\}$ respectively, compute the BIC scores of $x \cup y$ for H_1 and H_2 .

$$\begin{aligned}
\text{BIC}(H_1) &= \ln P(x \cup y | H_1) - \lambda \cdot \mathbf{P}_{H_1} \\
&= \ln P(x \cup y | H_1) - \lambda \cdot \frac{1}{2} \#(H_1) \ln N_{total} \\
&= \ln \{p(x|f_X; \theta_{f_X}) \cdot p(y|f_Y; \theta_{f_Y})\} - \lambda \cdot \frac{1}{2} \{\#(\theta_{f_X}) + \#(\theta_{f_Y})\} \ln N_{total} \\
&= \ln \left\{ p(x|f_X; \tilde{\theta}_x) \cdot p(y|f_Y; \tilde{\theta}_y) \right\} - \lambda \cdot \frac{1}{2} \left[2 \left\{ n + \frac{1}{2} n(n+1) \right\} \right] \ln N_{total}.
\end{aligned} \tag{2.13}$$

$$\begin{aligned}
\text{BIC}(H_2) &= \ln P(x \cup y | H_2) - \lambda \cdot \mathbf{P}_{H_2} \\
&= \ln P(x \cup y | H_2) - \lambda \cdot \frac{1}{2} \#(H_2) \ln N_{total} \\
&= \ln \{p(x|f_Z; \theta_{f_Z}) \cdot p(y|f_Z; \theta_{f_Z})\} - \lambda \cdot \frac{1}{2} \#(\theta_{f_Z}) \ln N_{total} \\
&= \ln \left\{ p(x|f_Z; \tilde{\theta}_z) \cdot p(y|f_Z; \tilde{\theta}_z) \right\} - \lambda \cdot \frac{1}{2} \left\{ n + \frac{1}{2} n(n+1) \right\} \ln N_{total}.
\end{aligned} \tag{2.14}$$

In Eqs. (2.13) and (2.14), λ is the parameter that should be tuned a priori for minimizing averaged speaker clustering performance (i.e., speaker error time rate) with a development set of data sources (which will be explained more in detail later), N_{total} is the total number of feature vectors for the entire clusters given as an input to AHSC, and n is the dimension of feature vectors.

2. Compute $\Delta\text{BIC}(C_x, C_y) = \text{BIC}(H_1) - \text{BIC}(H_2)$.

$$\begin{aligned}
& \Delta\text{BIC}(C_x, C_y) \\
&= \ln \left\{ p(x|f_X; \tilde{\theta}_x) \cdot p(y|f_Y; \tilde{\theta}_y) \right\} - \lambda \cdot \frac{1}{2} \left[2 \left\{ n + \frac{1}{2}n(n+1) \right\} \right] \ln N_{total} - \\
& \quad \ln \left\{ p(x|f_Z; \tilde{\theta}_z) \cdot p(y|f_Z; \tilde{\theta}_z) \right\} + \lambda \cdot \frac{1}{2} \left\{ n + \frac{1}{2}n(n+1) \right\} \ln N_{total} \\
&= \ln \frac{p(x|f_X; \tilde{\theta}_x) \cdot p(y|f_Y; \tilde{\theta}_y)}{p(x|f_Z; \tilde{\theta}_z) \cdot p(y|f_Z; \tilde{\theta}_z)} - \lambda \cdot \frac{1}{2} \left\{ n + \frac{1}{2}n(n+1) \right\} \ln N_{total} \\
&= \ln \text{GLR}(C_x, C_y) - \lambda \cdot \frac{1}{2} \left\{ n + \frac{1}{2}n(n+1) \right\} \ln N_{total} \tag{2.15} \\
& \underset{H_2}{\overset{H_1}{\geq}} 0.
\end{aligned}$$

3. If $\Delta\text{BIC}(C_x, C_y) < 0$ or $\text{BIC}(H_1) < \text{BIC}(H_2)$, decide that C_x and C_y are homogeneous and merge them. Otherwise, do not merge them and stop AHSC.

The stopping criterion mentioned above can be re-written as

$$\ln \text{GLR}(C_x, C_y) \underset{H_2}{\overset{H_1}{\geq}} \lambda \cdot c \cdot \ln N_{total}, \tag{2.16}$$

where $c = \frac{1}{2} \left\{ n + \frac{1}{2}n(n+1) \right\}$ is a constant. This criterion could be replaced by

$$\ln \text{GLR}(C_x, C_y) \underset{H_2}{\overset{H_1}{\geq}} \lambda \cdot c \cdot \ln(M+N). \tag{2.17}$$

This modified criterion was introduced in [8] based on its better performance for estimating the optimal stopping point for AHSC than Eq. (2.16). In this chapter, we will consider Eq. (2.17) as a baseline stopping criterion for the BIC-based stopping point estimation method for this reason. From this point on, the stopping criterion that we are mentioning throughout the chapter thus points out Eq. (2.17), not Eq. (2.16).

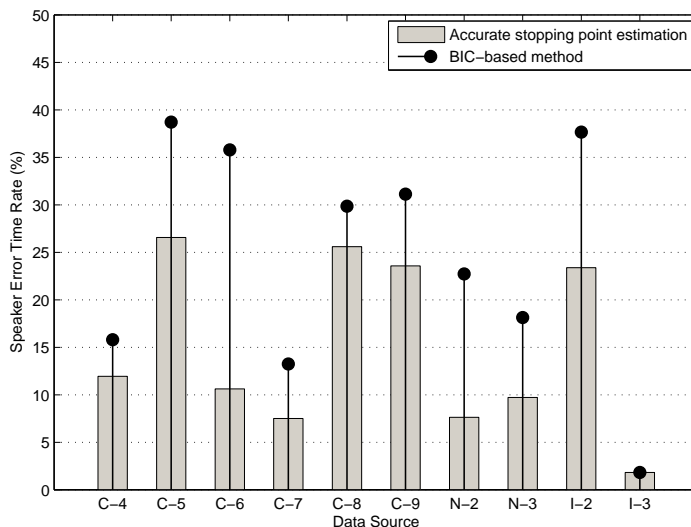


Figure 2.2: Comparison of speaker clustering performance (for the evaluation data set described in Section 2.2) with and without accurate stopping point estimation. For the BIC-based stopping point estimation method, we tuned λ to be 12.0. Average speaker error time rate degradation by incorrect estimation of the optimal stopping point is about 9.65% (absolute) per data source.

2.3.4 Tuning Parameter λ

An important aspect to note for this BIC-based stopping point estimation method is the use of the tuning parameter λ in Eqs. (2.13) and (2.14). This parameter is not included in the original BIC score computation as shown in Eq. (2.12), which means that the parameter was intentionally introduced when applying BIC to devise a stopping point estimation method for AHSC. Unfortunately, there is no explicit explanation in [13] of why λ is necessary and how it can be optimally chosen. In the research field of speaker clustering, however, the parameter is widely considered as a weighting factor to lift up the level of the whole right side term of Eq. (2.17), and is generally tuned so as for the stopping criterion to provide the minimum averaged speaker error time rate for a development data set. (In this chapter, we set λ to be 12.0 because $\lambda = 12.0$ minimized averaged speaker error time rate for our development data set.)

A problem is that λ does not work globally because it is tuned only based on a development data set. Such a tuned parameter cannot guarantee the stopping criterion to correctly estimate the optimal stopping points for data sources in a different data domain, due to its dependency upon the data set used for tuning. This problem is clearly confirmed in Figure 2.2⁸. We can see from this figure that with $\lambda = 12.0$ the BIC-based stopping point estimation method does not reliably estimate the optimal stopping point for the evaluation data set. In our experiments, the impact of incorrect estimation of the optimal stopping point is detrimental specifically for C-5, C-6, N-2, and I-2 while it is not the case for C-4, C-8, and I-3. Average speaker error time rate degradation due to such incorrect estimation is about 9.65% (absolute) per data source.

In order to handle this problem, one interesting approach was proposed in [3] based on the idea of [2], which is to automatically erase λ by equalizing $\#(H_1)$ to $\#(H_2)$ in the computation of BIC scores for H_1 and H_2 . For this, a Gaussian mixture model (GMM) with m model parameters for each cluster considered (C_x and C_y) in H_1 and another GMM with $2m$ model parameters for a hypothetically merged cluster (C_z) in H_2 were utilized respectively. By doing so, this approach can avoid parameter tuning. However, it has some side effects such as increased computing time for training GMMs at every recursion step of AHSC. Moreover, the approach does not directly take care of a fundamental cause for the robustness problem of the BIC-based stopping point estimation method, which is the stopping criterion itself being not robust to the variation of input speech data.

2.3.5 Stopping Criterion under the Variation of Input Speech Data

The stopping criterion of the BIC-based stopping point estimation method, Eq. (2.17), has an intrinsic flaw in terms of robustness to the variation of input speech data because it utilizes GLR. As aforementioned in Section 2.3.1, GLR is sensitive to the numbers of feature vectors within the clusters considered. As a result, the left side term of Eq. (2.17),

⁸In this experiment, GLR was used as an inter-cluster distance measure for AHSC to select the closest pair of clusters at every recursion step.

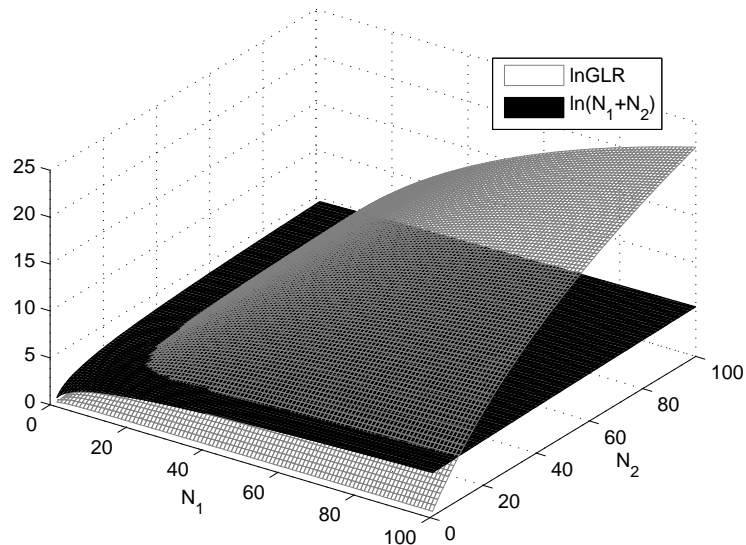


Figure 2.3: $\ln \text{GLR}$ and $\ln(M + N)$ ($= \ln(N_1 + N_2)$ in this case) for the same clusters considered in Figure 2.1 along with the number of feature vectors in each cluster, with the fixed second order statistics of $\mu_1 = 0$, $\mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$.

$\ln \text{GLR}$, is affected by several aspects in the entire speech segments given as an input to AHSC beyond statistical difference between the clusters considered. This is because the size of the clusters considered by the BIC-based stopping point estimation method at a certain recursion step of AHSC is determined jointly by the total amount of time for the entire speech utterances (or speech segments) in the given input data, the distributions of the speech segments in length and speaker identity, and merging procedures at the previous recursion steps of AHSC. One might claim that the right side term of Eq. (2.17) is also decided by the numbers of feature vectors within the clusters considered due to $\ln(M + N)$, so the stopping criterion looks robust to the variation of input speech data. However, $\ln \text{GLR}$ grows in a linear fashion⁹ in proportion to M and N while $\ln(M + N)$ increases in a logarithmic fashion, which is well shown in Figure 2.3. $\ln \text{GLR}$ is fast increasing along with M and N , but $\ln(M + N)$ looks relatively flat in the figure. This indicates that the right side term of Eq. (2.17) cannot compensate the data dependency

⁹We confirmed in Section 2.3.1 that GLR exponentially increased in proportion to the numbers of feature vectors within the clusters considered.

of the left side term fully enough, and the stopping criterion is thus highly likely to vary across input speech data sources. For this reason, it is too difficult to set a global λ .

2.4 ICR-based Stopping Point Estimation for AHSC

In the previous section, we investigated the BIC-based stopping point estimation method for AHSC and underscored that a fundamental reason for the robustness problem of the method is the stopping criterion being not robust to the variation of input speech data. In this section, based on the analysis in Section 2.3, we propose a new stopping point estimation method for AHSC that is more robust to the variation of input speech data than the BIC-based one.

2.4.1 Information Change Rate (ICR)

First, we propose a new statistical distance measure between clusters, *information change rate* (ICR), which is defined as follows for a pair of clusters C_x and C_y consisting of feature vectors $x = \{x_1, x_2, \dots, x_M\}$ and $y = \{y_1, y_2, \dots, y_N\}$, respectively:

$$\text{ICR}(C_x, C_y) \triangleq \frac{1}{M+N} \ln \text{GLR}(C_x, C_y). \quad (2.18)$$

As shown above, ICR is the normalized version of $\ln \text{GLR}$. This simple idea of normalizing $\ln \text{GLR}$ with the total number of feature vectors within a pair of clusters under consideration was inspired by analyzing GLR with an information-theoretic perspective. Let us consider Eq. (2.9) in Section 2.3.1 again. Considering that entropy can be regarded as average description length for a random sample from a given PDF, we can separate the right side term of the equation into the following two parts:

$$\begin{aligned} \ln \text{GLR}(C_x, C_y) &= \underbrace{(M+N) \cdot h(Z)}_{\text{Total description length for } z=x \cup y \text{ under } H_2} - \underbrace{\{M \cdot h(X) + N \cdot h(Y)\}}_{\text{Total description length for } z \text{ under } H_1}. \end{aligned} \quad (2.19)$$

This means that $\ln \text{GLR}$ equals difference between the *total description lengths* for the whole feature vectors considered under the two hypotheses H_1 (Unmerging) and H_2 (Merging). That is, $\ln \text{GLR}$ represents how much amount of information would be *totally* changed by merging the clusters considered. This is why GLR is sensitive to cluster size. Thus, it is natural to expect that a certain distance measure, if it represents how much amount of information would be changed *on average* over feature vectors by merging the clusters considered, could avoid being affected by the size of the clusters. ICR satisfies such an expectation.

From Eqs. (2.18) and (2.19), we can obtain a different version of ICR :

$$\text{ICR}(C_x, C_y) = h(Z) - \frac{M \cdot h(X) + N \cdot h(Y)}{M + N}. \quad (2.20)$$

In this form, ICR can show inter-cluster relation as follows, for two extreme examples:

- Ex 1: $C_x = C_y$ or $x = y$.

$$\begin{aligned} \text{ICR}(C_x, C_y) &= \text{ICR}(C_x, C_y) \\ &= h(X) - \frac{M \cdot h(X) + M \cdot h(X)}{M + M} \\ &= h(X) - h(X) \\ &= 0 \end{aligned}$$

- Ex 2: C_x and C_y are mutually independent.

$$\begin{aligned} \text{ICR}(C_x, C_y) &= h(X) + h(Y) - \frac{M \cdot h(X) + N \cdot h(Y)}{M + N} \\ &= \frac{(M + N) \cdot h(X) + (M + N) \cdot h(Y)}{M + N} - \frac{M \cdot h(X) + N \cdot h(Y)}{M + N} \\ &= \frac{N \cdot h(X) + M \cdot h(Y)}{M + N} \end{aligned}$$

Table 2.3: Comparison of ICR with other measures utilizing the idea of normalizing GLR. C_x and C_y : two clusters consisting of M and N feature vectors respectively, α : parameter empirically determined, and n : dimension of feature vectors.

ICR (C_x, C_y)	PLR in [40]	NLLR in [66]
$\frac{1}{M+N} \ln \text{GLR} (C_x, C_y)$	$\frac{1}{(M+N)^\alpha} \text{GLR} (C_x, C_y)$	$\frac{1}{(M+N) \cdot n} \ln \text{GLR} (C_x, C_y)$

2.4.2 Comparison of ICR with ICR-like Measures

In fact, there have been several ICR-like inter-cluster distance measures to normalize GLR in the research field of speaker clustering. Table 2.3 compares two of such measures, i.e. penalized likelihood ratio (PLR) [40] and normalized log-likelihood ratio (NLLR) [66], with ICR. PLR normalizes GLR with the α -th power of the total number of feature vectors within the clusters considered. However, it does not appear promising in terms of mitigating the effect of cluster size on distance measurement, because

$$\ln \text{PLR} (C_x, C_y) = \ln \text{GLR} (C_x, C_y) - \alpha \cdot \ln (M + N). \quad (2.21)$$

As shown in Section 2.3.5, $\ln (M + N)$ cannot compensate the dependency of $\ln \text{GLR}$ on cluster size entirely. Thus, it is difficult to set a global α . On the other hand, NLLR is very similar to ICR and its relation to ICR is shown as follows:

$$\text{NLLR} (C_x, C_y) = \frac{1}{n} \text{ICR} (C_x, C_y). \quad (2.22)$$

But it has a different physical meaning from that of ICR because it further normalizes $\ln \text{GLR}$ with the dimension of feature vectors, n .

2.4.3 ICR as a Homogeneity Decision Measure for Clusters

Since ICR represents how much amount of information would be changed on average over feature vectors by merging the clusters considered, it is natural to expect ICR to be very small when the clusters considered are homogeneous in terms of speaker-specific

characteristics and each cluster is large enough to fully cover the intra-speaker variance of corresponding speaker identity. In other words, ICR will be small when the clusters considered have the same speaker identity source and do not need additional information for representing full speaker-specific characteristics. On the contrary, ICR will be relatively large when the clusters considered are heterogeneous, or when they are homogeneous but contain small feature vectors to cover only a part of speaker-specific characteristics. Thus, ICR could properly work as a measure to decide homogeneity for clusters if every cluster considered were large enough to fully represent the characteristics of the corresponding speaker identity.

We assume that a cluster containing feature vectors which correspond to more than 30 seconds in amount of time is such a large enough cluster. This assumption is based on the fact that it requires long speech utterances (at least longer than 20 seconds) to derive reliable speaker characteristics [51–53]. Figure 2.4 displays distributions for ICR between homogeneous clusters and for ICR between heterogeneous clusters. The distributions were assumed to be Gaussian, and their sample means and sample variances were respectively obtained based on our development data set. The number of feature vectors in all the clusters considered here corresponded to more than 30 seconds in amount of time. Using the distributions in the figure, we set a threshold $\eta = \text{Th}_{\text{ICR}}$ to be 0.18603, with which classification error between the two distributions can be minimized. We can thus regard a pair of clusters having ICR less than $\eta = 0.18603$ as homogeneous in terms of speaker-specific characteristics.

2.4.4 ICR-based Stopping Point Estimation Method for AHSC

Based on ICR and its applicability to inter-cluster homogeneity decision in terms of speaker-specific characteristics, we now introduce an ICR-based stopping point estimation method for AHSC. This method is distinct from the BIC-based one in terms of 1) stopping criterion and 2) the order of the clusters considered. Its details are as follows:

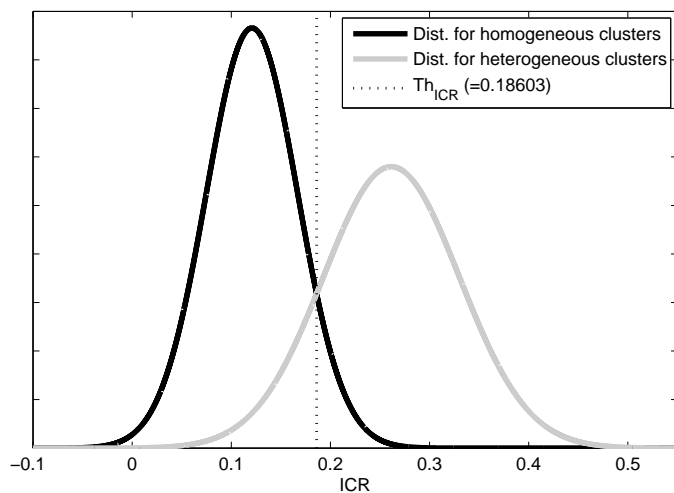


Figure 2.4: Distributions for correct and incorrect merging in terms of ICR. The threshold η is set so as to minimize classification error between the two distributions. The distributions were obtained based on our development data set, and feature vectors in every cluster considered corresponded to more than 30 seconds in amount of time.

1. Wait until AHSC reaches the end of its merging procedures, i.e., wait until all the clusters given as an input to AHSC are merged to one big cluster.
2. For the pair of clusters merged at the last recursion step of AHSC, C_x and C_y , consisting of feature vectors $x = \{x_1, x_2, \dots, x_M\}$ and $y = \{y_1, y_2, \dots, y_N\}$ respectively, compute ICR.
3. Compare ICR with η :

$$\text{ICR}(C_x, C_y) \underset{H_2}{\overset{H_1}{\gtrless}} \eta. \quad (2.23)$$

If $\text{ICR}(C_x, C_y) > \eta$, decide that C_x and C_y are heterogeneous in terms of speaker-specific characteristics and move on to consider the pair of clusters merged at the next latest recursion step of AHSC. Otherwise, stop considering more merging recursions and select the recursion step previously considered as the estimated optimal stopping point.

Table 2.4: ICR-based stopping point estimation method vs. BIC-based stopping point estimation method. $c = \frac{1}{2} \{n + \frac{1}{2}n(n+1)\}$, where n is the dimension of feature vectors. $n = 12$, $\eta = 0.18603$, and $\lambda = 12.0$.

	ICR-based method	BIC-based method
Criterion	$\text{ICR}(C_x, C_y) \underset{H_2}{\overset{H_1}{\geq}} \eta$	$\underset{H_2}{\overset{H_1}{\geq}} \ln \text{GLR}(C_x, C_y) \geq \lambda \cdot c \cdot \ln(M + N)$
Right side term in criterion	Fixed during AHSC	Floating along with M and N during AHSC
Computing complexity	Complexity for $\ln \text{GLR}(C_x, C_y)$ and $\eta \cdot (M + N)$	Complexity for $\ln \text{GLR}(C_x, C_y)$ and $\lambda \cdot c \cdot \ln(M + N)$
Order of clusters considered	From the pair of clusters merged at the last recursion step	From the pair of clusters merged at the 1 st recursion step

The ICR-based stopping point estimation method depends upon the reasoning¹⁰ that all the merging occurring after the optimal stopping point would occur between heterogeneous clusters. The reason why this stopping point estimation method starts its consideration from the pair of clusters merged at the last recursion step of AHSC is that such a strategy can make the stopping criterion, Eq. (2.23), consider large clusters only. As mentioned earlier, ICR can properly work as a homogeneity decision measure only for large enough clusters to represent full speaker-specific characteristics respectively. Eq. (2.23) can be re-written as follows:

$$\ln \text{GLR}(C_x, C_y) \underset{H_2}{\overset{H_1}{\geq}} \eta \cdot (M + N). \quad (2.24)$$

Comparing this criterion with Eq. (2.17) for the BIC-based stopping point estimation method, we can see that the difference of computational complexity between the two

¹⁰The BIC-based stopping point estimation method also relies on this same reasoning.

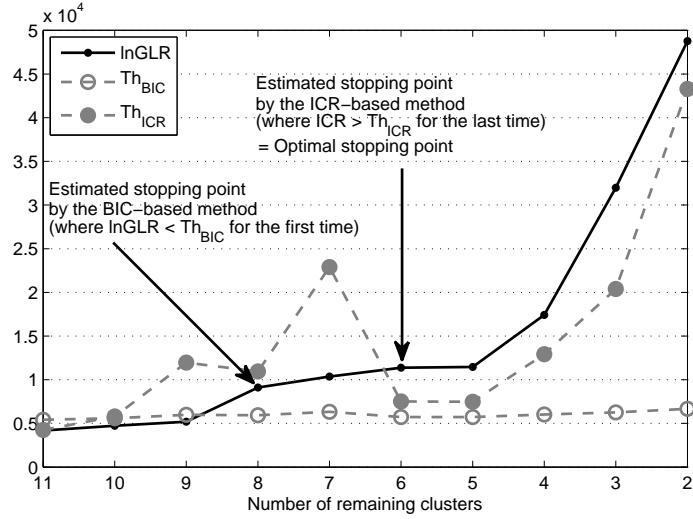


Figure 2.5: $\ln \text{GLR}$, $\text{Th}_{\text{BIC}} = \lambda \cdot c \cdot \ln(M + N)$, and $\text{Th}_{\text{ICR}} = \eta \cdot (M + N)$ for C-6, where $\lambda = 12.0$ and $\eta = 0.18603$. The stopping point estimated by the ICR-based stopping point estimation method is identical to the optimal one in this case.

stopping point estimation methods is negligible. For easier understanding of the ICR-based stopping point estimation method for AHSC, Table 2.4 is presented.

Figure 2.5 shows $\ln \text{GLR}$, $\text{Th}_{\text{BIC}} = \lambda \cdot c \cdot \ln(M + N)$, and $\text{Th}_{\text{ICR}} = \eta \cdot (M + N)$ for the data source C-6 in our evaluation data set, where $\lambda = 12.0$ and $\eta = 0.18603$. This figure focuses on the variations of the three terms at the final 10 merging recursions during AHSC for C-6. From the figure, we can see that Th_{ICR} varies along with $\ln \text{GLR}$ while Th_{BIC} does not. The observation that Th_{BIC} looks almost flat compared to $\ln \text{GLR}$ is consistent with what was shown in Figure 2.3 in Section 2.3.5, and verifies that Eq. (2.17) is not robust to the variation of input speech data. In contrast, the robustness of the criterion in Eq. (2.23) or Eq. (2.24) to the variation of input speech data is demonstrated through the figure above.

Figure 2.6¹¹ presents AHSC performance using the ICR-based stopping point estimation method ($\eta = 0.18603$) for the evaluation data set. In the figure, we can observe that the proposed stopping point estimation method exactly detected the optimal stopping points for all the data sources except C-4, C-8, and C-9. Even for the three data

¹¹GLR was used as an inter-cluster distance measure for AHSC in this experiment.

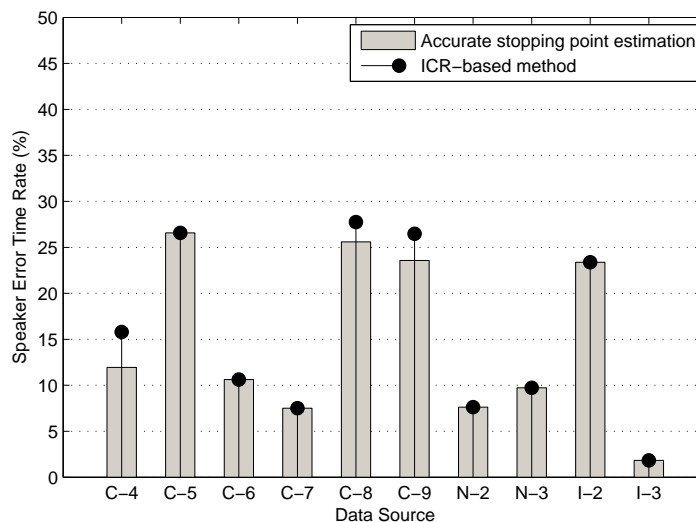


Figure 2.6: Comparison of speaker clustering performance for the evaluation data set with accurate stopping point estimation and with the ICR-based stopping point estimation method, for which $\eta = 0.18603$. Average speaker error time rate degradation by incorrect estimation of the optimal stopping point is less than 1% (absolute) per data source.

sources, gaps between speaker error time rates at the estimated stopping points and those at the optimal ones are shown to be insignificant. Compared to the results (in Figure 2.2) obtained using AHSC with the BIC-based stopping point estimation method for the same data set, the results in this figure are much improved overall, and indicate that the ICR-based method is superior to the BIC-based one in terms of robustness to the variation of input speech data. Consequently, the ICR-based method for AHSC led to average performance improvement by 8.76% (absolute) and 35.77% (relative).

2.5 Conclusions

In this chapter, we addressed the robustness problem of the BIC-based stopping point estimation method for AHSC to the variation of input speech data. For this, we proposed a novel ICR-based alternative. Through experimental results on the excerpts obtained from meeting speech corpora, AHSC with the ICR-based stopping point estimation method was shown to outperform and be more robust to the variation of input speech data than

Table 2.5: Global comparison (averaged speaker error time rate for the evaluation data set) of AHSC with the BIC-based stopping point estimation method and AHSC with the ICR-based stopping point estimation method

AHSC (BIC)	AHSC (ICR)
24.49%	15.73%

basic AHSC with the BIC-based stopping point estimation method. Table 2.5 presents performance comparison for AHSC with the BIC-based stopping point estimation method and AHSC with the ICR-based stopping point estimation method. A reason for the improvements achieved by our proposed method in terms of averaged speaker error time rate across the data sources in the evaluation data set is that the undesirable tendency of GLR, i.e., GLR tends to get larger as the total number of feature vectors within a pair of clusters under consideration increases, was removed.

One potential future direction is to identify the lower bound for cluster size that guarantees ICR to be reliable as a statistical distance measure, more specifically as a homogeneity decision measure, between the clusters considered. In this chapter, we avoided the possibility that ICR would not work properly, by checking ICR-based inter-cluster homogeneity starting from the pair of clusters merged at the last recursion step of AHSC under the assumption that clusters at the late recursion steps of AHSC would be large enough for reliable ICR. This assumption worked for the meeting conversation excerpts used for the experiments presented in the chapter because most of the speaker sources involved in the conversations generated enough speech utterances of which the total length in time was longer than at least 30 seconds, respectively. Thus, at the late recursion steps of AHSC where the ICR-based stopping point estimation method was usually applied, ICR could be reliable as an inter-cluster homogeneity measure as expected. The assumption could be however broken for other data sources which have a preponderance of short speech segments that are inadequate to reveal the corresponding speaker-specific characteristics completely.

Chapter 3

Robust Inter-Cluster Distance Measurement for AHSC

3.1 Introduction

In this chapter we handle the robustness problem of the GLR-based inter-cluster distance measure for AHSC under the variation of input speech data. Like the robustness problem of the BIC-based stopping point estimation method, this problem is caused mainly by the undesirable tendency of GLR, which has been described in Chapter 2 (Section 2.3.1), contributing to incorrect merging between heterogeneous clusters (in terms of speaker-specific characteristics) throughout the entire merging recursions in AHSC. In this chapter, we particularly focus on and tackle the negative effect of this problem on both early and late recursion steps of AHSC.

This chapter is organized as follows. In Section 3.2, we investigate the reason why the reliability of the GLR-based inter-cluster distance measure for AHSC severely varies across data sources, from a viewpoint of early AHSC recursion steps. Based on this investigation, in Section 3.3, we propose modified versions of AHSC, which are verified through experiments to enhance the reliability of the GLR-based inter-cluster distance measure at the early recursion steps of AHSC. As a result, all the modified AHSCs proposed in this section obtain better performance than basic AHSC in terms of inter-cluster distance measurement and thus speaker error time rate (assuming perfect estimation of the optimal stopping point for AHSC). In Section 3.4, based on the investigation done in Section 3.2,

we also propose a new method to measure distance between clusters at the late recursion steps of AHSC, which is to combine the advantages of GLR and ICR (proposed in Chapter 2). This novel method is demonstrated through experimental results to be better than the conventional GLR-based inter-cluster distance measure at later merging recursions in AHSC (assuming perfect estimation of the optimal stopping point for AHSC). One issue that needs to be addressed in those modified speaker clustering strategies in Sections 3.3 and 3.4 is that they are beneficial only when the optimal stopping point is accurately detected, which is not the case all the time in real situations. In this regard, in Section 3.5, we propose another modified version of AHSC, called selective AHSC, which offers better speaker clustering performance than the strategies dealt in Sections 3.3 and 3.4 under ICR-based stopping point estimation. In Section 3.6, we conclude this chapter with comments on future research work with regard to those handled in the entire chapter.

3.2 GLR at Early AHSC Recursion Steps

As examined in Section 2.3.1, GLR tends to get larger as the total number of feature vectors within a pair of clusters under consideration increases. Figure 3.1 explicitly shows this tendency. During AHSC, the tendency of GLR causes a pair of homogeneous clusters (in terms of speaker-specific characteristics) of small size to have a smaller GLR value and be regarded as mutually closer than those of large size.

This tendency of GLR leads AHSC with the GLR-based inter-cluster distance measure to preferentially select short speech segments (among the entire speech segments given as an input to AHSC) as the closest for merging at the early recursion steps of AHSC. This can be well noticed in Table 3.1. From the fourth row (‘sub-total’) of the table, we can observe that the speech segments shorter than 3 seconds are involved in more than a half of the first quarter of the entire merging recursions during AHSC for all the data sources in the development data set presented in Section 2.2. This trend is particularly distinct for C-1, C-2, and I-1 (92.38%, 88.57%, and 90.00% respectively), which seems

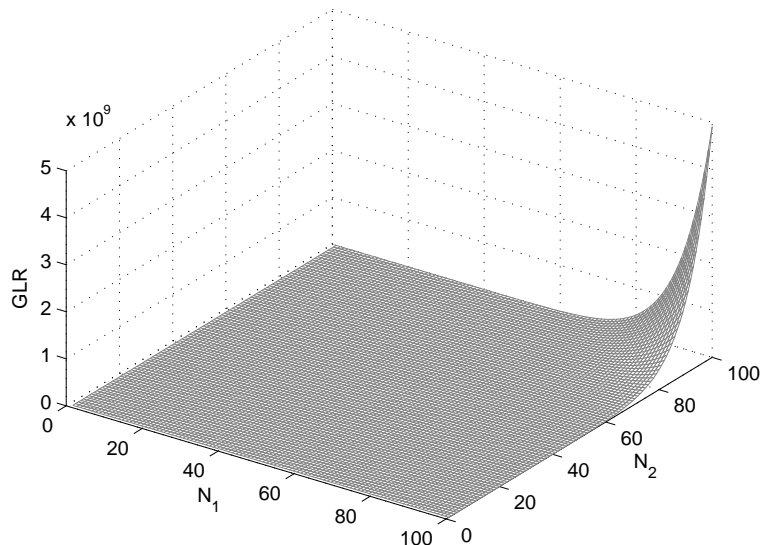


Figure 3.1: Figure 2.1 revisited. This figure displays GLR for two clusters C_1 and C_2 along with the number of feature vectors in each cluster. The second order statistics of the corresponding cluster models are fixed at $\mu_1 = 0$, $\mu_2 = 1$, and $\Sigma_1 = \Sigma_2 = 1$.

reasonable because these data sources contain a large number of short speech segments as shown in Figure 3.2. The interesting point that we observed from the first quarter of the entire merging recursions during AHSC for every data source in the development data set is that the accuracy of the GLR-based inter-cluster distance measure for AHSC stayed perfect for merging between the speech segments longer than or equal to 3 seconds (M_{ll}), which is shown in Table 3.2. In contrast, the accuracy became lower for the other merging types (M_{ss} and M_{sl}). In other words, only short speech segments¹ were involved in all the incorrect merging. In this context, we can say that incorrect merging at the early recursion steps of AHSC are more likely to occur for the data sources having a large number of short speech segments.

Considering that AHSC has a recursive structure and thus any incorrect merging during AHSC becomes a potential seed for other incorrect merging recursions, such incorrect merging at the early recursion steps of AHSC due to the aforementioned tendency of

¹From this point on, let us call the speech segments shorter than 3 seconds short speech segments. Accordingly, let us call the speech segments longer than or equal to 3 seconds long speech segments.

Table 3.1: Distribution of three different merging types (M_{ss} , M_{sl} , and M_{ll}) at the first quarter of the entire merging recursions during AHSC for every data source in the development data set in Section 2.2. M_{ss} : merging between the speech segments shorter than 3 seconds, M_{sl} : merging between one speech segment shorter than 3 seconds and the other longer than or equal to 3 seconds, and M_{ll} : merging between the speech segments longer than or equal to 3seconds.

	C-1	C-2	C-3	N-1	I-1
M_{ss}	60.95%	52.86%	21.32%	13.33%	50.00%
M_{sl}	31.43%	35.71%	39.34%	42.22%	40.00%
sub-total	92.38%	88.57%	60.66%	55.55%	90.00%
M_{ll}	7.62%	11.43%	39.34%	44.45%	10.00%

Table 3.2: Accuracy of the GLR-based inter-cluster distance measure for AHSC depending on the merging types defined in Table 3.1. These accuracies were obtained only based on the first quarter of the entire merging recursions during AHSC for every data source in the development data set in Section 2.2.

	M_{ss}	M_{sl}	M_{ll}
Accuracy	88.89%	93.81%	100.00%

GLR can be regarded as one of direct causes for high speaker error time rate. This is confirmed in an indirect way in Figure 3.3, which compares speaker error time rate (under the assumption of perfect estimation of the optimal stopping point estimation) for each data source in the development data set with that for the corresponding subset containing long speech segments only. From this figure, we can observe that performance improvement would be achieved for most of the data sources without short speech segments. The improvement is considerable for C-1, C-2 and I-1, where short speech segments have a relatively large portion compared to the other data sources (C-3 and N-1).

Based on all of these, we can conclude that the portion of short speech segments in the entire speech segments given as an input to AHSC can affect speaker error time

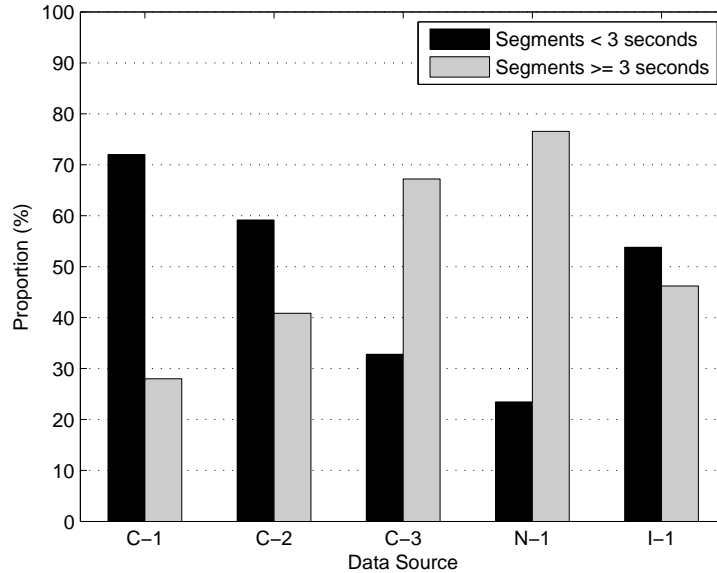


Figure 3.2: Segment length distributions for the development data set in Section 2.2.

rate, because of the undesirable tendency of GLR that it tends to get larger as the total number of feature vectors within a pair of clusters under consideration increases. For better AHSC performance, we thus need to mitigate this negative effect of short speech segments on the GLR-based inter-cluster distance measure, which will be handled in the next section.

3.3 Modification of AHSC

In this section, in order to address the problem (mentioned in the previous section) of incorrect merging between short speech segments at the early recursion steps of AHSC, we propose three modified versions of AHSC to constrain merging between short speech segments especially at the early recursion steps of AHSC so as to minimize its effect on speaker error time rate under GLR-based inter-cluster distance measurement. The first two modified clustering strategies try to avoid merging between short speech segments (M_{ss}) because the accuracy of the GLR-based inter-cluster distance measure for M_{ss} is relatively worse than that for the other merging types (M_{sl} and M_{ll}). The third modified

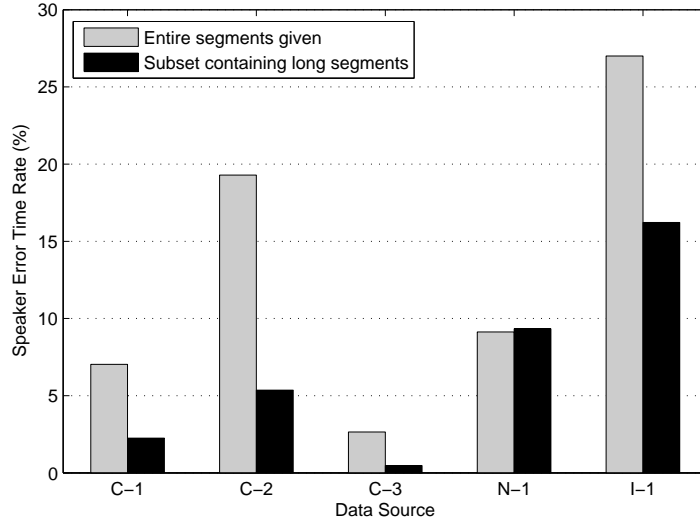


Figure 3.3: Speaker error time rate by AHSC with perfect detection of the optimal stopping point for the development data set in Section 2.2. This figure compares performance for the entire speech segments given as an input to AHSC with that for the corresponding subset containing the segments longer than or equal to 3 seconds only.

AHSC tries to preferentially consider M_{ll} other than M_{ss} or M_{sl} to utilize the high accuracy of the GLR-based inter-cluster distance measure for M_{ll} at the early recursion steps of AHSC. In the next three sub-sections, we will explain those strategies more in detail, respectively.

3.3.1 Constrained Cluster Selection for Merging

The first modified version of AHSC is to prevent M_{ss} by allowing only M_{sl} or M_{ll} during the entire AHSC procedures. If the pair of clusters selected for merging at a certain recursion step of AHSC are both short speech segments, the next closest pair of clusters at the recursion step are considered for merging until the pair of clusters considered are not both short speech segments. (See Algorithm 2.) This idea is based on the results in Table 3.2, showing that the accuracy of the GLR-based inter-cluster distance measure for M_{ss} is worse than that for M_{sl} or M_{ll} .

Algorithm 2 Modified Version 1 of AHSC

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: speech segments
 $\hat{C}_i, i = 1, \dots, \hat{n}$: initial clusters
Ensure: $C_i, i = 1, \dots, n$: finally remaining clusters
1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$
2: **do**
3: $i, j \leftarrow \arg \min_{k, l = 1, \dots, \hat{n}, k \neq l} \text{GLR}(\hat{C}_k, \hat{C}_l)$ such that either $\{\mathbf{x}_k\}$ or $\{\mathbf{x}_l\}$ is a long speech segment,
4: merge \hat{C}_i and \hat{C}_j
5: $\hat{n} \leftarrow \hat{n} - 1$
6: **until** optimal stopping point
7: **return** $C_i, i = 1, \dots, n$

Table 3.3: Comparison of basic AHSC and its first modified version in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. Both of the clustering strategies use the GLR-based inter-cluster distance measure to select clusters for merging at every recursion step of AHSC, and perfect stopping point estimation is assumed. (For each result in the table, the corresponding standard deviation is presented as well.)

	Basic AHSC	Modified Version 1
Dev.	13.02% (± 9.92)	10.90% (± 6.80)
Eval.	14.84% (± 9.00)	13.60% (± 8.71)

This modified version of AHSC, as shown in Table 3.3, provides better clustering performance (in terms of averaged speaker error time rate over data sources) than basic AHSC for both of the development and evaluation data set in Section 2.2 by 2.12% and 1.24% (absolute) respectively. This overall improvement in speaker clustering performance is achieved by the enhancement of the reliability of the GLR-based inter-cluster distance measure in the modified AHSC, which is supported by the reduced standard deviation of the performance results by the modified AHSC shown in the right column of the table. We can confirm from the results in this table that preventing merging between short speech segments from occurring at the early recursion steps of AHSC would improve the reliability of AHSC performance as a consequence, as expected.

Algorithm 3 Modified Version 2 of AHSC

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: speech segments
 $\hat{C}_i, i = 1, \dots, \hat{n}', \hat{n}' \leq \hat{n}$: initial clusters

Ensure: $C_i, i = 1, \dots, n$: finally remaining clusters

- 1: sort $\{\mathbf{x}_i\}$ in the descending order of length
- 2: $\hat{C}_j \leftarrow \{\mathbf{x}_i\}$ such that $\{\mathbf{x}_i\}$ is a long speech segment, $i = 1, \dots, \hat{n}$ and $j = 1, \dots, \hat{n}'$
- 3: $m = \hat{n}' + 1$
- 4: **do**
- 5: $\hat{C} \leftarrow \{\mathbf{x}_m\}$
- 6: $i \leftarrow \arg \min \text{GLR}(\hat{C}, \hat{C}_k), k = 1, \dots, \hat{n}'$
- 7: merge \hat{C} to \hat{C}_i
- 8: $m \leftarrow m + 1$
- 9: **until** $m > \hat{n}$
- 10: **do**
- 11: $i, j \leftarrow \arg \min \text{GLR}(\hat{C}_k, \hat{C}_l), k, l = 1, \dots, \hat{n}', k \neq l$
- 12: merge \hat{C}_i and \hat{C}_j
- 13: $\hat{n}' \leftarrow \hat{n}' - 1$
- 14: **until** optimal stopping point
- 15: **return** $C_i, i = 1, \dots, n$

Table 3.4: Comparison of basic AHSC and its second modified version in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. The same distance measure and assumption for stopping point estimation in AHSC as ones in Table 3.3 are applied.

	Basic AHSC	Modified Version 2
Dev.	13.02% (± 9.92)	11.67% (± 9.72)
Eval.	14.84% (± 9.00)	14.82% (± 9.87)

3.3.2 Pre-Classification of Short Speech Segments

The second modified version is to merge every short speech segment to a long speech segment prior to AHSC. It has the same basic idea as the first modified AHSC does in the sense of preventing M_{ss} during the entire AHSC procedures, but is a different approach to implementing the idea. This modified version of AHSC first has each of short speech segments merged to the closest long speech segment in terms of GLR, and then runs AHSC on the remaining set of speech segments containing long ones only. (See Algorithm 3.)

Algorithm 4 Modified Version 3 of AHSC

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: speech segments, ζ : threshold
 $\hat{C}_i, i = 0, \dots, \hat{n}', \hat{n}' \leq \hat{n}$: intermediate clusters
Ensure: $C_i, i = 1, \dots, n$: finally remaining clusters

- 1: sort $\{\mathbf{x}_i\}$ in the descending order of length
- 2: $\hat{C}_1 \leftarrow \{\mathbf{x}_1\}, \hat{n}' = 1, m = 2$
- 3: **do**
- 4: $\hat{C} \leftarrow \{\mathbf{x}_m\}$
- 5: $i \leftarrow \arg \min \text{GLR}(\hat{C}, \hat{C}_k), k = 1, \dots, \hat{n}'$
- 6: **if** $\min \text{GLR}(\hat{C}, \hat{C}_i) > \zeta$
- 7: $\hat{n}' = \hat{n}' + 1$
- 8: $\hat{C}_{\hat{n}'} = \hat{C}$
- 9: **else**
- 10: merge \hat{C} to \hat{C}_i
- 11: $m \leftarrow m + 1$
- 12: **until** $m > \hat{n}$
- 13: **do**
- 14: $i, j \leftarrow \arg \min \text{GLR}(\hat{C}_k, \hat{C}_l), k, l = 1, \dots, \hat{n}', k \neq l$
- 15: merge \hat{C}_i and \hat{C}_j
- 16: $\hat{n}' \leftarrow \hat{n}' - 1$
- 17: **until** optimal stopping point
- 18: **return** $C_i, i = 1, \dots, n$

This second modified version of AHSC, as shown in Table 3.4, also provides better clustering performance (in terms of averaged speaker error time rate over data sources) than basic AHSC for both of the development and evaluation data set in Section 2.2 by 1.24% and 0.02% (absolute) respectively, which is however a bit worse than the performance improvement of the first modified version of AHSC in the previous sub-section. One interesting point is that the standard deviation of the modified AHSC performance for the evaluation data set is higher than that of the basic AHSC performance. This suggests that this version of AHSC cannot provide better reliability in terms of AHSC performance although it can offer better overall speaker time error rate than basic AHSC.

3.3.3 Sequential Clustering prior to AHSC

The third modified version is a bit different from the other two versions previously proposed. Instead of pre-screening M_{ss} , this modified version of AHSC just reduces the proportion of M_{ss} (and M_{sl} as well) at the early recursion steps of AHSC by letting long speech segments be preferentially considered for merging through sequential clustering

prior to AHSC. Specifically, it first sorts the entire speech segments (given as an input to AHSC) in the descending order of length, runs leader-follower clustering² (LFC) [18] on the sorted segment set, and performs AHSC on the clusters provided by LFC. (See Algorithm 4.) The threshold ζ used in LFC was empirically set to be 250.0.

This third modified version of AHSC, as shown in Table 3.5, also provides better clustering performance (in terms of averaged speaker error time rate over data sources) than basic AHSC for both of the development and evaluation data set in Section 2.2 by 3.41% and 1.04% (absolute) respectively, which is the best overall performance improvement among those proposed thus far as shown in Figure 3.4. Comparing the standard deviations of the results in the table, we can see that this modified AHSC provides more reliability for clustering performance across data sources like the first modified AHSC did in Section 3.3.1.

3.4 Combination of GLR and ICR

In the previous two sections, we have focused on the early recursion steps of AHSC regarding GLR-based inter-cluster distance measurement. In this section, we move our attention to the GLR-based inter-cluster distance measure at the late recursion steps of AHSC. Let us start the section from bringing back the undesirable tendency of GLR mentioned in Section 3.2, i.e.,

- A pair of homogeneous clusters of small size are likely to have a smaller GLR value and be regarded as mutually closer than those of large size.

This tendency leads to the following:

- A pair of heterogeneous clusters of small size might have a smaller GLR value and be regarded as mutually closer than a pair of homogeneous clusters of large size.

²In this sequential clustering strategy, input data are classified in the order of incoming without any pre-trained class model. Thus, the first incoming datum automatically becomes the first class and every datum thereafter either is merged to one of existing class(es) or becomes another new class.

Table 3.5: Comparison of basic AHSC and its third modified version in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. The same distance measure and assumption for stopping point estimation in AHSC as ones in Table 3.3 are applied.

	Basic AHSC	Modified Version 3
Dev.	13.02% (± 9.92)	9.61% (± 8.41)
Eval.	14.84% (± 9.00)	13.80% (± 8.24)

In other words, the tendency could cause incorrect merging during AHSC. Incorrect merging is more detrimental to speaker error time rate when it occurs at the late recursion steps of AHSC than elsewhere. This is because average cluster size increases as merging recursions continue during AHSC, and thus any incorrect merging at the late recursion steps of AHSC is likely to occur between large size clusters. Such an incorrect merging at the late recursion steps of AHSC would generally raise speaker error time rate much more than that between small size clusters at any other recursion steps. Therefore, inter-cluster distance measurement needs to be more accurate at the late recursion steps of AHSC, for which in this section we propose a novel alternative to the GLR-based inter-cluster distance measure that we can apply to the late recursion steps of AHSC. This alternative distance measurement method is to consider both GLR and ICR (proposed in Section 2.4.1) in selection of clusters for merging at the late recursion steps of AHSC, and is motivated by the idea that ICR could be utilized as a complement inter-cluster distance measure to GLR in the sense that it could possibly compensate for the aforementioned undesirable tendency of GLR if we are able to manipulate it to handle large clusters only. As mentioned in Section 2.4.3, ICR would properly work as a sort of distance measure between clusters only if it handled large clusters.

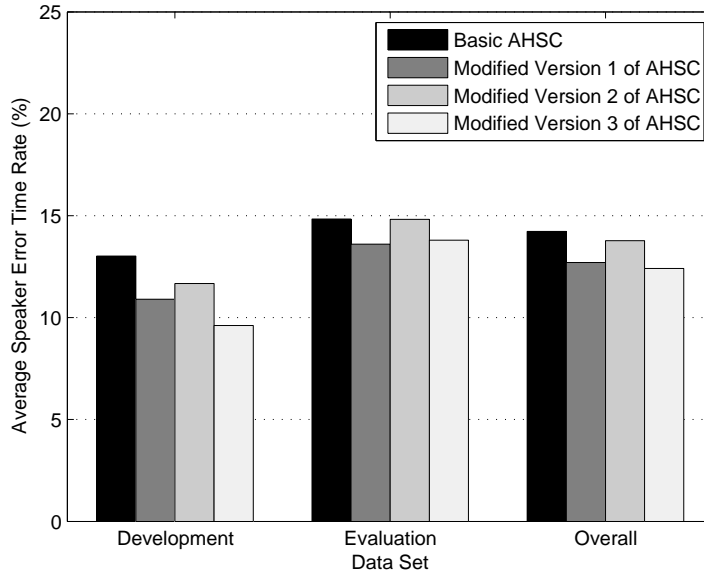


Figure 3.4: Comparison of basic AHSC and its three modified versions proposed in terms of average speaker error time rate.

3.4.1 (GLR+ICR)-based Inter-Cluster Distance Measurement

The new method to measure distance between clusters that we propose here basically depends upon GLR at every recursion step of AHSC, but starts to additionally consider ICR from a certain recursion step of AHSC where all remaining clusters contain data samples of more than 30 seconds in amount of time. (See Algorithm 5.) Since co-consideration of ICR begins only at such a recursion step, this method is naturally applicable to the late recursion steps of AHSC. The reason why 30 seconds is specifically chosen here is, as aforementioned above, that ICR could properly work as an inter-cluster distance measure if every cluster considered were large enough to fully represent the corresponding speaker-specific characteristics. In this section, we conservatively assume that a cluster containing feature vectors which correspond to more than 30 seconds in amount of time is a large enough cluster to represent speaker-specific characteristics completely, as we did in Section 2.4.3.

Algorithm 5 AHSC with combination of GLR and ICR as an inter-cluster distance measure

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: speech segments
 $\hat{C}_i, i = 1, \dots, \hat{n}$: initial clusters

Ensure: $C_i, i = 1, \dots, n$: finally remaining clusters

- 1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$
- 2: **do**
- 3: **if** all $\{\hat{C}_i\}_{i=1}^{\hat{n}}$ contain data of more than 30 seconds
- 4: $i, j \leftarrow \arg \min \left[\text{SR}_{\text{GLR}}(\hat{C}_k, \hat{C}_l) + \text{SR}_{\text{ICR}}(\hat{C}_k, \hat{C}_l) \right],$
 SR_{GLR} (or SR_{ICR}) : soft ranking of cluster pairs in terms of GLR (or ICR),
 $k = 1, \dots, \hat{n}$, and $l = k + 1, \dots, \hat{n}$
- 5: **else**
- 6: $i, j \leftarrow \arg \min \text{GLR}(\hat{C}_k, \hat{C}_l),$
 $k = 1, \dots, \hat{n}$, and $l = k + 1, \dots, \hat{n}$
- 7: merge \hat{C}_i and \hat{C}_j
- 8: $\hat{n} \leftarrow \hat{n} - 1$
- 9: **until** optimal stopping point
- 10: **return** $C_i, i = 1, \dots, n$

In order to consider both GLR and ICR when selecting clusters for merging at the late recursion steps of AHSC where the aforementioned condition is satisfied, the proposed method utilizes the sum of rankings (in terms of GLR and ICR) for the entire pairs of clusters at the recursion step considered, as a means of information fusion. Specifically, each pair of clusters at the recursion step of AHSC considered is ranked in two ways, one of which is in terms of GLR and the other is in terms of ICR. The smaller GLR (or ICR) value a certain pair of clusters have, the higher they are ranked in terms of GLR (or ICR). The proposed method selects the pair of clusters having the smallest summed ranking for merging. We use such a high level fusion strategy to exploit ‘ranking’ because GLR is empirically shown to have much wider variance than ICR for any given cluster pair, and thus low level fusion strategies like score normalization could cause GLR to be extremely dominant over ICR in selection of clusters for merging in our case.

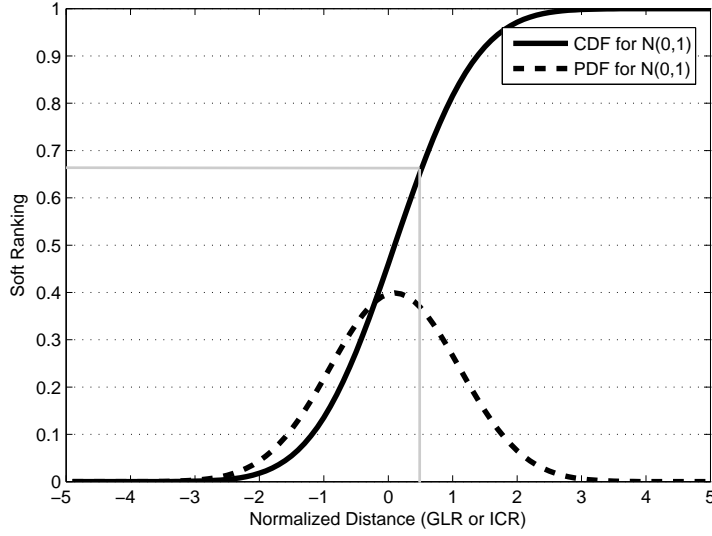


Figure 3.5: Soft ranking used in the proposed inter-cluster distance measurement method. If a certain pair of clusters have the normalized distance of 0.5, their soft ranking becomes 0.69 (grey line) in this system.

As for information fusion in the proposed method, we use ‘soft ranking’ when ranking clusters in terms of GLR and ICR. Each soft ranking is defined as follows:

$$\text{SR}_{\text{GLR}}(\hat{C}_k, \hat{C}_l) \triangleq F_{\mathcal{N}} \left\{ \frac{\text{GLR}(\hat{C}_k, \hat{C}_l) - \mu_{\text{GLR}}}{\sigma_{\text{GLR}}} \right\} \quad (3.1)$$

$$\text{SR}_{\text{ICR}}(\hat{C}_k, \hat{C}_l) \triangleq F_{\mathcal{N}} \left\{ \frac{\text{ICR}(\hat{C}_k, \hat{C}_l) - \mu_{\text{ICR}}}{\sigma_{\text{ICR}}} \right\}, \quad (3.2)$$

where μ_{GLR} and σ_{GLR} (or μ_{ICR} and σ_{ICR}) are mean and standard deviation for the GLR (or ICR) values of the entire cluster pairs at the recursion step of AHSC considered, and $F_{\mathcal{N}}(\cdot)$ is a normal cumulative density function with zero mean and unit variance. This soft ranking approach normalizes inter-cluster distances, assuming that they are normally distributed, and transforms them through a monotonic increasing function, so as to provide a sort of relative information between clusters. (See Figure 3.5.)

AHSC with our proposed method to measure distance between clusters, as shown in Table 3.6, provides better clustering performance (in terms of averaged speaker error time

Table 3.6: Comparison of AHSC with the GLR-based inter-cluster distance measure and that with our proposed method, in terms of average speaker error time rate for the development and evaluation data set in Section 2.2. Perfect stopping point estimation for AHSC is assumed.

	GLR	GLR+ICR
Dev.	13.02%	10.20%
Eval.	14.84%	14.64%

rate over data sources) than that with the conventional GLR-based inter-cluster distance measure for both of the development and evaluation data set in Section 2.2 by 2.82% and 0.20% (absolute) respectively. This improvement comes from, as expected, the reduced number of incorrect merging occurrences at the late recursion steps of AHSC. Based on these results, we can expect that applying this method to the late recursion steps of the modified AHSC approaches proposed in Section 3.3 would result in extra performance improvement.

3.4.2 Proposed Measure in Modified AHSC Approaches

Figure 3.6 explicitly displays the extra performance improvement that would be achieved if the proposed (GLR+ICR)-based inter-cluster distance measure were applied to the late recursion steps of the three modified versions of AHSC introduced in Section 3.3. The overall results in this figure indicate that the proposed, supplement inter-cluster distance measure does not degenerate the merits of the modified AHSC approaches at the early recursion steps, retaining its merit at the late recursion steps. The most outstanding improvement is 2.94% (absolute) for both of the modified versions 1 and 2 on the evaluation data set, while performance improvement for the modified version 3 is not significant.

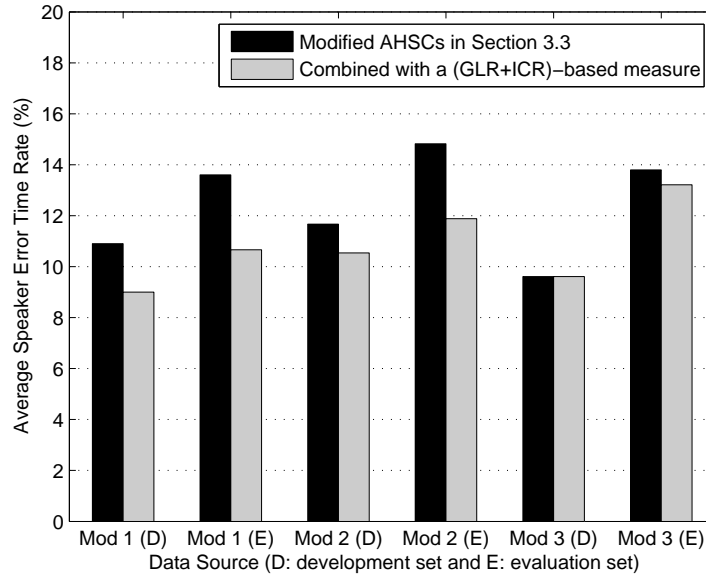


Figure 3.6: Extra performance improvement achieved if the proposed (GLR+ICR)-based inter-cluster distance measure were applied to the late recursion steps of the three modified versions of AHSC introduced in Section 3.3. The data sources used in this experiment are the development and evaluation data set in Section 2.2. Perfect estimation of the optimal stopping point for AHSC is assumed.

3.5 Selective AHSC

In the previous sections, we have proposed three modified versions of AHSC to tackle incorrect merging between heterogenous clusters (in terms of speaker-specific characteristics) at the early recursion steps of AHSC. Furthermore, we have introduced a new supplement inter-cluster distance measure to handle incorrect merging at the late recursion steps of AHSC. From those approaches, we could obtain performance improvement for the evaluation data set (in Section 2.2) in terms of average speaker error time rate by up to 4.18% (absolute) and 28.17% (relative), which is re-organized in Table 3.7. However, they work only under the assumption of perfect estimation of the optimal stopping point in AHSC.

In this section, we test how badly the clustering strategies proposed in Sections 3.3 and 3.4 work with the ICR-based stopping point estimation method (proposed in Chapter

Table 3.7: Average speaker error time rate for the evaluation data set in Section 2.2. This table compares AHSC and its three modified versions with both GLR-based and (GLR+ICR)-based inter-cluster distance measurement. Perfect estimation of the optimal stopping point for AHSC is assumed.

	AHC	Mod 1	Mod 2	Mod 3
GLR	14.84%	13.60%	14.82%	13.80%
GLR+ICR	14.64%	10.66%	11.88%	13.21%

Table 3.8: Average speaker error time rate for the evaluation data set in Section 2.2 when the ICR-based stopping point estimation method is applied. This table compares AHSC and its three modified versions with GLR-based and (GLR+ICR)-based inter-cluster distance measurement, as Table 3.7 does.

	AHC	Mod 1	Mod 2	Mod 3
GLR	15.73%	15.65%	18.48%	16.51%
GLR+ICR	22.42%	16.00%	14.18%	19.18%

2). In this regard, we propose a new clustering strategy, i.e., selective AHSC, utilizing (relatively) accurate stopping point estimation by the ICR-based stopping point estimation method and high reliability by the GLR-based inter-cluster distance measure for long speech segments. This clustering approach is empirically verified to be one of possible combinations that can well coordinate our results in Chapters 2 and 3.

3.5.1 Modified AHSCs with Stopping Point Estimation

The clustering performance that would be achieved from the three modified versions of AHSC (proposed in Section 3.3) if the ICR-based stopping point estimation method were applied instead of the assumption of perfect estimation of the optimal stopping point is shown in Table 3.8. Considering Table 3.7 together, we can see that incorrect stopping point estimation by the ICR-based method mostly erases the advantages of the

modified versions of AHSC that were obtained under the assumption of perfect stopping point estimation. A noticeable thing in Table 3.8 is that the (GLR+ICR)-based inter-cluster distance measure and the ICR-based stopping point estimation method do not work well together, which can be easily shown when we compare the results in the second row and their counterparts in the third row. Except for the second modified version of AHSC, clustering performance is observed to be degraded in every case when the two approaches are applied together. This is caused because the ICR-based stopping point estimation method starts its estimation process from the pair of clusters merged at the last recursion step of AHSC by comparing the ICR value of the pair with a pre-set threshold (η in Chapter 2). However, the (GLR+ICR)-based inter-cluster distance measure is likely to select the clusters having a small ICR (and GLR) value for merging, especially at the late recursion steps of AHSC. Thus, the ICR-based stopping point estimation method is more likely to confuse itself in estimating the optimal stopping point in AHSC (and its modified versions) under (GLR+ICR)-based inter-cluster distance measurement. Therefore, in practical applications, the modified versions of AHSC with the (GLR+ICR)-based inter-cluster distance measure introduced in the previous section need a better stopping estimation method using a different (or independent) measure from ICR. We do not further take care of this issue in this dissertation, remaining it as a future research topic.

3.5.2 Selective AHSC

In order to better coordinate our results in Chapters 2 and 3, we propose selective AHSC in this section. This proposed method is motivated by the same reason for the other modified versions of AHSC introduced in Section 3.3, which is well shown in Figure 3.7. We can see from the figure that the accuracy of the GLR-based inter-cluster distance measure for AHSC would mostly rise up and thus result in better clustering performance when an input to AHSC contains only long speech segments. Selective AHSC can utilize this advantage that could be obtained when dealing with long speech segments only, by

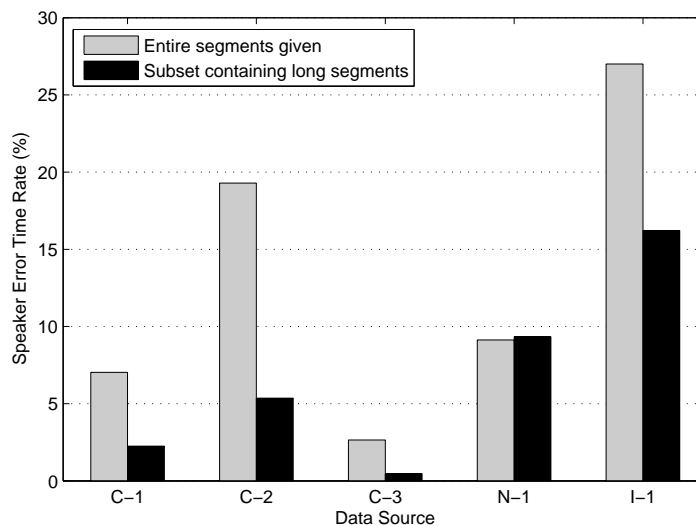


Figure 3.7: Figure 3.3 revisited, showing speaker error time rate by AHSC with perfect detection of the optimal stopping point for the development data set in Section 2.2. This figure compares performance for the entire speech segments given as an input to AHSC with that for the corresponding subset containing the segments longer than or equal to 3 seconds only.

first running basic AHSC with the GLR-based inter-cluster distance measure and the ICR-based stopping point estimation method only on long speech segments among the entire given speech segments, and then classifying the rest (i.e., short speech segments) into one of the clusters provided by the initial AHSC step. (See Algorithm 6.) By selective classification of speech segments in terms of length, selective AHSC can mitigate the negative effect of short speech segments on GLR-based inter-cluster distance measurement during AHSC, especially at the early recursion steps.

Note that robust stopping point estimation to the variation of input speech data, like by the ICR-based stopping point estimation method, is critical to selective AHSC due to selective consideration of speech segments at the initial AHSC step. Such selective consideration causes the variability of an input to AHSC, so selective AHSC would not work properly if it were with any other stopping point estimation method not robust to the variation of data sources. How badly the BIC-based stopping point estimation method, which has been verified in Chapter 2 to be not robust to the variation of input

Algorithm 6 Selective AHSC

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: speech segments
 $\hat{C}_i, i = 1, \dots, \hat{n}', \hat{n}' \leq \hat{n}$: initial clusters

Ensure: $C_i, i = 1, \dots, n$: finally remaining clusters

- 1: permute $\{\mathbf{x}_i\}$ in the descending order of length
- 2: $\hat{C}_j \leftarrow \{\mathbf{x}_i\}$ such that $\{\mathbf{x}_i\}$ is a long speech segment, $i = 1, \dots, \hat{n}$ and $j = 1, \dots, \hat{n}'$
- 3: $m = \hat{n}'$
- 4: **do**
- 5: $i, j \leftarrow \arg \min \text{GLR}(\hat{C}_k, \hat{C}_l), k, l = 1, \dots, m, k \neq l$
- 6: merge \hat{C}_i to \hat{C}_j
- 7: $m \leftarrow m - 1$
- 8: **until** optimal stopping point (detected by ICR-based stopping point estimation)
- 9: **return** $C_i, i = 1, \dots, n$
- 10: $m = \hat{n}' + 1$
- 11: **do**
- 12: $\hat{C} \leftarrow \{\mathbf{x}_m\}$
- 13: $i \leftarrow \arg \min P(\hat{C}|\hat{C}_k), k = 1, \dots, n$
- 14: merge \hat{C} to \hat{C}_i
- 15: $m \leftarrow m + 1$
- 16: **until** $m > \hat{n}$
- 17: **return** $C_i, i = 1, \dots, n$

speech data, would break selective AHSC performance is given in Figure 3.8. The results in this figure suggest that without robust stopping point estimation we cannot get any benefit from this novel approach to speaker clustering.

Figure 3.9 shows the performance of selective AHSC with robust stopping point estimation for the evaluation data set in Section 2.2. From this figure, we can see that selective AHSC is a reasonable strategy to coordinate our results in Chapters 3 and 4. The performance of selective AHSC with ICR-based stopping point estimation is shown to be better than that of basic AHSC with perfect stopping point estimation for every data source except C-4, C-5, and I-3. Even for the three data sources, performance gap is negligible. The result that average speaker error time rate by selective AHSC for the evaluation data set (in Section 2.2) is even better than that by AHSC with perfect stopping point estimation can be regarded as promising. Table 3.9 explicitly indicates the superiority of selective AHSC over all the counterparts that have been dealt with in this dissertation.

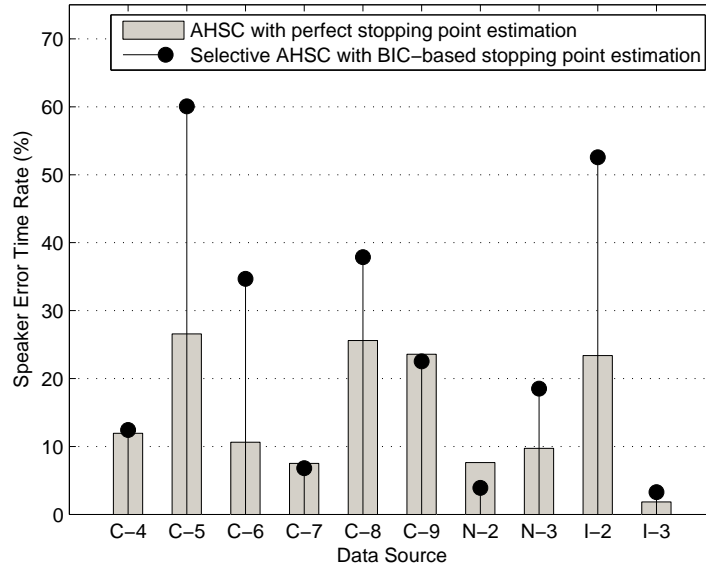


Figure 3.8: Comparison of basic AHSC with the assumption of perfect estimation of the optimal stopping point and selective AHSC (including the BIC-based stopping point estimation method), in terms of speaker error time rate on the evaluation data set in Section 2.2.

3.6 Conclusions

In this chapter, we addressed the robustness problem of the GLR-based inter-cluster distance measure for AHSC to the variation of input speech data. For this, we proposed 1) three modified versions of AHSC so as to enhance the reliability (or accuracy) of the GLR-based inter-cluster distance measure at the early recursion steps AHSC and 2) a (GLR+ICR)-based inter-cluster distance measure so as to improve reliability in measuring inter-cluster distance at the late recursion steps of AHSC. Through experimental results on the excerpts obtained from meeting corpora, all the proposed ones are demonstrated to provide better clustering performance than basic AHSC with the GLR-based inter-cluster distance measure in terms of averaged speaker error time rate over data sources. Furthermore, we proposed selective AHSC to better coordinate the merits of our research results in this chapter with ICR-based stopping point estimation (proposed in Chapter 2)

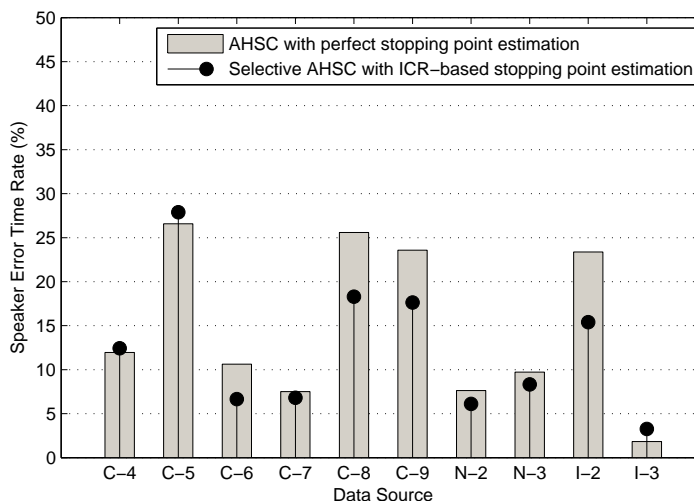


Figure 3.9: Comparison of basic AHSC with the assumption of perfect estimation of the optimal stopping point and selective AHSC (including the ICR-based stopping point estimation method), in terms of speaker error time rate on the evaluation data set in Section 2.2.

in more realistic situations for speaker clustering. This novel clustering strategy provided AHSC performance with even better reliability over input speech data.

There are several directions for future work including further refinements to the proposed solutions. For instance, in the third modified version of AHSC, the threshold parameter ζ determines the number of intermediate clusters, which is directly linked to the final speaker error time rate. It was chosen empirically in this chapter, but finding ways for optimally setting ζ would be beneficial in further enhancing clustering performance. As another example, we might have to consider how to optimally fuse two different statistical information on the same object for (GLR+ICR)-based inter-cluster distance measurement at the late recursion steps of AHSC. In this chapter, we used soft rankings in terms of GLR and ICR for that purpose, but it is not theoretically proven to be optimal to the task considered. Establishing more systematic frameworks for selection of information fusion methods could be one of valuable future research directions.

In addition, as mentioned earlier in the middle part of this chapter, it would be a good research topic to find out a stopping point estimation method for AHSC with the

(GLR+ICR)-based inter-cluster distance measure, other than the ICR-based one. A new stopping point estimation method should be comparable to our proposed ICR-based one in terms of estimation accuracy, but needs to use an inter-cluster homogeneity decision measure independent of ICR. Then it could keep the advantages of the modified versions of AHSC and the (GLR+ICR)-based inter-cluster distance measure valid even in practical applications.

Table 3.9: Average speaker error time rate for the evaluation data set in Section 2.2. This table compares selective AHSC with all the counterparts that have been dealt with in this dissertation.

	Key Components	Performance
AHSC	Distance Measure: GLR Stopping Method: BIC	<i>24.49%</i>
AHSC	Distance Measure: GLR Stopping Method: ICR	<i>15.73%</i>
Modified Version 1 of AHSC	Distance Measure: GLR Stopping Method: ICR	15.65%
Modified Version 1 of AHSC	Distance Measure: (GLR+ICR) Stopping Method: ICR	16.00%
Modified Version 2 of AHSC	Distance Measure: GLR Stopping Method: ICR	18.48%
Modified Version 2 of AHSC	Distance Measure: (GLR+ICR) Stopping Method: ICR	<i>14.18%</i>
Modified Version 3 of AHSC	Distance Measure: GLR Stopping Method: ICR	16.51%
Modified Version 3 of AHSC	Distance Measure: (GLR+ICR) Stopping Method: ICR	19.18%
Selective AHSC	Distance Measure: GLR Stopping Method: ICR	<i>12.28%</i>

Chapter 4

Robust Cluster Modeling for Inter-Cluster Distance Measurement in AHSC

4.1 Introduction

Thus far we have tried to address the robustness problem of AHSC performance under the variation of input speech data in this dissertation. To overcome the problem in the perspective stopping point estimation, we proposed a more reliable way to determine the optimal (recursion) stopping point for AHSC across a variety of input speech data than the conventional method [13] utilizing Bayesian information criterion (BIC) [58] (Chapter 2). Specifically we defined a new statistical distance measure between clusters, i.e., information change rate (ICR), and applied it to stopping point estimation for AHSC based on its superiority to BIC in terms of robustness to input data variation. To tackle the robustness problem of AHSC performance from the viewpoint of inter-cluster distance measurement, on the other hand, we claimed and verified in Chapter 3 that short speech segments ($< 3s$, in general) among input speech data degraded the accuracy of picking up the closest cluster pair especially at the early recursion steps of AHSC, and proposed a variety of schemes to prevent short input speech segments from negatively affecting distance measurement of clusters. All of the proposed schemes were empirically verified to offer clustering performance improvement particularly for the input data suffering from the negative effect of short speech segments, meaning that they can enhance the

reliability of AHSC performance against short input speech segments. In addition, we introduced a new inter-cluster distance measure by combining generalized likelihood ratio (GLR) [26] and ICR. This metric mitigated the accuracy degradation of GLR-based inter-cluster distance measurement at the late recursion steps of AHSC, caused by unbalanced speaking time distribution over speaker sources in input speech data.

In this chapter we tackle the robustness problem of AHSC performance across input speech data, in terms of statistical cluster modeling for inter-cluster distance measurement. This work was motivated by the reasoning that ideal cluster modeling for inter-cluster distance measurement within the framework of AHSC should account for variable cluster size, which grows when clusters are merged, and be dynamic enough to represent the statistical changes of data in clusters throughout the entire AHSC procedures. Since such changes in clusters during AHSC largely depend upon a number of input data characteristics, cluster modeling without dynamic representation capability would be affected by input data variation, which is undesirable for reliable AHSC performance under the variation of input speech data. Conventional cluster modeling approaches using either single Gaussian distributions or Gaussian mixture models (GMMs) are not ideal in this regard. We introduce a novel cluster modeling approach with dynamic representation capability in this chapter. In this regard, the chapter is organized as follows. In Section 4.2, we re-investigate GLR-based inter-cluster distance measurement, which is a general method to statistically choose the closest cluster pair at every recursion step of AHSC. This investigation leads us to why reliable statistical cluster modeling is important for inter-cluster distance measurement in AHSC. Then we examine the aforementioned conventional cluster modeling approaches for this GLR-based distance measurement framework. Through this we show the merits and demerits of the conventional approaches in terms of cluster representation capability and computational complexity. In Section 4.3, we propose a new cluster modeling approach using *incremental Gaussian mixture models* (IGMMs) and compare it with the conventional approaches. The comparison verifies that the proposed method not only provides improved clustering performance but

also has moderate computational cost so that it is feasible in practice. In Section 4.4, we provide concluding remarks and future research directions regarding speaker-specific data modeling.

4.2 Inter-Cluster Distance Measurement for AHSC

Inter-cluster distance measurement is a critical part in AHSC of selecting the closest pair of clusters (in terms of speaker-specific characteristics) for merging at every recursion step. Because of the recursiveness of AHSC, erroneous selection of merging clusters at any given recursion step would affect subsequent recursion steps and might result in the significant degradation of overall clustering performance in the end. Therefore, precise selection of merging clusters is desirable at every recursion step of AHSC.

In general, cluster distance is statistically measured within the framework of AHSC. A typical method [26], i.e., GLR described in Section 2.3.1, calculates inter-cluster distance by comparing likelihoods for two hypotheses on the clusters considered. The details of this method are re-presented in the next subsection, for better understanding of the rest of this chapter.

4.2.1 GLR-based statistical inter-cluster distance measurement

Let us consider a certain recursion step during AHSC. Suppose that a pair of clusters $\mathbf{x} = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_M\}$ and $\mathbf{y} = \{\underline{y}_1, \underline{y}_2, \dots, \underline{y}_N\}$ are given for distance measurement. Then, GLR for the given pair is computed as follows:

$$\text{GLR}(\mathbf{x}, \mathbf{y}) = \frac{P(\mathbf{x}, \mathbf{y} | \mathcal{H}_1)}{P(\mathbf{x}, \mathbf{y} | \mathcal{H}_2)}, \quad (4.1)$$

where

- \mathcal{H}_1 (unmerging hypothesis): \mathbf{x} and \mathbf{y} are hypothesized to be left unmerged,

- \mathcal{H}_2 (merging hypothesis): \mathbf{x} and \mathbf{y} are hypothesized to be merged so as to become a new, larger cluster \mathbf{z} , where $\mathbf{z} = \mathbf{x} \cup \mathbf{y}$.

By this way distance for every pair of clusters at the recursion step considered can be measured in terms of GLR, and the cluster pair having the smallest GLR value is chosen to be merged.

4.2.2 Conventional cluster modeling approaches

Note that each hypothesis in this method of measuring distance between clusters is modeled by some probabilistic distribution. This is called *hypothesis* or *cluster modeling*, by which all the clusters considered for distance measurement (\mathbf{x} , \mathbf{y} , and \mathbf{z}) are represented by PDFs, respectively. Thus, proper distribution selection for cluster modeling is very important for precise distance measurement of clusters. In this subsection, we examine two conventional distributions for cluster modeling in the research field of speaker clustering.

4.2.2.1 Single Gaussian cluster modeling

One of conventional selection of PDFs for cluster modeling is to use a single Gaussian distribution. In this approach, all the three clusters aforementioned are modeled by multivariate normal PDFs, $\theta_{\mathbf{x}}^{\mathcal{N}} = \mathcal{N}(\underline{m}_{\mathbf{x}}, \Sigma_{\mathbf{x}})$, $\theta_{\mathbf{y}}^{\mathcal{N}} = \mathcal{N}(\underline{m}_{\mathbf{y}}, \Sigma_{\mathbf{y}})$, and $\theta_{\mathbf{z}}^{\mathcal{N}} = \mathcal{N}(\underline{m}_{\mathbf{z}}, \Sigma_{\mathbf{z}})$. The sample mean vectors ($\underline{m}_{\mathbf{x}}$, $\underline{m}_{\mathbf{y}}$, and $\underline{m}_{\mathbf{z}}$) and (full) covariance matrices ($\Sigma_{\mathbf{x}}$, $\Sigma_{\mathbf{y}}$, and $\Sigma_{\mathbf{z}}$) are determined by way of maximizing the likelihoods of \mathbf{x} , \mathbf{y} , and \mathbf{z} for $\theta_{\mathbf{x}}^{\mathcal{N}}$, $\theta_{\mathbf{y}}^{\mathcal{N}}$, and $\theta_{\mathbf{z}}^{\mathcal{N}}$, respectively. As a result, Eq. (4.1) can be rewritten as follows:

$$\begin{aligned}
 \text{GLR}(\mathbf{x}, \mathbf{y}) &= \ln \frac{p(\mathbf{x}, \mathbf{y} | \mathcal{H}_1)}{p(\mathbf{x}, \mathbf{y} | \mathcal{H}_2)} \\
 &= \ln \frac{p(\mathbf{x} | \theta_{\mathbf{x}}^{\mathcal{N}}) \cdot p(\mathbf{y} | \theta_{\mathbf{y}}^{\mathcal{N}})}{p(\mathbf{z} | \theta_{\mathbf{z}}^{\mathcal{N}})} \\
 &= \ln \frac{p(\mathbf{x}; \underline{m}_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \cdot p(\mathbf{y}; \underline{m}_{\mathbf{y}}, \Sigma_{\mathbf{y}})}{p(\mathbf{z}; \underline{m}_{\mathbf{z}}, \Sigma_{\mathbf{z}})}. \tag{4.2}
 \end{aligned}$$

This single Gaussian cluster modeling approach has been popular mainly due to its small computational cost. For example, the right-hand side in Eq. (4.2) can be simplified by the following relation [13]:

$$\ln \frac{p(\mathbf{x}; \underline{m}_{\mathbf{x}}, \Sigma_{\mathbf{x}}) \cdot p(\mathbf{y}; \underline{m}_{\mathbf{y}}, \Sigma_{\mathbf{y}})}{p(\mathbf{z}; \underline{m}_{\mathbf{z}}, \Sigma_{\mathbf{z}})} = \frac{M + N}{2} \ln \det(\Sigma_{\mathbf{z}}) - \frac{M}{2} \ln \det(\Sigma_{\mathbf{x}}) - \frac{N}{2} \ln \det(\Sigma_{\mathbf{y}}), \quad (4.3)$$

where $\det(\cdot)$ stands for a matrix determinant operator. Hence we can avoid direct computation of the likelihoods, i.e., $p(\mathbf{x}; \underline{m}_{\mathbf{x}}, \Sigma_{\mathbf{x}})$, $p(\mathbf{y}; \underline{m}_{\mathbf{y}}, \Sigma_{\mathbf{y}})$, and $p(\mathbf{z}; \underline{m}_{\mathbf{z}}, \Sigma_{\mathbf{z}})$, which would require more processing time as the cardinalities of \mathbf{x} , \mathbf{y} , and \mathbf{z} increase. Another advantageous factor of this cluster modeling approach in terms of computational complexity is relatively simple parameter estimation for Gaussian PDFs. Specifically, $\underline{m}_{\mathbf{z}}$ and $\Sigma_{\mathbf{z}}$ can be simply calculated from $\underline{m}_{\mathbf{x}}$, $\underline{m}_{\mathbf{y}}$, $\Sigma_{\mathbf{x}}$ and $\Sigma_{\mathbf{y}}$ using the following closed-form relations, instead of direct maximum likelihood estimation from \mathbf{z} :

$$\underline{m}_{\mathbf{z}} = \frac{M \cdot \underline{m}_{\mathbf{x}} + N \cdot \underline{m}_{\mathbf{y}}}{M + N} \quad (4.4)$$

and

$$\Sigma_{\mathbf{z}} = \frac{M \cdot \Sigma_{\mathbf{x}} + N \cdot \Sigma_{\mathbf{y}}}{M + N} + \frac{M \cdot \underline{m}_{\mathbf{x}} \underline{m}_{\mathbf{x}}^T + N \cdot \underline{m}_{\mathbf{y}} \underline{m}_{\mathbf{y}}^T}{M + N} - \underline{m}_{\mathbf{z}} \underline{m}_{\mathbf{z}}^T. \quad (4.5)$$

Therefore, in this cluster modeling approach, there is no need to estimate model parameters for merging-hypothesized clusters at distance measurement during AHSC. This reduces a lot of computational cost over the entire clustering procedures, especially as cluster size increases.

However, this approach has a critical issue in terms of representation capability. Single Gaussian distributions are known to have limited capability in representing the statistical characteristics of large speech data in terms of speaker-specific properties [51–53].

Considering that the average size of the clusters handled by AHSC increases as merging recursions continue, one-mode PDFs like normal PDFs could degenerate inter-cluster discernibility in terms of speaker-specific characteristics especially at the late recursion steps of AHSC, and hence cause overall clustering performance to degrade severely.

4.2.2.2 GMM cluster modeling

The other conventional approach for cluster modeling is to utilize GMMs as cluster models. In this approach, all the aforementioned three clusters (\mathbf{x} , \mathbf{y} , and \mathbf{z}) considered at distance measurement are modeled by GMMs, $\theta_{\mathbf{x}}^\lambda = \lambda(\{\underline{m}_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i, w_{\mathbf{x}}^i\}_{i=1}^\nu)$, $\theta_{\mathbf{y}}^\lambda = \lambda(\{\underline{m}_{\mathbf{y}}^i, \Sigma_{\mathbf{y}}^i, w_{\mathbf{y}}^i\}_{i=1}^\nu)$, and $\theta_{\mathbf{z}}^\lambda = \lambda(\{\underline{m}_{\mathbf{z}}^i, \Sigma_{\mathbf{z}}^i, w_{\mathbf{z}}^i\}_{i=1}^\nu)$. The mean vectors ($\underline{m}_{\mathbf{x}}^i$, $\underline{m}_{\mathbf{y}}^i$, and $\underline{m}_{\mathbf{z}}^i$), (diagonal) covariance matrices ($\Sigma_{\mathbf{x}}^i$, $\Sigma_{\mathbf{y}}^i$, and $\Sigma_{\mathbf{z}}^i$), and weights ($w_{\mathbf{x}}^i$, $w_{\mathbf{y}}^i$, and $w_{\mathbf{z}}^i$) for Gaussian mixture components are estimated by the expectation-maximization (EM) procedures [18]. The number of component mixtures in GMMs, ν , is empirically fixed at 8, 16 or 32¹ in general. As a consequence, Eq. (4.1) can be rewritten as follows:

$$\begin{aligned}
 \text{GLR}(\mathbf{x}, \mathbf{y}) &= \ln \frac{p(\mathbf{x}, \mathbf{y} | \mathcal{H}_1)}{p(\mathbf{x}, \mathbf{y} | \mathcal{H}_2)} \\
 &= \ln \frac{p(\mathbf{x} | \theta_{\mathbf{x}}^\lambda) \cdot p(\mathbf{y} | \theta_{\mathbf{y}}^\lambda)}{p(\mathbf{z} | \theta_{\mathbf{z}}^\lambda)} \\
 &= \ln \frac{\sum_{i=1}^\nu w_{\mathbf{x}}^i \cdot p(\mathbf{x}; \underline{m}_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i) \cdot \sum_{i=1}^\nu w_{\mathbf{y}}^i \cdot p(\mathbf{y}; \underline{m}_{\mathbf{y}}^i, \Sigma_{\mathbf{y}}^i)}{\sum_{i=1}^\nu w_{\mathbf{z}}^i \cdot p(\mathbf{z}; \underline{m}_{\mathbf{z}}^i, \Sigma_{\mathbf{z}}^i)}. \quad (4.6)
 \end{aligned}$$

This GMM cluster modeling approach has better representation capability in terms of speaker-specific characteristics because of multiple modes (or component mixtures) and the respective weights for them, compared to the previous single Gaussian approach. Thus, this approach can provide better clustering performance overall. However, there exist some problems as well with using GMMs as cluster models.

¹These values come from Reynolds' work [51–53] saying that GMMs with those numbers of mixture components well represent speaker-specific characteristics for speaker identification tasks.

First, GMMs with a fixed number² of mixture components cannot consider variations in cluster size throughout the entire AHSC procedures. For example, GMMs with a lot of mixture Gaussians could be overfitted for small clusters at the early recursion steps of AHSC because most initial clusters handled by AHSC usually do not contain sufficient data to train multi-component GMMs properly. On the other hand, GMMs with a small number of mixture components might not be able to fully represent the speaker-specific characteristics of large clusters at the late recursion steps as in the single Gaussian case. To tackle this problem, there has been some research [14,45,60,62] to adjust the number of mixture components during AHSC in proportion to cluster size based on certain criteria, but they cost a lot of computational burden for mixture number selection for each GMM in addition to the EM procedures that have already high computational complexity. In this regard, it is necessary to dynamically represent clusters during AHSC with low (or at least moderate) computational complexity.

The second issue in the GMM cluster modeling approach is that, although they require a significant amount of processing time in proportion to cluster size and mixture number, the EM procedures might degrade discernibility between clusters in terms of speaker-specific characteristics. Let us consider some examples shown in Figure 4.1, which presents simple test results about the effectiveness of the EM procedures in the GMM cluster modeling approach with 16 mixture components in each GMM. In this figure, each subfigure compares GLR distances between two pairs of clusters along with the number of iterations in the EM procedures. One pair comes from the same speaker source (black curve), meaning that cluster distance should be relatively close in terms of speaker-specific characteristics, while the other is from different sources (grey curve). Interestingly, we can observe from the leftmost subfigure that distance for the heterogeneous cluster pair, presented by the grey curve, gets smaller than that for the homogeneous one as iterations continue, which is undesirable because distance between homogeneous clusters (in terms of speaker-specific characteristics) should be always less than that between

²It is common that the number of mixture components is universally set throughout the whole AHSC procedures as an empirically reasonable value like 8, 16 or 32.

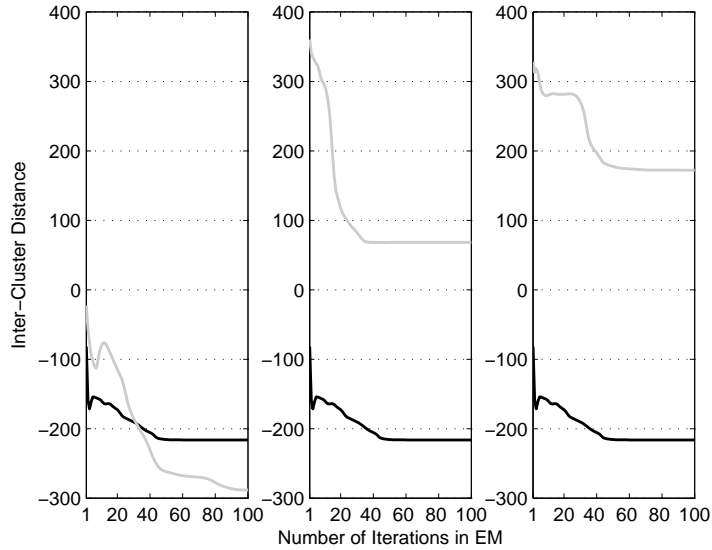


Figure 4.1: Effectiveness of the EM procedures in the GMM cluster modeling approach for GLR-based inter-cluster distance measurement. Each subfigure compares distances between two pairs of clusters along with the number of iterations in the EM procedures for GMMs with 16 mixture components. One pair comes from the same speaker source (black curve) while the other is from different sources (grey curve).

heterogeneous ones in the ideal case. Even from the other subfigures, the EM procedures do not significantly widen distance between the two pairs of clusters compared, indicating that the EM procedures do not much improve inter-cluster discernibility in terms of speaker-specific characteristics. These observations can be explained as follows: the EM procedures iteratively adapt the parameters of any GMM toward optimization in terms of maximum likelihood and thus increase $p(\mathbf{z}|\theta_{\mathbf{z}}^{\lambda})$ in Eq. (4.6) regardless of speaker-specific homogeneity between the clusters considered. However, this does not help increase inter-cluster discernibility and may even make it worse, as shown in the leftmost subfigure in Figure 4.1.

Another issue in this cluster modeling approach is that the EM procedures are affected by random initialization in the beginning and result in different estimation of model parameters for GMMs every session, which might cause the variation of clustering performance for the same input speech data.

4.2.2.3 Experimental comparison

We have thus far examined the two conventional cluster modeling approaches within the framework of GLR-based inter-cluster distance measurement for AHSC, and listed the merits and demerits of each approach. In this subsection, we verify our examination by comparing the two approaches empirically.

1) *Experimental setup*

Before we start, we need to take a look at speech data and setup for the entire experiments in this chapter. Table 4.1 presents input speech data for AHSC. These data sources are 15 sets of speech segments with approximately 4hr-long total duration, and were randomly chosen from the ICSI, NIST, and ISL meeting corpora. They are distinct from one another in terms of the number of speaker sources (N_s), gender distribution over speaker sources, total utterance time (T_s), number of speech segments (N_t), and average segment length (T_a). For preparing each set of speech segments, we manually segmented each audio clip at every point of speaking turn changes according to the given reference transcription beforehand. In order to avoid any potential confusion in performance analysis that might result from overlaps between segments, we excluded all the segments involved in any overlap during data preparation.

In all experiments in this chapter, we assume that stopping point estimation for AHSC is optimal, i.e, the optimal stopping point where extra merging in AHSC would not improve speaker error time rate any further can be exactly estimated for every data source. For this assumption to get realized, we manually stopped AHSC where the lowest speaker error time rate would be achieved, because we only focus on inter-cluster distance measurement in this chapter, not stopping point estimation.

2) *Comparison*

Table 4.2 shows performance comparison of the two conventional cluster modeling approaches in terms of speaker error time rate. For the GMM cluster modeling approach,

Table 4.1: Data source. N_s : number of speaker sources (male:female), T_s : total utterance time (sec.), N_t : number of speech segments, and T_a : average segment length (sec.).

	Data Source							
	1	2	3	4	5	6	7	8
N_s	7 (5:2)	7 (5:2)	7 (6:1)	6 (4:2)	5 (1:4)	6 (5:1)	5 (5:0)	4 (4:0)
T_s	1064.9	931.3	2336.3	1148.5	805.1	1664.9	1609.1	1475.9
N_t	418	279	611	244	228	532	591	478
T_a	2.5	3.3	3.8	4.7	3.5	3.1	2.7	3.1

	Data Source							
	9	10	11	12	13	14	15	
N_s	9 (7:2)	4 (3:1)	4 (3:1)	6 (4:2)	8 (4:4)	4 (2:2)	4 (0:4)	
T_s	659.7	443.4	835.7	624.1	272.4	477.7	429.1	
N_t	159	75	179	144	93	119	95	
T_a	4.1	5.9	4.7	4.3	2.9	4.0	4.5	

we chose 4 different values, i.e., 4, 8, 16, and 32, as the number of mixture components in GMMs. The lowest error rate in each column (or each data source) is bold-faced.

From this table we can first observe that the GMM approach is better than the single Gaussian approach in terms of overall performance, except for the case³ of 4 mixture components in the GMM approach. Other than a few data sources (Data 1, 4, 5, and 11) GMMs provide better clustering accuracy and, even for Data 1, 4, 5, and 11, difference between the clustering error rates of AHSC by the two cluster modeling approaches is not that significant, which verifies our previous statement that the GMM approach has better representation capability for modeling clusters and thus provides better clustering performance overall.

However, the results in this table also show the difficulty to set the proper number of mixture Gaussians in the GMM cluster modeling approach. 8 mixture Gaussians fit to Data 7, 9, 10, 13, and 14 while 16 mixtures are better for Data 3, 6, and 8. For

³In this case, we can see that only 4 mixture Gaussians (with diagonal covariance matrices) in GMMs are not enough to represent speaker-specific characteristics in cluster data, which can be supported by the previous research in [53].

Table 4.2: Performance comparison of the two conventional cluster modeling approaches in terms of speaker error time rate (%). \mathcal{N} : single Gaussian cluster modeling and λ_x : GMM cluster modeling with x mixture components.

	Data Source							
	1	2	3	4	5	6	7	8
\mathcal{N}	7.0	19.3	10.6	2.7	15.1	7.5	13.2	25.6
λ_4	30.9	20.8	25.8	9.3	25.2	21.5	14.0	20.4
λ_8	13.7	13.8	8.1	8.6	29.5	7.4	7.0	21.5
λ_{16}	12.9	16.9	6.2	4.1	20.2	7.0	10.9	16.3
λ_{32}	10.3	10.0	18.4	4.7	15.9	9.5	9.1	26.4

	Data Source							
	9	10	11	12	13	14	15	Avg.
\mathcal{N}	23.6	7.6	9.1	9.7	23.4	27.0	29.2	15.4
λ_4	20.6	13.1	27.8	12.4	29.4	27.6	38.7	22.5
λ_8	10.9	2.5	9.9	13.4	10.5	22.8	30.2	14.0
λ_{16}	12.3	7.8	10.3	9.1	33.3	27.1	29.2	14.9
λ_{32}	12.0	5.2	10.0	8.7	28.7	30.7	28.7	15.2

Data 2, 12, and 15, 32 mixtures are relatively superior to 8 or 16. This suggests that in order to obtain the best clustering performance in the GMM cluster modeling approach we need to optimize the number of mixture components for every data source. This is impractical because there is no theoretical way yet to find out the optimal number of mixture Gaussians in GMMs depending upon input data sources. Thus, there is still a need for a more adaptive and reliable way of modeling clusters during AHSC.

Another demerit of the GMM cluster modeling approach is well depicted in Figure 4.2, which shows comparison of the processing time⁴ of AHSC depending upon cluster modeling approaches. From this figure it is so easy to confirm that the GMM approach requires by far more time than the single Gaussian approach. For example, the processing time for Data 3 by the GMM approach with 32 mixture components is more than

⁴For this experiment, an MS-Windows machine with the Intel Pentium-4 3.2GHz CPU was used. The number of iterations in the EM procedures for each GMM parameter estimation was fixed at 15. After 15 iterations we can empirically assume that there is no significant change in likelihoods for GMMs.

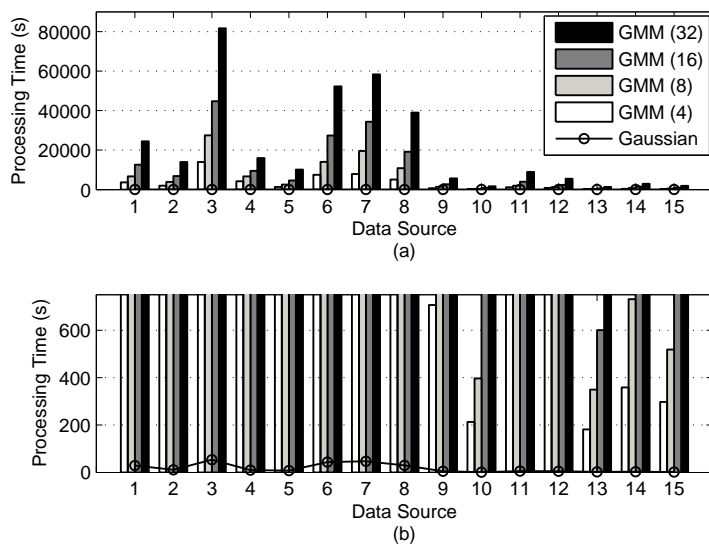


Figure 4.2: Processing time comparison of the two conventional cluster modeling approaches. (For the GMM approach, four different mixture numbers are compared, i.e., 4, 8, 16, and 32.) (a) Full-shot version. (b) Zoomed-in version.

22hrs and is approximately 1500 times that by the single Gaussian approach, which is prohibitive in practice. This high computational cost is mainly due to the EM procedures for GMM parameter estimation and is unavoidable in the GMM cluster modeling approach, because in the EM procedures there are no closed form relations like Eqs. (4.4) and (4.5) and thus the parameters of the merging-hypothesized clusters used for GLR-based inter-cluster distance measurement should be newly estimated for every possible cluster pair. Therefore, the processing time of the GMM approach exponentially increases in proportion to the total amount of input speech data.

Figure 4.3 presents another minor issue in the GMM cluster modeling approach, i.e. performance variation due to initial randomness in the EM procedures for GMMs. In this figure clustering performance for Data 13 is shown, and we can clearly see the session-to-session variation of speaker error time rate in every number of mixture Gaussians considered. The variations of the error rates in all the GMM approaches are so large that we cannot claim that the GMM approach is in general better than the single Gaussian

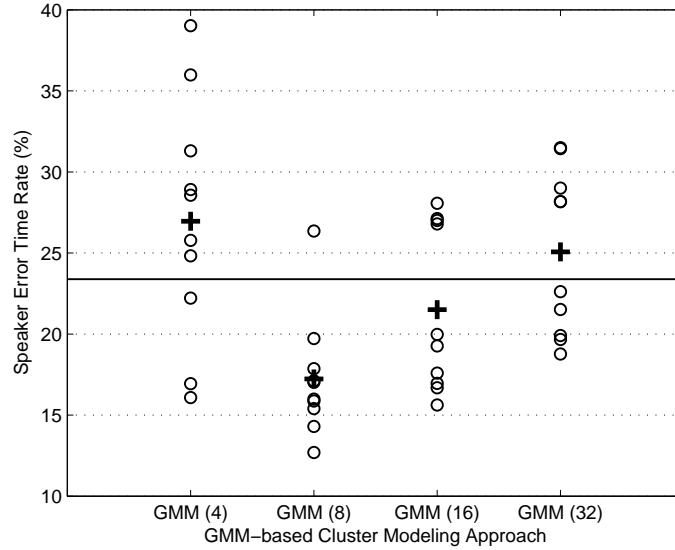


Figure 4.3: Clustering performance variation for Data 13 in the GMM cluster modeling approach. The circles denote speaker error time rates for the respective 10 sessions of the GMM approach with four different numbers of mixture components (i.e., 4, 8, 16, and 32), and the bold crosses are the corresponding mean values. The horizontal line presents the speaker error time rate obtained from the single Gaussian cluster modeling approach, which is 23.4%.

approach in terms of performance. Only the GMM approach with 8 mixture components shows lower error rates than the single Gaussian approach despite such performance variation⁵. However, as aforementioned, the value of 8 mixture components is not universally optimal across data sources.

From all these examinations and comparisons of the two conventional cluster modeling approaches for GLR-based inter-cluster distance measurement within the framework of AHSC in this section, we are able to confirm a need for a new cluster modeling approach, which not only requires moderate computational cost but also has dynamic representation capability. In the next section, we propose such an alternative method to overcome the disadvantages of the conventional cluster modeling approaches.

⁵Nevertheless we can still see one outlier worse than the speaker error time rate by the single Gaussian approach in Figure 4.3.

4.3 Incremental Gaussian Mixture Cluster Modeling

For GLR-based inter-cluster distance measurement within the framework of AHSC, ideal cluster modeling should:

- Keep clusters well represented in terms of speaker-specific characteristics throughout the whole AHSC procedures as cluster sizes continue to increase due to merging recursions.
- Be reliable in terms of performance across data sources or sessions.
- Have moderate computational complexity so that it is feasible in practice.

To achieve all these, we introduce a novel cluster modeling approach in this section, named *incremental Gaussian mixture* cluster modeling.

4.3.1 Proposed cluster modeling approach

For this new cluster modeling approach, we devise a simple but dynamic distribution for AHSC, called an incremental Gaussian mixture model (IGMM), which increments mixture components from one Gaussian to multiple Gaussians by summing the PDFs of the respective distributions for merging clusters to represent a newly merged cluster. The details of the incremental Gaussian mixture or IGMM cluster modeling approach are as follows:

- Every initial cluster is modeled by a multivariate normal distribution.
- Any newly merged or merging-hypothesized cluster is modeled by a distribution whose PDF is determined by the weighted sum of the PDFs of the respective distributions for the two clusters involved in (potential) merging. The weights for the two PDFs are the normalized cardinalities of the clusters considered, respectively.

In this approach, all the three clusters (i.e., two clusters under consideration: \mathbf{x} and \mathbf{y} , and a merging-hypothesized cluster: \mathbf{z}) considered within the framework of GLR-based inter-cluster distance measurement are thus represented by IGMMs, $\theta_{\mathbf{x}}^{\text{IGMM}} =$

IGMM($\{\underline{m}_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i, w_{\mathbf{x}}^i\}_{i=1}^{\nu_{\mathbf{x}}}\}$), $\theta_{\mathbf{y}}^{\text{IGMM}} = \text{IGMM}(\{\underline{m}_{\mathbf{y}}^i, \Sigma_{\mathbf{y}}^i, w_{\mathbf{y}}^i\}_{i=1}^{\nu_{\mathbf{y}}}\}$, and $\theta_{\mathbf{z}}^{\text{IGMM}} = \text{IGMM}(\{\underline{m}_{\mathbf{z}}^i, \Sigma_{\mathbf{z}}^i, w_{\mathbf{z}}^i\}_{i=1}^{\nu_{\mathbf{z}}}\}$). The IGMM parameters are given as below:

- The numbers of mixture Gaussians in $\theta_{\mathbf{x}}^{\text{IGMM}}$ and $\theta_{\mathbf{y}}^{\text{IGMM}}$, $\nu_{\mathbf{x}}$ and $\nu_{\mathbf{y}}$, are equal to the numbers of the initial clusters that have been merged to make \mathbf{x} and \mathbf{y} , respectively. In general, $\nu_{\mathbf{x}} \neq \nu_{\mathbf{y}}$. Since we can express \mathbf{x} and \mathbf{y} as $\mathbf{x} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^{\nu_{\mathbf{x}}}\}$ and $\mathbf{y} = \{\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^{\nu_{\mathbf{y}}}\}$, respectively, where $\{\mathbf{x}^i\}_{i=1}^{\nu_{\mathbf{x}}}$ and $\{\mathbf{y}^i\}_{i=1}^{\nu_{\mathbf{y}}}$ are all initial clusters, and $\mathbf{z} = \mathbf{x} \cup \mathbf{y} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^{\nu_{\mathbf{x}}}, \mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^{\nu_{\mathbf{y}}}\}$, then the number of mixture components in $\theta_{\mathbf{z}}^{\text{IGMM}}$ is $\nu_{\mathbf{z}} = \nu_{\mathbf{x}} + \nu_{\mathbf{y}}$.
- The weights $w_{\mathbf{x}}^i$ and $w_{\mathbf{y}}^i$ are $\frac{|\mathbf{x}^i|}{\sum_{j=1}^{\nu_{\mathbf{x}}} |\mathbf{x}^j|}$ and $\frac{|\mathbf{y}^i|}{\sum_{j=1}^{\nu_{\mathbf{y}}} |\mathbf{y}^j|}$, respectively, where $|\cdot|$ denotes cardinality. Thus, $\{w_{\mathbf{z}}^i\}_{i=1}^{\nu_{\mathbf{z}}} = \left\{ \frac{|\mathbf{x}^i|}{\sum_{j=1}^{\nu_{\mathbf{x}}} |\mathbf{x}^j| + \sum_{j=1}^{\nu_{\mathbf{y}}} |\mathbf{y}^j|} \right\}_{i=1}^{\nu_{\mathbf{x}}}$ and $\{w_{\mathbf{z}}^i\}_{i=\nu_{\mathbf{x}}+1}^{\nu_{\mathbf{z}}} = \left\{ \frac{|\mathbf{y}^i|}{\sum_{j=1}^{\nu_{\mathbf{x}}} |\mathbf{x}^j| + \sum_{j=1}^{\nu_{\mathbf{y}}} |\mathbf{y}^j|} \right\}_{i=1}^{\nu_{\mathbf{y}}}$.
- The mean vectors and (full) covariance matrices in $\theta_{\mathbf{x}}^{\text{IGMM}}$ and $\theta_{\mathbf{y}}^{\text{IGMM}}$ are those of model distributions for the constituent initial clusters of \mathbf{x} and \mathbf{y} , respectively, i.e., $\text{IGMM}(\{\underline{m}_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i\}_{i=1}^{\nu_{\mathbf{x}}}) = \{\mathcal{N}(\underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i})\}_{i=1}^{\nu_{\mathbf{x}}}$ and $\text{IGMM}(\{\underline{m}_{\mathbf{y}}^i, \Sigma_{\mathbf{y}}^i\}_{i=1}^{\nu_{\mathbf{y}}}) = \{\mathcal{N}(\underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i})\}_{i=1}^{\nu_{\mathbf{y}}}$. Thus, $\text{IGMM}(\{\underline{m}_{\mathbf{z}}^i, \Sigma_{\mathbf{z}}^i\}_{i=1}^{\nu_{\mathbf{z}}}) = \{\mathcal{N}(\underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i})\}_{i=1}^{\nu_{\mathbf{x}}}$ and $\text{IGMM}(\{\underline{m}_{\mathbf{z}}^i, \Sigma_{\mathbf{z}}^i\}_{i=\nu_{\mathbf{x}}+1}^{\nu_{\mathbf{z}}}) = \{\mathcal{N}(\underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i})\}_{i=1}^{\nu_{\mathbf{y}}}$.

As a consequence, Eq. (4.1) can be rewritten in this cluster modeling approach as follows:

$$\begin{aligned}
\text{GLR}(\mathbf{x}, \mathbf{y}) &= \ln \frac{p(\mathbf{x}, \mathbf{y} | \mathcal{H}_1)}{p(\mathbf{x}, \mathbf{y} | \mathcal{H}_2)} \\
&= \ln \frac{p(\mathbf{x} | \theta_{\mathbf{x}}^{\text{IGMM}}) \cdot p(\mathbf{y} | \theta_{\mathbf{y}}^{\text{IGMM}})}{p(\mathbf{z} | \theta_{\mathbf{z}}^{\text{IGMM}})} \\
&= \ln \frac{\sum_{i=1}^{\nu_{\mathbf{x}}} w_{\mathbf{x}}^i \cdot p(\mathbf{x}; \underline{m}_{\mathbf{x}}^i, \Sigma_{\mathbf{x}}^i) \cdot \sum_{i=1}^{\nu_{\mathbf{y}}} w_{\mathbf{y}}^i \cdot p(\mathbf{y}; \underline{m}_{\mathbf{y}}^i, \Sigma_{\mathbf{y}}^i)}{\sum_{i=1}^{\nu_{\mathbf{z}}} w_{\mathbf{z}}^i \cdot p(\mathbf{z}; \underline{m}_{\mathbf{z}}^i, \Sigma_{\mathbf{z}}^i)} \\
&= \ln \frac{\sum_{i=1}^{\nu_{\mathbf{x}}} w_{\mathbf{x}}^i \cdot p(\mathbf{x}; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) \cdot \sum_{i=1}^{\nu_{\mathbf{y}}} w_{\mathbf{y}}^i \cdot p(\mathbf{y}; \underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i})}{\sum_{i=1}^{\nu_{\mathbf{x}}} w_{\mathbf{z}}^i \cdot p(\mathbf{z}; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) + \sum_{i=1}^{\nu_{\mathbf{y}}} w_{\mathbf{z}}^{\nu_{\mathbf{x}}+i} \cdot p(\mathbf{z}; \underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i})}. \quad (4.7)
\end{aligned}$$

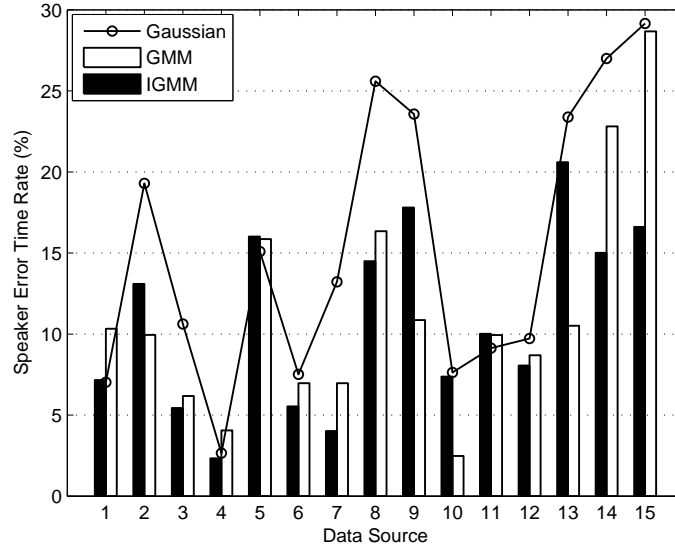


Figure 4.4: Performance comparison of the proposed and two conventional cluster modeling approaches in terms of speaker error time rate (%). For this comparison, the best performance of the GMM approach for each data source was chosen among the 4 candidates (4, 8, 16, and 32 mixture components).

4.3.2 Comparison and analysis

The proposed IGMM cluster modeling approach has several merits compared to the two conventional approaches. The first advantage is that the numbers of mixture components in IGMMs keep increasing during AHSC. This is because the number of component mixtures in the IGMM representing any newly merged cluster is determined by the sum of the numbers of mixture Gaussians in the IGMMs representing the clusters involved in merging. In other words, smooth transition from a single Gaussian distribution for modeling every initial cluster to multiple Gaussian mixtures for larger clusters generating from merging recursions occurs during AHSC. For this reason, the IGMM cluster modeling approach can provide dynamic cluster representation capability throughout the whole AHSC procedures. Considering that both of the conventional cluster modeling approaches have limitation in this regard, i.e., limited representation capability for large clusters in the single Gaussian approach and overfitting for small clusters in the GMM approach, we

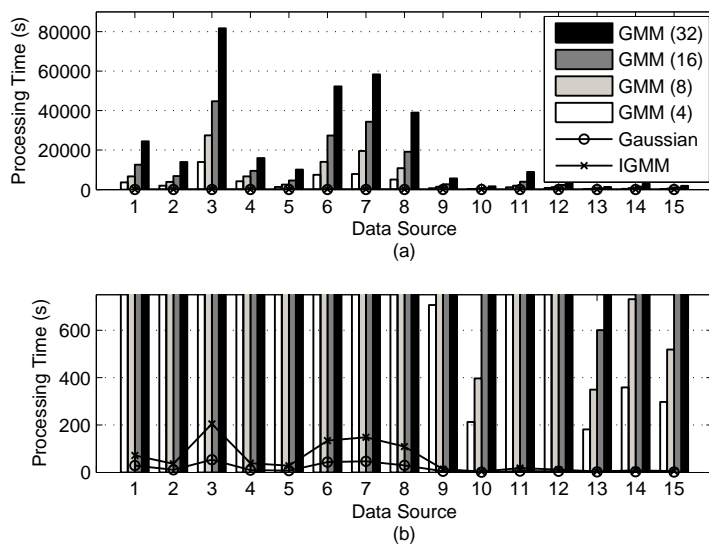


Figure 4.5: Processing time comparison of the proposed and two conventional cluster modeling approaches. (For the GMM approach, four different mixture numbers are compared, i.e., 4, 8, 16, and 32.) (a) Full-shot version. (b) Zoomed-in version.

can say that this new cluster modeling approach compromises the two conventional approaches efficiently. As a consequence, it can provide better clustering performance than the two conventional approaches. This claim is verified by experimental results in Figure 4.4, which compares the proposed and two conventional cluster modeling approaches in terms of speaker error time rate. From this figure, we can easily observe that the proposed approach provides much better clustering performance than the single Gaussian approach by around 30% (relative) on average over the entire 15 data sources while it gives as comparable error rate as the GMM approach. Note that, in this comparison, we chose the best performance among the 4 candidates (4, 8, 16, and 32 mixture components) for each data source in the case of the GMM approach. Considering this, we can insist that the proposed approach has even better performance than the GMM approach. According to our comparison test (not shown here), average speaker error time rate by the IGMM cluster modeling approach is lower than that by the highest-performing GMM approach with 8 mixture components by approximately 20% (relative).

The second merit of the proposed cluster modeling approach is that despite many mixture Gaussians in IGMMs the approach requires only moderate computational complexity, which makes it much more feasible in practice than the GMM approach. Figure 4.5 shows this advantage of the IGMM approach by comparing processing time. It is clear from this figure that the computational cost of the IGMM approach is a lot less than that of the GMM approach, and as comparably small as that of the single Gaussian approach. For instance, the processing time of the IGMM cluster modeling approach for Data 3 is impressively about 3.5mins while that of the GMM approach with 32 mixture components is more than 22hrs (also mentioned in Section 4.2.2.3). The interesting fact is that according to our observation there are 5 clusters finally remaining at the optimal stopping point for AHSC on Data 3, and the numbers of mixture Gaussians in the corresponding IGMMs for the clusters are 343, 91, 78, 68, and 31, respectively. The main reason why the IGMM approach has relatively low computational complexity although there are much more mixture components involved than the GMM approach is that there is no complex process for parameter estimation like the EM procedures. Instead, like the single Gaussian approach, the right-hand side in Eq. (4.7) can be simplified into a closed form based on cross-likelihood⁶ between initial clusters. To verify this, let us go back to Eq. (4.7). From the definition of cross-likelihood, the first term of the numerator in Eq. (4.7) can be rewritten as

$$\begin{aligned}
& \sum_{i=1}^{\nu_{\mathbf{x}}} w_{\mathbf{x}}^i \cdot p(\mathbf{x}; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) \\
&= \sum_{i=1}^{\nu_{\mathbf{x}}} w_{\mathbf{x}}^i \cdot p(\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^{\nu_{\mathbf{x}}}; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) = \sum_{i=1}^{\nu_{\mathbf{x}}} w_{\mathbf{x}}^i \prod_{j=1}^{\nu_{\mathbf{x}}} p(\mathbf{x}^j; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) = \sum_{i=1}^{\nu_{\mathbf{x}}} w_{\mathbf{x}}^i \prod_{j=1}^{\nu_{\mathbf{x}}} \mathcal{L}_{\mathbf{x}^j|\mathbf{x}^i}.
\end{aligned} \tag{4.8}$$

⁶We define cross-likelihood between initial clusters as follows. Suppose that we have two initial clusters \mathbf{x}^i and \mathbf{x}^j , and the respective single Gaussian models $\theta_{\mathbf{x}^i}^{\mathcal{N}} = \mathcal{N}(\underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i})$ and $\theta_{\mathbf{x}^j}^{\mathcal{N}} = \mathcal{N}(\underline{m}_{\mathbf{x}^j}, \Sigma_{\mathbf{x}^j})$. The cross-likelihoods of the two clusters, $\mathcal{L}_{\mathbf{x}^i|\mathbf{x}^j}$ and $\mathcal{L}_{\mathbf{x}^j|\mathbf{x}^i}$, are defined as $p(\mathbf{x}^i; \underline{m}_{\mathbf{x}^j}, \Sigma_{\mathbf{x}^j})$ and $p(\mathbf{x}^j; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i})$, respectively.

Similarly, the second term of the numerator can be simplified into

$$\sum_{i=1}^{\nu_y} w_{\mathbf{y}}^i \cdot p(\mathbf{y}; \underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i}) = \sum_{i=1}^{\nu_y} w_{\mathbf{y}}^i \prod_{j=1}^{\nu_y} \mathcal{L}_{\mathbf{y}^j | \mathbf{y}^i}. \quad (4.9)$$

The denominator can be also rewritten in the similar way as

$$\begin{aligned} & \sum_{i=1}^{\nu_x} w_{\mathbf{z}}^i \cdot p(\mathbf{z}; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) + \sum_{i=1}^{\nu_y} w_{\mathbf{z}}^{\nu_x+i} \cdot p(\mathbf{z}; \underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i}) \\ &= \sum_{i=1}^{\nu_x} w_{\mathbf{z}}^i \cdot p(\mathbf{x}^1, \dots, \mathbf{x}^{\nu_x}, \mathbf{y}^1, \dots, \mathbf{y}^{\nu_y}; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) + \\ & \quad \sum_{i=1}^{\nu_y} w_{\mathbf{z}}^{\nu_x+i} \cdot p(\mathbf{x}^1, \dots, \mathbf{x}^{\nu_x}, \mathbf{y}^1, \dots, \mathbf{y}^{\nu_y}; \underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i}) \\ &= \sum_{i=1}^{\nu_x} w_{\mathbf{z}}^i \prod_{j=1}^{\nu_x} p(\mathbf{x}^j; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) \prod_{j=1}^{\nu_y} p(\mathbf{y}^j; \underline{m}_{\mathbf{x}^i}, \Sigma_{\mathbf{x}^i}) + \\ & \quad \sum_{i=1}^{\nu_y} w_{\mathbf{z}}^{\nu_x+i} \prod_{j=1}^{\nu_x} p(\mathbf{x}^j; \underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i}) \prod_{j=1}^{\nu_y} p(\mathbf{y}^j; \underline{m}_{\mathbf{y}^i}, \Sigma_{\mathbf{y}^i}) \\ &= \sum_{i=1}^{\nu_x} w_{\mathbf{z}}^i \prod_{j=1}^{\nu_x} \mathcal{L}_{\mathbf{x}^j | \mathbf{x}^i} \prod_{j=1}^{\nu_y} \mathcal{L}_{\mathbf{y}^j | \mathbf{x}^i} + \sum_{i=1}^{\nu_y} w_{\mathbf{z}}^{\nu_x+i} \prod_{j=1}^{\nu_x} \mathcal{L}_{\mathbf{x}^j | \mathbf{y}^i} \prod_{j=1}^{\nu_y} \mathcal{L}_{\mathbf{y}^j | \mathbf{y}^i}. \end{aligned} \quad (4.10)$$

Eq. (4.7) thus can be rewritten as

$$\begin{aligned} & \text{GLR}(\mathbf{x}, \mathbf{y}) \\ &= \ln \frac{\sum_{i=1}^{\nu_x} w_{\mathbf{x}}^i \prod_{j=1}^{\nu_x} \mathcal{L}_{\mathbf{x}^j | \mathbf{x}^i} \cdot \sum_{i=1}^{\nu_y} w_{\mathbf{y}}^i \prod_{j=1}^{\nu_y} \mathcal{L}_{\mathbf{y}^j | \mathbf{y}^i}}{\sum_{i=1}^{\nu_x} w_{\mathbf{z}}^i \prod_{j=1}^{\nu_x} \mathcal{L}_{\mathbf{x}^j | \mathbf{x}^i} \prod_{j=1}^{\nu_y} \mathcal{L}_{\mathbf{y}^j | \mathbf{x}^i} + \sum_{i=1}^{\nu_y} w_{\mathbf{z}}^{\nu_x+i} \prod_{j=1}^{\nu_x} \mathcal{L}_{\mathbf{x}^j | \mathbf{y}^i} \prod_{j=1}^{\nu_y} \mathcal{L}_{\mathbf{y}^j | \mathbf{y}^i}}. \end{aligned} \quad (4.11)$$

If we calculated the cross-likelihoods of every pair of initial clusters beforehand, there would be thus no additional computational cost for direct parameter estimation and likelihood calculation at every inter-cluster distance measurement in the IGMM cluster modeling approach. The cross-likelihood computation does not take relatively long as empirically verified by Figure 4.5 although its complexity increases in proportion to the number of initial clusters.

Another merit of the proposed approach is that, because of no randomization, performance variation would not occur from session to session in contrast to the GMM approach. This advantage can boost the reliability of AHSC performance.

4.4 Conclusions

In this chapter we proposed the incremental Gaussian mixture cluster modeling approach for GLR-based inter-cluster distance measurement within the framework of AHSC. The proposed approach addressed the limitations of the two conventional cluster modeling approaches, i.e., single Gaussian and GMM cluster modeling, by smoothly updating cluster models from normal distributions to GMMs with multiple mixture components during AHSC. Apart from this, the IGMM approach requires only moderate computational cost compared to the GMM approach. By this low complex and dynamic cluster modeling approach, we obtained clustering performance improvement in terms of speaker error time rate by approximately 30% and 20% (relative) against the single Gaussian and GMM approaches, respectively. These performance improvements obtained from our work suggest that the proposed cluster modeling approach enhanced the reliability of AHSC performance across input speech data.

The results of this work could be extended to speaker modeling in the research field of speaker recognition. Currently, speaker modeling is performed based on GMMs with a fixed number of mixture components like 16 or 32, but we still do not know how many mixture Gaussians would be necessary for the optimal modeling of speaker-specific characteristics. Based on our intuition it should be speaker-dependent, but there is no canonical method yet to derive the proper number of mixture components in GMMs for speaker-specific representation of data. The cluster modeling approach proposed in this chapter does not require any fixed number for mixture Gaussians beforehand, so it might be able to be a good alternative to the conventional GMM-based speaker modeling.

Chapter 5

Reliable Speaker Diarization based on Robust AHSC

5.1 Introduction

In this chapter we extend our research results on robust speaker clustering under the variation of input speech data toward one application domain, by applying them to one of main speaker clustering applications, i.e., speaker diarization. In the research field of speaker diarization, the robustness problem of diarization performance has been a big issue as well and is definitely caused by speaker clustering, which plays a decisive role in current state-of-the-art speaker diarization systems. This chapter is organized as follows. In Section 5.2, we first take a general, but closer look at speaker diarization and its various uses. In Section 5.3, based on the previous research results in Chapters 2, 3, and 4, we propose our own speaker diarization system equipped with sequential clustering for speaker change detection, and ICR-based stopping point estimation and IGMM cluster modeling for speaker clustering. In Section 5.4, we also propose clustering performance refinement schemes in the framework of speaker diarization, which can enhance the reliability of diarization performance across data sources. We conclude this chapter in Section 5.5 with the final remarks on robust speaker clustering in a speaker diarization perspective.

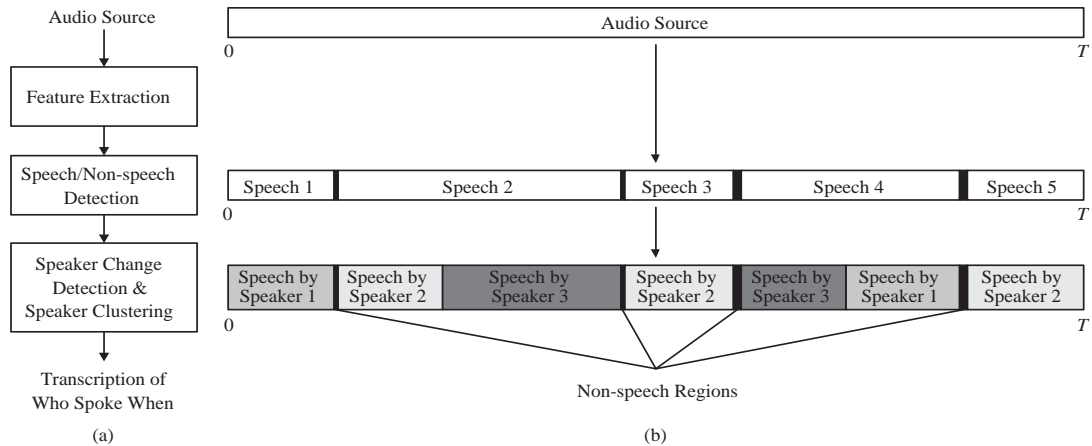


Figure 5.1: Speaker diarization: (a) Block diagram of a speaker diarization system. (b) Step-by-step graphical interpretation of how a given audio source is transcribed (in terms of “who spoke when”) by speaker diarization.

5.2 Speaker Diarization

Speaker diarization refers to the automatic process of dividing a given audio source, predominantly using speech, into speaker-specific segments by transcribing it in terms of “who spoke when” [63]. Such speaker-specific segmentation done by speaker diarization can be beneficial and have many applicable areas, such as for automatic speech recognition. For instance, speaker diarization enables selecting speaker-specific data that can be utilized for unsupervised speaker adaptation. It also can help provide statistics that rely on speaker-specific information, such as frequency of speaking turn change, average speaking time per turn, number of speakers, speaking time distribution over speakers, and so on. These statistics are useful for multimedia content analysis. Because of its broad significance, speaker diarization is currently regarded as one of the main categories evaluated in the Rich Transcription (RT) Evaluation led by NIST.

Many state-of-the-art speaker diarization systems have a basic structure in common as shown in Figure 5.1, consisting of three main steps following audio feature extraction. One is speech/non-speech detection, which separates target speech regions from a given audio source. The others are speaker change detection and speaker clustering. Speaker

change detection identifies potential speaker changing points in each speech region, and further divides the speech region into smaller speaker-specific segments. Speaker clustering classifies the resultant segments by speaker identity to append a unique label to the segments belonging to the same speaker class. These two steps are in general performed in the order mentioned, i.e., speaker change detection followed by speaker clustering.

The overall performance of speaker diarization systems is evaluated by diarization error rate (DER). This performance indicator for speaker diarization is defined by NIST as the sum of three constituent error rates: false alarm speaker time rate, missed speaker time rate, and speaker error time rate. The first two jointly indicate how precisely speech/non-speech detection is performed, while the last one solely tells how well speaker change detection and speaker clustering coordinate. Recent research papers demonstrated the dominance of speaker error time rate over the other error rates in deciding DER, and besides its severe variability across data sources [63], [30].

Among those two steps in speaker diarization, speaker clustering is more critical than speaker change detection in terms of impact on DER. Furthermore, serial concatenation of speaker change detection and speaker clustering in typical speaker diarization systems could require speaker clustering to be more precise in terms of performance. Speaker change detection is typically tuned not to miss speaker changing points in given speech regions at the cost of false alarms because, if actual speaker changing points were missed during speaker change detection, there would be no chance for them to be recovered (or re-detected) by speaker clustering. The segments unnecessarily divided due to false alarms during speaker change detection, on the other hand, could be possibly merged through speaker clustering. As a consequence, such a tuning pattern for speaker change detection in typical speaker diarization systems results in not so many detection errors, but cannot help burdening speaker clustering with a large number of short speech segments. (It is generally more difficult to classify short speech segments by speaker-specific characteristics, as we have seen in Chapter 3, than to classify long ones because speaker-specific identification requires long speech utterances (at least longer than 3 seconds) [51–53].)

Thus, it is very important to make speaker clustering work properly for reliable speaker diarization.

In this regard, in the next section, we implement a more reliable speaker diarization system based on robust AHSC approaches, which have been dealt with in this dissertation thus far. Specifically, we utilize ICR-based stopping point estimation and IGMM cluster modeling for this purpose. In addition, we also exploit the merit of the third modified version of AHSC in Section 3.3.3 for speech/non-speech detection and speaker change detection in this system. We call this proposed system the *SAIL speaker diarization system*.

5.3 SAIL Speaker Diarization System

In this section we propose a novel speaker diarization system based on robust speaker clustering. The proposed SAIL speaker diarization system has the same structure as other state-of-the-art speaker diarization systems have: speech/non-speech detection, speaker change detection, and speaker clustering. It first applies a sequential clustering concept to segmentation of a given audio data source, and then performs AHSC for speaker-specific classification (or speaker clustering) of speech segments. The speaker clustering algorithm utilizes an IGMM cluster modeling strategy for inter-cluster distance measurement, and ICR-based stopping point estimation to properly stop the recursion of AHSC. Before explaining the details of each step in the system, let us describe the data sets and experimental setup used in this chapter.

5.3.1 Data Description and Experimental Setup

Tables 5.1 and 5.2 present the two data sets (training and testing) used for the experiments reported in this chapter. The training data set is used for tuning the whole speaker diarization system, while the testing data set is used for performance evaluation. All the data sources in the data sets were chosen from the ICSI, NIST, and USC meeting

Table 5.1: Training data set.

	Source	Name	Length (min:sec)	No. of Speakers
1	ICSI	Bmr018	20:01	7
2	ICSI	Bro003	20:00	7
3	ICSI	Bsr001	20:00	8
4	NIST	20020214	19:59	6
5	NIST	20030925	19:59	4

Table 5.2: Testing data set.

	Source	Name	Length (min:sec)	No. of Speakers
1	ICSI	Bdb001	19:57	5
2	ICSI	Bed015	20:00	6
3	ICSI	Bmr013	20:01	7
4	ICSI	Bed002	20:01	6
5	NIST	20011115	17:52	4
6	NIST	20030702	20:00	4
7	NIST	20031215	19:57	5
8	USC	200804011207	17:23	5
9	USC	200804011259	6:28	4
10	USC	200804011325	19.41	4

speech corpora, and are distinct from one another in terms of the number of speakers and meeting topics (not given in the tables).

In order to measure DER, we use a scoring tool distributed by NIST, i.e., md-eval-v21.pl¹. This tool calculates DER as the sum of missed speaker time rate, false alarm speaker time rate, and speaker error time rate.

5.3.2 Speech/Non-Speech Detection

The proposed speech/non-speech detection step in the SAIL speaker diarization system is based on leader-follower clustering (LFC) [18], which is a well-known sequential clustering

¹Available at <http://www.nist.gov/speech/tests/rt/2006-spring>.

Algorithm 7 Leader-Follower Clustering (LFC)

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: data sequentially incoming
 θ : threshold

Ensure: $C_i, i = 1, \dots, n$: clusters finally remaining

```
1:  $C_1 \leftarrow \{\mathbf{x}_1\}, n \leftarrow 1, m \leftarrow 1$ 
2: do  $m \leftarrow m + 1$ 
3:    $\hat{C} \leftarrow \{\mathbf{x}_m\}$ 
4:    $i \leftarrow \arg \min d(C_j, \hat{C}), j = 1, \dots, n$ 
5:   if  $d(C_i, \hat{C}) > \theta$ 
6:      $n \leftarrow n + 1$ 
7:      $C_n \leftarrow \hat{C}$ 
8:   else
9:     merge  $\hat{C}$  into  $C_i$ 
10: until  $m = \hat{n}$ 
11: return  $C_i, i = 1, \dots, n$ 
```

strategy. As shown in Algorithm 7, LFC sequentially classifies incoming data, either by having them merged to existing clusters or by generating new clusters for them. Decision is made by comparing the minimum distance between each of incoming data and the existing clusters with a pre-set threshold, and continues until there are no more data available. The speech/non-speech detection step utilizes this sequential process of LFC, as follows:

1. We divide the data source given for speaker diarization into 2s-long frames² without overlap, and perform LFC on all the frames.
2. LFC decides which cluster every incoming frame is the closest to, choosing from 1) the silence cluster, 2) the universal background cluster, and 3) one of the existing speaker clusters.
 - If 1) is selected, the frame considered is labeled as silence.
 - If 2) is chosen, a new speaker cluster for the frame is generated. (The frame is newly labeled as well.)

²The reason that we select 2s as a frame length is that 2s is widely known to be the minimum window length for reasonable segmentation results [13], [64], [17].

Table 5.3: Performance comparison of the proposed speech/non-speech detection process with and without updating the silence cluster, in terms of the two detection error rates for the training data set.

	Without Update	With Update
False-Alarm Rate	2.80%	2.56%
Missed-Detection Rate	3.28%	4.46%
Total Detection Error Rate	6.08%	7.02%

- In case 3), the frame is merged to the corresponding speaker cluster. (It comes to have the same label as the other frames in the cluster.)

3. The previous step is repeated until there remain no more incoming frames.

For this process, the silence and the universal background cluster should be generated prior to LFC. (For reference, there is no speaker cluster initially other than these two clusters. Speaker clusters are generated during LFC.) For the silence cluster, we gather a total of 15s of 25ms-long audio chunks with the lowest energy from the entire data source given for speaker diarization, assuming that silence spreads over the given data source with various lengths at least longer than 25ms, and that the total length of such silence chunks in the data source is at least longer than 15s overall. Empirically, 15s is considered as enough amount to fully represent the spectral characteristics of silence. For the universal background cluster, we use the given data source entirely. This huge cluster works as if it is a source-dependent threshold for LFC, and thus we do not need to tune such a certain threshold value prior to the process as shown (as θ) in Algorithm 7 in the previous page.

Note that the silence cluster is not updated during the proposed sequential speech/non-speech detection process, while the speaker clusters keep being updated through merging. This is to preserve the initial purity of the silence cluster, which might be damaged by incorrectly merging it with speech frames. Such contamination in the silence cluster could be propagated over the entire process and thus result in a lower rate of speech detection. As shown in Table 5.3, the proposed speech/non-speech detection process with updating

the silence cluster would reduce the false-alarm rate at the relatively high cost of the missed-detection rate. As a result, the sum of the two error rates would increase overall in this case.

In the proposed speech/non-speech detection process, distance between the frame considered and all the clusters is measured by GLR [26]. For a frame F and one of the clusters C , GLR for the two objects is given as

$$\text{GLR}(F, C) = \frac{p(F|\Theta_F) \cdot p(C|\Theta_C)}{p(F \cup C|\Theta_{F \cup C})}. \quad (5.1)$$

Each object and the union of the objects are modeled by single Gaussian distributions with full covariance matrices to compute the likelihoods in the equation above, and Θ is a set of parameters in each normal distribution and is estimated toward maximizing the likelihoods of data in F , C , and $F \cup C$ for the respective model distributions.

5.3.3 Speaker Change Detection

For speaker change detection, we use the result of the previous process for speech/non-speech detection. As shown in the previous subsection, every 2s-long incoming frame to LFC is labeled as silence or one of the speaker tags assigned to the speaker clusters, respectively. In other words, all the frames except silence frames have the respective speaker tags, which means that we already have the boundary information of potential speaker changing points in the given data source. Therefore, using this information, we can further divide the data source into speaker-specific segments, each of which is surrounded by two consecutive boundaries. Every resultant segment becomes an initial cluster for AHSC in the next step.

5.3.4 Speaker Clustering

In this subsection, we apply our work in Chapters 2 and 4 to the framework of SAIL speaker diarization. Let us start this section by briefly investigating how AHSC works

Algorithm 8 Agglomerative Hierarchical Speaker Clustering (AHSC) revisited

Require: $\{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$: speaker-specific segments

$\hat{C}_i, i = 1, \dots, \hat{n}$: initial clusters

Ensure: $C_i, i = 1, \dots, n$: clusters finally remaining

1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\}, i = 1, \dots, \hat{n}$

2: **do**

3: $i, j \leftarrow \arg \min d(\hat{C}_k, \hat{C}_l), k, l = 1, \dots, \hat{n}, k \neq l$

4: merge \hat{C}_i and \hat{C}_j

5: $\hat{n} \leftarrow \hat{n} - 1$

6: **until** DER is estimated to reach the lowest level

7: **return** $C_i, i = 1, \dots, n$

in the SAIL speaker diarization system. As shown in Algorithm 8, AHSC considers the speaker-specific segments given from speaker change detection as individual initial clusters, and recursively merges the closest pair of clusters in terms of speaker-specific characteristics. Its recursive process continues until it is decided that extra cluster merging would not improve speaker clustering performance any further, i.e. until DER is estimated to reach its lowest level. All the segments in each of the clusters finally remaining are identically labeled, and every cluster label is unique.

In order for AHSC to achieve reliable performance, two critical questions need to be answered properly: 1) how to select the closest pair of clusters for merging at every recursion step and 2) how to decide the optimal (recursion) stopping point where the lowest DER would be achieved. In this context, our proposed speaker clustering method utilizes two novel approaches to address the questions, respectively: IGMM cluster modeling and ICR-based stopping point estimation.

5.3.4.1 IGMM Cluster Modeling

The inter-cluster distance measurement to select the closest pair of clusters at every recursion step of AHSC is done by comparing GLR values for all possible cluster pairs.

(Once such comparison is done, the cluster pair having the smallest GLR value is picked for merging.) For two clusters C_x and C_y , GLR is presented as follows:

$$\text{GLR}(C_x, C_y) = \frac{p(C_x|\Theta_{C_x}) \cdot p(C_y|\Theta_{C_y})}{p(C_x \cup C_y|\Theta_{C_x \cup C_y})}, \quad (5.2)$$

where Θ is a set of parameters in each cluster model distribution. Unlike speech/non-speech detection in Section 5.3.2, single Gaussian cluster modeling is not appropriate for inter-cluster distance measurement in AHSC as discussed in Chapter 4. Therefore, we utilize the IGMM cluster modeling approach in Chapter 4, which works as follows:

- Each initial cluster is modeled by a normal distribution with a full covariance matrix,
- For GLR computation in Eq. (5.2), the union of the clusters considered is modeled by the distribution³ whose PDF is the weighted sum of the PDFs of the distributions representing the clusters, respectively, and
- Any newly merged cluster is modeled by the distribution whose PDF is the weighted sum of the PDFs of the respective distributions representing merging-involved clusters, for GLR computation with other clusters at the subsequent recursion steps of AHSC.

This approach during AHSC enables not only the smooth transition of cluster models from single Gaussian distributions to GMMs, but also a gradual increase in the complexity of GMMs (or the number of mixture components in GMMs). In this cluster modeling method, Eq. (5.2) is thus written as below:

$$\text{GLR}(C_x, C_y) = \frac{p(C_x|\Lambda_{C_x}) \cdot p(C_y|\Lambda_{C_y})}{p(C_x \cup C_y|\Lambda_{C_x \cup C_y})}, \quad (5.3)$$

³As a consequence, this distribution has a mixed form of weighted normal distributions, which is a GMM.

where Λ_{C_x} , Λ_{C_y} , and $\Lambda_{C_x \cup C_y}$ are sets of parameters in the IGMMs representing the clusters considered, and the PDF of the distribution representing $C_x \cup C_y$ is simply determined as follows:

$$f_{\Lambda_{C_x \cup C_y}} = \frac{N_{C_x}}{N_{C_x} + N_{C_y}} f_{\Lambda_{C_x}} + \frac{N_{C_y}}{N_{C_x} + N_{C_y}} f_{\Lambda_{C_y}}. \quad (5.4)$$

In the above equation, N is the cardinality of the clusters, and f is the PDF of a model distribution with Λ .

5.3.4.2 ICR-based Stopping Point Estimation

A conventional stopping point estimation method, which is based on BIC, checks if GLR for the closest pair of clusters is greater than 0 using Eq. (5.3) at every recursion step of AHSC [13]. However, as discussed in Chapter 2, this method is known to be unreliable (across data sources) in terms of estimation accuracy. In order to overcome such unreliability, we utilize ICR-based stopping point estimation (proposed in Chapter 2) here.

ICR for two clusters C_x and C_y is defined as

$$\text{ICR}(C_x, C_y) \triangleq \frac{1}{N_{C_x} + N_{C_y}} \ln \text{GLR}(C_x, C_y). \quad (5.5)$$

From a viewpoint of information theory, this statistical measure between clusters represents how much entropy would be increased by merging the clusters considered. Thus, it is natural to expect ICR to be small when the clusters considered are homogeneous in terms of speaker-specific characteristics and each cluster is large enough to fully cover the intra-speaker variance of the corresponding speaker identity. In other words, ICR would be small when the clusters considered have the same speaker source and do not need additional information in representing full speaker-specific characteristics. On the contrary, ICR would be relatively large when the clusters considered are heterogeneous, or when they are homogeneous but contain small size data to cover only a part of the

Table 5.4: Comparison of 1) IGMM cluster modeling + ICR-based stopping point estimation, and 2) single Gaussian cluster modeling + BIC-based stopping point estimation, in terms of speaker-error-time rate for the testing data set. $\lambda = 25.0$ (for BIC-based stopping point estimation) and $\eta = 0.225$ (for ICR-based stopping point estimation), which are tuned based on the training data set.

	1)	2)
Speaker-Error-Time Rate	17.79%	22.75%

whole speaker-specific characteristics. As a consequence, ICR could properly work as a measure to decide homogeneity for clusters only if every cluster considered were large enough to fully represent the characteristics of the corresponding speaker identity.

Based on this, the ICR-based stopping point estimation method for AHSC in the SAIL speaker diarization system

1. Waits until AHSC reaches the end of its process, i.e., until all the initial clusters are merged to one big cluster.
2. For the pair of clusters merged at the last recursion step of AHSC, C_x and C_y , computes ICR.
3. Compares ICR with a pre-set threshold η . If $\text{ICR}(C_x, C_y) > \eta$, decides that C_x and C_y are heterogeneous in terms of speaker-specific characteristics and considers the pair of clusters merged at the next latest recursion step. Otherwise, stops considering the merged clusters and selects the recursion step previously considered as the final stopping point.

Like the conventional BIC-based one, this stopping point estimation method depends upon the reasoning that every merging after the optimal stopping point would occur only between heterogeneous clusters. The reason why its consideration of the merged clusters starts from the pair of clusters merged at the last recursion step of AHSC (i.e., the opposite direction to the one used in the BIC-based method) is that such a strategy can make ICR properly work as a homogeneity measure by handling large clusters only.

5.3.4.3 Comparison

Table 5.4 shows comparison of our proposed approaches versus the conventional ones to cluster modeling and stopping point estimation for AHSC. The proposed techniques resulted in improvement of 4.96% (absolute) and 21.80% (relative) in terms of speaker clustering performance (i.e., speaker error time rate) in the end-to-end speaker diarization system. This improvement is directly connected to the enhancement of the proposed speaker clustering strategies in terms of performance reliability.

5.3.5 Experimental Results

Figure 5.2 presents the overall performance of the proposed SAIL speaker diarization system on non-overlapped speech in the testing data set, in terms of DER. The lowest DER (6.77%) was achieved for Data 9 while the highest one (40.32%) was obtained for Data 10. Average DER is 21.90%. These results are quite comparable with those in the recent RT evaluations. (However, fair comparison with other state-of-the-art speaker diarization systems is practically impossible in this dissertation because system performance varies across data sources and our training/testing data sets are different from those used for the RT evaluations, and the best way to do such a fair comparison would be to join in any RT evaluation and compete with the other systems.)

One interesting observation is that, despite our proposed approaches to robust speaker clustering, speaker error time rate for some data sources such as Data 4, 6, and 10 still show a huge difference from that for the others, which means that there exists a room for further development in the reliability of AHSC performance. A main reason for such relatively bad results at Data 4, 6 and 10 was a lot of wrong merging between heterogeneous clusters (in terms of speaker-specific characteristics) during AHSC. This also caused mismatch between the optimal and the estimated stopping point, which led to severe DER degradation overall compared to the DERs for the other test data sources. The biggest contributor to this phenomena in speaker clustering in the framework of

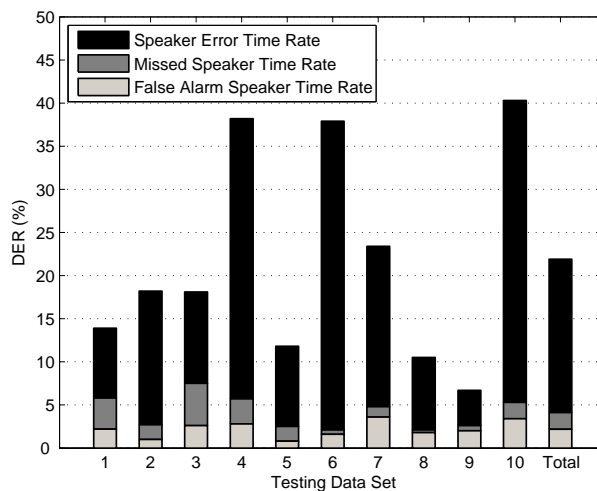


Figure 5.2: Performance of the proposed SAIL speaker diarization system on non-overlapped speech in the testing data set, in terms of DER.

speaker diarization is incorrect speaker change detection, which causes many speech segments (i.e., individual initial clusters in AHSC) to have more than one speaker sources in them. Due to their mixed statistical characteristics, those segments can confuse inter-cluster distance measurement and result in a series of incorrect merging during AHSC. In addition, considering that we handle spontaneous meeting conversations speech as data sources for speaker diarization, some segments cannot help containing overlapped speech parts, which naturally happen in real-life conversations. These kinds of ‘impure’ speech segments also can cause confusion in inter-cluster distance measurement. In the next section, we introduce a method to overcome this problem in the framework of SAIL speaker diarization for more reliable speaker diarization performance, as well as a high-level dialogue pattern modeling approach for better AHSC performance under speaker diarization of meeting conversations speech.

5.4 Refined Speaker Clustering

In this section, we propose two approaches to making AHSC better and more refined in terms of DER in the framework of SAIL speaker diarization. The first approach selects

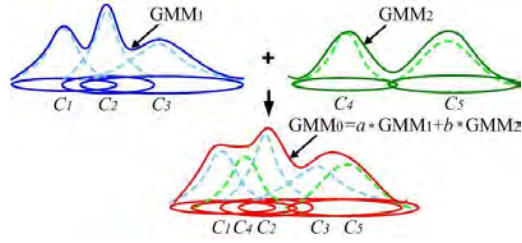


Figure 5.3: IGMM cluster modeling. $\{C_i\}_{i=1}^5$ are initial clusters for AHSC, and a and b ($a + b = 1$) are weights for the respective constituent GMMs. The weights are determined by the cardinalities of $\{C_1, C_2, C_3\}$ and $\{C_4, C_5\}$, respectively. This figure illustrates how IGMMs grow through merging during AHSC.

representative speech segments when modeling clusters with IGMMs instead of using all the segments available. This can avoid the negative effect of the aforementioned 'impure' speech segments naturally generated throughout speaker change detection onto cluster modeling and thus clustering/diarization performance. The second approach utilizes interaction patterns between speakers in a given audio source for speaker diarization. By modeling such a high-level dialogue pattern, it can provide more robust diarization performance under the variation of input audio data. Let us start this section from the first approach, by re-considering IGMM cluster modeling for inter-cluster distance measurement in AHSC.

5.4.1 Selection of Representative Speech Segments

In IGMM cluster modeling, clusters are modeled as follows:

- Every (initial) cluster in the beginning of AHSC is represented by a normal PDF with a sample mean vector and (full) covariance matrix.
- After merging during AHSC, a newly merged cluster is represented by the weighted sum of the PDFs for the clusters being merged.
- The weights are determined by the normalized cardinalities of the merged clusters.

In this way, the PDFs of cluster models not only have smooth transitions from normal PDFs to the PDFs of GMMs but also obtain a gradual increase in the number of Gaussian mixtures in the PDFs of GMMs. Computational complexity for this cluster modeling approach is quite low because there are no training sessions in IGMMs like the (EM) procedures used for conventional GMM training.

Figure 5.3 presents how the PDFs of IGMMs grow through merging in AHSC. In this figure, GMM_1 and GMM_2 represent two clusters $\{C_1, C_2, C_3\}$ and $\{C_4, C_5\}$, respectively. Each C_i is an initial cluster (i.e., individual input speech segment to AHSC). In the top row of the figure, the two clusters that have gone through merging between the initial clusters twice and once, respectively, are illustrated. Now suppose that these two clusters are merged and a newly merged cluster $\{C_1, C_2, C_3, C_4, C_5\}$ is represented by GMM_0 , depicted in the bottom part of the figure. The PDF of GMM_0 is formed by the weighted sum of the PDFs of GMM_1 and GMM_2 .

In this cluster modeling framework, therefore, every initial cluster is modeled by the PDF of a single Gaussian distribution, and once any initial cluster is merged into a larger cluster during AHSC then the PDF of its cluster model contributes to the respective IGMM by providing an individual Gaussian component. A problem is that some initial clusters might contain data from more than one speaker source due to imprecise speaker change detection or inherently overlapped speech in a given audio data source. The Gaussian mixtures generated based on those ‘impure’ initial clusters degrade the capability of IGMMs that represents the statistical characteristics of the corresponding data clusters. In this subsection, we propose a novel idea to address this problem in IGMM cluster modeling: *representative speech segment selection*. The basic idea is that, when modeling a certain, large cluster during AHSC, selecting representative initial sub-clusters from the cluster would help because they can represent the cluster statistically better.

Our way of choosing representative speech segments from a cluster is as follows. Let us consider a cluster \mathbf{C} . Suppose that the cluster has gone through merging and contains

n initial clusters, i.e., $\mathbf{C} = \{C_1, C_2, \dots, C_n\}$, where $\{C_i\}_{i=1}^n$ are initial clusters. Then, $\text{IGMM}\{\mathbf{C}\} = \text{IGMM}\{C_1, C_2, \dots, C_n\} = \lambda(\underline{m}^i, \underline{\Sigma}^i, w^i)_{i=1}^n$, where $\lambda(\cdot)$ is a GMM, \underline{m}^i and $\underline{\Sigma}^i$ are the sample mean vector and (full) covariance matrix estimated from C_i , respectively, and w^i is a weight for the Gaussian component representing C_i in this GMM.

1. Compute the likelihood of the entire data in the cluster \mathbf{C} for the PDF of every single Gaussian component, i.e.,

$$\{p(\mathbf{C}; \underline{m}^i, \underline{\Sigma}^i)\}_{i=1}^n.$$

Note that we exclude weights $\{w^i\}_{i=1}^n$ in likelihood computation. Otherwise, Gaussian components with large weights in i GMMs would tend to have high likelihood values, which is not desirable for a fair comparison in the next step.

2. Select N -best components in terms of likelihood, where N is less than the total number of Gaussian mixtures in the respective IGMM. The initial clusters (or speech segments) corresponding to the chosen N Gaussian components are considered *representative*. The N components form a new GMM (with N mixtures), which we call a refined IGMM for the cluster \mathbf{C} .
3. During AHSC, repeat 1) and 2) for every newly merged cluster whose IGMM has the number of Gaussian components greater than N . This can keep updating representative speech segments for clusters throughout AHSC.

This is simply illustrated in Figure 5.4 where we reconsider $\{C_1, C_2, C_3, C_4, C_5\}$ and its IGMM_0 in Figure 5.3. Assuming that $N = 3$, $\{C_2, C_4, C_5\}$ are selected as representative speech segments in this case and form a new, refined GMM with 3 Gaussian mixtures.

Note that our interest in this method is to see how universally individual Gaussian components in the IGMM considered represent the entire cluster data. This is because it is reasonable to regard speech segments that correspond to the Gaussian components selected in terms of such universality as representative. This selective approach for cluster

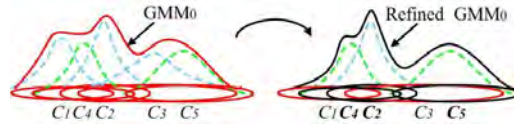


Figure 5.4: Selection of representative speech segments for improved IGMM cluster modeling. In this case, C_2 , C_4 , and C_5 are selected as representative speech segments to model $\{C_i\}_{i=1}^5$.

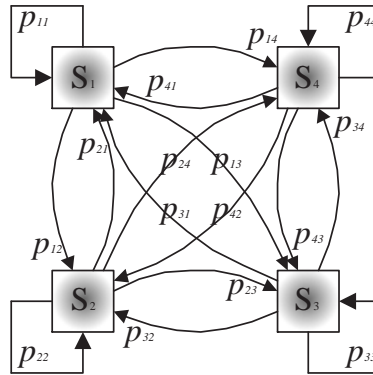


Figure 5.5: 1st-order Markov chain model for participant interaction patterns when the estimated number of speakers is 4, where p_{ij} is the transition probability from the speaker S_i to the speaker S_j for $1 \leq i, j \leq m$ ($m = 4$ in this case).

modeling using a portion of the entire cluster data can refine representation capability in cluster models in terms of not only keeping statistically representative speech segments but also excluding potentially unnecessary or even degenerate speech segments.

5.4.2 Participant Interaction Pattern Modeling

In this section, we propose another idea to draw improvement in SAIL speaker diarization of meeting conversations speech by refining speaker clustering performance regarding *interaction patterns between meeting participants*. This idea was motivated by the expectation that temporal dynamics between participants in meeting conversations are informative from a diarization perspective [11]. Modeling such dynamics would help in understanding the whole meeting speech and would reduce DER as a consequence.

We estimate participant interaction patterns, which are meeting-dependent, based on diarization results. For this purpose, we use an m -state 1st-order Markov chain model, illustrated in Figure 5.5 as an example when the number of states is 4. The number of states in this interaction pattern model is set to the number of clusters that remain after AHSC. This number means the estimated number of speakers in the given meeting speech. Each transition probability is decided as follows:

1. “Who spoke when” resulting from speaker diarization is used to count the number of speaking turn transitions (N_{ij}) from the speaker S_i to the speaker S_j , where $1 \leq i, j \leq m$. Every 2s-long segment, which is the smallest unit handled in our speaker diarization system, is considered for transition number counting.
2. Average N_{ij} with N_i , where $N_i = \sum_{j=1}^m N_{ij}$. Thus, each transition probability p_{ij} ($1 \leq i, j \leq m$) is determined by

$$p_{ij} = \frac{N_{ij}}{N_i} = \frac{N_{ij}}{\sum_{j=1}^m N_{ij}}.$$

The estimated transition probabilities in this model are used as *a priori* information for refinement of diarization results.

The refinement step performs a simple speaker identification task with considering m GMMs⁴ as pre-trained speaker models. Specifically, it refines diarization results by classifying every 2s-long segment into one of the clusters that remain after AHSC based on maximum *a posteriori*. Suppose that GMMs for the clusters that remain after AHSC are λ_i ($1 \leq i \leq m$) and the entire input meeting speech \mathbf{x} can be split into L 2s-long segments, i.e., $\mathbf{x} = \{x_1, x_2, \dots, x_L\}$. The refinement step computes the likelihood of x_l ($1 \leq l \leq L$) for each λ_i and assigns the argument i providing the highest *a posteriori* to x_l as a speaker label, i.e.,

$$\arg \max_i p(x_l | \lambda_i) p_{ji},$$

⁴These GMMs are trained by the EM procedures over representative speech segments in the respective clusters. The number of Gaussian mixtures is empirically set to 32.

Table 5.5: Improved speaker diarization performance with the two approaches proposed in this section, i.e., representative speech segment selection and participant interaction pattern modeling. For the former approach we empirically set $N = 32$. Performance comparison is given in terms of average DER (%) across 10 data sources in the testing data set in Section 5.3.1.

	DER
Baseline Performance	21.90
+ Representative Speech Segment Selection	16.76
+ Participant Interaction Pattern Modeling	14.49

where p_{ji} is the transition probability from the speaker S_j to the speaker S_i in the estimated interaction pattern model and it is assumed that the speaker label j is assigned to x_{l-1} .

5.4.3 Experimental Results

Table 5.5 presents speaker diarization performance by our original SAIL speaker diarization system in Section 5.3 and the modified system with the two approaches proposed in this section, in terms of average DER across data sources in the testing data set given in Section 5.3.1. The main reason for the diarization performance increase in the modified system with selection of representative speech segments for IGMM cluster modeling (21.90% \rightarrow 16.76%, 23.47% relative improvement) is that the proposed approach helped not only in choosing the closest pair of clusters at every recursion step of AHSC properly but also in estimating the optimal stopping point for AHSC accurately. This indicates that selecting speech segments with representativeness is better for IGMM cluster modeling than using the entire data in clusters. This claim is reasonable because clusters could contain unnecessary or defective data from a cluster representation perspective due to incorrect speaker change detection or wrong merging during AHSC and there is, therefore, a significant need to keep purifying such clusters throughout AHSC for better clustering performance. From the table, we can also see that the second approach contributed to DER reduction as well (13.54% relative improvement), as expected. It is especially

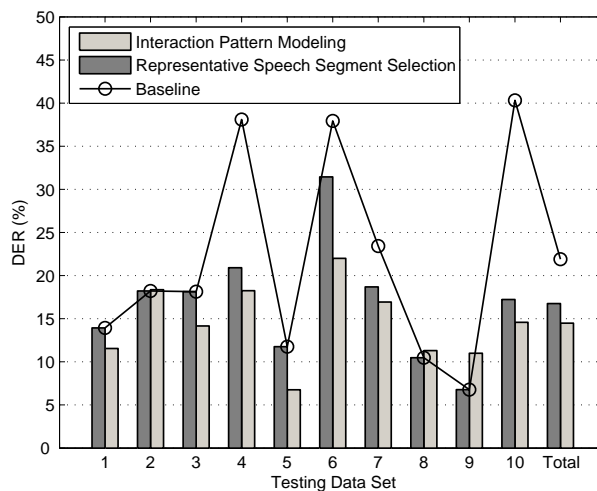


Figure 5.6: Performance of the modified SAIL speaker diarization system on non-overlapped speech in the testing data set, in terms of DER.

meaningful in this high-level modeling approach that interaction patterns between participants, which are hard to be universally modeled due to their data-dependency, can be mathematically represented in an unsupervised fashion based on diarization results. Note that a very accurate stopping point estimation for AHSC is required in the proposed approach because the number of states, m , in the 1st-order Markov chain model for interaction patterns is determined by the number of the clusters that remain after AHSC. This is already bolstered in the modified speaker diarization system by the first approach proposed in this section as well as the ICR-based stopping point estimation method.

Figure 5.6 shows how much the proposed approaches improve the reliability of AHSC/diarization performance across data sources in the testing data set more explicitly. Being compared with baseline performance (which was also shown in Figure 5.2 in Section 5.3.5), the improved performance by the modified speaker diarization system, particularly for Data 4, 6, and 10, indicates that the SAIL speaker diarization system with the two refinement methods for robust speaker clustering performance can further enhance the reliability of diarization performance under the variation of input data sources.

5.5 Conclusions

In this chapter, we implemented the SAIL speaker diarization system not only with our research results from work on robust speaker clustering in the previous chapters, but also with refinement schemes to further improve speaker clustering performance, based on 1) representative speech segment selection for IGMM cluster modeling and 2) interaction pattern modeling. The proposed speaker diarization system showed much improvement in terms of performance reliability under the variation of input data sources.

Important future work includes finding a way of robustly dealing with overlapped speech in this framework of speaker diarization. Although the first approach proposed in Section 5.4 provided some prototypical ideas in this regard, i.e., selective clustering of data can maintain or even boost representativeness in cluster models, there must be a long way to a state-of-the-art level of handling overlapped speech from a speaker clustering/diarization viewpoint. We keep working on this topic.

Chapter 6

Conclusions

6.1 Contributions

In this dissertation, we dealt with one big, yet unsolved, issue in the research field of speaker clustering: *unreliable clustering performance under the variation of input speech data*. For this, we focused on two main perspectives in the framework of agglomerative hierarchical speaker clustering (AHSC): stopping point estimation and inter-cluster distance measurement. In Chapter 2, we addressed the robustness problem of the BIC-based stopping point estimation method for AHSC under the variation of input speech data. For this, we first took a short review of GLR and BIC, and then investigated a main reason for the problem considered. This investigation led to understanding why a new statistical distance measure between clusters is needed for more robust stopping point estimation in AHSC under the variation of input speech data, which resulted in our proposal of ICR. In addition, we introduced a stopping point estimation method for AHSC based on ICR in this chapter. This stopping point estimation method was verified through experimental results to be more robust to the variation of input speech data than the conventional BIC-based one. In Chapter 3, we tackled the robustness problem of the GLR-based inter-cluster distance measure from both viewpoints of early and late AHSC recursion steps. For this, we first examined why the reliability of the GLR-based inter-cluster distance measure severely varies across input data sources. Based on this

examination, we proposed several modified versions of AHSC approaches to improve the accuracy of the GLR-based inter-cluster distance measure, particularly at the early recursion steps of AHSC. Then we proposed a supplement inter-cluster distance measure to utilize the advantages of GLR and ICR in order to tackle the robustness problem of the GLR-based inter-cluster distance measure at the late recursion steps of AHSC. All the methods proposed in this chapter were compared with original AHSC in terms of averaged performance across data sources, and were proven to provide benefit to the reliability of the GLR-based inter-cluster distance measure and thus the overall speaker clustering performance. In Chapter 4, we introduced incremental Gaussian mixture cluster modeling for inter-cluster distance measurement in AHSC. This dynamic cluster modeling approach not only provided AHSC with as comparable clustering performance as the conventional GMM-based one does, but also had a lot more feasibility in computational complexity. In Chapter 5, we applied our research results to speaker diarization. For this, we implemented our own speaker diarization system and further modified it with two clustering performance refinement schemes.

6.2 Possible Future Research Topics

One potential future research direction is to identify the lower bound for cluster size that guarantees ICR to be reliable as a statistical distance measure, more specifically as a homogeneity decision measure, between the clusters considered. In Chapter 2, we avoided the possibility that ICR would not work properly, by checking ICR-based inter-cluster homogeneity starting from the pair of clusters merged at the last recursion step of AHSC under the assumption that clusters at the late recursion steps of AHSC would be large enough for reliable ICR. This assumption worked for the meeting conversation excerpts used for the experiments presented in the chapter because most of the speaker sources involved in the conversations generated enough speech utterances of which the total length in time was longer than at least 30 seconds, respectively. Thus, at the

late recursion steps of AHSC where the ICR-based stopping point estimation method was usually applied, ICR could be reliable as an inter-cluster homogeneity measure as expected. The assumption could be however broken for other data sources which have a preponderance of short speech segments that are inadequate to reveal the corresponding speaker-specific characteristics completely.

There are also several directions for future work regarding what was handled in Chapter 3, including further refinements to the proposed, modified AHSC approaches. For instance, in the third modified version of AHSC in the chapter, the threshold parameter ζ determines the number of intermediate clusters, which is directly linked to the final speaker error time rate. It was chosen empirically in this chapter, but finding ways for optimally setting ζ would be beneficial in further enhancing clustering performance. As another example, we might have to consider how to optimally fuse two different statistical information on the same object for (GLR+ICR)-based inter-cluster distance measurement at the late recursion steps of AHSC. In this chapter, we used soft rankings in terms of GLR and ICR for that purpose, but it is not theoretically proven to be optimal to the task considered. Establishing more systematic frameworks for selection of information fusion methods could be one of valuable future research directions. In addition, as mentioned in the middle part of this chapter, it would be a good research topic to find out a stopping point estimation method for AHSC with the (GLR+ICR)-based inter-cluster distance measure, other than the ICR-based one. A new stopping point estimation method should be comparable to our proposed ICR-based one in terms of estimation accuracy, but needs to use an inter-cluster homogeneity decision measure independent of ICR. Then it could keep the advantages of the modified versions of AHSC and the (GLR+ICR)-based inter-cluster distance measure valid even in practical applications.

The research results in Chapter 4 could be extended to speaker modeling in the research field of speaker recognition. Currently, speaker modeling is performed based on GMMs with a fixed number of mixture components like 16 or 32, but we still do not know

how many mixture Gaussians would be necessary for the optimal modeling of speaker-specific characteristics. Based on our intuition it should be speaker-dependent, but there is no canonical method yet to derive the proper number of mixture components in GMMs for speaker-specific representation of data. The cluster modeling approach proposed in this chapter does not require any fixed number for mixture Gaussians beforehand, so it might be able to be a good alternative to the conventional GMM-based speaker modeling.

6.3 Final Remarks

My Ph.D. research work on this topic is a tiny part of vast research effort now being conducted within the research field of pattern classification, but I hope and believe that it is a meaningful contribution to this field because the reliability issue of speaker clustering performance across data sources has not been significantly tackled thus far although there has been much recognition on the seriousness of this issue. The entire results in this dissertation could be utilized for other data domains beyond speech data, particularly where there exists similar data-dependency in clustering performance.

Bibliography

- [1] J. Ajmera, C. Wooters, B. Peskin, and C. Oei. Speaker segmentation and clustering. In *NIST RT-03S Workshop*, Boston, MA, USA, May 2003.
- [2] Jitendra Ajmera, Iain McCowan, and Herve Bourlard. Robust speaker change detection. *IEEE Signal Processing Letter*, 11(8):649–651, 2004.
- [3] Jitendra Ajmera and Chuck Wooters. A robust speaker clustering algorithm. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 411–416, St. Thomas, VI, USA, November 2003.
- [4] X. Anguera, C. Wooters, and J. Pardo. Robust speaker diarization for meetings: ICSI RT06s evaluation system. In *International Conference on Spoken Language Processing*, pages 1674–1677, Pittsburgh, PA, USA, September 2006.
- [5] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo. Robust speaker segmentation for meetings: the ICSI-SRI spring 2005 diarization system. In *Multimodal Interaction and Related Machine Learning Algorithms*, pages 402–414, Edinburgh, UK, July 2005.
- [6] Raimo Bakis, Scott Chen, Ponani Gopalakrishnan, Ramesh Gopinath, Stephane Maes, Lazaros Polymenakos, and Martin Franz. Transcription of broadcast news shows with the IBM large vocabulary speech recognition system. In *DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.
- [7] P. Balaram. Information overload. *Current Science*, 78(5):533–534, 2000.
- [8] Claude Barras, Xuan Zhu, Sylvain Meignier, and J. Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1505–1512, 2006.
- [9] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, Dietrich Klakow, A. Wendemuth, Sirko Molau, Michael Pitz, and A. Sixtus. The Philips/RWTH system for transcription of broadcast news. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [10] Peter Beyerlein, Xavier Aubert, Reinhold Haeb-Umbach, Dietrich Klakow, Meinhard Ullrich, Andreas Wendemuth, and Patricia Wilcox. Automatic transcription of English broadcast news. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.

- [11] Carlos Busso, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Real-time monitoring of participants interaction in a meeting using audio-visual sensors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 685–688, Honolulu, HI, USA, April 2007.
- [12] Scott S. Chen, Ellen Eide, M. J. F. Gales, Ramesh A. Gopinath, D. Kanevsky, and P. Olsen. Recent improvements to IBM’s speech recognition system for automatic transcription of broadcast news. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [13] Scott S. Chen and Ponani S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [14] Scott S. Chen and Ramesh A. Gopinath. Model selection in acoustic modeling. In *European Conference on Speech Communication and Technology*, pages 1087–1090, Budapest, Hungary, September 1999.
- [15] Scott Shaobing Chen, M. J. F. Gales, P. S. Gopalakrishnan, R. A. Gopinath, H. Printz, D. Kanevsky, P. Olsen, and L. Polymenakos. IBM’s LVCSR system for transcription of broadcast news used in the 1997 Hub-4 English evaluation. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [16] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 1991.
- [17] P. Delacourt and C. J. Wellekens. DISTBIC: a speaker-based segmentation for audio data indexing. *Speech Communication*, 32(1-2):111–126, 2000.
- [18] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2001.
- [19] D. Eichmann, Miguel Ruiz, Padmini Srinivasan, Nick Street, Chris Culy, and Filippo Menczer. A cluster-based approach to tracking, detection, and segmentation of broadcast news. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [20] J. Gauvain, G. Adda, L. Lamel, and M. Adda-Decker. Transcribing broadcast news: the LIMSI Nov96 Hub4 system. In *DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.
- [21] J. Gauvain, L. Lamel, G. Adda, L. Chen, and H. Schwenk. The LIMSI RT03 BN systems. In *NIST RT-03S Workshop*, Boston, MA, USA, May 2003.
- [22] J. Gauvain, Lori Lamel, and G. Adda. The LIMSI 1997 Hub-4E transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.

- [23] J. Gauvain, Lori Lamel, and Gilles Adda. Partitioning and transcription of broadcast news data. In *International Conference on Spoken Language Processing*, pages 1335–1338, Sydney, Australia, November 1998.
- [24] J. Gauvain, Lori Lamel, Gilles Adda, and Michele Jardino. The LIMSI 1998 Hub-4E transcription system. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [25] Jean Gauvain, Lori Lamel, and Gilles Adda. Transcribing broadcast news for audio and video indexing. *Communications of the ACM*, 43(2):64–70, 2000.
- [26] Herbert Gish, M. Siu, and Robin Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 873–876, Toronto, Ontario, Canada, May 1991.
- [27] XueFeng Guo, WeiBin Zhu, Qin Shi, Scott S. Chen, and Ramesh A. Gopinath. The IBM LVCSR system used for 1998 Mandarin broadcast news transcription evaluation. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [28] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S.J. Young. Segment generation and clustering in the HTK broadcast news transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [29] Kyu J. Han, Samuel Kim, and Shrikanth S. Narayanan. Robust speaker clustering strategies to data source variation for improved speaker diarization. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 262–267, Kyoto, Japan, December 2007.
- [30] Kyu J. Han, Samuel Kim, and Shrikanth S. Narayanan. Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1590–1601, 2008.
- [31] Kyu J. Han and Shrikanth S. Narayanan. A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system. In *European Conference on Speech Communication and Technology*, pages 1853–1856, Antwerp, Belgium, August 2007.
- [32] Kyu J. Han and Shrikanth S. Narayanan. A novel inter-cluster distance measure combining GLR and ICR for improved agglomerative hierarchical speaker clustering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4373–4376, Las Vegas, NV, USA, March 2008.
- [33] Jing Huang, Etienne Marcheret, Karthik Visweswariah, and Gerasimos Potamianos. The IBM RT07 evaluation systems for speaker diarization on lecture meetings. In *International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 497–508, Baltimore, MD, USA, May 2007.

- [34] Juan M. Huerta, Stanley Chen, and Richard M. Stern. The 1998 Carnegie Mellon University Sphinx-3 Spanish broadcast news transcription system. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [35] Hubert Jin, Francis Kubala, and Rich Schwartz. Automatic speaker clustering. In *DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.
- [36] F. Kubala, T. Anastasakos, H. Jin, L. Nguyen, and R. Schwartz. Transcribing radio news. In *International Conference on Spoken Language Processing*, pages 598–601, Philadelphia, PA, USA, October 1996.
- [37] Francis Kubala, Jason Davenport, Hubert Jin, Daben Liu, Tim Leek, Spyros Matsoukas, David Miller, Long Nguyen, Fred Richardson, Richard Schwartz, and John Makhoul. The 1997 BBN Byblos system applied to broadcast news transcription. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [38] Francis Kubala, Hubert Jin, Spyros Matsoukas, Long Nguyen, Rich Schwartz, and John Makhoul. The 1996 BBN Byblos Hub-4 transcription system. In *DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.
- [39] D. Liu, A. Srivastava, F. Kubala, D. Kiecza, A. Ann, J. Maguire, R. Schwartz, M. Snover, and B. Dorr. BBN+UMD Rich Transcription system for broadcast news. In *NIST RT-03F Workshop*, Washington, DC, USA, October 2003.
- [40] Daben Liu and Francis Kubala. Fast speaker change detection for broadcast news transcription and indexing. In *European Conference on Speech Communication and Technology*, pages 1031–1034, Budapest, Hungary, September 1999.
- [41] P. Maes. Agents that reduce work and information overload. *Communications of the ACM*, 37(7):30–40, 1994.
- [42] D. Mararu, L. Besacier, S. Meignier, C. Fredouille, and J. Bonastre. ELISA, CLIPS and LIA NIST 2003 segmentation. In *NIST RT-03S Workshop*, Boston, MA, USA, May 2003.
- [43] Spyros Matsoukas, Long Nguyen, Jason Davenport, Jay Billa, Fred Richardson, Man-hung Siu, Daben Liu, and Richard Schwartz. The 1998 BBN BYBLOS primary system applied to English and Spanish broadcast news transcription. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [44] L. Nguyen, N. Duta, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xing, and D. Xu. The BBN RT03 BN English system. In *NIST RT-03S Workshop*, Boston, MA, USA, May 2003.
- [45] Masafumi Nishida and Tatsuya Kawahara. Speaker model selection based on the Bayesian information criterion applied to unsupervised speaker indexing. *IEEE Transactions on Speech and Audio Processing*, 13(4):583–592, 2005.

- [46] J. M. Noyes and P. J. Thomas. Information overload: an overview. *IEE Colloquium on Information Overload*, pages 0–6, 1995.
- [47] Katsutoshi Ohtsuki, Sadaoki Furui, Naoyuki Sakurai, Atushi Iwasaki, and Z. Zhang. Improvements in Japanese broadcast news transcription. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [48] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny. Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 701–704, Atlanta, GA, USA, May 1996.
- [49] N. Reeves, S. Mills, and J. Noyes. Information retrieval from a user perspective. *IEE Colloquium on Information Overload*, pages 3/1–3/7, 1995.
- [50] D. Reynolds, P. Torres, and R. Roy. EARS RT03S diarization. In *NIST RT-03S Workshop*, Boston, MA, USA, May 2003.
- [51] Douglas A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1-2):91–108, 1995.
- [52] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.
- [53] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.
- [54] Douglas A. Reynolds and Pedro A. Torres-Carrasquillo. The MIT Lincoln laboratory RT-04F diarization systems: applications to broadcast news and telephone conversations. In *NIST RT-04F Workshop*, Palisades, NY, USA, November 2004.
- [55] Douglas A. Reynolds and Pedro A. Torres-Carrasquillo. Approaches and applications of audio diarization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 953–956, Philadelphia, PA, USA, March 2005.
- [56] R. H. Rockland. Reducing the information overload: a method on helping students research engineering topics using the Internet. *IEEE Transactions on Education*, 43(4):420–425, 2000.
- [57] Ananth Sankar, Ramana Rao Gadde, and Fuliang Weng. SRI’s 1998 broadcast news system - toward faster, better, smaller speech recognition. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [58] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

- [59] K. Seymore, Stanley Chen, S. Doh, M. Eskenazi, E. Gouvea, B. Raj, Mosur Ravishankar, Ronald Rosenfeld, M. A. Siegler, Richard M. Stern, and Eric Thayer. The 1997 CMU Sphinx-3 English broadcast news transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [60] Koichi Shinoda and Takao Watanabe. Acoustic modeling based on the MDL principle for speech recognition. In *European Conference on Speech Communication and Technology*, pages 99–102, Berlin, Germany, September 1997.
- [61] Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern. Automatic segmentation, classification, and clustering of broadcast news audio. In *DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.
- [62] Hiroshi Tenmoto, Mineichi Kudo, and Masaru Shimbo. MDL-based selection of the number of components in mixture models for pattern classification. In *Joint IAPR International Workshop on Structural and Syntactic Pattern Recognition, and Statistical Pattern Recognition*, pages 831–836, Sydney, NSW, Australia, August 1998.
- [63] S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 43(5):1557–1565, 2006.
- [64] A. Tritschler and R. Gopinath. Improved speaker segmentation and segments clustering using the Bayesian information criterion. In *European Conference on Speech Communication and Technology*, pages 679–682, Budapest, Hungary, September 1999.
- [65] David A. van Leeuwen. The TNO speaker diarization system for NIST RT05s meeting data. In *Multimodal Interaction and Related Machine Learning Algorithms*, pages 440–449, Edinburgh, UK, July 2005.
- [66] An Vandecatseye and J. Martens. A fast, accurate and stream-based speaker segmentation and clustering algorithm. In *European Conference on Speech Communication and Technology*, pages 941–944, Geneva, Switzerland, September 2003.
- [67] Steven Wegmann, Francesco Scattone, Ira Carp, Larry Gillick, Robert Roth, and Jonathan P. Yamron. Dragon systems’ 1997 broadcast news transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [68] P. Woodland, G. Evermann, M. Gales, T. Hain, R. Chan, B. Jia, D. Y. Kim, A. Liu, D. Mrva, D. Povey, K. C. Sim, M. Tomalin, S. Tranter, L. Wang, and K. Yu. CU-HTK STT systems for RT03. In *NIST RT-03S Workshop*, Boston, MA, USA, May 2003.
- [69] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young. The development of the 1996 HTK broadcast news transcription system. In *DARPA Speech Recognition Workshop*, Chantilly, VA, USA, February 1997.

- [70] P. C. Woodland, T. Hain, S. E. Johnson, T. R. Niesler, A. Tuerk, E. W. D. Whittaker, and S. J. Young. The 1997 HTK broadcast news transcription system. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [71] P. C. Woodland, T. Hain, G. L. Moore, T. R. Niesler, D. Povey, Andreas Tuerk, and E. W. D. Whittaker. The 1998 HTK broadcast news transcription system: development and results. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [72] C. Wooters, J. Fung, B. Peskin, and X. Anguera. Towards robust speaker segmentation: the ICSI-SRI fall 2004 diarization system. In *NIST RT-04F Workshop*, Palisades, NY, USA, November 2004.
- [73] Chuck Wooters and Marijn Huijbregts. The ICSI RT07s speaker diarization system. In *International Evaluation Workshops CLEAR 2007 and RT 2007*, pages 509–519, Baltimore, MD, USA, May 2007.
- [74] Xintian Wu, Chaojun Liu, Yonghong Yan, Doughwa Kim, Seth Cameron, and Randy Parr. The 1998 OGI-Fonix broadcast news transcription system. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [75] Rui Xu and Donald C. Wunsch. *Clustering*. IEEE Press, 2009.
- [76] Yonghong Yan, Xintian Wu, Johan Schalkwyk, and Ron Cole. Development of the CSLU LVCSR: the 1997 DARPA Hub-4 evaluation system. In *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, February 1998.
- [77] Puming Zhan, Steven Wegmann, and Larry Gillick. Dragon Systems’ 1998 broadcast news transcription system for Mandarin. In *DARPA Broadcast News Workshop*, Herndon, VA, USA, February 1999.
- [78] Xuan Zhu, Claude Barras, Lori Lamel, and J. Gauvain. Speaker diarization: from broadcast news to lectures. In *Machine Learning for Multimodal Interaction*, pages 396–406, Bethesda, MD, USA, May 2006.