SPEAKER CLUSTERING BASED ON MINIMUM RAND INDEX

Wei-Ho Tsai[1] *and Hsin-Min Wang*[2]

[1]Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[1]whtsai@en.ntut.edu.tw, [2]whm@iis.sinica.edu.tw

ABSTRACT

This paper presents an effective method for clustering unknown speech utterances based on their associated speakers. The proposed method jointly optimizes the generated clusters and the number of clusters by estimating and minimizing the Rand index of the clustering. The Rand index, which reflects clustering errors that utterances from the same speaker are placed in different clusters, or utterances from different speakers are placed in the same cluster, reaches its minimal value only when the number of clusters is equal to the true speaker population size. We approximate the Rand index by a function of the similarity measures between utterances and employ the genetic algorithm to determine the cluster where each utterance should be located, such that the overall clustering errors are minimized. The experimental results show that the proposed speaker-clustering method outperforms the conventional method based on hierarchical agglomerative clustering in conjunction with the Bayesian information criterion to determine the number of clusters.

***Index Terms*—**Clustering methods, Speech processing, Speaker recognition

1. INTRODUCTION

With the burgeoning availability of digital audio material, speaker clustering is gaining importance as a means of indexing the voluminous spoken data accumulated daily for archival use [1-14]. Given N speech utterances produced by P speakers, the goal of speaker clustering is to partition N utterances into M clusters, such that $M = P$ and each cluster consists exclusively of utterances from only one speaker. Since no prior information regarding the speakers involved and the speaker population size is available in most practical applications, solving the speaker-clustering problem usually involves characterizing the voice similarities between utterances, generating clusters based on those similarities, and determining the optimal number of clusters.

Currently, the most popular method of speaker clustering generates a cluster tree by sequentially merging the utterances deemed similar to each other, and then cuts the tree via a Bayesian information criterion (BIC) [5,8,10-12,15], in order to retain an appropriate number of clusters. During the agglomeration procedure, the nearest neighborhood selection rule is usually employed in an attempt to maximize the similarities between all the utterances within each cluster. Since the interaction between clusters is not considered, this method can only make each individual cluster as homogeneous as possible; however it cannot guarantee that the homogeneity for all the clusters can finally be summed to reach a maximum. In particular, mis-clustering errors arising from grouping different-speaker utterances together can propagate down the whole process, and hence limit the clustering performance. In addition, the cluster tree is generated separately from the determination of the optimal number of clusters. Since the latter trusts the former completely, the inevitable errors from the former can propagate to the latter, which may lead to a poor estimation of the speaker population size.

To overcome the above-mentioned limitations of the conventional method, we propose a new clustering method that jointly optimizes the generated clusters and the number of clusters by estimating and minimizing a metric called the Rand index [16,17]. This metric indicates the clustering errors that place utterances from the same speaker in different clusters, or place utterances from different speakers in the same cluster. We approximate the Rand index by a function of the similarity measures between utterances, and employ the genetic algorithm [18] to determine the cluster where each utterance should be located. The resulting clusters are thus optimized in a global fashion, rather than a pair-by-pair manner used in the conventional method. In addition, by exploiting a characteristic of the Rand index that it only reaches the minimal value when the number of clusters equals the true speaker population size, speaker clustering based on the minimization of the estimated Rand index also enables the resulting number of clusters to approach the optimum.

2. PROBLEM FORMULATION

For convenience of discussion, we begin by defining the following symbols.

$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$: N speech utterances to be clustered;
$s_1, s_2, \ldots, s_P$: P unknown speakers involved in N utterances;
$c_1, c_2, \ldots, c_M$: M clusters to be generated;
o_n : index of the speaker producing utterance $\mathbf{X}_n$;
h_n : index of the cluster that utterance $\mathbf{X}_n$ is assigned to;
n_{m*} : number of utterances in c_m;
n_{*p} : number of utterances spoken by s_p;
n_{mp} : number of utterances in c_m spoken by s_p.

The goal of speaker clustering is to produce a set of indices $\mathbf{H} = \{h_1, h_2, \ldots, h_N\}$ that satisfy $h_i = h_j$ for any $\mathbf{X}_i$ and $\mathbf{X}_j$ from the same speaker, and $h_i \neq h_j$ for any $\mathbf{X}_i$ and $\mathbf{X}_j$ from different speakers.

Depending on the application, there are a number of ways to evaluate the performance of speaker clustering. This study uses two metrics: cluster purity [4] and the Rand index [4,16,17]. Cluster purity represents the probability that if we pick any utterance from a cluster twice at random, with replacement, both of the selected utterances will be from the same speaker. Specifically, the average purity for M clusters is computed by

$$\bar{\rho} = \frac{1}{N}\sum\nolimits_{m=1}^{M} n_{m*}\ \rho_m, \tag{1}$$

where ρ_m is the purity of cluster c$_m$:

$$\rho_m = \sum\nolimits_{p=1}^{P}\left(n_{mp}\big/n_{m*}\right)^2. \tag{2}$$

Apparently, a perfect clustering should produce an average purity of one. However, this does not work both ways. The value of the average purity generally increases as the number of clusters increases, since the metric does not consider errors that place utterances from the same speaker in different clusters. Hence, the cluster purity is only suitable for comparing the performance of different clustering methods under a specified number of clusters

In contrast, the Rand index indicates the number of utterance pairs from the same speaker that are in different clusters, or from different speakers that are in the same cluster. Specifically, the Rand index is computed by

$$R(M) = \sum_{m=1}^{M} n_{m*}^2 + \sum_{p=1}^{P} n_{*p}^2 - 2\sum_{m=1}^{M}\sum_{p=1}^{P} n_{mp}^2. \tag{3}$$

Obviously, the smaller the value of $R(M)$, the better the clustering performance will be. Unlike the cluster purity, which favors a large value of M, the Rand index generally decreases with an increase in the value of M initially, and reaches the minimum at $M = P$. When $M > P$, the Rand index starts to increase as the value of M increases.

To illustrate why the minimal value of $R(M)$ occurs only at $M = P$, let us consider the following cases.

(i) The clustering is perfect, which satisfies

$$\begin{pmatrix} n_{11} & n_{21} & \dots & n_{M1} \\ n_{12} & n_{22} & \dots & n_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{1P} & n_{2P} & \dots & n_{MP} \end{pmatrix} = \begin{pmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_P \end{pmatrix}_{P\times P}, \tag{4}$$

where $n_i = n_{*i} = n_{i*}$, $1 \le i \le P$. Then, the resulting Rand index is

$$R^*(P) = \sum_{m=1}^{P} n_{m*}^2 + \sum_{p=1}^{P} n_{*p}^2 - 2\sum_{m=1}^{P}\sum_{p=1}^{P} n_{mp}^2 = \sum_{m=1}^{P} n_m^2 + \sum_{p=1}^{P} n_p^2 - 2\sum_{k=1}^{P} n_k^2 = 0. \tag{5}$$

(ii) Let $M = P + 1$, and modify Eq. (4) by splitting cluster c_k into two clusters, c_k and c_{P+1}, i.e.,

$$\begin{pmatrix} n_{11} & n_{21} & \dots & n_{M1} \\ n_{12} & n_{22} & \dots & n_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{1P} & n_{2P} & \dots & n_{MP} \end{pmatrix} = \begin{pmatrix} n_1 & 0 & \dots & 0 & \dots & 0 & 0 \\ 0 & n_2 & \dots & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & n_{kk} & \dots & 0 & n_{(P+1)k} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & \dots & n_P & 0 \end{pmatrix}_{(P+1)\times P}, \tag{6}$$

where $n_{kk} + n_{(P+1)k} = n_k$. Then, the resulting Rand index is

$$\begin{aligned} &R(P+1) \\ &= \sum_{m=1}^{P+1} n_{m*}^2 + \sum_{p=1}^{P} n_{*p}^2 - 2\sum_{m=1}^{P+1}\sum_{p=1}^{P} n_{mp}^2 \\ &= \left(\sum_{m=1}^{P} n_m^2 - n_k^2 + n_{kk}^2 + n_{(P+1)k}^2\right) + \sum_{p=1}^{P} n_p^2 - 2\left(\sum_{m=1}^{P} n_m^2 - n_k^2 + n_{kk}^2 + n_{(P+1)k}^2\right) \\ &= n_k^2 - n_{kk}^2 - n_{(P+1)k}^2 \\ &= n_k^2 - n_{kk}^2 - (n_k - n_{kk})^2 = 2n_{kk}(n_k - n_{kk}) > 0. \end{aligned} \tag{7}$$

(iii) Let $M = P - 1$, and modify Eq. (4) by merging cluster c_P into cluster c_k, i.e.,

$$\begin{pmatrix} n_{11} & n_{21} & \dots & n_{M1} \\ n_{12} & n_{22} & \dots & n_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ n_{1P} & n_{2P} & \dots & n_{MP} \end{pmatrix} = \begin{pmatrix} n_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & n_k & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \dots & n_{P-1} \\ 0 & 0 & \dots & n_P & \dots & 0 \end{pmatrix}_{P\times(P-1)}. \tag{8}$$

Then, the resulting Rand index is

$$\begin{aligned} R(P-1) &= \sum_{m=1}^{P-1} n_{m*}^2 + \sum_{p=1}^{P} n_{*p}^2 - 2\sum_{m=1}^{P-1}\sum_{p=1}^{P} n_{mp}^2 \\ &= \left(\sum_{m=1}^{P} n_m^2 - n_P^2 - n_k^2 + (n_k + n_P)^2\right) + \sum_{p=1}^{P} n_p^2 - 2\sum_{m=1}^{P} n_m^2 \\ &= 2n_k n_P > 0. \end{aligned} \tag{9}$$

We observe from these three cases that, in general, $R(M) > R(P)$ if $M \neq P$. Therefore, the Rand index can be used not only to examine if each generated cluster is homogeneous in terms of the speaker, but also to serve as a criterion to determine the true speaker population size. This property motivates us to develop a clustering method that jointly optimizes the generated clusters and the number of clusters by estimating and minimizing the Rand index.

3. MINIMUM RAND INDEX CLUSTERING (MRIC)

Our basic strategy is to find a set of indices $\mathbf{H}^{(M)} = \{h_1^{(M)}, h_2^{(M)}, \dots, h_N^{(M)}\}$ for the N utterances to be clustered, such that the resulting Rand index is minimized, where $h_i^{(M)}$, $1 \le i \le N$, is an integer between 1 and M, and the value of M is to be determined. Since in Eq. (3)

$$\begin{aligned} \sum_{m=1}^{M} n_{m*}^2 &= \sum_{m=1}^{M}\left[\ \sum_{i=1}^{N} \delta(h_i^{(M)}, m)\ \right]^2 \\ &= \sum_{m=1}^{M}\left[\sum_{i=1}^{N} \delta(h_i^{(M)}, m)\right]\left[\ \sum_{j=1}^{N} \delta(h_j^{(M)}, m)\right] \\ &= \sum_{m=1}^{M}\sum_{i=1}^{N}\sum_{j=1}^{N} \delta(h_i^{(M)}, m)\delta(h_j^{(M)}, m) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \delta(h_i^{(M)}, h_j^{(M)}), \end{aligned} \tag{10}$$

$$\begin{aligned} &\sum_{m=1}^{M}\sum_{p=1}^{P} n_{mp}^2 = \sum_{m=1}^{M}\sum_{p=1}^{P}\left[\sum\nolimits_{i=1}^{N} \delta(h_i^{(M)}, m)\delta(o_i, p)\ \right]^2 \\ &= \sum_{m=1}^{M}\sum_{p=1}^{P}\left[\ \sum_{i=1}^{N} \delta(h_i^{(M)}, m)\delta(o_i, p)\ \right]\left[\ \sum_{j=1}^{N} \delta(h_j^{(M)}, m)\delta(o_j, p)\ \right] \\ &= \sum_{m=1}^{M}\sum_{p=1}^{P}\sum_{i=1}^{N}\sum_{j=1}^{N} \delta(h_i^{(M)}, m)\delta(o_i, p)\delta(h_j^{(M)}, m)\delta(o_j, p) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} \delta(h_i^{(M)}, h_j^{(M)})\delta(o_i, o_j), \end{aligned} \tag{11}$$

and $\Sigma_{p=1}^{P} n_{*p}^2 = \Omega$ is a constant that is irrelevant to the clustering, the optimal set of cluster indices can be determined by

$$\mathbf{H}^* = \underset{\mathbf{H}^{(M)}, 1\le M\le N}{\arg\min}\ \hat{R}(\mathbf{H}^{(M)}), \tag{12}$$

and

$$\hat{R}(\mathbf{H}^{(M)}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\delta(h_i^{(M)}, h_j^{(M)}) + \Omega - 2\sum_{i=1}^{N}\sum_{j=1}^{N}\delta(h_i^{(M)}, h_j^{(M)})\delta(o_i, o_j), \tag{13}$$

where $\delta(\cdot)$ in Eqs. (10)–(13) is a Kronecker Delta function.

However, as the computation of $\delta(o_i, o_j)$ requires that the true speaker of each utterance be known in advance, it is impossible to find $\mathbf{H}^*$ directly from Eqs. (12) and (13). To solve this problem, we propose estimating $\delta(o_i, o_j)$ by means of the similarity measure between $\mathbf{X}_i$ and $\mathbf{X}_j$. Specifically,

$$\delta(o_i, o_j) \leftarrow \begin{cases} 1 & , \text{ if } i = j \\ S(\mathbf{X}_i, \mathbf{X}_j) \big/ S_{\max} , & \text{if } i \neq j, \text{and } S_{\max} > 0 , \\ S_{\max} \big/ S(\mathbf{X}_i, \mathbf{X}_j) , & \text{if } i \neq j, \text{and } S_{\max} < 0 \end{cases} \tag{14}$$

where $S(\mathbf{X}_i, \mathbf{X}_j)$ denotes a certain similarity measure between $\mathbf{X}_i$ and $\mathbf{X}_j$ that could be either positive or negative, but cannot be zero, and $S_{\max}$ is the maximum among the similarities $S(\mathbf{X}_i, \mathbf{X}_j)$, $\forall\ i \neq j$. In our implementation, $S(\mathbf{X}_i, \mathbf{X}_j)$ is computed by the generalized likelihood ratio (GLR) [1,4]:

$$S(\mathbf{X}_i, \mathbf{X}_j) = \log\Pr(\mathbf{X}_{ij} \,|\lambda_{ij}) - \log\Pr(\mathbf{X}_i \,|\lambda_i) - \log\Pr(\mathbf{X}_j \,|\lambda_j), \tag{15}$$

where $\mathbf{X}_{ij}$ is the concatenation of $\mathbf{X}_i$ and $\mathbf{X}_j$, and λ_i, λ_j, and λ_{ij} are parametric models trained using $\mathbf{X}_i$, $\mathbf{X}_j$, and $\mathbf{X}_{ij}$, respectively. Using this estimation, we can solve Eq. (12) by further assigning to Ω an arbitrary positive constant that ensures $\hat{R}(\mathbf{H}^{(M)}) \geq 0$.

Given that neither a gradient-based optimization method nor an exhaustive search is applicable in this scenario, we propose using the genetic algorithm (GA) [18] to find $\mathbf{H}^*$ by virtue of its global scope and parallel searching power. The basic operation of the GA is to explore a given search space in parallel by means of iterative modifications of a population of chromosomes. Each chromosome, encoded as a string of alphabets or real numbers called genes, represents a potential solution to a given problem. In our task, a chromosome is exactly a legitimate $\mathbf{H}^{(M)}$, and a gene corresponds to a cluster index associated with an utterance. However, since the index of one cluster can be interchanged with that of another cluster, multiple chromosomes may amount to an identical clustering result. For example, the chromosomes {1 1 1 2 2 3 3}, {1 1 1 3 3 2 2}, {2 2 2 1 1 3 3}, and {1 1 1 5 5 4 4} represent the same clustering result derived by grouping seven utterances into three clusters. Such a non-unique representation of the solution would significantly increase the GA search space, and may lead to an inferior clustering result. To avoid this problem, we limit the inventory of chromosomes to conform to a baseform representation defined as follows.

Let $I(c_m)$ be the lowest index of the utterance in cluster c_m. Then, a chromosome is a baseform

$$\text{iff } \forall\ c_m, c_l \neq \{\phi\}, \text{ if } m < l, \text{ then } I(c_m) < I(c_l), \tag{16}$$

where $\{\phi\}$ indicates that a cluster does not contain any utterance. Among the above chromosomes, {1 1 1 2 2 3 3} is a baseform, since the lowest index of the utterance in clusters c_1, c_2, and c_3 is 1, 4, and 6, respectively, which satisfies Eq. (16). In contrast, chromosomes {1 1 1 3 3 2 2} and {2 2 2 1 1 3 3} are not baseforms, since the lowest index of the utterance in clusters c_1, c_2, and c_3 does not satisfy Eq. (16). In addition, chromosome {1 1 1 5 5 4 4} implies that clusters c_2 and c_3 do not contains any utterance; hence it is not a baseform, either. However, it is conceivable that all the non-baseform chromosomes can be converted into a unique baseform representation by re-arranging the cluster indices.

GA optimization starts with a random generation of chromosomes according to a certain population size, Z. Then, the fitness of all chromosomes is evaluated via the inverse of the estimated Rand index, i.e., $F(\mathbf{H}^{(M)}) = 1\big/\hat{R}(\mathbf{H}^{(M)})$. Based on this evaluation, a particular group of chromosomes is selected from the population to generate offspring by subsequent recombination. To prevent premature convergence of the population, the selection is performed with the linear ranking scheme described in [19]. Next, crossover among the selected chromosomes proceeds by exchanging the substrings of two chromosomes between two randomly selected crossover points. A crossover probability is assigned to control the number of offspring produced in each generation. After crossover, a mutation operator is used to introduce random variations into the genetic structure of the chromosomes. This is done by generating a random number and then replacing one gene of an existing chromosome with a mutation probability. The resulting chromosomes that do not conform to the baseform representations are converted into their baseform counterparts.

The procedure of fitness evaluation, selection, crossover, and mutation is repeated continuously, in the hope that the overall fitness of the population will increase from generation to generation. When the maximum number of generations is reached, the best chromosome in the final population is taken as the solution, $\mathbf{H}^*$. Note that the estimated speaker population size can be obtained by selecting the maximal value of the cluster index in $\mathbf{H}^*$. For example, if $\mathbf{H}^*$ = {1 2 1 3 4 3 1}, the estimated number of speakers in a seven-utterance collection is 4.

4. EXPERIMENTAL RESULTS

The speech data used in this study consisted of six excerpts of broadcasts from the evaluation set of the *2002 Rich Transcription Broadcast News and Conversational Telephone Speech Corpus* [20]. Each excerpt was segmented into speaker-homogeneous utterances, according to the annotation files in the corpus. Speaker clustering was then applied to each excerpt separately. Prior to the experiments, every speech utterance was converted from its digital waveform representation into a sequence of feature vectors, each of which consisted of 12 Mel-scale frequency cepstral coefficients (MFCCs) and 12 delta MFCCs. Then, the similarities between the utterances were computed using Eq. (15), in which all the parametric models are of a uni-Gaussian model with a full covariance matrix.

In GA optimization, the parameter values used for the maximum number of generations, the population size, the crossover probability, and the mutation probability were empirically determined to be 2000, 5000, 0.5, and 0.1, respectively. For the performance comparison, we also implemented a baseline speaker-clustering system based on hierarchical agglomerative clustering (HAC) in conjunction with the Bayesian information criterion (BIC) to determine the optimal number of clusters [5]. In the agglomeration procedure, the similarities between clusters were computed using the *complete linkage* of the GLR-based inter-utterance similarities. In addition, in using the BIC, the penalty weight was set to one.

Table 1 shows the speaker-clustering results. First, we evaluated the performance of the proposed minimum Rand index clustering (MRIC) by specifying the number of clusters *a priori* as the true number of speakers. This served as an upper bound of the performance that could be achieved by the automatic determination of the speaker population size. We can see from Table 1 that MRIC consistently yielded larger values of purity and smaller values of the Rand index compared with the baseline HAC method. This shows the superiority of global optimization applied in MRIC over pairwise optimization used in HAC.

Next, we examined the speaker-clustering performance of both systems under the practical condition that the true speaker population size is unknown and must be estimated. It can be seen from Table 1 that the number of speakers estimated by MRIC for each excerpt was very close to the true speaker population size. It is also clear that, for the estimated speaker population sizes, MRIC consistently yielded smaller values of the Rand index compared with the baseline system. Some of the values were even smaller than the counterparts obtained by specifying the true speaker population sizes in the baseline system. The results confirm the validity of the proposed method.

5. CONCLUSIONS

We have investigated techniques for clustering speech data, whereby utterances from the same speaker can be grouped into a single cluster. This requirement is formulated as a problem of estimating and minimizing the clustering errors characterized by the Rand index. We represent the Rand index as a function of the inter-utterance similarities and apply the genetic algorithm to determine the index of the cluster where each utterance should be located. As a result, we have demonstrated a noticeable improvement in the speaker-clustering performance, compared to the conventional method based on hierarchical agglomerative clustering and the Bayesian information criterion for the estimation of the speaker population size.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC95-2218-E-027-020, NSC95-2221-E-001-034, and NSC94-2213-E-001-009.

7. REFERENCES

[1] Gish, H., Siu, M. H., and Rohlicek, R. “Segregation of speakers for speech recognition and speaker identification,” *ICASSP’91*.

[2] Jin, H., Kubala, F., and Schwartz, R. “Automatic speaker clustering,” *DARPA Speech Recognition Workshop’97*.

[3] Siegler, M. A., Jain, U., Raj, B. and Stern, R. M. “Automatic segmentation, classification and clustering of broadcast news audio,” *DARPA Speech Recognition Workshop’97*.

[4] Solomonoff, A., Mielke, A., Schmidt, M., and Gish, H. “Clustering speakers by their voices,” *ICASSP’98*.

[5] Chen, S. S. and Gopalakrishnan, P. S. “Clustering via the Bayesian information criterion with applications in speech recognition,” *ICASSP’98*.

[6] Reynolds, D. A., Singer, E., Carson, B. A., O’Leary, G. C., McLaughlin, J. J., and Zissman, M. A. “Blind clustering of speech utterances based on speaker and language characteristics,” *ICSLP’98*.

[7] Johnson, S. E. “Who spoke when?−Automatic segmentation and clustering for determining speaker turns”, *Eurospeech’99*.

[8] Zhou, B., and Hansen, J. H. L. “Unsupervised audio stream segmentation and clustering via the Bayesian information criterion,” *ICSLP’00*.

[9] Moh, Y., Nguyen, P., and Junqua, J. C. “Towards domain independent speaker clustering,” *ICASSP’03*.

[10] Ben, M., Betser, M., Bimbot, F., and Gravier, G. “Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs,” *ICSLP’04*.

[11] Tranter, S. E. “Two-way Cluster Voting to Improve Speaker Diarisation Performance,” *ICASSP’05*.

[12] Zhu, X., Barras, C., Meignier, S., and Gauvain, J. L. “Combining speaker identification and BIC for speaker diarization,” *Interspeech’05*.

[13] Sinha, R., Tranter, S. E., Gales, M. J. F., and Woodland, P. C. “The Cambridge University March 2005 Speaker Diarisation System,” *Interspeech’05*.

[14] Tsai, W. H. and Wang, H. M. “Speaker clustering of unknown utterances based on maximum purity estimation,” *Interspeech’05*.

[15] Schwarz, G. “Estimating the Dimension of a Model,” *The Annals of Statistics* 6:461-464, 1978.

[16] Rand, W. M. “Objective criteria for the evaluation of clustering methods,” *J. Amer. Stat. Assoc.*, 66:846-850, 1971.

[17] Hubert, L., and Arabie, P. “Comparing Partitions,” *Journal of Classification*, 2:193-218, 1985.

[18] Goldberg, D. E. Genetic Algorithm in Search, Optimization and Machine Learning. New York: Addison-Wesley, 1989.

[19] Baker, J. E. “Adaptive selection methods for genetic algorithm,” *International Conference on Genetic Algorithms and Their Applications*, 1985.

[20] http://www.nist.gov/speech/tests/rt/rt2002/

Table 1: Speaker-clustering results.

Excerpt	# Utterances	True # Speakers	# Clusters = True # Speakers				# Clusters = Estimated # Speakers			
			Baseline Method (HAC)		Proposed Method (MRIC)		Baseline Method (HAC-BIC)		Proposed Method (MRIC)	
			Purity	Rand Index	Purity	Rand Index	Estimated # Speakers	Rand Index	Estimated # Speakers	Rand Index
bn02en_1	44	16	0.89	80	0.93	41	8	100	17	56
bn02en_2	29	9	0.94	24	0.95	16	13	52	11	20
bn02en_3	13	6	1.00	0	1.00	0	6	0	6	0
bn02en_4	43	16	0.90	84	0.91	77	18	98	15	80
bn02en_5	26	10	0.72	78	0.76	72	11	80	11	75
bn02en_6	45	14	0.86	66	0.88	58	15	136	15	87