Investigation of Cross-show Speaker Diarization

Qian Yang[1], *Qin Jin*[2], *Tanja Schultz*[12]

[1]Cognitive Systems Lab, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Language Technologies Institute, Carnegie Mellon University, USA
{qian.yang, tanja.schultz}@kit.edu, qjin@cs.cmu.edu

Abstract

The goal of cross-show diarization is to index speech segments of speakers from a set of shows, with the particular challenge that reappearing speakers across shows have to be labeled with the same speaker identity. In this paper, we introduce three cross-show diarization systems namely *Global-BIC-Seg*, *Global-BIC-Cluster*, and *Incremental*. We compared the three systems on a set of 46 English scientific podcast shows. Among the three systems, the Global-BIC-Cluster achieves the best performance with 15.53% and 13.21% cross-show diarization error rate (DER) on the dev and test set, respectively. However, an incremental approach is more practical since data and shows are typically collected over time. By applying T-Norm on our incremental system, we obtain 13.18% and 10.97% relative improvements in terms of cross-show DER on dev and test set. We also investigate the impact of the show processing order on cross-show diarization for the incremental system.

Index Terms: speaker diarization, cross-show diarization, conversational podcast shows

1. Introduction

Given a spoken recording with multiple speakers involved, the goal of speaker diarization is to partition the input audio stream into homogeneous segments according to the speaker identity. Many speaker diarization systems [1] [2] [3] [4] [5] have been proposed so far. However, these systems treat each show independently and do not take into account that in reality, speakers, such as news show hosts and famous politicians, may appear in multiple audio files. In such cases, in addition to producing speaker clusters within each audio file, a diarization system should be able to automatically find the linkages of speakers across the audio files as well. We address this task as *cross-show speaker diarization*. Different from conventional single-show speaker diarization which produces local speaker labelings within each single show, cross-show diarization is to assign the global IDs for the speakers who appear in multiple shows. Cross-show diarization is a crucial task for the French-German project Quaero, and is carried out by LIMSI [6] and our group. Some systems have been proposed to provide similar functionality like cross-show diarization. [7] applies open-set speaker identification on the output of conventional diarization to find the speakers' true IDs across shows. But the speaker models are built in the enrollment stage as prior knowledge. [8] uses linguistic information in the transcription to predict speaker IDs based on a set of pre-defined rules. [9] introduces an online speaker diarization system based on Never-Ending Learning principle.

In this paper, three cross-show diarization systems will be presented to combine both intra-audio, also known as single-show diarization and inter-audio connection, i.e. cross-show diarization. The remaining paper is organized as follows: in Section 2, we describe the single-show diarization system. In Section 3, we introduce the three cross-show diarization systems in detail. In Section 4, we first describe the dataset and evaluation metrics used for cross-show diarization task, and then present the inital experiments and results on the cross-show diarization task. Section 5 concludes the paper and suggests the directions for future work.

2. Single-show Speaker Diarization

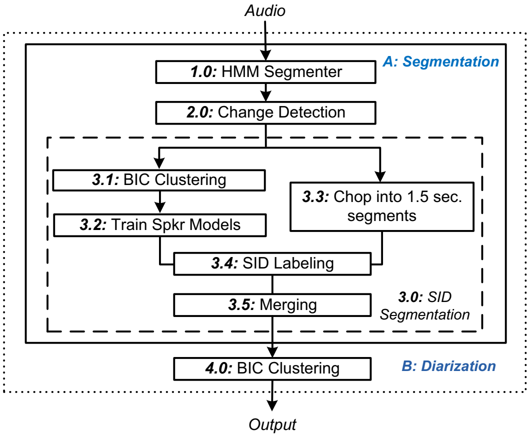

Figure 1: *Architecture of the Single-show Diarization System*

Figure 1 illustrates the system components of our single show diarization system:

- ***1.0: HMM segmenter*** discriminates speech from non-speech segments parts by performing Viterbi decoding on the audio data. The system uses 4 GMMs with 64 Gaussians which are trained on 3 hours manually annotated HUB4 English broadcast news. These GMMs represent 4 acoustic classes: Speech, Noise, Silence and Music. 13 dimensional MFCC plus their first and second derivatives are extracted as features.
- ***2.0: Change detection*** is applied on segments longer than 5 seconds to detect speaker turn changes that are missed by the HMM segmenter. A distance based approach is used [10] by calculating the Generalized Likelihood Ratio (GLR) between two neighboring windows with fixed size and by determining the local maximum of distance to locate the turn changes.

This work was realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

- ***3.0: Speaker identification (SID) segmentation*** refines the segment boundaries generated by the first two steps using speaker identification techniques [11]. The following steps are carried out within this component:
  - *3.1: Bayesian information criterion (BIC) clustering* is applied to the output of change detection based on a Tied Gaussian Mixture Model (TGMM). We perform agglomerative hierarchical clustering using GLR as distance between two clusters and BIC as stopping criteria [10].
  - *3.2: speaker models are trained* on the output of the first-pass BIC clustering by doing maximum a posteriori (MAP) adaptation on a universal background model (UBM) [12].
  - *3.3:* In parallel to *BIC clustering* and *speaker model training*, we *chop the segments* generated by the change detection into 1.5 seconds. The length of the chopped segments can be tuned on the development data. However, in this work, we did not tune this number, simply use the optimal number from our old system.
  - *3.4: SID labeling* assigns each 1.5 second segment a label by using the trained speaker models.
  - *3.5: Segments Merging* concatenates the adjacent speech segments belonging to the same speaker and generates the final SID segmentation results.
- ***4.0: Second-pass BIC clustering*** is applied after the SID segmentation to produce the final diarization results.

3. Cross-show Speaker Diarization

Given a set of shows, a conventional speaker diarization system provides speaker segmentation and clustering. In addition, a cross-show diarization system associates the speakers who reappear across shows by assigning them the same speaker identities. In this paper we propose three strategies, namely: Global-BIC-Seg, Global-BIC-Cluster and Incremental cross-show diarization.

3.1. System 1: Global-BIC-Seg Cross-show Diarization

Obviously, the most straightforward way for cross-show diarization is to concatenate all the segments from all the shows together and run clustering on them. We call this strategy as Global-BIC-Seg system: the segmentation module (see inner box with solid line in Figure 1) is first applied to each single show separately. After the SID segmentation, we concatenate segments from all the single shows into one file and then perform BIC clustering on this file. Figure 2 shows the system architecture of Global-BIC-Seg system.

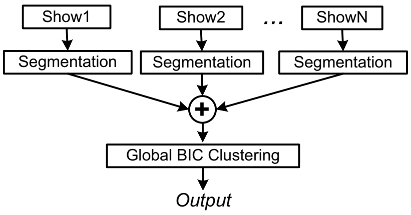

Figure 2: *Global-BIC-Seg Cross-show Diarization System*

The advantage of Global-BIC-Seg is that it looks for the speaker linkage among the shows from a global point of view. However, the memory requirements and computation time for this strategy grow significantly with an increasing number of initial segments. Besides, with a large number of initial segments, the confusability between segments also increases, therefore more errors may occur during the clustering step.

3.2. System 2: Global-BIC-Cluster Cross-show Diarization

As shown in Figure 3, the Global-BIC-Cluster system is similar to Global-BIC-Seg. The only difference is that we apply the complete diarization (see outer box with dotted line in Figure 1) on each single show separately, instead of only segmentation. The global BIC clustering is carried out on the concatenated diarization result of each single show to find the connection of the speakers between the shows.

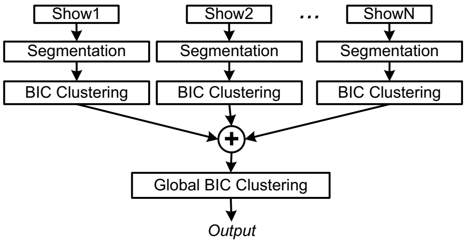

Figure 3: *Global-BIC-Cluster Cross-show Diarization System*

The Global-BIC-Cluster strategy partially solves the computation issues in Global-BIC-Seg strategy. The disadvantage of this approach is that local BIC clustering errors will be carried over without the chance to be corrected during the global clustering.

3.3. System 3: Incremental Cross-show Diarization

The first two approaches have their limitations in reality when the number of shows to be processed is very large and the computing resources are limited. Also in the real world, shows are collected in a multimedia database over time and it is not practical to repeat the global clustering procedure every time when new data is available. Therefore, we seek a new approach which could link the speakers of the former audio archives to those from the new incoming shows without the need to perform any global inter-audio clustering. Our proposed incremental approach satisfies these requirements by applying state-of-the-art speaker recognition. In this iterative system, diarization is first applied to each shows independently. Let the incremental process start with show i. We train the speaker models on the diarization results of show i using UBM-MAP and perform speaker tracking, i.e. open-set SID, on the clusters of show j. If speakers of show i exist in the show j as well, their data will be accumulated and their models will be retrained. For speakers which appear in show j for the first time, new speaker models are trained and added to the database. The retrained old speaker models and new speaker models are used to index the next show. This incremental procedure continues until the last show is processed. Figure 4 illustrates the flow chart of the incremental approach.

4. Experiments

4.1. Data Set and Evaluation Metrics

All the experiments are performed on the Naked Scientists Shows [13], which are English podcast data. This data set was produced within the Quaero project and was used for the internal project evaluation. All speakers have labels assignments

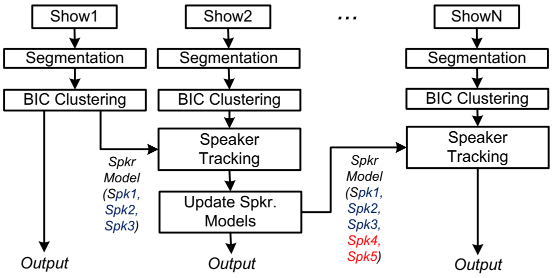

Figure 4: *Incremental Cross-show Diarization System*

which are globally defined across all shows in the manually annotated reference. A 10-minute snippet is extracted from each show for the evaluation purpose. There are 46 shows in total and they are divided into two sets, dev and test sets. The dev set contains 23 shows and 49 speakers, 9 out of the 49 speakers reappear across shows. The test set contains 23 shows and 49 speakers, 10 out of the 49 speakers reappear across shows. We calculate the speaker speaking time entropy $H(Show)$ for each show in dev and test set respectively based on [10]. As $H(Show)$ gets close to zero, it becomes more likely that there is only one dominant speaker in the show. We expect to have better performance when the $H(Show)$ is low. Figure 5 shows the speaker speaking time entropy per show on dev and test set respectively. The mean and standard deviation of $H(Show)$ on the dev set is 1.229 and 0.2312, on the test set is 1.229 and 0.3475. From these numbers, we would expect better performance on the dev set than on the test set for the single show speaker diarization task. We also calcuate the speaking time entropy for cross-show, which is 1.423 and 1.201 on dev and test set respectively. We would expect better performance on the test set than on the dev set for the cross show speaker diarization task. The experiment results presented in later sections generally match these expectations.

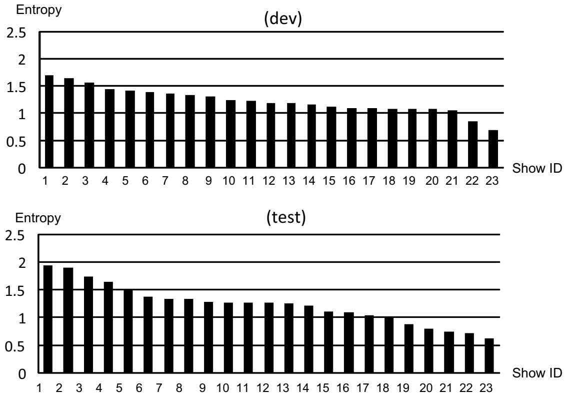

Figure 5: *Speaking Time Entropy of Each Show*

We use the diarization error rate (DER) to evaluate the performance of our system. After finding the optimal mapping between the hypothesized speaker and reference speaker, three types of errors can be calculated. They are *miss* (speaker exists in the reference but not in the hypothesis), *false alarm* (hypothesized speaker does not appear in reference) and *speaker error* (the mapped speaker in the reference and hypothesis is not the same). DER is the sum of these three errors. In order to distinguish from conventional DER, we calculate the cross-show DER by considering all shows as one single show and optimizing the mapping between reference and hypothesis globally. We use the NIST scoring tool [14] to evaluate our system.

4.2. Experiments on Cross-show Diarization

The front-end used for SID segmentation and BIC clustering is composed of 20-dimensional MFCC features. In SID segmentation, the UBM model with 128 Gaussians is trained on the whole test data itself, instead of using other data sources. In the *speaker tracking* and *speaker model training* component of the incremental system, 20-dimensional MFCCs plus their first derivatives are extracted. Feature warping [15] is used to compensate channel effects. In the incremental system, we train two gender-dependent UBMs with 256 Gaussians on the subset of ESTER2 set [16], which contains about 3 hours and 2.5 hours speech for male and female respectively. The speaker models are trained by MAP adaptation on the corresponding UBM.

The performance of the single-show system is shown in Table 1. We have 6.82% and 9.49% in terms of DER on dev and test set respectively. Table 2 compares the performance of our three cross-show diarization systems. The first two Global-BIC approaches are superior to the incremental approach, because they look for connection among audios from a global point of view. The Global-BIC-Cluster achieves the best cross-show DER and has 15.53% and 13.21% DER on dev and test set respectively.

Set	MISS	FA	DER
dev	2.3%	1.1%	6.82%
test	3.5%	0.8%	9.49%

Table 1: *Single-show performance on dev & test set*

Set	System	MISS	FA	Cross-show DER
dev	Global-BIC-Seg	2.3%	1.1%	17.12%
	Global-BIC-Cluster	2.3%	1.1%	**15.53%**
	Incremental	2.3%	1.1%	17.90%
test	Global-BIC-Seg	3.5%	0.8%	14.3%
	Global-BIC-Cluster	3.5%	0.8%	**13.21%**
	Incremental	3.5%	0.8%	20.24%

Table 2: *Cross-show performance on dev & test set*

As the incremental system includes an open-set SID component, we can use state-of-the-art speaker recognition technology to improve the overall performance. T-Norm [17] is applied to the speaker tracking step. 32 females and 82 males are chosen from ESTER2 data [16] as T-Norm speakers. The speech of these speakers are excluded from the UBM training in the previous stage. Table 3 shows that by using T-Norm, the DER reduces by 13.18% and 10.97% relatively on dev and test set respectively.

Set	System	Cross-show DER	*Imprv.*
dev	Incremental	17.90%	-
	Incremental+TNorm	**15.54%**	*13.18%*
test	Incremental	20.24%	-
	Incremental+TNorm	**18.02%**	*10.97%*

Table 3: *Incremental system with & without T-Norm*

4.3. Experiments on the show order

Given a set of shows, the incremental approach is applied show by show. The overall diarization performance will vary based on different show processing order. The following experiments are designed to investigate this impact based on the oracle reference. We use the incremental system with T-Norm to carry out all the show order experiments.

Our first assumption is that, by starting with shows that have fewer errors on the single-show diarization stage, we can achieve better performance on the cross-show diarization. Therefore, we process the shows in the order of single-show DER sorted increasingly contrasted to decreasing order. From Table 4, we can see that the show order indeed has large impact on the overall performance. The difference between the maximal and minimal DER on the three permuations is 3.87% and 3.69% for dev and test set respectively. The results also demonstrate a conflict with our initial assumption. In the dev set, the show order sorted on decreasing DER gives the best results, while DER increasingly shows the worst. In the test set, although DER increasing order is worse than DER decreasing order, both are worse than random order. From this experiment, we learn that the DER is not a good indicator for the optimal show order.

By applying the incremental approach, we expect that the training data of a speaker is accumulated show by show. More reliable and robust model can be obtained over time and we can make more precise decisions. From the data analysis, we know that only some of the speakers appear across shows. If the starting shows contains more such kind of speakers, can we achieve better overall performance? Our second experiment is designed to test this assumption. We sort the shows based on the number of speakers who appear more than once in our data set increasingly and decreasingly and perform the incremental approach on the show orders respectively. The results in Table 4 indicate once more the large impact of the show order on the cross-show DER. By changing the show order, the cross-show DER reduces to 11.32% and 16.81% on dev and test set respectively. The experiment also shows that the number of speakers appear across shows is not a good indicator for optimal show order either.

We also process the shows based on the entropy value calculated in section 4.1. We assume that the shows with higher entropy and more homogeneous speaking time distribution is more interactive and of more challenges. If we start with easier shows, we may obtain more reliable models for detecting speakers in the rest of shows. The results are shown in Table 4. Still we could not find a certain pattern that associates the entropy value with optimal show order.

Set	Sorting criterion	random	decreasing	increasing
dev	DER	15.54%	18.07%	**14.20%**
	#spk reappear	15.54%	13.12%	**11.32%**
	entropy	15.54%	17.13%	**14.62%**
test	DER	**18.02%**	19.67%	21.71%
	#spk reappear	18.02%	19.34%	**16.81%**
	entropy	**18.02%**	19.55%	19.55%

Table 4: *Results of Show Order Experiments*

5. Conclusions and Future work

In this paper, we compared three systems for cross-show diarization and present our initial results on this task. The two Global-BIC approaches are superior to the incremental approach. Global-BIC-Cluster achieves 15.53% and 13.21% on the dev and test set. However, the incremental approach is more realistic in real life applications. By applying T-Norm, we gain 13.18% and 10.97% relatively in terms of cross-show DER for the incremental system. Our experiments also show that for the incremental approach, the show order has a large impact on the overall performance. The results indicate that performance improves if the incremental procedure starts out with easy shows (low DER, low entropy, process speakers early that appear only once) and tackles the hardest shows last. However, our experiments on the show order are oracle experiments since we derived from the reference all information about the amount of speakers across shows. Therefore we will investigate in the future how to find the optimal show order without information given by the reference.

6. References

[1] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proc. of Interspeech*, Lisbon, Portugal, 2005.

[2] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Proc. of MLMI/NIST Workshop*, Edinburgh, UK, July 2005.

[3] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT lincoln laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, November 2004.

[4] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, 2005.

[5] A. G. Friedland, B. O. Vinyals, C. Yan Huang, and D. C. Muller, "Fusing short term and long term features for improved speaker diarization," *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 4077–4080, 2009.

[6] V.-A. Tran, V. B. Le, C. Barras, and L. Lamel, "Comparing multi-stage approaches for cross-show speaker diarization," in *submitted to Interspeech 2011*, Florence, Italy, 2011.

[7] J. Žibert, B. Vesnicer, and F. Mihelič, "A system for speaker detection and tracking in audio broadcast news," *Informatica(Slovenia)*, pp. 51–61, August 2008.

[8] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts," in *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korear, October 2004.

[9] K. Markov and S. Nakamura, "Never-ending learning system for on-line speaker diarization," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, Kyoto, Japan, December 2007.

[10] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, "Speaker segmentation and clustering in meetings," in *ICSLP*, Jeju, South-Korea, 2004.

[11] R. Li, T. Schultz, and Q. Jin, "Improving speaker segmentation via speaker identification and text segmentation," in *Proc. of Interspeech*, Brighton, UK.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[13] "The naked scientists online," http://www.thenakedscientists.com/.

[14] NIST, "The 2001 NIST speaker recognition evaluation," http://www.nist.gov/speech/tests/spk/2001, 2001.

[15] J. Pelecanos and S. Sridharan, "Feauture warping for robust speaker verification," in *Proc. Speaker Odyssey 2001 conference*, June 2001.

[16] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proc. of Interspeech*, Brighton, UK, 2009.

[17] R. Auckenthaler, M. Careya, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, pp. 42–54, 2000.