Comparing Multi-Stage Approaches for Cross-Show Speaker Diarization

Viet-Anh Tran[1]*, Viet Bac Le*[3]*, Claude Barras*[1,2] *and Lori Lamel*[1]

[1]LIMSI-CNRS, Spoken Language Processing Group, BP 133, 91403, Orsay, France
[2]Univ Paris-Sud, 91405, Orsay, France
[3] Vocapia Research, 3 rue Jean Rostand, Parc Orsay Université, 91400 Orsay, France
{tranviet,barras,lamel}@limsi.fr, levb@vocapia.com

Abstract

Acoustic speaker diarization is investigated for situations where a collection of shows from the same source needs to be processed. In this case, the same speaker should receive the same label across all shows. We compare different architectures for cross-show speaker diarization: the obvious concatenation of all shows, a hybrid system combining first a local clustering stage followed by a global clustering stage, and an incremental system which processes the shows in a predefined order and updates the speaker models accordingly. This latter system being best suited to real applicative situations. These three strategies were compared to a baseline single-show system on a set of 46 ten-minutes samples of British English scientific podcasts.
Index Terms: speaker diarization, speaker segmentation and clustering, cross-show diarization

1. Introduction

Automatic speaker diarization can improve the readability of an automatic transcription by structuring the audio stream into speaker turns, and help multimedia indexation [1]. As defined in the context of NIST evaluations on rich transcriptions (RT), the usual diarization task is relative to an individual show and does not make use of *a priori* knowledge of the speaker's voice or even of the number of speakers [2]. Despite being useful for assessing the performance of the underlying technologies, this definition may not fit to all applications. In some situations, it is not possible to process the whole show globally but instead, a decision has to be performed after a limited delay, e.g. for a streaming input; it was measured for contrastive systems in the NIST RT 2009 by the sample processing latency. In the framework of the Quaero program[1], we are considering another situation, where a collection of shows from the same source has to be processed. This is a frequent situation for digital library and multimedia archives and it is likely that in this case some speakers (journalists, actors, frequent guests...) will occur in several shows. It would be convenient that the same speaker is associated with the same identifier across all the shows. In our work, we have addressed this cross-show diarization task and have built upon a standard diarization system different architectures of cross-show diarization. A similar evaluation framework has been explored at the Karlsruhe Institute of Technology [3].

In the following section, we describe the baseline LIMSI speaker diarization system and the proposed cross-show architectures. Then, we present the data set, the system configuration and the experimental results.

This work has been partially financed by OSEO, the French State Agency for Innovation, under the Quaero program.

[1]www.quaero.org

2. Multi-Stage Cross-Show Speaker Diarization

2.1. Baseline diarization system

The system used for our experiments is based on the LIMSI multi-stage speaker diarization system, which was developed for NIST RT-04F evaluation on English broadcast news data [4]. After splitting the signal into acoustically homogeneous segments, the clustering into speaker classes is performed in two steps: a first agglomerative clustering stage uses the BIC criterion with single full-covariance Gaussians [5] and is optimized for providing pure clusters; then, a second clustering stage takes advantage of an increased amount of data per cluster and uses more complex models and a cross-likelihood ratio (CLR) as cluster distance [6].

2.2. Cross-show diarization schemes

To deal with the cross-show condition, an obvious solution is to simulate the single-show condition by concatenating all the shows into a single, large show as presented in the scheme 1 on the left of Figure 1. The diarization system is then run without any modification. However, this architecture is limited by the memory capacity when several hours of signal are concatenated and processed together, and the computation time for the agglomerative clustering grows quadratically with the number of initial segments.

The resource limitation issue of scheme 1 can be partially solved by the hybrid architecture presented on the right of Figure 1: in the scheme 2, the BIC clustering stage is performed for each show independently, followed by a CLR clustering across all shows on the merged outputs of the BIC stages. This architecture is therefore faster than the first scheme because the number of clusters received from the BIC stages is limited.

Although the two schemes presented above can deal with the cross-show condition, they are not realistic from an application point of view, and are presented only in a contrastive perspective. With a large collection of shows, even scheme 2 will exceed memory capacities with our diarization system. Also, all shows may not be available at the time of processing, as is necessary in these two configurations; a more realistic situation is that new shows are recorded and added to the collection over time.

The third architecture simulates an incremental presentation of the shows, where only the information from the shows already processed can help the diarization of the current show. In the scheme 3 presented in Figure 2, the audio segmentation and the BIC clustering are performed independently for each show. For the first show, the system has no prior information and the clustering result from the BIC stage is passed directly to

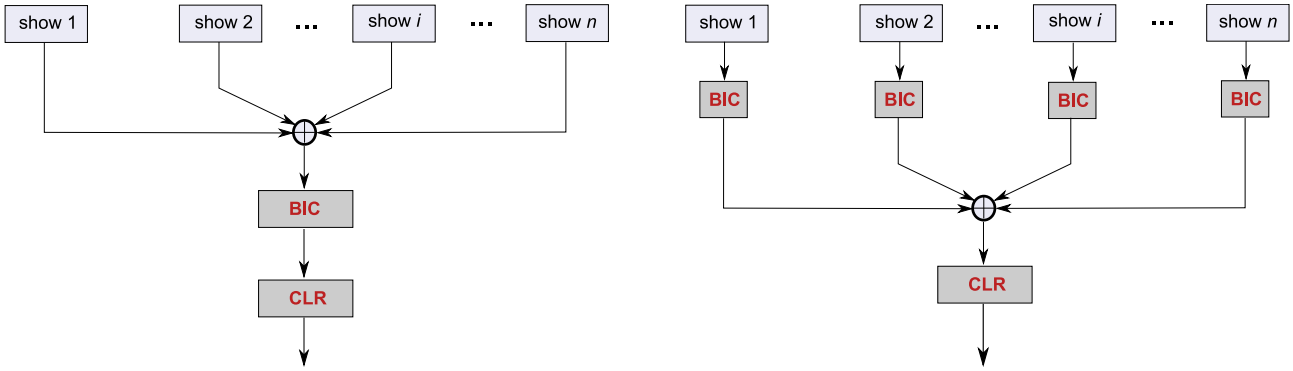

Figure 1: *Cross-show speaker diarization using a global approach by concatenation (scheme 1, to the left) and a hybrid local BIC + global CLR approach (scheme 2, to the right).*

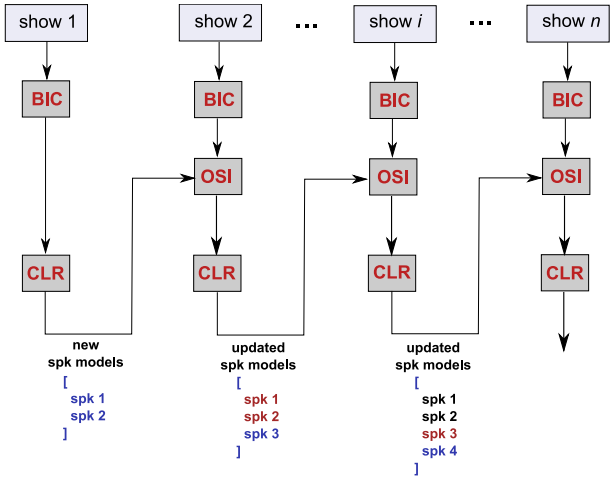

Figure 2: *Cross-show speaker diarization with an incremental presentation of the shows (scheme 3).*

the CLR stage like in the baseline system. Starting with the second show, the existing speaker models extracted from the preceding shows can be used. An intermediate module of open-set speaker identification (OSI) is inserted between the BIC stage and the final CLR clustering stage, identifying the speakers appeared in previous shows. After processing of the current show, the speaker list is updated by adding new speaker models and retraining the existing speaker models on the additional data from the current show plus the data available from the previous shows. This process continues until the last show in the dataset. Alternatively, the OSI stage could be performed after the CLR clustering; but our previous results on speaker tracking using a similar combination of speaker diarization and speaker identification did not show a large difference [7].

3. Experiments

3.1. Data

English radio talk-show "The Naked Scientists" covering scientific topics and available as podcasts[2] were recorded and annotated for the Speaker Diarization task in the Quaero program

[2] http://www.thenakedscientists.com/

Table 1: *Statistics (minimum, maximum, mean and standard deviation) on the number of speakers per show, speaking time per speaker and per segment (in second) in the development and test sets.*

	NS-Dev		NS-Test	
	min - max	μ / σ	min - max	μ / σ
# Speaker	2 - 8	4.7 / 1.4	3 - 9	4.7 / 1.6
Spklen	0.8 - 3964	276 / 629	4.2 - 2368	281 / 398
Seglen	0.2 - 160	15 / 21.3	0.2 - 154	19 / 23

Table 2: *Statistics (minimum, maximum, mean and standard deviation) on the number of shows per speaker, speaking time per speaker and per segment (in second) in the development and test sets, restricted to the recurrent speakers.*

	NS-Dev		NS-Test	
	min - max	μ / σ	min - max	μ / σ
# Show	2 - 19	7.4 / 6	2 - 15	6.9 / 4.3
Spklen	107 - 3964	1059 / 1215	223 - 2368	757 / 672
Seglen	0.2 - 160	13.3 / 21	0.2 - 146	14.8 / 20.9

and were used in these experiments. For each show, a 10 minutes extract was selected as being interactive and was annotated into speaker turns. A set of 46 shows was selected and divided into two sets. The development set (NS-Dev) contains 23 shows for a total of about 4 hours and 49 different speakers, among which 9 appear in several shows. The test set (NS-Test) contain another 23 shows, which has a total duration of 4 hours and 10 out of 49 speakers in this set are present in multiple shows. 8 speakers are found in both sets. Table 1 shows some statistics on the speaker count per show, speaking time per segment and per speaker in the two sets. Table 2 is dedicated to statistics on the "cross-shows" only (or recurrent) speakers.

3.2. System description

Acoustic features are extracted from the speech signal on the 0-8kHz bandwidth for studio speech segments and 0-3.8kHz for telephone speech segments every 10ms using a 30ms window. For the OSI step (in scheme 3) and the CLR clustering stage, 15 cepstral coefficients plus 15 delta coefficients and delta energy, for a total of 31 features are used. Feature warping normalization [8], which reshapes the short-term histogram of the coefficients into a Gaussian distribution is performed using a sliding

window of 3 seconds in order to reduce the effect of the acoustic environment. For each gender and channel condition (studio, telephone) combination, a Multilingual Universal Background Model (UBM) [9] with 128 diagonal Gaussians was trained on a Multilingual Broadcast Corpus which contains broadcast data in Arabic, Chinese, English, French, Italian, Russian and Spanish. Then, for each speaker cluster c_i, a speaker model λ_i is derived by MAP adapting the channel and gender matched *UBM*'s parameters using the acoustic frames X_i found in the cluster c_i. Segments shorter than 3 seconds are discarded for training the speaker models.

For each system, a specific CLR clustering threshold was optimized on the development set and applied to the test set. This was also the case for the OSI identification threshold δ of the incremental system. Furthermore, given that the order of presentation of the shows can have a significant impact on the speaker models, several random permutations were performed on the development and test sets, and the mean μ and standard deviation σ of the resulting error is reported. This allows to assess the variability induced by the incremental process.

3.3. Results

The chosen evaluation metric is the overall Diarization Error Rate (DER) defined by NIST as the fraction of speaker time that is not attributed to the correct speaker, given an optimum one-to-one mapping between the reference speaker labels and the hypothesis speaker labels. When the mapping is show-specific as is usually the case, we refer to this as the average single-show DER over all the shows. Inversely, the cross-show DER results from a speaker mapping that takes into account all shows simultaneously.

In order to assess the complexity of the cross-show task, we scored the references of the development with a local prefix added to the speaker labels in order to simulate a perfect but only local diarization. Table 3 shows that, even when we do not have any local error on each show (0% single-show DER), the cross-show DER is very high, at 52.9%. The baseline system which processes each show independently has a 6.9% single-show DER and a 54.7% cross-show DER which is not very far from the score obtained with the local references. Our aim is to reduce this cross-show DER with our diarization system and try to approach the single-show value.

The results of concatenated and hybrid systems are also presented in Table 3. There is no significant difference between performing a global or a local BIC clustering, probably because the BIC clustering stage only gathers the most acoustically similar segments which are often in the same show, however the local BIC clustering performs 10 times faster than the global one[3]. Compared to the baseline system, the single-show DER increases from 6.9% to 8.2%, a relative degradation of 18.4%. However, as expected, these global architectures in contrast achieve a much lower cross-show error at 15.2%, allowed by the global clustering of the speakers, roughly twice the single-show DER of the baseline system.

For the incremental system, the OSI identification threshold δ was varied from 0.5 to 1.5 and the performance of the system was measured for 25 permutations of the shows found in the development set. Figure 3 shows the cross-show DER as a function of δ , with each point representing the mean of the 25 values and vertical bars their standard deviation. For any fixed δ, the overall DER largely depends on the order of the shows.

[3] 12 min. instead of 130 min. for the BIC clustering of all 23 development shows on an 3GHz Intel Xeon processor

Table 3: *Single-show and cross-show errors on NS-Dev for the local reference, the baseline and the cross-show diarization systems, given as the first quartile, median and third quartile over 25 permutations for the incremental system.*

System	Single-show DER	Cross-show DER
Local reference	0.0	52.9
Baseline system	6.9	54.7
Concatenated	8.2	15.2
Hybrid	8.1	15.4
Incremental	[6.9 6.9 6.9]	[17.8 19.1 20.8]

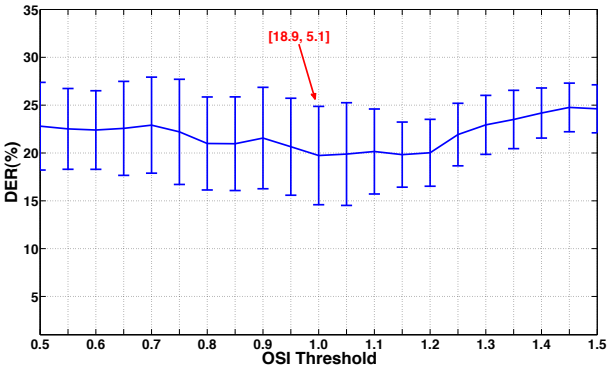

Figure 3: *Mean of cross-show DER ($\pm\sigma$) over 25 permutations on NS-Dev as a function of the identification threshold δ.*

The chosen threshold value $\delta = 1.0$ has the minimum mean of the DER ($\mu = 18.9\%$, $\sigma = 5.1\%$). The last line in Table 3 shows the performance of the incremental system with the chosen δvalue on the 25 permutations of the development set. The result, presented in quartile form, shows that 25% of the permutations have a cross-show DER lower than 17.8%, 50% of these permutations have a cross-show DER lower than 19.1% and 25% have a cross-show DER higher than 20.8%. The mean cross-show DER increases compared to the concatenated and hybrid schemes (almost 24% rel. from 15.4% to 19.1%). This can be explained by the limitation of the prior knowledge to the only shows presented before the current show. On the other hand, the single-show DER decreases back to the one of the baseline system (6.9%) and is very stable across the permutations. Compared to the other systems, the incremental system seems to offer a good balance, providing a cross-show DER slightly higher than the concatenated and hybrid systems without any degradation of the single-show DER compared to the baseline system.

Figure 4 displays the evolution of the cross-show DER with an increased number of shows in the development set for the hybrid and incremental systems for 25 permutations. The mean error tends to be stable in the global scheme while in the incremental scheme, this error monotonically increases when the number of shows augments. Results for the concatenated system are not shown but were very similar to the hybrid system. The hybrid system is insensitive to the order of the shows, this explains why the standard deviation of the error decreases to zero when all the 23 shows of the development set are used.

Using the thresholds optimized on the development set, we observe the same trend for the test set in Table 4. The difficulty

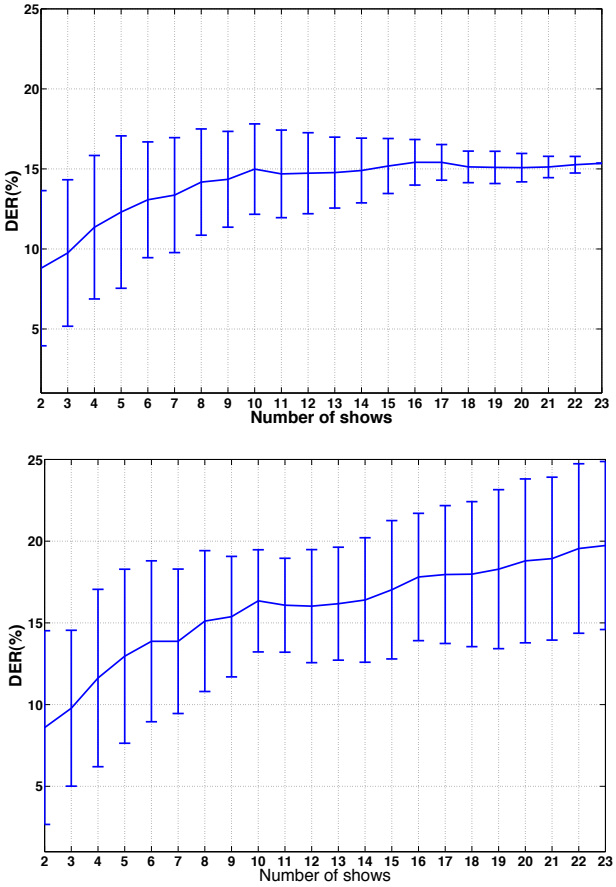

Figure 4: *Mean of cross-show DER ($\pm\sigma$) of the hybrid system (above) and incremental system (below) over 25 permutations on NS-Dev as a function of the cumulated number of shows.*

of the cross-show task on this set measured with the local references seems lower than for the development set (39.2% compared to 52.9% for the cross-show DER and 3.7% compared to 6.9% for the single-show DER), due to a different distribution of the speakers across the shows (c.f Table 1). By performing a global clustering with the concatenated system, the single-show DER increases from 3.7% to 4.1% while the cross-show DER is reduced to 6.1%. The hybrid system has slightly worse performances than the concatenated system, both for single-show and cross-show DER. The incremental system was evaluated over 15 permutations and reduces the single-show DER back to 3.7%, still performing better than the global systems; however, the cross-show DER increases significantly compared to them, with a mean value of 15.5% across the 15 permutations. This degradation is possibly due to a threshold problem.

4. Conclusions

We have considered different architectures for a cross-show speaker diarization system, either global or incremental, and compared them with a baseline diarization system. The hybrid system, performing a first clustering stage locally before a global second clustering, obtains almost the same performance as the trivial concatenation of all shows for the cross-show evaluation while being computationally more efficient but degrades slightly compared to the baseline for the single-show evaluation. The incremental system appears more realistic from an application point of view; it is very similar to the baseline system for the single-show evaluation, but presented better results on the development data than on the test for the cross-show evaluation. Also, the order of presentation of the shows has a significant impact on the performance, as was shown by testing random permutations of the shows.

Table 4: *Single-show and cross-show errors on NS-Test, given as quartiles over 15 permutations for the incremental system.*

System	Single-show DER	Cross-show DER
Local reference	0.0	39.2
Baseline system	3.7	40.9
Concatenated	4.1	6.1
Hybrid	4.8	6.7
Incremental	[3.5 3.7 3.7]	[14.8 15.5 18.5]

5. Acknowledgements

The cross-show diarization task and data set were defined in coordination with KIT (Karlsruhe Institute of Technology). The data collection and annotation were supervised by Vecsys in the context of the Quaero program. The cross-show diarization scoring tool was provided by Ludovic Quintard and Olivier Galibert from the LNE (Laboratoire national de métrologie et d'essais, France).

6. References

[1] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] NIST, "Rich transcription evaluation project," 2002-2009, `http://www.nist.gov/speech/tests/rt/`.

[3] Q. Yang, T. Schultz, and Q. Jin, "Investigation of cross-show speaker diarization," in *Proc. Interspeech 2011*, Florence, Italy, August 2011.

[4] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multi-Stage Speaker Diarization of Broadcast News," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1505–1512, September 2006.

[5] S. Chen and P. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *ICASSP 98*, Seattle, Washington, USA, May 1998.

[6] D. A. Reynolds, E. Singer, B. A. Carlson, G. C. O'Leary, J. J. McLaughlin, and M. A. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proc. ICSLP*, Sydney, Australia, November 1998.

[7] V. B. Le, C. Barras, and M. Ferras, "On the use of GSV-SVM for Speaker Diarization and Tracking," in *Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 146–150.

[8] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey 2001 - The Speaker Recognition Workshop*, Chania, Crete, June 2001, pp. 213–218.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.