Active correction for speaker diarization with human in the loop

Yevhenii Prokopalo[1], *Meysam Shamsi*[1], *Loïc Barrault*[2], *Sylvain Meignier*[1], *Anthony Larcher*[1]

[1] LIUM, Le Mans Université, [2]The University of Sheffield

firstname.lastname@univ-lemans.fr, l.barrault@sheffield.ac.uk

Abstract

State of the art diarization systems now achieve decent performance but those performances are often not good enough to deploy them without any human supervision. In this paper we propose a framework that solicits a human in the loop to correct the clustering by answering simple questions. After defining the nature of the questions, we propose an algorithm to list those questions and two stopping criteria that are necessary to limit the work load on the human in the loop. Experiments performed on the ALLIES dataset show that a limited interaction with a human expert can lead to considerable improvement of up to 36.5% relative diarization error rate (DER) compared to a strong baseline.

Index Terms: Speaker diarization, Active learning, Clustering

1. Introduction

Speaker diarization answers the question ”Who speaks when?” along an audio recording[1, 2]. Being important for audio indexing, it is also a pre-processing step for many speech tasks such as speech recognition, spoken language understanding or speaker recognition. For an audio stream that involves multiple speakers, diarization is usually achieved in two steps: i) a segmentation of the audio stream into segments involving a single acoustic event (speech from one speaker, silence, noise...); ii) a clustering that groups segments along the stream when they belong to the same class of event. A last step could be added to name the resulting speakers but this step is out of the scope of this paper.

Modern diarization systems achieve decent performance depending on the type of data they process [3] but those performances are often not good enough to deploy such systems without any human supervision [4, 5]. Human assisted learning offers a way to achieve better performance by engaging an interaction between the automatic system and a human expert in order to correct or guide the automatic diarization process [6, 7]. Amongst the different modes of human assisted learning, our work focuses on active learning where the automatic system, while processing an incoming stream of audio, is allowed to ask simple questions to the human expert [8].

We propose in this study a system architecture depicted in Figure 1. Given an audio file, the human assisted speaker diarization system (HASDS) first produces an hypothesis based on which a questioning module sends a request to the human expert. The expert's answer is taken into account to correct the hypothesis and possibly adapt the diarization system. This process iterates until reaching a stopping criteria out of those three: (i) the system has no more question (ii) the human expert stops answering (iii) a maximum interaction cost is reached. In this work, we define a binary question that allows a user/system interaction and propose two questioning methods with the associated correction module. The scope of this paper doesn't encompass the system adaptation that will be studied in a future work.

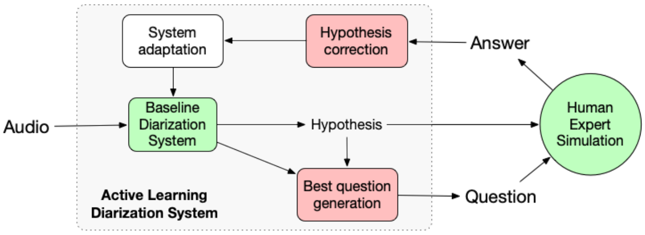


Figure 1: *Life-cycle of a human assisted speaker diarization system.*

Section (2) describes the related works. Section 3 introduces the HASDS, whose evaluation is described and discussed in Section 4. The outcomes and the perspectives of this study are summarized in Section 5.

2. Related work

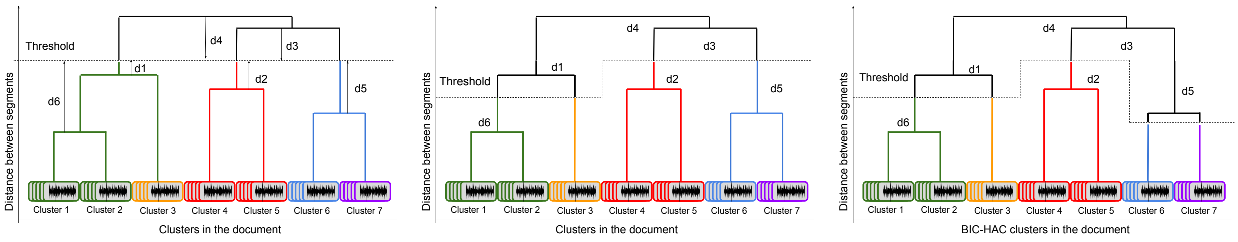
Literature on active learning for speaker diarization is very sparse and existing approaches are complementary to our work more than competitive. In [9], active learning is used to find the initial number of speaker models in a collection of documents. This information is used to perform speaker diarization without involving the human expert anymore. In [10], multi-modal active learning is proposed to process speech segments according to their length to add missing labels, task that is out of the scope of our study. [4] proposed an active learning framework to apply different types of corrections together with metrics to evaluate the cost of human-computer interactions. Unlike previous cited papers, in our work, one interaction with the human expert can lead to correcting a whole cluster of segments (obtained with first of two clustering steps) instead of correcting a single segment only.

In [11], active learning is used to leverage training data and improve a speaker recognition system similar to the one we use for clustering. Active learning based approaches have been developed for other speech processing tasks including speech recognition [12, 13, 14], language recognition[15], speech activity detection [16] or speech emotion recognition [17] but are not directly applicable to speaker diarization.

Active learning literature for clustering is much wider [18, 19] but mostly focuses on K-mean clustering [20, 21] or spectral clustering [22, 23]. Hierarchical agglomerative clustering, that is used in many speaker diarization systems including our baseline, has also been studied for semi-supervised clustering [24, 25].Those studies propose to use predefined constraints to modify the clustering tree. In our work, instead of modifying the dendrogram, we propose a dynamic approach to update the threshold used to merge and split the clusters.

This work has been funded by the CHIST-ERA project ALLIES (ARN-17-CHR2-0004-01)

10.21437/IberSPEECH.2021-55



(a) *Initial HAC*

(b) *Stopping criteria : 2c criteria*

(c) *Stopping criteria : All*

Figure 2: *The change of HAC dendrogram after interaction with human expert*

Regarding evaluation of the active learning process, multiple approaches have been proposed [26, 4]. In [4], systems are evaluated by DER together with an estimate of the human work load to correct the hypotheses. We make the choice to use a penalized version of DER described in our previous work [27]. The human correction effort is computed to be in the same unit and thus added to the DER in order to provide a single performance estimator reflecting both the final performance and the cost of interacting with human.

3. Human Assisted Diarization System

The proposed Human Assisted Speaker Diarization System (HASDS) is depicted in Figure 1 and includes 4 modules. A fully automatic baseline diarization system, a question generation module, a correction module and an adaptation module. This section describes the baseline diarization system that is considered fixed in this work before introducing the proposed question generation and correction modules. The adaptation module is out of the scope of this work and will be considered in future work.

Given an audio stream, diarization consists of producing a segmentation hypothesis, i.e. a list of segments and speaker IDs with each segment allocated to a single speaker ID (segments might overlap). Diarization errors can be due to errors in the segment borders or in a wrong label allocation. The former error being the most harmful in terms of performance [4], this work only focuses on correcting labeling errors.

3.1. Baseline automatic diarization module

To provide a fair study, the chosen baseline system is the best automatic diarization system available at LIUM for the given task. This system performs the diarization in two steps: a segmentation process, that splits the audio stream into (possibly overlapping) segments and a clustering process, that groups the segments into clusters: one cluster per speaker. Since this study only focuses on labeling errors, the segmentation step is considered perfect, i.e., border of the speech segments are taken from the reference. The clustering is performed in four steps: (i) a first hierarchical agglomerative clustering (HAC) is performed on vectors of 13 MFCC using the BIC criteria [28]; (ii) a Viterbi decoding is then used to smooth the segment borders along the audio stream; (iii) x-vectors are extracted from each segment and averaged to provide a single x-vector per BIC-HAC cluster; (iv) a second (final) HAC clustering is done by using x-vectors. The distance matrix used for this clustering is computed using a PLDA scoring [29]. X-vectors are extracted using the SincNet extractor described in [30] and their dimension is 100. A simplified-PLDA [29] is trained using an EigenVoice matrix of rank 100.

Applying two consecutive clustering makes the application of active correction more complex but removing one of the steps degrade the performance of the baseline system, thus we chose to keep the two consecutive clustering but to only apply active correction to the second clustering step while considering the BIC-HAC clusters as frozen. This choice has the advantage to reduce the correction to a simpler HAC-tree correction process. Another drawback is that errors from the BIC-HAC clustering will not be corrected and purity of those clusters is thus very important.

3.2. Question generation module

Assuming that correcting the clustering provides higher gains than segmentation and considering an HAC clustering algorithm, we propose to limit the human/system interaction to a simple binary question that can be asked for each node of the HAC dendrogram (Figure 2a). HAC clustering is done with no prior on the number of clusters and the threshold is empirically determined on a development set. Once this threshold is set, it separates the dendrogram in two parts (above and below the threshold). From this point, the same question can be asked to the human expert for each node of the dendrogram: *"Do the two branches of the node belong to the same speaker?"* . A "yes" answer from the human expert requires either to join the two branches of a node above the threshold (merging operation) or to leave as it is the branches of a node below (no splitting required). In case of a "no" answer, a node above the threshold would not be modified (no merging required) and the two branches of a node below the threshold would be separated (split operation).

One must now determine which node to ask about and when to stop asking. To do so, we rely on the distance between the threshold and the nodes, referred to as *delta* to differentiate with distance between x-vectors. Examples of those *delta* are labeled $d1$ to $d6$ on Figure 2a. Nodes are ranked in increasing order according to their absolute *delta* value. We propose to ask questions about the nodes in this order, and consider two different stopping criteria. First, a **Two confirmation criteria (*2c criteria*)** illustrated in Figure 2b, in which we assume that if a node above the threshold is confirmed by the human expert to be separated ("no" answer) then other nodes above it, with higher *deltas* will not be investigated. Similarly, if one node below the threshold is confirmed by the human expert to be merged, the other nodes, lower in the dendrogram, will not be investigated. Second, a **criteria exploring the tree per branch (*All*)** that is illustrated in Figure 2c. Nodes are still considered according to their ranked *delta* but the dendrogram is explored in more details. If the human expert confirms a merge on a node (”yes” answer), the lower nodes in the two branches will not be investigated for splitting. If the human experts confirms a separation on a node (”no” answer), the upper nodes will not be investigated for grouping (but can be investigated for splitting). The *2c criteria* relies on a high confidence in the *delta* ranking (the estimation of the distance between x-vectors) and strongly limits the number of questions, while the *All* criteria leads to more questions and thus a finer correction of the dendrogram.

To facilitate the work of the user answering the question, we consider that the HASDS proposes two audio segments (*samples*), for the user to listen to; one for each branch of the current node. Each branch, can link several segments, even for nodes located at the very bottom of the tree (remember that, due to the sequential HAC clustering process, leaves of the dendrogram are clusters linked by the BIC-HAC clustering). The system must select the two most representative or informative *samples*. We investigate 5 *sample* selection methods:

Longest selects the longest segment from each cluster. It assumes that x-vectors from those segments are more robust and that the gain provided by the correction would lead to higher improvement of DER.

Cluster center selects the closest segment to cluster center assuming this is the best representation of this cluster. The center is selected according to the euclidean distance between segments’ x-vectors.

Max / Min selects the couple of segments, one from each branch, with the lowest (max) or highest (min) similarity in terms of PLDA score (distance).

Random as a contrastive criteria, a random segment is selected from each cluster (statistics from this method are consolidated by repeating experiments 20 times).

3.3. User simulation and correction module

The correction module simply remembers the successive corrections provided by the human expert. The human expert is simulated for reproducibility and makes use of the ground truth reference to provide a correct answer to each question. To establish a lower bound, we also consider an **ideal** correction method. When a node has been chosen to be investigated, the optimal correction (merging or splitting) is found by looking at the ground truth (reference) to maximize the gain in terms of DER.

4. Experiment: protocol and results

Experiments are performed on the ALLIES dataset[1], an extension of previously existing corpora [31, 32, 33], that includes a collection of 1,008 French TV and Radio shows partitioned in three non-overlapping parts whose statistics are provided in Table 1. The performances are reported as weighed diarization error rate (DER) [34], averaged over all documents of the collection according to their annotated duration. Penalized DER [27] described in Equation 1 is reported as a unique performance indicator including both final DER and human interaction cost.

$$DER_{pen} = \frac{T_{miss} + T_{false} + T_{confusion} + N \cdot t_{pen}}{T_{total}} \quad (1)$$

T_{miss}, T_{false} and $T_{confusion}$ are respectively the duration of missed speech, non-speech considered as speech and wrongly classified speech. T_{total} is the total speech duration in the document. N is the number of corrections applied to the document and t_{pen} is the estimated time a human spend answering one question (see [27]). The quality of the questioning module is estimated by computing the correction/question ratio (CQR) between the number of corrections (question that leads to a modification of the clustering) and the number of questions asked to the human expert. The training set is used to train the x-vector extractor and the PLDA model while the Dev is used to set the clustering threshold. performance are reported on the Eval set.

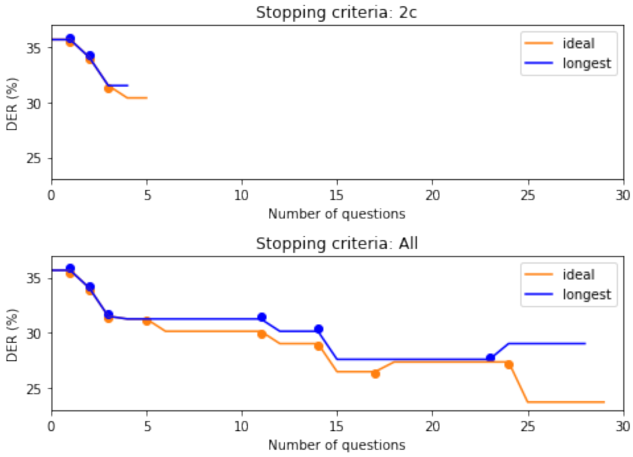


Figure 3: *Tracking of the DER corresponding to a single show file (with duration of 1 hour and 11 minutes) by applying question-correction with different methods. Points in each question indicate that it has resulted in a correction.*

Table 1: *ALLIES dataset description, all duration are given in hh:mm:ss, speakers are considered recurrent when they appear in 3 episodes or more across the dataset.*

Partition	Total Duration	Annotated Duration	#speaker	#recurrent speaker	#shows
Training	223:37:17	175:22:04	3,680	355	475
Dev	105:51:06	33:54:15	983	105	200
Eval	282:36:16	118:53:34	1,720	220	333

4.1. Experiments

Figure 3 illustrates the evolution of DER for an audio file for both stopping criteria. As expected, *All* leads to more questions and achieve a better final DER (lower) than *2c criteria*. This example shows the necessity of taking into account the cost of human interaction to fairly compare HASDS systems.

A first experiment is performed with **ideal** correction to compare the benefit of merging or splitting clusters. Results in Table 2 reveal that for both stopping criteria, DER reduces more when splitting clusters than merging them, but also that the CQR is higher when splitting. Both kind of correction can lead to conflicts. For instance, a node could be merged first while one of its child nodes would requires to be split due to non-purity of the clusters (even with **ideal** correction). For this reason and considering the higher benefits of splitting compared to merging, we chose to prioritize splitting to merging. So if a node has been split into two clusters, its parent nodes will not

[1] Database and protocols will be made publicly available after the ALLIES 2021 challenge

be investigated for merging. It helps to avoid investigating the nodes that will not be used for correction.

Table 2: *Performance of the HASDS using* ***ideal*** *correction when applying only one type of correction. Second column is the average number of questions per hour and last column reflects is the quality of interaction (CQR).*

Stopping criteria		DER	Avg. #Question / hour	CQR
2c criteria	Merging	15.58	3.75	16.20
	Splitting	11.36	5.94	61.02
All	Merging	15.02	19.86	6.50
	Splitting	10.72	9.49	45.65

A second experiment is performed to compare the 5 *sample* selection methods for both stopping criteria. Results are presented in Table 3. As expected, *ideal* correction provides the lowest DER for both stopping criteria and all 5 proposed methods perform at least as well as the contrastive random selection process. It appears that the longest segments or cluster centers are the most representative from their cluster and that **Min/Max** provide the smaller improvement probably due to the similarity between those criteria and the clustering criteria used by the HAC algorithm. The 5 selection methods are comparable in terms of number of questions asked per hour of audio and CQR. This is visible on the penalized DER which preserves the conclusions drawn by observing the DER.

Comparing the two stopping criteria, we observe that **Longest** and **Cluster center** selection method using the *All* criteria achieves better performance than the *2c criteria* but both criteria achieve similar performance for **Min/Max**. Penalized DER shows that although the *All* criteria achieves lower DER than *2c criteria*, the cost of human interaction (for a t_{pen} empirically set to 6s) for the *All* criteria is much higher and that *2c criteria* might be a better compromise to reduce human interaction. The proposed approach considering *2c criteria* and a selection of the longest segment leads to a reduction of DER from 16.46% (without human in loop) to 11.18% while asking less than 10 questions to the human expert perhour of speech processed. However, we found that only 43.56% of the questions asked lead to a correction (i.e., in 56.44% of the cases, the human validates the decision of the automatic system) which will be investigated in future work.

Table 3: *DER improvement using different stopping criteria and segment selection methods*

Method	Stopping criteria	DER	Avg. #Q / h	CQR	DER_{pen}
Baseline	-	16.46	-	-	-
Ideal	*2c criteria*	10.57	9.69	43.68%	12.18
	All	9.65	28.21	19.86%	14.35
Random (20 times)	*2c criteria*	12.77±0.13	9.66	44.15%	14.38
	All	12.99±0.16	28.26	22.01%	17.75
Longest	***2c criteria***	**11.18**	9.66	43.56%	**12.79**
	All	**10.45**	28.14	19.97%	15.14
Cluster center	*2c criteria*	11.28	9.64	42.66%	12.89
	All	10.52	28.10	20.17%	15.21
Max	*2c criteria*	12.52	9.69	43.98%	14.13
	All	12.99	28.12	21.40%	17.68
Min	*2c criteria*	12.74	9.61	43.58%	14.35
	All	12.80	28.14	22.33%	17.49

4.2. Analysis

In order to further improve our approach, we analysed the correlation between the benefit of human active correction (in terms of DER reduction or number of questions asked) and the characteristics of the processed audio files (number of speakers, duration of the file...) based on Pearson correlation coefficient, no strong correlation has been found (all less than 0.4).

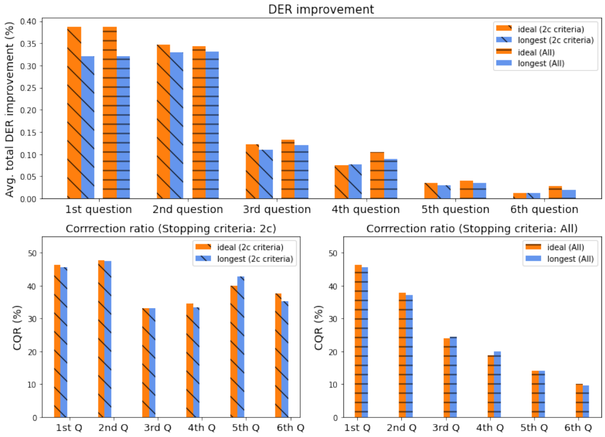


Figure 4: *The performance of questioning based on DER improvement and ratio of correction number to question number.*

We then evaluated the usefulness of successive questions in an ordinal way for each audio file. The total DER improvement and the ratio: number of corrections over the number of questions asked (CQR) are compared base on the question order on Figure 4. For both stopping criteria, the first questions lead to larger DER reductions (upper figure). Interestingly, we can see that the questions asked when using the *2c criteria* have a similar CQR ratio, meaning that successive questions keep contributing to the DER reduction (bottom left). On the other hand, we observe that the CQR reduces for the *All* criteria, meaning that the system tends to ask less useful questions to the expert.

5. Conclusion

The benefit of human active correction for speaker diarization has been investigated. This preliminary study has focused on an active correction of HAC clustering errors. Starting from a strong automatic baseline, we proposed two criteria to ask questions to a human expert. 5 methods to select samples for auditory tests have been proposed and evaluated using a large and challenging dataset that will be publicly released.

Performance of our human assisted speaker diarization system have been evaluated by using a penalized DER proposed in [27] and shows that it can decrease by up to 22,29% relative when applying active correction with the *2c criteria*. This leads to a reduction of 32,07% relative without taking into account the cost of human interaction. The second proposed stopping criteria (*All*) can achieve a relative reduction of 36,51% of DER but requires a higher and less efficient human effort.

This preliminary study is very promising and opens large avenues for future studies. More analyses are ongoing to understand and refine the stopping criteria depending on the nature of the processed audio file and its difficulty for diarization systems. Current studies are conducted to improve the question generation module by estimating the quality of the question before soliciting the human expert. We are also developing the adaptation process in order to improve the automatic system using the information provided by the human expert.

A limitation of this work comes from the restriction to HAC clustering when many works in the literature have been exploring active learning for other clustering algorithms. So far, we have only considered diarization of isolated files but it is very likely that a higher benefit can be expected from active learning when applied to the diarization of a collection of audio files.

6. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Multistage speaker diarization of broadcast news," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1505–1512, 2006.

[3] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[4] P.-A. Broux, D. Doukhan, S. Petitrenaud, S. Meignier, and J. Carrive, "Computer-assisted speaker diarization: How to evaluate human corrections," in *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, 2018.

[5] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The second dihard diarization challenge: Dataset, task, and baselines," *Proc. Interspeech 2019*, pp. 978–982, 2019.

[6] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *Ai Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[7] L. Jiang, S. Liu, and C. Chen, "Recent research advances on interactive machine learning," *Journal of Visualization*, vol. 22, no. 2, pp. 401–417, 2019.

[8] M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 2, pp. 1–21, 2011.

[9] C. Yu and J. H. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.

[10] B. Mateusz, J. Poignant, L. Besacier, and G. Quénot, "Active selection with label propagation for minimizing human effort in speaker annotation of tv shows," in *Workshop on Speech, Language and Audio in Multimedia (SLAM 2014)*.

[11] S. H. Shum, N. Dehak, and J. R. Glass, "Limited labels for unlimited data: Active learning for speaker recognition," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[12] G. Riccardi and D. Hakkani-Tur, "Active learning: Theory and applications to automatic speech recognition," *IEEE transactions on speech and audio processing*, vol. 13, no. 4, pp. 504–511, 2005.

[13] H. Jiaji, C. Rewon, R. Vinay, L. Hairong, S. Sanjeev, and C. Adam, "Active learning for speech recognition: The power of gradients," in *The 30th Conference on Neural Information Processing Systems, NIPS. Barcelona, Spain*, 2016, pp. 1–5.

[14] J. Bang, H. Kim, Y. Yoo, and J.-W. Ha, "Efficient active learning for automatic speech recognition via augmented consistency regularization," *arXiv preprint arXiv:2006.11021*, 2020.

[15] E. Yilmaz, M. McLaren, H. van den Heuvel, and D. A. van Leeuwen, "Language diarization for semi-supervised bilingual acoustic model training," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 91–96.

[16] D. G. Karakos, S. Novotney, L. Z. 0002, and R. M. Schwartz, "Model adaptation and active learning in the bbn speech activity detection system for the darpa rats program." in *INTERSPEECH*, 2016, pp. 3678–3682.

[17] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[18] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[19] M. Wang, F. Min, Z.-H. Zhang, and Y.-X. Wu, "Active learning through density clustering," *Expert systems with applications*, vol. 85, pp. 305–317, 2017.

[20] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, 2004, pp. 333–344.

[21] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[22] Q. Xu, K. L. Wagstaff *et al.*, "Active constrained clustering by examining spectral eigenvectors," in *International Conference on Discovery Science*. Springer, 2005, pp. 294–307.

[23] X. Wang and I. Davidson, "Active spectral clustering," in *2010 IEEE International Conference on Data Mining*. IEEE, 2010, pp. 561–568.

[24] S. Miyamoto and A. Terami, "Semi-supervised agglomerative hierarchical clustering algorithms with pairwise constraints," in *International Conference on Fuzzy Systems*. IEEE, 2010, pp. 1–6.

[25] I. Davidson and S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2005, pp. 59–70.

[26] E. Geoffrois, "Evaluating interactive system adaptation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 256–260.

[27] Y. Prokopalo, S. Meignier, O. Galibert, L. Barrault, and A. Larcher, "Evaluation of lifelong learning systems," in *International Conference on Language Resources and Evaluation*, 2020.

[28] P.-A. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, and S. Meignier, "S4d: Speaker diarization toolkit in python," 2018.

[29] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7673–7677.

[30] A. Larcher, A. Mehrish, M. Tahon, S. Meignier, J. Carrive, D. Doukhan, O. Galibert, and N. Evans, "Speaker embeddings for diarization of broadcast data in the allies challenge," in *submitted to 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[31] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The repere corpus: a multimodal corpus for person recognition." in *LREC*, 2012, pp. 1102–1107.

[32] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ester phase ii evaluation campaign for the rich transcription of french broadcast news," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[33] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *International Conference on Language Resources, Evaluation and Corpora*, 2012.

[34] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech." in *INTERSPEECH*, 2013, pp. 1131–1134.