Towards Lifelong Human Assisted Speaker Diarization

Meysam Shamsi*, Anthony Larcher, Loic Barrault, Sylvain Meignier, Yevheni Prokopalo, Marie Tahon, Ambuj Mehrish, Simon Petitrenaud

LIUM, Le Mans Université, Avenue Olivier Messiaen, 72085 LE MANS CEDEX 9, France

Olivier Galibert

LNE

Samuel Gaist, André Anjos, Sebastien Marcel

Idiap Research Institute, rue Marconi 19, 1920, Martigny, Switzerland

Marta R. Costa-jussà

TALP Research Center - Universitat Politècnica de Catalunya, Campus Nord C/Jordi Girona 31, 08034 Barcelona, Spain

Abstract

This paper introduces the resources necessary to develop and evaluate human assisted lifelong learning speaker diarization systems. It describes the ALLIES corpus and associated protocols, especially designed for diarization of a collection audio recordings across time. This dataset is compared to existing corpora and the performances of three baseline systems, based on x-vectors, i-vectors and VBxHMM, are reported for reference. Those systems are then extended to include an active correction process that efficiently guides a human annotator to improve the automatically generated hypotheses. An open-source simulated human expert is provided to ensure reproducibility of the human assisted correction process and its fair evaluation. An exhaustive evaluation, of the human assisted correction shows the high potential of this approach. The ALLIES corpus, a baseline system including the active correction module and all evaluation tools are made freely available to the scientific community.

Keywords: Speaker diarization, Lifelong learning, human assisted learning, evaluation

1. Introduction

Speaker diarization is the task of answering the question "Who speaks when?" along an audio recording [1]. The result of speaker diarization is essential for indexing and analysing various types of audio data, such as audio/video broadcasts, conference speeches, lectures, court proceedings or business meetings. It is also required as a pre-processing step to guaranty optimal performance for tasks like speech recognition, spoken language understanding or speaker recognition [2, 3, 4, 5, 6, 7].

Given an audio stream, speaker diarization systems address the segmentation and clustering problem in two separated stages [1, 8, 9] or in an integrated stage [10, 11]. The segmentation of the audio stream into homogeneous segments (overlapping or not) classically involves voice activity detection (VAD) [12, 13] and speaker change detection [14]. Upon this segmentation, the speech segments are cluster into homogeneous speaker groups.

*This work has been funded by the CHIST-ERA project ALLIES (ARN-17-CHR2-0004-01), the French ANR Extensor (ANR-19-CE23-0001-01) and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666, the Agency is not responsible for this results or use that may be made of the information. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011012565)

*Corresponding author

Email address: meysam.shamsi@univ-lemans.fr (Meysam Shamsi)

While a majority of speaker diarization systems are based on those two stages [1, 8, 9] some also involve an additional re-segmentation step [15] and recent approaches are attempting at solving the task by using an End-to-End neural architecture [16, 17].

This work focuses on speaker diarization to produce speaker annotations on large audio corpora collected across years. This task differs from the classical speaker diarization in three main aspects. First, the audio stream is collected in a discontinuous manner, for instance by recording daily TV or radio shows. This discontinuity, that is actually inherent to TV and radio broadcasts or web contents, severely affects the performance of automatic systems by introducing strong acoustic mismatches between shows [18, 19, 20, 21]. Moreover, the variability across shows involves changes of speakers or topics. Second, the collection of data across years implies aging of the speakers and evolution of the recording channel (new compression codecs, various quality, style changes...). This generates a dataset shift [22, 23, 24] that data-driven automatic systems have difficulties to compensate [25, 26, 21]. The third difference with classical speaker diarization is due to the size of the audio data to process. Archivers and content managers are collecting thousands of hours of audio a day, every day. As the collection of audio shows increases endlessly at a very fast paste, it is thus necessary to process the stream of data in an efficient manner.

In short, an automatic speaker diarization system processing such a data collection has to deal with a discontinuous stream of audio shows presenting a high cross-show variability and a constantly fast increasing volume affected by a temporal dataset shift. We assume in this work that the quantity and complexity of the task does not allow the speaker diarization system to re-process already processed data and require an incremental processing. The proposed work flow consists thus of processing each audio show on arrival, by performing within show diarization and then to link the speakers from this show with previously seen speakers in an incremental cross-show diarization process. The performance of such a system would strongly depend on two factors: the quality of the within-show diarization and the robustness of the system to cross-show variability.
Across time, the quality of the within-show diarization might degrade due to the dataset shift that will quickly make the data-driven models of the automatic systems obsolete. It is therefore essential

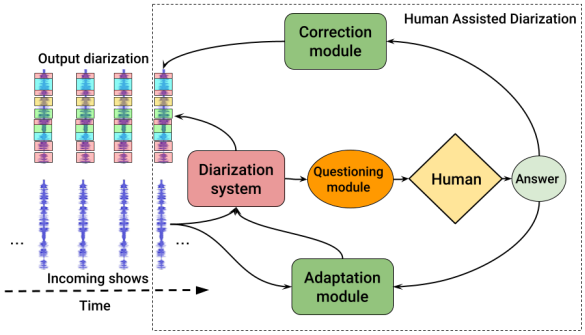

Figure 1: Proposed structure of a Human Assisted Lifelong Learning Speaker Diarization System. While processing an incoming audio file, the automatic diarization system generates question for a human expert, whose answers are used to correct the output of the automatic system and to adapt the system across time.

to adapt those models to cope with the incoming stream of data and we propose to address this challenge via lifelong learning. Lifelong learning is the process of continuously learn or adapt while performing a sequence of tasks so that the knowledge leveraged in the past will help performing the future tasks [27, 28]; in our case we consider that continuously learning on incoming data will contribute to the robustness of the system to the dataset shift.
Lifelong learning might not be enough to compensate for the abrupt discontinuities across audio shows and the quality of the overall process rests upon the performance of within show diarization. For this reason, we propose to involve a human-in-the-loop to correct the within-show diarization and address new events appearing across time. Amongst possible scenarios, the human-in-the-loop can initiate interactions with the system by providing feedback or, in an active learning scenario [29], the system can itself initiate this interaction by asking questions to the human.

Figure 1 depicts our vision of a Human-Assisted Lifelong Learning Speaker Diarization System. This system embeds an automatic diarization system which incrementally processes the incoming data and asks questions to a human expert. The answers provided by the human expert are immediately exploited by a correction module to improve the system's output while an adaptation module leverage information from the incoming data and the human answers to sustain the performance of the automatic system across time. Our motivation in this work is to prepare the ground for research on human assisted lifelong learning speaker diariza-

tion.

As a first major contribution, we provide the necessary framework and materials to develop and evaluate human assisted lifelong learning speaker diarization systems (HALSDS). This includes the necessary data, protocols and metrics. In a second major contribution, we mark a step forward the development of a new protocol for HALSDS by introducing three human assisted speaker diarization systems. Faced with the magnitude of the task and the lack of previous work in the literature, we limit the scope of this paper to human assisted within-show diarization systems that only employ a correction module (no adaptation). Moreover, we only consider, clustering correction without modifying segmentation borders as errors due to clustering mistakes are often more harmful than segmentation errors in terms of performance [30]. The article is organized as follows. Section 2 gives a review of related works and resources. Section 3 describes in details the ALLIES corpus and its associated protocols that will be made public and freely available for scientific purposes. Section 4 describes three automatic speaker diarization systems and their augmentation with an active-correction module, as depicted on figure 1. Section 5 describes classical metrics used for speaker diarization evaluation and introduces new ones especially developed to evaluate the specific features of human-assisted lifelong learning speaker diarization. Results and analyses are reported in Section 6 while Section 7 proposes a deeper discussion on the impact of the data on the different metrics proposed. The outcomes and perspectives of this study are eventually summarized in Section 8.

2. Related works

In this section, we review the wide scope of elements that are necessary to develop and evaluate human assisted speaker diarization across time. The following sections are providing a brief overview of existing speaker diarization systems; their use for diarization of collection across time; the previous attempts to involve a human in the process; the available corpora and eventually the existing protocols and metrics.

2.1. A brief overview of speaker diarization systems

It is possible to consider that all speaker diarization system are taking as input a sequence of audio samples or segments and produce as output a stream of labels which naturally leads to describe those systems in two stages: segmentation and clustering.

Segmentation aims at producing homogeneous audio segments that can be represented with a compact, robust, representation to be later clustered. Intrinsically, shorter segments are more likely to be homogeneous. However, longer segments provide more robust and discriminative representations to improve the classification performance. This is the reason why segmentation historically involves speech activity detection and speaker change detection to produce audio segments as long as possible. Speech activity detection is a well studied task [31, 32, 33] that has achieved recent improvement with neural approaches [34, 35, 36]. In our study, we choose a standard approach based on the work from [37, 14].
Speaker change detection can be achieved by using statistical models such as in [38, 39, 40] or neural approaches as in [41, 14] as our work focuses on clustering we used a well-known statistical approach that has shown robust performance in the past on the type of data included in the ALLIES corpus [42].

Different clustering methods can be implemented after the segmentation step. Some methods like K-Means [43, 44] require a preliminary estimation of the number of speakers, while other like spectral clustering [45] or Hierarchical Agglomerative Clustering (HAC) [46, 47, 48, 49] can automatically estimate this number. Based on the work from [50], it appears that HAC performs slightly better than spectral clustering when applied after a classical segmentation process. This is the reason why two of our baseline systems use an HAC clustering as described in section 4.1. Additionally, our choice of HAC is motivated by its convenience to estimate the confidence of a clustering choice by measuring the difference between clustering threshold and distance between clusters.

Due to recent improvements in acoustic modeling, since the rise of i-vectors [51] and then neural-based embeddings [52, 45, 53, 54], it is now possible to obtain very robust representation of audio segments as short as a few seconds. Recent works such as [8] exploit the accuracy of x-vector representations to reduce the speaker diarization process to the clustering of x-vectors extracted on a sliding window of 3 seconds. The clustering is efficiently performed by first estimating the number of speaker with a spectral clustering and then clustering them

with a simple k-means algorithm. This work opens avenues in human assisted learning speaker diarization but the involvement of a human-in-the-loop within the k-means algorithm is not obvious. This is the reason why our first attempt makes use of HAC clustering. In [10, 11], the authors present another approach that clusters x-vectors extracted on a sliding window. After automatically estimating the number of speakers in a file, a Variational Bayesian HMM (VB-HMM) is used to cluster the x-vectors by applying an iterative re-segmentation process. This method has shown excellent performance in the latest benchmarking evaluations [55] and is used as one of the systems for our study. Its description is given in section 4.1.3.

End-to-End approaches also rely on the fact that neural networks can directly produce their own local representation of the audio signal starting from very short segments that can be reduced to tens of milliseconds when processing MFCC or filterbanks [16, 56, 17] or be as short as the sampling period. if feeding the network with raw speech signal [57, 58, 59]. End-to-end systems have shown good performance in the latest benchmarking evaluations [55] and offer large avenues for improvement. Since the purpose of this work is to introduce human assisted learning, we decide to first develop human assisted systems starting from two types of approach: one based on x-vectors and HAC clustering and the other based on VB-HMM that has shown to be the best stand-alone system in recent evaluations.

2.2. Diarization across time

In the literature, a few works have addressed cross-show diarization and its challenges [21, 60, 61, 62] but to our knowledge, none of them has released a complete protocol on publicly available data.

2.3. Corpora for speaker diarization

Most of the existing corpora for speaker diarization do not include cross-show speaker IDs and are thus not usable for cross show diarization [63, 64, 65, 66, 9, 67]. Other corpora in which cross-show speaker IDs exist are too small, in terms of number of shows [68] or speaker [69] to be used for cross-show diarization. Additionally, those corpora which include cross-show speaker IDs are not provided with the time steps including the date that are necessary for lifelong learning diarization [70, 71].

The number of speakers and the ratio of speech per file is an important variability factor that can affect speaker diarization performance. The number of speakers can be either fixed [63, 65, 70], or variable as in [71] where it ranges from 1 to 21 speakers per file. Corpora collected from telephone conversations, TV or Radio and meetings usually offer a high ratio of speech duration. On the opposite, the very special context of the *Fearless Steps* corpus [68], extracted from the APOLLO-11 mission, provides about 100 hours of recordings including only 36% of speech. The amount of overlapped speech also strongly affects the performance and is intrinsically linked to the type of data; focusing on overlapped speech, the AISHELL-4 [72] corpus shows an overlap ratio of 18.2% while, by construction, there is no overlap in the CHiME-5 [70].

Most of the corpora only include English speech [64, 70, 68, 65]. Amongst the few other languages available for speaker diarization, one can cite AISHELL-4 [72] in Chinese Mandarin or Albayzin [69] in several Spanish languages.

Although there is a variety of speech corpora which can be used for speaker diarization, we found that no existing corpus gathers all required characteristics to enable lifelong learning speaker diarization, especially. Existing corpora lack the chronological time stamps (dates) that are necessary for lifelong learning and speakers appearing in several files with a unique ID that enables linking speakers across the entire collection. Those are the reasons why we introduce the ALLIES corpus, an extension of existing French corpora released for the ESTER [73], REPERE [18] and ETAPE [20] benchmarking campaigns.

2.4. Human assisted learning for diarization

Modern diarization systems achieve decent performance depending on the type of data they process [14] but those performances are often not good enough to deploy such systems without any human supervision [9, 30].

Human assisted approaches have been developed for other speech processing tasks, including speech recognition [75, 76, 77], language recognition [78], speech activity detection [79] or speech emotion recognition [80], but are not directly applicable to speaker diarization. Literature on human assisted speaker diarization is very sparse and existing approaches are complementary to our work more than competitive. In [81], active learning is used to find

Table 1: Comparison of existing diarization corpora for the purpose of human assisted lifelong learning speaker diarization.

Name	Language	Duration	# Speaker	Cross-show speaker ID	Recording time for Lifelong	Detail
CALLHOME [63]	Multilingual	20 h	2–7 Spk./file	No	No	Telephone conversations
AMI [64]	English	100 h	3-5 Spk./file	No	No	Meeting
Voxconvers [71]	Mostly English	74 h	1-21 Spk./file	Yes	No	Conversation from YouTube video
CHiME-5 [70]	English	50 h	4 Spk./file	Yes	No	Conversations in the home environment
APOLLO-11 [68]	English	100 h	34 Spk./hour	Yes, but only 30 files	No	Only 36% speech, speaker turn duration 0.5 s
AISHELL-4 [72]	Mandarin	118 h	4-8 Spk./file	unknown	unknown	Conference venues, 18% overlap ratio
DIHARD3 [67]	Multilingual	67 h	1-7 Spk./file	No	No	Contains different styles
LibriCSS [65]	English	10 h	8 Spk./file	No	No	Simulated dialogs
Albaizin [69]	Spanish	569 h	Avg. 27 Spk./file	Yes, but only 166 speakers	Yes	TV broadcast
MGB [74]	English	1600 h	unknown	Yes	Yes (Only 1m:20d)	TV broadcast

the initial number of speakers in a collection of documents. The human is then not involved anymore after this preliminary step. In [82], multi-modal active learning is proposed to process speech segments according to their length and obtain missing labels; a task that is out of the scope of our study. In [83], active learning is used to leverage training data and improve a speaker recognition system and could be similarly used for clustering. In [30], the authors propose an active learning framework to apply different types of corrections together with metrics to evaluate the cost of human-computer interactions. This work is based on a chronological correction process that can strongly limit the efficiency of the process. Our work is thus complementary to existing ones as we focus on on-line clustering correction where questions can be asked by the system regardless of any chronological constraint.

2.5. Evaluation and analysis of automatic and human assisted diarization

Performance of speaker diarization systems is usually reported in terms of diarization error rate (*DER*) [84] and more recently of Jaccard error rate *JER* [67] that gives a higher weight on clustering errors. Segmentation is evaluated by combining purity and coverage additionally with detection error rate that provides information on the quality of speech activity detection as in [85].

The performance of diarization systems has been investigated in several studies [86, 87, 88] to find the features that affect the diarization error. These features can be related to speakers or to the acoustic and transmission conditions of the recordings. These analyses provide an insight on the performance of automatic systems and directions for future improvements. In [87], the authors report the predominance of speech activity and speech overlap detection errors. In [88], the authors propose a performance prediction paradigm assuming that information related to the speech duration of a speaker and speaker turn duration can help the systems to recognize this speaker. In [86], two evaluation criteria are proposed: *Nuttiness* that measures characteristics causing high DER and *Flakiness* that measures the stability of performance of different systems on a given audio file. The authors calculate the Spearman correlation coefficient between various features of the input audio and diarization performance in order to have better understanding of diarization output.

In the context of human assisted diarization, the cost and the quality of human interaction must be evaluated. Keystroke Saving Rate (KSR) [89] which is the number of keyboard strokes made by the user can be used in some platform but is highly dependent on the user interface and does not allow easy comparison between systems. In order to propose a metric that enables a fair comparison between a wide range of systems and that can be expressed in the same unit as DER we propose to focus on the amount of time required for human interaction. This idea, first introduced in [30], has been refined and extended in [90] where we proposed a penalized error that integrates the interaction cost together with the diarization error rate. In this work, we investigate this metric and report the *Penalized DER* [91] for our experiments.

3. The ALLIES corpus and associated protocol

In this section we introduce the ALLIES corpus, designed to enable development and evaluation of human assisted lifelong learning speaker diarization systems. For this purpose, the ALLIES corpus includes the date of each audio recording and a large number of recurrent speakers over the years (i.e., speakers who speak in many shows over the years and are consistently labeled). In this article, we consider a show as a session of a given show title (i.e. BFM Story) and shows are processed independently from the show title itself. As shown in Table

1, those characteristics are unique amongst publicly available corpora for speaker diarization.

3.1. An extension of existing corpora

The ALLIES corpus has been designed to gather and extend existing French corpora collected for ESTER [73], REPERE [18] and ETAPE [20] projects. The ALLIES data comes from 1,008 French TV and radio shows collected from 7 Radio stations and 4 TV channels for a total amount of audio data consisting of 25 days, 12 hours and 46 minutes out of which 53% are annotated (13 days, 16:21:17). Table 2 displays the statistics of ESTER, REPERE, ETAPE compared with the ALLIES corpus. This section details the history of the ALLIES corpus, its statistics in terms of speakers and gives a special focus on temporal statistics before describing the corpus partition proposed for experimental purposes.

Table 2: Statistics of the ALLIES corpus and previously released part of this corpus.

Corpus	ESTER	ETAPE	REPERE	ALLIES
# shows	157	73	291	1,008
# spks per show	28.4	10.5	9.6	11.6
# unique spks	3,059	688	1,518	5,901
Annotated time (h:m:s)	110:40:48	34:09:26	52:37:26	328:21:17
Speech ratio	0.97	0.96	0.94	0.96
Overlap ratio (%)	<1	3	4	3
Start date	1998-12-17	2010-02-03	2011-07-06	1998-12-07
End date	2008-12-02	2011-05-26	2013-04-24	2014-12-01

Note that the current ALLIES corpus is already being extended with new audio material but also with more annotation on the available audio. This material will be added to the corpus over the years. The new data collected for the ALLIES corpus has been precisely annotated for overlapping speech. In the new set of data, overlapping speech segments (3.2% of the total time) involving two speakers are annotated with the name of the speakers, while segments involving three speakers or more are labeled *"+3"*.

3.2. Speaker statistics

The ALLIES corpus includes 5,901 unique speakers recorded over 16 years which ensures an important intra-speaker variability due to the aging of recurrent speakers. Note that TV and Radio shows are labeled with the date of their first broadcast, which means that the age of most speakers in a recording is consistent with the date of recording (of course it is possible that a minor part of the shows include archived recordings). The longest period of a speaker appearance is longer than 15 years (from 1998-12-10 to 2014-06-15. This speaker appears in 12 shows within the ALLIES corpus.

Some speakers appear more than hundred times in different shows of the corpus over months or years. Table 3 and Figure 2 provide a more detailed picture of the top recurrent speakers appearance across years.

Table 3: Appearance of the most recurrent speakers in the ALLIES corpus across years

#Speakers	Min #occurrences	Avg. recording period
1	146	1107 days
10	27	965 days
50	5	1502 days
1018	2	785 days

Recurrent speakers are necessary to evaluate the performance of incremental cross-show diarization. On average, 49% of the speakers encountered in a show have already been seen in the past (in a show with older recording date). Additionally, the detection of recurrent speakers is very important for practical reasons (evaluation of fairness before elections, sociological studies across time...) and in terms of performance as they count for 42% of annotation time. Recurrent speakers are not only presenters or journalists who appear in a single series of shows with homogeneous acoustic condition and speaking style. In average, odds are higher than 7% (resp. 2%) for a recurrent speaker to have been seen in the past in a show from a different series (resp. channel).

Speaker turn duration can have a huge impact on diarization performance and is strongly dependent on the type of show. Analyses given in [1] show that speaker turn duration in broadcast news is longer than in other contexts. The average duration of speaker turn in ALLIES is 14.1 seconds with a large standard deviation of 27.46 seconds which highlights the wide diversity of show genres that is covered by the ALLIES corpus.

3.3. Partitioning of the ALLIES corpus

To enable fair comparisons of systems on the ALLIES dataset we propose an evaluation protocol, that will be used for the ALLIES challenge[1]. The dataset is chronologically split into three disjoint parts: a **Training** set, a **Development** set and an

[1] `https://git-lium.univ-lemans.fr/Larcher/allies-evaluation`

Figure 2: Chronological appearance of all recurrent speakers in the ALLIES corpus according to the recording date. Each horizontal line corresponds to a unique speaker and each dot represents one occurrence of this speaker at this date (x-axis)

Evaluation set (see Figure 3). The partition has been done so that the three sets include respectively 40%/30%/30% of the annotated data (in terms of annotated speech duration). Table 4 lists channels, shows and durations for each partition of the ALLIES corpus.

To get a better sense of the chronology of the ALLIES corpus, Figure 3 displays the cumulative duration of annotated data across time together with the time limits of the **Training** , **Development** and **Evaluation** sets.

Due to historical reasons in the collection process, the duration of annotated data is highly variable across shows; for instance: the only show from the *Culture* radio channel includes 1h01m04s of annotations, while the average duration of annotations for *Planete Show Biz* is less than 2 minutes (see Table 2). For the same historical reasons, the sampling of TV and Radio shows is not uniform across time. It explains why the **Training** set runs over 12 years while **Development** and **Evaluation** sets spread over 16 and 34 months respectively.

The number of speakers in the three partitions is also very different. In Figure 4, a Venn diagram displays the number of speakers for the three parts of the corpus with details of speakers overlapping in the different partitions. The ALLIES corpus contains 66 speakers who appear in the three parts of the corpus and 261 speakers who appear both in **Development** and **Evaluation** parts.

3.4. The ALLIES protocols

Designed for human assisted lifelong-learning diarization, the ALLIES corpus can also be used for a classic speaker diarization task as reported in Section 6. The protocol for this task is given in the following section and extended later for lifelong learning.

3.4.1. Human assisted diarization protocol

In this protocol, data from the **Training** set can only be used to train the initial automatic system, possibly with additional data. After this step, **Training** data is discarded. In the following steps (i.e., development and evaluation), each show is processed with the exact same initial automatic system. For each single show, the system is free to use data from this unique show in any way without using the **Training** data.

Additionally, using a questioning and correction module (see Figure 1), the system can interact with the simulated human expert by asking two types of questions: (i) questions related to the clustering: "Are the speakers speaking at time t_0 and t_1 the same?" (ii) and questions related to the segmentation: "What are the borders of the speaker turn around time t?". The answers of these questions can be retrieved from the reference annotation to adapt the system or correct the system hypothesis. After processing a show, the system is reset to its initial state (this scenario does not include any sequential processing).

Table 4: Global partitioning of the ALLIES corpus recorded from 4 channels and 19 show titles with their corresponding annotated duration (and number of shows). All durations are given in hh:mm:ss format.

Type	Channel	Show title	Total	Training	Development	Evaluation
TV (204:14:03)	BFM	BFM Story	26:40:33 (49)	2:28:29 (3)	12:45:46 (25)	11:26:18 (21)
		Planete Showbiz	2:24:14 (73)	-	2:24:14 (73)	-
		Ruthel Krief	0:21:06 (4)	-	-	0:21:06 (4)
	LCP	Ca Vous Regarde	24:22:29 (45)	1:32:18 (2)	14:58:13 (27)	7:51:58 (16)
		Culture Et Vous	2:45:12 (87)	-	0:16:49 (8)	2:28:23 (79)
		Entre Les Lignes	25:32:35 (62)	0:52:47 (2)	10:36:20 (29)	14:03:28 (31)
		LCP Actu	21:34:51 (80)	-	-	21:34:51 (80)
		LCP Info	46:40:14 (156)	-	28:54:48 (97)	17:45:26 (59)
		Pile Et Face	25:57:08 (76)	2:13:07 (5)	14:40:09 (46)	9:03:52 (25)
		Top Questions	24:08:38 (104)	-	9:59:32 (46)	14:09:06 (58)
	TVME	-	2:09:23 (8)	2:09:23 (8)	-	-
	TV8	-	1:37:40 (4)	-	1:37:40 (4)	-
Radio (124:07:14)	Africa1	-	3:47:36 (18)	3:47:36 (18)	-	-
	Classique	-	1:00:04 (1)	1:00:04 (1)	-	-
	Culture	-	1:01:21 (1)	1:01:21 (1)	-	-
	France Info	-	12:00:43 (13)	12:00:43 (13)	-	-
	France Inter	-	54:56:52 (86)	52:59:09 (79)	1:57:43 (7)	-
	RFI	-	28:49:18 (38)	28:49:18 (38)	-	-
	RTM	-	22:31:20 (103)	22:31:20 (103)	-	-
Total			328:21:17 (1008)	131:25:35 (273)	98:11:14 (362)	98:44:28 (373)

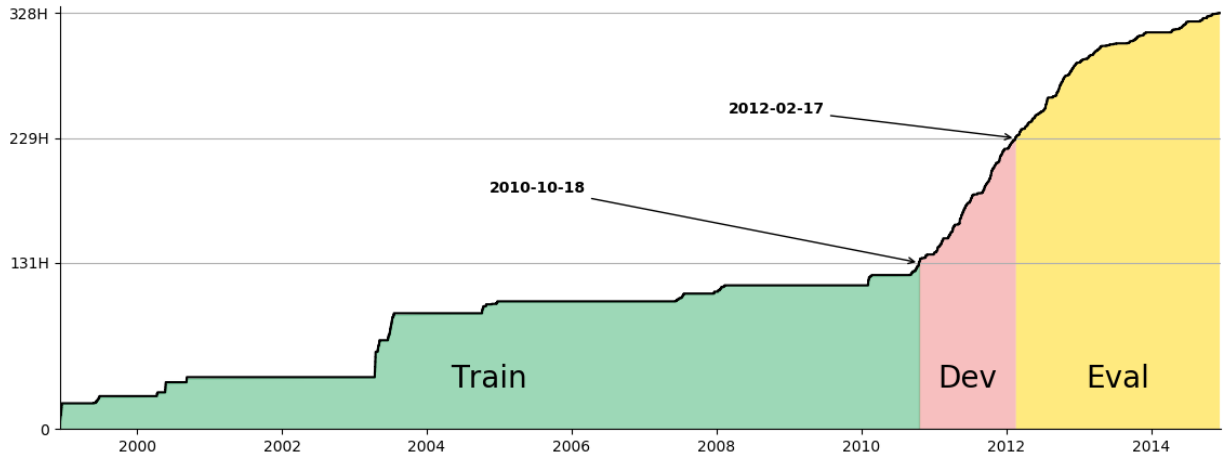

Figure 3: Cumulative duration of annotated signal across time. The shows recorded before the 18^{th} of October 2010 are used as **Training** data, shows between the 18^{th} of October 2010 and the 17^{th} of February 2012 are used as **Development** data and the remaining shows (recorded after the 17^{th} of February 2012 are part of the **Evaluation** data

Data from the **Development** set can be used to tune the hyper-parameters of the human assisted systems but not to retrain or adapt the automatic system itself. **Evaluation** set is then provided to fairly evaluate the systems. While processing the **Evaluation** set, adaptation of the automatic system and tuning of the hyper-parameters is forbidden. Optimal results of three standard systems using this protocol are provided in Section 6 for the **Development** set. Results on the **Evaluation** set will be published after the ALLIES Challenge.

3.4.2. Lifelong-Learning protocols

Compared to the previous scenario, this one considers the sequential processing of shows across time to evaluate human assisted lifelong-learning systems. In this scenario, the **Training** set can be used the exact same way as previously to train an

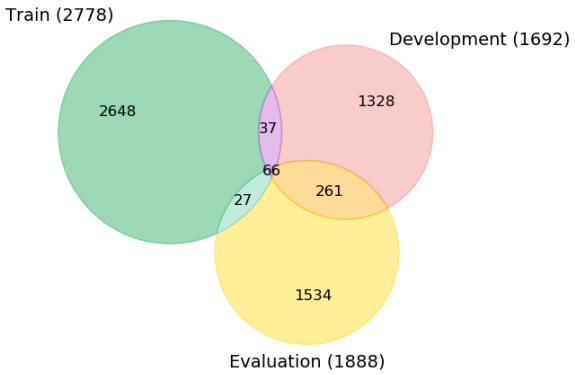

Figure 4: Number of speakers in different partitions of ALLIES corpus and the number of common speakers.

initial system. The **Training** data is then set aside to be re-used anytime by the system.

For **Development** and **Evaluation**, the extension of the previous protocol for human assisted lifelong-learning speaker diarization requires to strictly process the shows in chronological order. Each show is processed as described in the previous protocol with possible interaction with the human expert. After the system produces its final hypothesis for one show, the next show (in chronological order) is then processed without resetting the system, i.e., when processing one show, the system can make use of any information gathered on previously seen shows, including models of previously encountered speakers that are used for cross-show clustering (speaker linking across shows). **Development** data can be used to optimize the hyper-parameters of the system that are then fixed when processing the **Evaluation** set.

For **Evaluation**, two *lifelong-learning* protocols are proposed depending on the state of the human assisted diarization system when starting processing the **Evaluation** set. In a first scenario, named *ALLIES-reset-lifelong*, it is possible to use the initial system trained on **Training** set with hyper-parameters tuned on **Development** set. In a second scenario, named *ALLIES-lifelong*, one can start processing the **Evaluation** using a version of the human assisted diarization system that has already gathered knowledge by processing the **Development** set. In this former scenario, the system might have learned about the speakers encountered in the **Development** set.

The remaining of this paper focuses on within show human assisted diarization. This is a first step toward lifelong human assisted diarization. Cross show and incremental diarization are not performed in this work and will be published in the future.

4. Human Assisted Diarization Systems

This section first describes three baseline diarization systems that are used to provide a wider view of diarization performances on the ALLIES corpus. Second, it presents the proposed question generation module and correction module that are used on top of each of the three baseline systems. In the followings, only the human assisted diarization protocol is investigated. It means each baseline diarization system is considered fixed after its initial training. Highly dependent on the baseline system architecture, the adaptation module is let out of the scope of this work. The adaptation module will be investigated in future works.

Diarization errors can be due to wrong segment borders or wrong label allocation. The former error being the most harmful in terms of performance [30], this work only focuses on correcting labeling errors, *i.e.* clustering errors.

4.1. Automatic diarization systems

Figure 5 provides an overview of the 4-step automatic diarization process. The automatic systems can be described as a succession of two phases: initial segmentation and clustering. The first phase is common to all systems while the clustering differs between systems.

4.1.1. Initial segmentation

The initial segmentation is illustrated as step 1 in Fig. 5. The LIUM Voice Activity Detection (*VAD*) system is used to segment the audio stream by discarding non-speech segments (silence, noise, breathing, etc.). This VAD, based on stacked LSTM [14], is implemented in the S4D open-source framework [92]. The output of the network is smoothed by removing non-speech segments shorter than 50ms and speech segments shorter than 25ms. In order to investigate the impact of clustering correction on diarization performance without interfering with segmentation errors, all results will also be provided by using an ideal speech activity detection obtained from the ground truth annotation and referred to as the reference segmentation (*ref*).

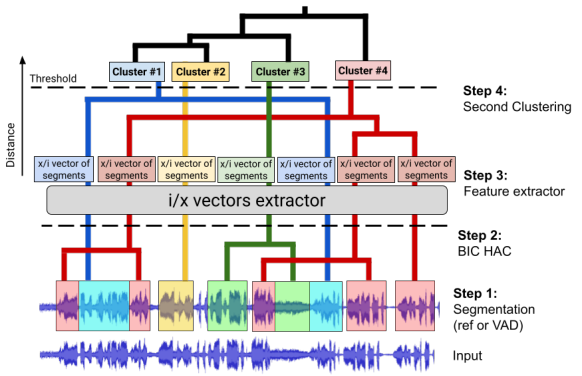

Figure 5: Diarization steps of the Evallies systems; 1. Initial segmentation that withdraws non speech part of the given audio stream is obtained by using the reference segmentation or an automatic Voice Activity Detection system, 2. BIC-HAC clustering, 3. Extraction of x/i vectors from each cluster of segment obtained from the previous step, 4. Second clustering step using PLDA

4.1.2. Evallies systems

The Evallies diarization systems are based on a hierarchical agglomerative clustering (HAC). Two flavours of this system are used: one with i-vectors and the other one with x-vectors. The three steps of this system are the following ones:

- BIC-HAC (step 2 in Fig. 5): A first HAC is performed on vectors of 13 MFCC using the BIC criteria [92] starting from the initial segmentation (VAD or *ref*). When using *VAD* segmentation, the initial clustering result is followed by a Viterbi decoding to smooth the segment borders along the audio stream. The threshold of BIC-HAC is optimized based on the final DER on a development set.

- i/x-vectors representation (setp 3 in Fig. 5): The i/x-vectors (*iv* or *xv*) are extracted from each speaker turn and averaged to provide a single i/x-vector per BIC-HAC cluster. The i-vectors [51] extractor, including a 256 component UBM and a total variability matrix of rank 128 is trained on the ALLIES **Training** set, while the Half-ResNet34 used for x-vector extraction [93, 94, 95] is trained on a larger set of data combining VoxCeleb1&2 [96, 97].

- HAC clustering (step 4 in Fig. 5): A second HAC clustering is performed by using i/x-vectors. The distance matrix used for this clustering is computed using a PLDA [98] trained on the **Training** set. The threshold of this clustering is also optimized based on the final DER on the development set.

In the remaining, the flavour of the Evallies system using i-vectors (respectively x-vectors) is named *Evallies iv* (respectively *Evallies xv*).

4.1.3. VBxHMM system

The third baseline system is the *VBxHMM* system proposed in [11]. Starting from the initial segmentation, x-vectors are extracted using a ResNet34 on a sliding window of 1.5s with a shift of 0.25s and then centered, whitened and length normalized. The x-vectors are pre-clustered using HAC (with cosine similarity) to obtain the initial speaker labels. Eventually, x-vectors are further clustered using the VBx model after applying a dimensionality reduction. The optimal HAC threshold and VBx hyper-parameters are tuned to optimize the DER on the development data. A deeper description of this system is given in [11]. All parameters from this system are trained on VoxCeleb 1 & 2 [96, 97].

4.2. Active correction process

The active correction module detailed in this section is composed of three parts: a confidence estimation module, a questioning module and a correction module. An initial hypothesis (illustrated on Fig. 6) is generated once using an automatic within-show diarization system (Evallies HAC clustering or VBxHMM). Based on the initial clusters obtained automatically, the confidence estimation module ranks the audio segments that are likely to be wrongly annotated. Exploiting the human expert answers, a correction module modifies the initial hypothesis. Our motivation in this work is to produce an active correction system that is independent of the baseline system. Note that the current implementation of the active correction system only modifies the clustering and does not modify borders of the segments.

4.2.1. Confidence estimation module

We propose to represent the initial hypothesis with a clustering tree that is obtained in a three-step process illustrated on Fig. 6.

1. For each initial cluster, a HAC is performed (on MFCCs with a BIC criteria). A threshold, set experimentally, is used to determine

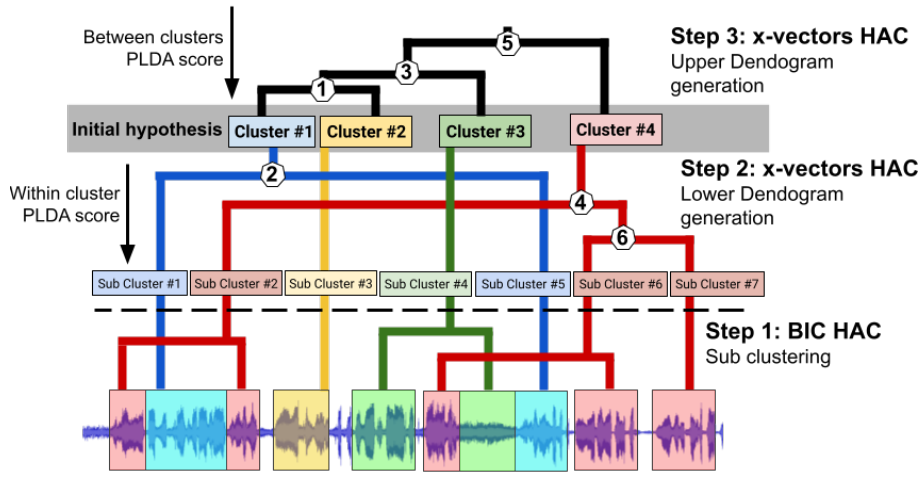

Figure 6: Active correction steps; 1. For each cluster of basline diarization system (*Evallies* or *VBxHMM* systems) run BIC-HAC to generate sub-clusters, 2. Extracting x-vector per sub-cluster and link them with HAC, 3. Extract x-vector per cluster and link them with HAC (active correction would be applied only on the links generated in step 2 and 3)

sub-clusters of segments. Those sub-clusters will be considered as non-separable during the human correction process. This BIC-HAC prevents from extracting speaker representations from segments too short to contain enough information (see Figure 6, step 1).

2. i/x-vectors are extracted for each sub-cluster generated after step (1), and linked by another HAC (based on PLDA log-likelihood) to create a clustering tree. After this step, each initial cluster is divided in sub-clusters linked in a clustering tree. Such a clustering tree will be now referred to as *sub-tree* (see Figure 6, step 2).
3. All *sub-trees* are finally linked to create a *between-cluster* tree. Similarly to step (2), one i/x-vector is extracted for each cluster created by the automatic diarization system and a third HAC (based on PLDA log-likelihood) generates a final between-cluster tree for the entire audio stream, which leaves are the non-separable sub-clusters (see Figure 6, step 3).

These three steps generate a dendrogram based on the initial hypothesis and the PLDA score. Figure 6 illustrates two types of node. *Within-cluster* nodes are below the initial hypothesis (2, 4 and 6), they merge two branches of the tree (two clusters) that are supposed to belong to the same speaker (according to the automatic system). *Between-cluster* nodes are above the initial hypothesis (1, 3 and 5) merge clusters that belong to two different speakers (according to the automatic system). For human assisted correction, we define a confidence metric c for each node in the dendrogram; the further from the initial hypothesis, the higher the confidence. For a *between-cluster* node, c is the inverse PLDA score. For a *within-cluster* node, c is the PLDA score. We propose to rank all nodes by their confidence (from lowest to highest) in order to obtain one single ranked list of nodes to investigate.

4.2.2. Questioning module

The numbers assigned to nodes in Figure 6 shows the order of questions (1 to 6) which follows the confidence ranking. For example node 1 is considered with less confidence than node 6.

Correction module. This work considers that the human expert can only be asked the following question: *"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 is not modified (no merging re-

quired) and the two branches of a node below the threshold are separated (splitting operation).

Limitations of the questioning module. Questioning each node of the clustering tree would minimize the error, but would induce a prohibitive cost of human interactions. To constrain this cost, we impose two limitations to reduce the number of questions. The first limitation aims at discarding questions addressing already known information. In case a human expert confirms that a node is correctly merged (for example node 4 in Fig. 6), we assume that asking about its descendants is useless (for example node 6 in Fig. 6). Similarly, in case a human experts answers that a node must be split, its ancestors won't be questioned. To avoid asking useless questions, we keep in memory human expert answers by considering two sets of nodes. The *Stop separation set* contains nodes that should not be investigated for separation (or splitting). The *Stop clustering set* contains nodes that should not be investigated for clustering (or merging). These sets are initially empty and updated after each interaction with the human expert. If a question related to a *within-cluster* node gets confirmed (respectively corrected) by the human, all descendant nodes will be added to the *Stop separation set* (respectively all ancestor nodes will be added to the *Stop clustering set*). In case a question related to a *between-cluster* node gets corrected (respectively confirmed), all ancestor nodes will be added to the *Stop clustering set* (respectively all descendant nodes will be added to the *Stop separation set*). For example, in Figure 6, if the human expert answers that node 1 should be merged, (*i.e.* cluster 1 and 2 belong to the same speaker), then node 2 will not be investigated for separation. Note that despite this first limitation, the number of possible questions can still be very high.

The second limitation aims at minimizing the number of questions for which the human expert is likely to confirm the system's decision. We propose to limit possible questions to questions with low confidence. To do so, we set an empirical early stopping criterion called confirmation to stop (*C2S*) . Once the human has confirmed a number *C2S* of decisions from the automatic system, this one stops asking questions. Reducing *C2S* reduces the risk of useless questions, while increasing *C2S* allows the system to explore a larger part of the tree.

One question asked for a given node of the clustering tree might relate to many audio segments; indeed, each branch of the node might correspond to several segments. To facilitate the work of the human expert, our system selects two audio examples of each branch of the node under investigation for the user to listen to. Based on our preliminary results [91], the longest segment in each branch are the best candidates for comparing two sub-clusters and will be used in all experiments.

5. Evaluation

Extending diarization to a human assisted lifelong learning task requires to extend metrics to take into account the cost of human interaction. In this section, we first describe the metrics used to evaluate our baseline systems; additionally to the classic ones, we introduce a metric that specifically evaluates the segmentation. Then we give a brief overview of the Penalized DER introduced in [90]. Eventually, we describe the simulated human expert module developed to enable reproducible research in the context of human assisted diarization.

5.1. Baseline systems assessment

Diarization results are reported using four metrics: the weighed diarization error rate (*DER*) [84], the weighed Jaccard error rate (*JER*) which, unlike DER, considers an equal weight for each speaker [67], *Purity* of clusters and *Coverage* of speakers [85].

Additionally, we introduce a Segmentation Error Rate (*SER*) to shed a light on the initial stage of diarization. Since our human assisted diarization process only focuses on clustering correction, we use the *SER* to differentiate between the effect of both steps of diarization (segmentation and clustering). The *SER* measures the mismatch between reference and hypothesis segmentation borders. For each segment, S_i^r in the reference (resp. hypothesis), the segment $\hat{S}_i^h$, in the hypothesis (resp. reference) that has maximum intersection duration with S_i^r is selected. The duration of S_i^r that does not have a match in $\hat{S}_i^h$, referred as $t_{S_i^r}$, is then divided by the total duration of S_i^r, noted $D_{S_i^r}$, to compute a segmentation error rate for the segment S_i^r. By applying this process in a symmetrical manner to the reference and the hypothesis, we compute SER_{ref} and SER_{hyp} as detailed in eq. 1 and 2.

$$SER_{ref} = \frac{1}{N_{ref}} \sum_{i=1}^{N_{ref}} \frac{t_{S_i^r}}{D_{S_i^r}} \quad (1)$$

$$SER_{hyp} = \frac{1}{N_{hyp}} \sum_{i=1}^{N_{hyp}} \frac{t_{S_i^h}}{D_{S_i^h}} \quad (2)$$

where N_{hyp} is the number of segments in the hypothesis and N_{ref} the number of segments in the reference. Finally, the Segmentation Error Rate, *SER*, is obtained by getting the average of SER_{ref} and SER_{hyp} as shown in eq. 3 :

$$SER = \frac{SER_{hyp} + SER_{ref}}{2}. \quad (3)$$

The measure SER reveals the mismatch of two segmentation sequences; it increases when segment borders in reference and hypothesis are not matching, or when several segments in reference (or hypothesis) correspond to a single segment in the other sequence.

5.2. Human interaction assessment

DER/JER improvement. The assessment of a human assisted system must take into account the cost of human interaction together with the quality of the interaction process. An optimal interaction reduces the work of the human while maximizing the gain in terms of performance. One way to estimate the cost of the human interaction for diarization is to measure the *DER* (resp. *JER*) improvement, defined as the absolute difference between *DER* (resp. *JER*) before and after the human correction. Penalized DER (DER_{pen}), a metrics introduced in [91], is used to merge the information about the final performance (after human interaction) with the cost of the interaction required to reach this result. This metric adds a constant amount of error time, called penalized time (t_{pen}), to the diarization error time, for each question asked to the human expert. Eq. 4 defines the DER_{pen}, where FA is false time, $Miss$ is missed time, $Conf$ is confusion time of diarization hypothesis, T_{total} is total duration of audio files and N is the number of human interactions.

$$DER_{pen} = \frac{FA + Miss + Conf + N \cdot t_{pen}}{T_{total}} \quad (4)$$

We also propose to use the effective number of corrections over number of questions ratio (CQR) as a questioning performance criteria to estimate the quality of the human interaction.. It is used to evaluate the early stopping criteria and the question generation module.

5.3. Simulated human expert

To enable fair and reproducible benchmarking, a human expert is simulated by using ground truth reference to provide a correct answer to each question. In the context of this study, the system can ask questions of the form: "Have the segments A and B been spoken by the same speaker?". Since segments A and B might not be pure (i.e., they can include speech from several speakers), the simulated human expert first assigns each segment to its dominant speaker in the reference. The dominant speaker of a segment is the one with maximum speech duration in the reference segmentation. Eventually, the simulated expert answers the question by comparing the dominant speakers from segments A and B.

To establish a lower bound, we also develop an *ideal* correction process that does not consider any limitation of interaction cost. In this approach, all segments from the hypothesis to be corrected are matched with their counterparts from the reference and labeled accordingly with the dominant speaker. This assignment is done based on maximum intersection time between reference and hypothesis segments. The *ideal* correction simulates a process in which all combinations of segments pairs would be questioned. The main difference of $C2S = inf$ and the *ideal* correction is that in the *ideal* correction, the protocol described in Section 4.2 is not applied and no limitation of the interaction costs is considered. This lower bound provides an opportunity to disentangle segmentation error from clustering error; the former one being the main focus of our proposed human correction process. This simulation is provided as part of the ALLIES evaluation package[2].

6. Experiments

In this study, only within show experiments are realized. Shows are processed independently and the average of DER is weighed based on their duration and referred to as total DER. The total JER also is computed as the weighed average JER of all speakers in ALLIES **Development** set. Cross-show experiments will be investigated in future work.

[2] `https://git-lium.univ-lemans.fr/Larcher/evallies`

6.1. Baseline diarization system performance

Performance of the three baseline diarization systems introduced in 4.1 are reported in Table 5 when using an automatic VAD, or the reference segmentation as initial segmentation.

One can expect the SER to be null when using the reference segmentation, but it can be observed in Table 5 that this is not the case. This non-zero SER can be explained by the merging of two consecutive segments or by the pre-processing step (the segmentation smoothing, and possibility of merging two segments in BIC-HAC, see 4.1) that is applied on top of the initial segmentation. Note that improving the SER mechanically improves Purity, Coverage and DER.

SER are eight times higher when using the VAD compared to the initial segmentation (it is not the case for *VBxHMM* due to re-segmentation inherent to this system) and this error is not corrected by our human assisted process that only focuses on clustering correction.

VBxHMM system performs significantly better than the two Evallies HAC-based systems for all metrics except SER where performance is equivalent. Obtaining a DER of 16.19% with VAD and 10.44% with the reference segmentation, *Evallies-xv* performs slightly better than *Evallies-iv* which achieves 17.24% and 11.08% in the same conditions.

Comparing the number of estimated clusters with the number of speakers in the reference shows that the use of an automatic VAD segmentation for the Evallies system results in under-clustering, while the VBxHMM system detects the correct number of speaker.

6.2. Human assisted correction performance

Table 6 shows the performance of the three baseline systems after applying the proposed active correction process with different configurations. The first conclusion is that this process improves the performances of all systems in terms of DER and JER when using VAD or reference segmentation.

As a reference, the bottom line performance obtained with *Ideal* correction is given for each system and both initial segmentations. As explained in Section 5.3, it corresponds to the optimal performance that can be obtained when applying clustering correction without any limitation of interaction.

As expected, increasing the confirmation to stop (*C2S*), i.e., increasing the possibility of having more corrections, leads to a lower DER. The observation of the correction to question ratio (*CQR*) reveals that the system asks more useless questions when increasing the *C2S* (i.e., *CQR* decreases when *C2S* increases). This result suggests that increasing *C2S* leads to more corrections but that more non-informative questions will be asked, leading to high human interaction cost with limited gain.

DER_{pen}, proposed in [90], includes the cost of questions together with the final DER after correction. As expected, DER_{pen} is minimum for low values of *C2S* (1 or 2) and increases for higher values, as shown in the last column of Table 6. For all systems using reference segmentation, DER_{pen} is lower than baseline DER for all *C2S* values, highlighting the importance of improving the segmentation process. A growing DER_{pen} means that corrected segments are shorter than t_{pen}, the time spent to correct it, or that the re-labeled clusters are not pure enough, bringing more degradation than benefit when re-labeled. This will be the topic of a future work on improving the correction process, especially focusing on segmentation errors.

Efficiency of the correction strongly depends on the design of the human-computer interface that is not the topic of this work. However, Figure 7 gives an overview of the benefit that can be obtain when reducing the interaction time (t_{pen}). On this figure, baseline DER (before correction) and final DER (after correction) are given by the dash line and the blue bars respectively, while penalization (time spent by the human expert to reach this final DER) is given by the upper colored bars. Remember that DER_{pen} is the sum of the final DER and the penalization. For all three systems, applying corrections leads to lower DER but one can observe that depending on the correction time (t_{pen}), the balance between time spent for correction and time of signal corrected during the process is not always positive. For instance, spending 8 seconds for each correction leads to a DER_{pen} that is higher than the baseline DER for all systems and *C2S* values. For *Evallies-iv* (resp. *Evallies-xv*) system, it is reasonable to dedicate up to 4 seconds (resp. 2 seconds) per correction while for *VBx* a correction time higher than 1 second leads to a DER_{pen} higher than DER baseline for any *C2S* value. Note that the comparison between DER_{pen} and DER only reflects a global correction benefit as more corrections always leads to lower DER. Ideally, the quality of the correction module should be compared to a fully manual correction, which will be the topic of a future work.

Table 5: Baseline systems performance: segmentation error rate (SER), False time ratio (FA), Miss time ratio (Miss), Confusion time ratio (Conf), Diarization error rate (DER), Jaccard error rate (JER), Purity, Coverage, and the average Number of Speakers ± 95% confidence interval (the number of speakers in the reference is 9.81±0.63) with two initial segmentations: reference or VAD

Segmentation	Systems	SER	FA	Miss	Conf	DER	JER	Purity	Coverage	Num. Spk.
VAD	*Evallies-iv*	0.16	2.12	4.54	10.57	17.24	32.13	80.50	88.04	21.06±1.09
	Evallies-xv	0.16	2.05	4.69	9.46	16.19	31.49	80.65	87.88	20.41±0.98
	VBxHMM	0.17	2.15	4.31	4.56	11.01	14.42	84.34	88.46	9.41±0.64
ref	*Evallies-iv*	0.02	0.72	1.99	8.37	11.08	17.51	99.98	99.84	8.31±0.58
	Evallies-xv	0.02	0.73	2.04	7.67	10.44	16.17	99.98	99.84	8.57±0.58
	VBxHMM	0.14	0.21	4.27	4.89	9.37	11.52	94.52	88.70	9.33±0.64

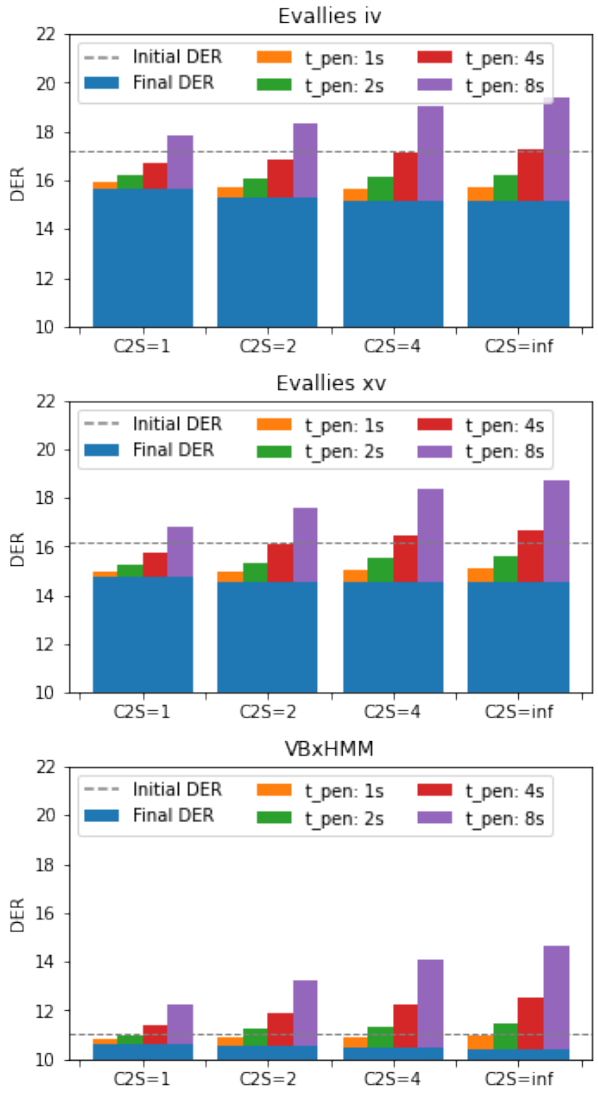

Figure 7: The impact of different penalization of asking questions for systems using *VAD* result as initial segmentation.

7. Discussion

Nuttiness (hard to crack) was introduced in [86] as the "exhibition of high DER" to understand why some audio files show unusual high DER. In this section, we investigate some speaker and show characteristics, and also the impact of the human correction process, together with system performances in order to try to predict the *nuttiness* of an audio file.

To identify shows that could lead to *nuttiness*, the Figure 8 summarizes DER and DER improvements for each series of shows obtained with the *Evallies-xv* system. It reveals that diarization performs best for some titles, e.g. the DER of *Planete Showbiz* is higher than *BMF Story*. We can also observe a degradation of DER with the human assisted correction on *TV8* title.

7.1. Analysis of the baseline diarization system

Assuming that some shows or speakers are complex, i.e., lead to higher DER or JER, we investigate different features of shows and speakers that could impact the diarization performance. Intuitively, complexity can be due to the type of show (e.g., number of speakers, background noise), the collection process (e.g., dates, compression) and the spontaneity of speech (e.g., intonation variations, overlaps, etc.).

In the following sections, *Evallies-xv* system using VAD result and reference diarization as initial segmentation is used as baseline system. Pearson correlation coefficients, R, are computed between various show features and 5 metrics which capture the performance of the baseline system: DER, JER, coverage, purity and SER. Considering the difficulty of finding significant correlations we only report correlations with a p-value lower than 0.01.

Table 6: Performance of 3 baseline systems on **Development** data given in terms of DER, JER, and the average Number of Speakers ± 95% confidence interval (the number of speakers in the reference is 9.81±0.63). Human assisted correction is evaluated in terms of averaged number of questions per hour (Avg. #Q/h), number of corrections over number of questions (CQR) and penalized DER (DER_{pen}). The collar is set to 0.250 seconds and penalized time to 4 seconds).

Segmentation	System	HAL	DER	JER	Num. Spk.	Avg. #Q / h	CQR	DER_{pen}
VAD	*Evallies-iv*	Baseline	17.24	32.13	21.06±1.09	-	-	-
		C2S=1	15.66	28.52	21.29±1.13	9.82	40.74	16.75
		C2S=2	15.31	28.25	21.27±1.13	13.75	33.14	16.84
		C2S=4	15.19	28.11	21.23±1.12	17.34	28.85	17.12
		C2S=inf	15.17	28.11	21.21±1.12	18.88	27.34	17.27
		Ideal	9.08	20.97	28.98±1.46	-	-	-
	Evallies-xv	Baseline	16.19	31.49	20.41±0.98	-	-	-
		C2S=1	14.75	29.03	20.72±1.02	9.16	32.41	15.77
		C2S=2	14.59	28.70	20.75±1.02	13.46	27.25	16.08
		C2S=4	14.58	28.66	20.75±1.01	17.03	22.79	16.47
		C2S=inf	14.57	28.66	20.74±1.01	18.65	20.96	16.64
		Ideal	9.05	20.73	29.33±1.46	-	-	-
	VBxHMM	Baseline	11.01	14.42	9.41±0.64	-	-	-
		C2S=1	10.60	13.92	9.63±0.66	7.39	18.33	11.42
		C2S=2	10.56	13.83	9.64±0.67	11.89	13.17	11.88
		C2S=4	10.45	13.72	9.61±0.67	16.29	11.29	12.26
		C2S=inf	10.41	13.67	9.58±0.66	19.08	10.70	12.53
		Ideal	6.81	8.94	10.76±0.69	-	-	-
ref	*Evallies iv*	Baseline	11.08	17.51	8.31±0.58	-	-	-
		C2S=1	9.01	14.19	8.39±0.59	8.64	34.97	10.00
		C2S=2	8.77	13.84	8.35±0.58	12.48	27.69	10.20
		C2S=4	8.68	13.75	8.32±0.58	15.85	22.75	10.47
		C2S=inf	8.63	13.72	8.3±0.58	17.47	20.99	10.59
		Ideal	2.46	1.33	9.73±0.61	-	-	-
	Evallies xv	Baseline	10.44	16.17	8.57±0.58	-	-	-
		C2S=1	8.59	13.46	8.36±0.58	7.62	25.73	9.26
		C2S=2	8.55	13.37	8.33±0.58	11.29	20.14	9.74
		C2S=4	8.29	13.21	8.3±0.57	14.79	16.67	10.11
		C2S=inf	8.26	13.20	8.3±0.57	16.57	14.94	10.31
		Ideal	2.46	1.33	9.73±0.61	-	-	-
	VBxHMM	Baseline	9.37	11.52	9.33±0.64	-	-	-
		C2S=1	8.97	11.11	9.55±0.67	7.53	20.27	9.80
		C2S=2	8.96	11.02	9.57±0.68	11.96	14.36	10.29
		C2S=4	8.85	10.90	9.53±0.67	16.23	12.20	10.66
		C2S=inf	8.83	10.80	9.5±0.67	19.15	11.35	10.96
		Ideal	4.72	4.25	9.24±0.59	-	-	-

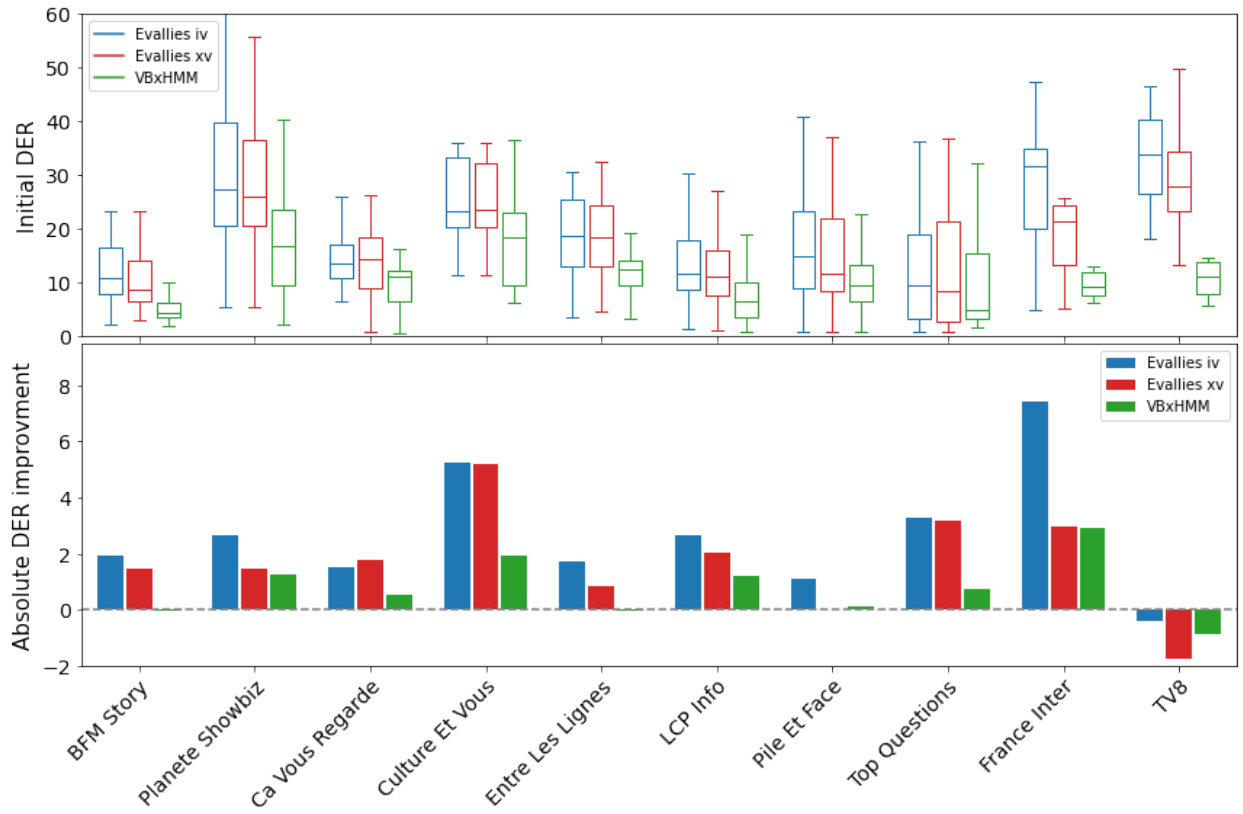

Figure 8: Initial DER and DER improvement after clustering correction (c2s=inf) according to title in Dev with VAD's result. Simple average without taking into account the show duration. Number of shows and duration of each title can be found in Table 4.

7.1.1. Segmentation and diarization errors

In Table 5, segmentation errors due to the VAD, are mostly reflected in the FA rate ($\simeq$2%) and probably also affect the Miss rate. This is clearly reflected by the SER which is 8 times higher when using the VAD. For both Evallies systems, this error affects the confusion rate which is higher by 2 points when using the automatic VAD. This result confirms previous works in the domain but also shows that x-vectors are relatively robust to the pollution caused by the segmentation error. Indeed, degradation in terms of confusion is in the same range as the FA rate.

In [86], the authors report that short speaker turn durations damages the DER. Indeed, short speaker turns, possibly due to frequent interruptions or short sentences, can impact the final diarization result in two ways. The authors assume it can make the segmentation task more difficult and increase the chances of error in VAD segmentation. In the ALLIES **Development** set, we observe a correlation of $R = -0.27$ ($p < 0.001$) between speaker turn and SER which weakly supports this hypothesis. The authors of the study also concluded that short segments will reduce the quality of speaker embeddings. We found no significant correlation between speaker turn durations and DER when using VAD segmentation ($R = -0.08$; $p = 0.11$). This correlation is slightly higher with the reference segmentation ($R = -0.23$; $p < 0.001$) but is not conclusive. Our finding seems to support the idea that speaker representation has made great progresses during the last decade and is now more robust to the perturbations introduced by segmentation errors.

According to our study, the major factor affecting segmentation errors is a compression mismatch that appears for some files in the ALLIES corpus. We found that the compression of audio file is slightly correlated with SER ($R = -0.36$; $p < 0.001$). However, it does not significantly affects the DER when considering the reference segmentation ($R = -0.07$;

$p = 0.21$), which demonstrates the robustness of x-vectors to the audio compression observed in this corpus.

7.1.2. Dominant vs. minor speakers

In order to investigate the correlation of JER and speakers, different features related to the role of speakers such as appearance duration and number of appearance in a given show have been computed.

We selected the 50 and 100 speakers from the **Development** set having the longest appearance duration within a show: for the top 50 speakers, they appear more than 750 sec. in each show where they are present while the 100 majors speakers appear more than 634 seconds. We found that JER and appearance duration are correlated with a coefficient $R = 0.53$ for the top 50 and $R = 0.60$ for the top 100 (both with a p-value of $p < 0.001$). When considering all speakers, these features are not correlated. This reveals that major speakers (in duration) are more difficult to diarize.

Similarly, for minor speakers, i.e., the 50 speakers and 100 speakers having the shortest duration appearance, JER and appearance duration are negatively correlated with a coefficient $R = -0.41$ ($p < 0.001$). Those observations indicate that minor and major speakers are the one exhibiting the highest JER and highlight the fact that automatic systems are developed to minimize averaged DER and JER, which probably implies an over-optimization for average speakers and raises the issue of out-layers (the tails of the speaker distribution). Taking into consideration the role of speakers within a show seems an interesting avenue to tackle this effect. The question of cluster purity for major speakers and the poor performances on minor speakers will be investigated in future work.

7.2. Analysis of the human correction process

In this section, the impact of the human correction process, measured using the DER improvement, the JER improvement and the number of generated questions (CQR) is also analysed with regard to show and speaker features.

Figure 9 shows the average DER improvement obtained for the ten first ranked questions asked during the human assisted correction process for each of the three baseline systems. For the two Evallies systems, the first questions bring the highest improvement in DER, which confirms the efficiency of the question generation module based on our clustering confidence measure. For *VBxHMM*, our question generation module is not able to rank most useful questions first, and it is the second question asked to the simulated human expert that brings the best improvement in terms of DER. This is probably due to the fact that our question generation module confidence criteria is the same as the one in Evallies systems but differs from the one in *VBxHMM* system.

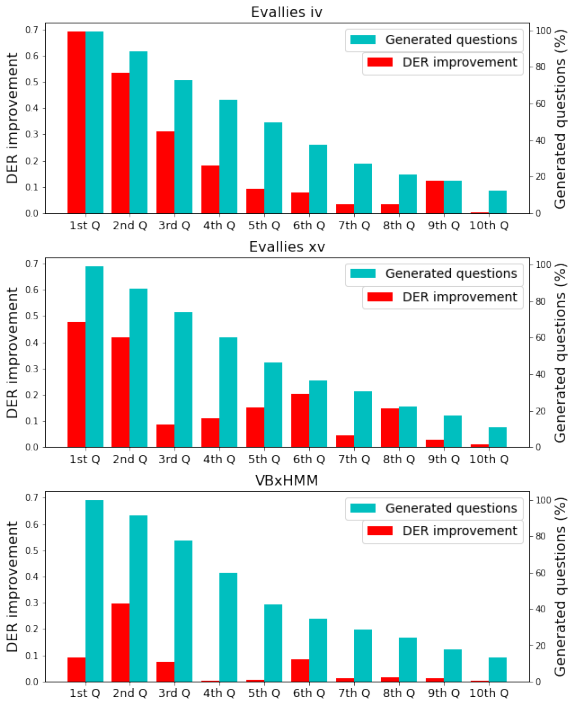

Figure 9: DER improvement and ratio of generated questions in question order. The first questions are generated more often and gain more DER improvement.

No correlation between show features and DER or JER improvement has been found in our study. Deeper investigations are required to understand the *nuttiness* of shows in human correction process.

8. Conclusion

In this article, we have prepared the ground for human assisted lifelong speaker diarization. We have proposed a complete set of resources to support the development and evaluation of human assisted speaker diarization systems including a new corpus, protocols and metrics for this task. The ALLIES corpus, build as an extension of previously

existing corpora, offers a large quantity of data including 1,008 TV and radio shows recorded over more than 14 years with many recurrent speakers. This characteristic makes it useful for speaker diarization within and across shows, but also for speaker verification as aging of speakers and various acoustic conditions makes it very challenging for this task. The release of a protocol for speaker verification on this data is part of an on-going work. Note that this corpus is currently used for the ALLIES challenge and will be publicly released for free after this challenge.

A simulated human expert has been developed to enable fair and reproducible evaluation of the human assisted learning process. This simulated human expert is able to answer different types of question from the automatic system and will be extended in the future to answer questions related to segmentation, cross-show speaker diarization and other types of questions. Note that this simulated human expert is released for the ALLIES challenge as part of an open-source package[3].

We have reported various performances on the ALLIES corpus with three automatic baseline systems without and with human assisted correction. To our knowledge, this work is the first attempt to develop an active speaker diarization system for which the human operator is involved along the diarization process through an active correction process. Amongst those systems, two are released in our *git* repository[3] together with the proposed confidence estimation module, and questioning module.

In this work, we have also extended our previous work on metrics for human assisted diarization by proposing a segmentation error rate that has been shown useful to better understand the performance of the different steps of the diarization system.

This work opens many avenues for future research on human assisted speech processing and many works are already engaged in this direction. The ALLIES corpus is currently being extended to extend its time coverage over 20 years and increase the number and sessions of recurrent speakers. The current human assisted speaker diarization system will be extended for incremental cross-show diarization across time. Impacts of TV versus radio channels will be investigated in futur work. Our aim is to develop an incremental adaptation module including an evolutive speaker embedding extractor to automatically adapt to new speakers and acoustic environments. Eventually, as discussed above, we believe that the ALLIES corpus can benefit the speaker verification community and plan to release a speaker verification protocol on this dataset.

[3] `https://git-lium.univ-lemans.fr/Larcher/evallies`

References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals, Speaker diarization: A review of recent research, IEEE Transactions on Audio, Speech, and Language Processing 20 (2) (2012) 356–370.

[2] S. E. Tranter, D. A. Reynolds, An overview of automatic speaker diarization systems, IEEE Transactions on audio, speech, and language processing 14 (5) (2006) 1557–1565.

[3] L. El Shafey, H. Soltau, I. Shafran, Joint speech recognition and speaker diarization via sequence transduction, Annual Conference of the International Speech Communication Association (INTERSPEECH) (2019) 396–400.

[4] D. A. Reynolds, P. Torres-Carrasquillo, Approaches and applications of audio diarization, in: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., Vol. 5, IEEE, 2005, pp. v–953.

[5] H. H. Mao, S. Li, J. McAuley, G. W. Cottrell, Speech recognition and multi-speaker diarization of long conversations, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2020, pp. 691–695.

[6] M. McLaren, L. Ferrer, D. Castan, A. Lawson, The speakers in the wild (sitw) speaker recognition database., in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2016, pp. 818–822.

[7] S. O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, J. Hernandez-Cordero, The 2019 nist speaker recognition evaluation cts challenge, in: The Speaker and Language Recognition Workshop (Odyssey, Vol. 2020, 2020, pp. 266–272.

[8] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, H. Na, Ecapa-tdnn embeddings for speaker diarization, arXiv preprint arXiv:2104.01466.

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

[10] M. Diez, L. Burget, P. Matejka, Speaker diarization based on bayesian hmm with eigenvoice priors., in: The Speaker and Language Recognition Workshop (Odyssey, 2018, pp. 147–154.

[11] F. Landini, J. Profant, M. Diez, L. Burget, Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: Theory, implementation and analysis on standard tasks, Computer Speech & Language 71 (2022) 101254.

[12] Z.-H. Tan, N. Dehak, et al., rvad: An unsupervised segment-based robust voice activity detection method, Computer speech & language 59 (2020) 1–21.

[13] J. Kim, M. Hahn, Voice activity detection using an adaptive context attention model, IEEE Signal Processing Letters 25 (8) (2018) 1181–1185.

[14] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, 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.

[15] R. Yin, H. Bredin, C. Barras, Neural speech turn segmentation and affinity propagation for speaker diarization, in: Annual Conference of the International Speech Communication Association, 2018.

[16] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, S. Watanabe, End-to-end neural speaker diarization with self-attention, in: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019, pp. 296–303.

[17] S. Maiti, H. Erdogan, K. Wilson, S. Wisdom, S. Watanabe, J. R. Hershey, End-to-end diarization for variable number of speakers with local-global networks and discriminative speaker embeddings (2021). arXiv:2105.02096.

[18] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, L. Quintard, The repere corpus: a multimodal corpus for person recognition., in: International Conference on Language Resources and Evaluation (LREC), 2012, pp. 1102–1107.

[19] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, 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.

[20] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, 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.

[21] G. Le Lan, D. Charlet, A. Larcher, S. Meignier, An adaptive method for cross-recording speaker diarization, IEEE/ACM Transactions on Audio, Speech, and Language Processing 26 (10) (2018) 1821–1832.

[22] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern recognition 45 (1) (2012) 521–530.

[23] H. Aronowitz, Inter dataset variability compensation for speaker recognition, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 4002–4006.

[24] P.-M. Bousquet, M. Rouvier, On robustness of unsupervised domain adaptation for speaker recognition, in: InterSpeech, 2019.

[25] H.-J. Chang, H.-y. Lee, L.-s. Lee, Towards lifelong learning of end-to-end asr, arXiv preprint arXiv:2104.01616.

[26] G. Le Lan, D. Charlet, A. Larcher, S. Meignier, Iterative plda adaptation for speaker diarization, in: Interspeech 2016, Vol. 2016, 2016, pp. 2175–2179.

[27] Z. Chen, B. Liu, Lifelong machine learning, Synthesis Lectures on Artificial Intelligence and Machine Learning 12 (3) (2018) 1–207.

[28] B. Liu, Lifelong machine learning: a paradigm for continuous learning, Frontiers of Computer Science 11 (3) (2017) 359–361.

[29] M. Wang, X.-S. Hua, Active learning in multimedia annotation and retrieval: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (2) (2011) 1–21.

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

[31] T. Pfau, D. P. Ellis, A. Stolcke, Multispeaker speech activity detection for the icsi meeting recorder, in: IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01., IEEE, 2001, pp. 107–110.

[32] E. Rentzeperis, A. Stergiou, C. Boukis, A. Pnevmatikakis, L. C. Polymenakos, The 2006 athens information technology speech activity detection and speaker diarization systems, in: International Workshop on Machine Learning for Multimodal Interaction, Springer, 2006, pp. 385–395.

[33] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, P. Matějka, Developing a speech activity detection system for the darpa rats program, in: Thirteenth annual conference of the international speech communication association, 2012.

[34] N. Ryant, M. Liberman, J. Yuan, Speech activity detection on youtube using deep neural networks., in: INTERSPEECH, Lyon, France, 2013, pp. 728–731.

[35] V. Pannala, B. Yegnanarayana, A neural network approach for speech activity detection for apollo corpus, Computer Speech & Language 65 (2021) 101137.

[36] S. Shahsavari, H. Sameti, H. Hadian, Speech activity detection using deep neural networks, in: 2017 Iranian Conference on Electrical Engineering (ICEE), IEEE, 2017, pp. 1564–1568.

[37] G. Gelly, J.-L. Gauvain, Minimum word error training of rnn-based voice activity detection, in: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[38] G. Schwarz, et al., Estimating the dimension of a model, The annals of statistics 6 (2) (1978) 461–464.

[39] J.-L. Gauvain, L. F. Lamel, G. Adda, Partitioning and transcription of broadcast news data, in: Fifth International Conference on Spoken Language Processing, 1998.

[40] L. V. Neri, H. N. Pinheiro, R. Tsang, G. D. d. C. Cavalcanti, A. G. Adami, Speaker segmentation using i-vector in meetings domain, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 5455–5459.

[41] M. Hrúz, Z. Zajíc, Convolutional neural network for speaker change detection in telephone speaker diarization system, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4945–4949.

[42] C. Barras, X. Zhu, S. Meignier, J.-L. Gauvain, Multistage speaker diarization of broadcast news, IEEE Transactions on Audio, Speech, and Language Processing 14 (5) (2006) 1505–1512.

[43] S. H. Shum, N. Dehak, R. Dehak, J. R. Glass, Unsupervised methods for speaker diarization: An integrated and iterative approach, IEEE Transactions on Audio, Speech, and Language Processing 21 (10) (2013) 2015–2028.

[44] D. Dimitriadis, P. Fousek, Developing on-line speaker diarization system, in: Annual Conference of the International Speech Communication Association (INTER-

SPEECH), 2017, pp. 2739–2743.
[45] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, I. L. Moreno, Speaker diarization with lstm, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5239–5243.
[46] S. Meignier, T. Merlin, Lium spkdiarization: an open source toolkit for diarization, in: CMU SPUD Workshop, 2010.
[47] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, A. McCree, Speaker diarization using deep neural network embeddings, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4930–4934.
[48] G. Sell, D. Snyder, A. McCree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, et al., Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge., in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2018, pp. 2808–2812.
[49] Y. Wang, M. He, S. Niu, L. Sun, T. Gao, X. Fang, J. Pan, J. Du, C.-H. Lee, Ustc-nelslip system description for dihard-iii challenge, arXiv preprint arXiv:2103.10661.
[50] J. Luque, J. Hernando, On the use of agglomerative and spectral clustering in speaker diarization of meetings, in: The Speaker and Language Recognition Workshop (Odyssey), 2012.
[51] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, P. Dumouchel, Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification, in: Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH), 2009, p. 1559–1562.
[52] E. Variani, X. Lei, E. McDermott, I. L. Moreno, J. Gonzalez-Dominguez, Deep neural networks for small footprint text-dependent speaker verification, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 4052–4056.
[53] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, C. Wang, Fully supervised speaker diarization, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 6301–6305.
[54] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust dnn embeddings for speaker recognition, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5329–5333.
[55] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, M. Liberman, The third dihard diarization challenge, arXiv preprint arXiv:2012.01477.
[56] S. Horiguchi, N. Yalta, P. Garcia, Y. Takashima, Y. Xue, D. Raj, Z. Huang, Y. Fujita, S. Watanabe, S. Khudanpur, The hitachi-jhu dihard iii system: Competitive end-to-end neural diarization and x-vector clustering systems combined by dover-lap, arXiv preprint arXiv:2102.01363.
[57] M. Ravanelli, Y. Bengio, Speaker recognition from raw waveform with sincnet, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 1021–1028.
[58] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, H.-J. Yu, Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms, Proc. Interspeech 2020 (2020) 1496–1500.
[59] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, A. Cristia, M.-P. Gill, L. P. Garcia-Perera, End-to-end domain-adversarial voice activity detection, in: Interspeech 2020, 2020.
[60] G. Dupuy, S. Meignier, Y. Esteve, Is incremental cross-show speaker diarization efficient for processing large volumes of data?, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2014, pp. 587–591.
[61] Q. Yang, Q. Jin, T. Schultz, Investigation of cross-show speaker diarization, in: Twelfth Annual Conference of the International Speech Communication Association, 2011.
[62] V.-A. Tran, V. Le, C. Barras, L. Lamel, Comparing multi-stage approaches for cross-show speaker diarization, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011.
[63] A. Canavan, D. Graff, G. Zipperlen, CALLHOME American English Speech LDC97S42, Linguistic Data Consortiumdoi:doi.org/10.35111/exq3-x930.
[64] S. Renals, T. Hain, H. Bourlard, Recognition and understanding of meetings the ami and amida projects, in: 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), IEEE, 2007, pp. 238–247.
[65] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, J. Li, Continuous speech separation: dataset and analysis, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 7284–7288.
[66] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The first dihard speech diarization challenge, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2018.
[67] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, M. Liberman, Third DIHARD challenge evaluation plan, arXiv preprint arXiv:2006.05815.
[68] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, C. Yu, Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon., in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2018, pp. 2758–2762.
[69] E. Lleida, A. Ortega, A. Miguel, V. Bazán-Gil, C. Pérez, M. Gómez, A. De Prada, Albayzin 2018 evaluation: the iberspeech-rtve challenge on speech technologies for spanish broadcast media, Applied Sciences 9 (24) (2019) 5412.
[70] J. Barker, S. Watanabe, E. Vincent, J. Trmal, The fifth CHiME speech separation and recognition challenge: dataset, task and baselines, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2018, pp. 1561–1565.
[71] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, A. Zisserman, Spot the conversation: speaker diarisation in the wild, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2020, pp. 299–303.
[72] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, et al., Aishell-4: An open source dataset for speech enhancement, separation, recognition

and speaker diarization in conference scenario, arXiv preprint arXiv:2104.03603.

[73] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, K. Choukri, The ester evaluation campaign for the rich transcription of french broadcast news., in: International Conference on Language Resources and Evaluation (LREC), 2004.

[74] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester, et al., The mgb challenge: Evaluating multi-genre broadcast media recognition, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 687–693.

[75] G. Riccardi, D. Hakkani-Tur, Active learning: Theory and applications to automatic speech recognition, IEEE transactions on speech and audio processing 13 (4) (2005) 504–511.

[76] H. Jiaji, C. Rewon, R. Vinay, L. Hairong, S. Sanjeev, 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.

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

[78] E. Yilmaz, M. McLaren, H. van den Heuvel, 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.

[79] D. G. Karakos, S. Novotney, L. Z. 0002, 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.

[80] M. Abdelwahab, 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.

[81] C. Yu, J. H. Hansen, Active learning based constrained clustering for speaker diarization, IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (11) (2017) 2188–2198.

[82] B. Mateusz, J. Poignant, L. Besacier, 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, 2014.

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

[84] 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.

[85] H. Bredin, pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems, in: 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 2017.

[86] N. Mirghafori, C. Wooters, Nuts and flakes: A study of data characteristics in speaker diarization, in: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Vol. 1, IEEE, 2006, pp. I–I.

[87] M. Huijbregts, C. Wooters, The blame game: Performance analysis of speaker diarization system components, in: Eighth Annual Conference of the International Speech Communication Association, 2007.

[88] D. Charlet, J. Poignant, H. Bredin, C. Fredouille, S. Meignier, What makes a speaker recognizable in tv broadcast? going beyond speaker identification error rate, 2015.

[89] M. E. Wood, E. Lewis, Windmill-the use of a parsing algorithm to produce predictions for disabled persons, Vol. 18, Citeseer, 1996, pp. 315–322.

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

[91] Y. Prokopalo, M. Shamsi, L. Barrault, S. Meignier, A. Larcher, Active correction for speaker diarization with human in the loop, in: Proc. IberSPEECH 2021, 2021, pp. 260–264.

[92] P.-A. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, S. Meignier, S4d: Speaker diarization toolkit in python, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2018, pp. 1368–1372.

[93] A. Mehrish, M. Tahon, N. Evans, A. Larcher, Incremental adaptation of speaker recognition system with membership and bias assessment, in: submitted to Workshop on Automatic Speech Recognition and Understanding (ASRU), 2021.

[94] A. Larcher, A. Mehrish, M. Tahon, S. Meignier, J. Carrive, D. Doukhan, O. Galibert, 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.

[95] D. Garcia-Romero, G. Sell, A. Mccree, MagNetO: X-vector Magnitude Estimation Network plus Offset for Improved Speaker Recognition, in: Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, 2020, pp. 1–8. doi:10.21437/Odyssey.2020-1.
URL http://dx.doi.org/10.21437/Odyssey.2020-1

[96] A. Nagrani, J. S. Chung, A. Zisserman, Voxceleb: A large-scale speaker identification dataset, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), 2017, pp. 2616–2620.

[97] A. Nagrani, J. S. Chung, W. Xie, A. Zisserman, Voxceleb: Large-scale speaker verification in the wild, Computer Speech & Language 60 (2020) 101027.

[98] A. Larcher, K. A. Lee, B. Ma, 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.