An active learning method for speaker identity annotation in audio recordings

Pierre-Alexandre Broux, David Doukhan, Simon Petitrenaud, Sylvain Meignier, Jean Carrive

An active learning method for speaker identity annotation in audio recordings

Broux Pierre-Alexandre[1,2] and **Doukhan David**[1] and **Petitrenaud Simon**[2] and **Meignier Sylvain**[1] and **Carrive Jean**[2]

Abstract. Given that manual annotation of speech is an expensive and long process, we attempt in this paper to assist an annotator to perform a speaker diarization. This assistance takes place in an annotation background for a large amount of archives. We propose a method which decreases the intervention number of a human. This method corrects a diarization by taking into account the human interventions. The experiment is done using French broadcast TV shows drawn from ANR-REPERE evaluation campaign. Our method is mainly evaluated in terms of KSR (Keystroke Saving Rate), and we reduce the number of actions needed to correct a speaker diarization output by 6.8% in absolute value.

1 Introduction

The work presented in this paper has been realized to meet the needs of the French national audiovisual institute[3] (INA). INA is a public institution in charge of the digitalization, preservation, distribution and dissemination of the French audiovisual heritage. Annotations related to speaker identity, together with speech transcription, meet several use-cases. Temporal localization of speaker interventions can be used to enhance the navigation within a media [12, 22]. It may also be used to perform complex queries within media databases [5, 11, 19].

This article focuses on the realization of human-assisted *speaker diarization* systems. Speaker diarization methods consist in estimating "who spoke when" in an audio stream [2]. This media structuring process is an efficient pre-processing step, for instance to help segmenting a broadcast news into anchors and reports before manual documentation processes. Speaker diarization algorithms are generally based on unsupervised machine learning methods [21], in charge of estimating the number of speakers, and splitting the audio stream into labelled speech segments assigned to hypothesized speakers. Speaker identity and temporal localization is known to be a pertinent information for the access and exploitation of speech recordings [5, 20]. However, the accuracy of automatic state-of-the-art speaker recognition methods is still inadequate to be embedded into INA's archiving or media enhancement applications, and a human intervention is required to obtain an optimal description of a speech archive.

Manual annotation of speech is a very expensive process. Nine hours are required to perform the manual annotation corresponding to one hour of spontaneous speech (speech transcription and speaker identity). Previous studies have shown that the speech annotation process may be sped-up using the output of automatic speech recognition systems (ASR) together with speech turn annotations [3]. The resulting annotation task consists in correcting the output of automatic systems, instead of doing the whole annotation manually.

The model proposed in this paper is an active-learning extension of this paradigm, applied to the *speaker diarization* task. Annotator corrections are used in real-time to update the estimations of the speaker diarization system. The aim of this update strategy is to lower the amount of manual corrections to be done, which impact the time spent in the interaction with the system. The quality of the annotations obtained through this process should be maximal, with respect to human abilities on speaker recognition tasks [13].

The paper is organized as follows: Section 2 presents the Human-assisted speaker diarization system. Section 3 presents the corpus, the metrics, whereas section 4 analyzes the results. Section 5 concludes with a discussion of possible directions for future works.

2 Human-assisted speaker diarization system

The proposed speaker diarization prototype is aimed at interacting in real-time with a human user, in charge of correcting the predictions of the system. This system is aimed at producing high quality diarization annotations with a minimal human cost. Such system could be used to speed-up the annotation process of any speech corpus requiring temporal speaker information.

2.1 System overview

In the following description, we assume that an easy-to-use interface is provided to the user, and that the speech segments are presented together with the speech transcription. We also assume that the feedback of the user is limited to three actions:

1. The validation, when the speech segment has a correct speaker label;
2. The speaker label modification, when the speech segment has an incorrect speaker label;
3. The speaker label creation: for speakers encountered for the first time in the recording.

Actions such as speech segment split, or speech segment boundaries modifications are not taken into account in the scope of this paper.

Annotated speech segments corresponding to the whole recording are presented to the annotator. The segment presentation order follows the temporal occurrence of the segments. This choice has been made in order to ease the manual speaker recognition task, with the assumption that the media chronology provides the annotator with a

[1] Computer science laboratory of the university of Maine (LIUM - EA 4023), Le Mans, France
[2] French National audiovisual institute (Ina), Paris, France
[3] `http://www.ina.fr`

better understanding of the speech material. The annotator has to correct, or validate the predictions of the diarization system. Our working paradigm is that a correction requires more time for the annotator than a validation.

Figure 1 describes the proposed active-learning system. The system consists in associating each annotator correction to a real-time re-estimation of the labels of the remaining speech segments to be presented. This method is aimed at improving the quality of the next diarization predictions, resulting in a lower amount of corrections to be done by the annotator, thus lowering the time required for the manual correction. The system is composed of three main steps, which will be detailed in the next sections. The two last steps are repeated until all the segments are checked. Let us give a brief description of these stages:

Initialization: an initial diarization is performed with a fully-automatic speaker diarization system. This step can be time consuming and is performed offline.

User input: the annotator checks each segment, and validates or corrects the speaker label before inspecting the next segment.

Real-time reassignment: the annotator modifications are associated to a re-evaluation of the speaker labels corresponding to the next speech segments to be presented. The computations realized during this step should be fast enough to allow real-time interaction with a human user.

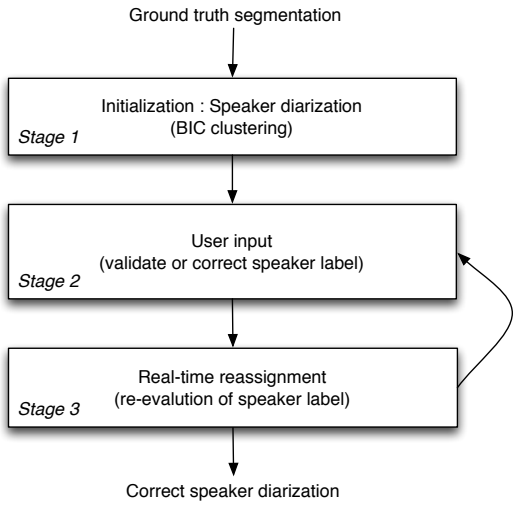

Figure 1: Active-learning system

2.2 Initialization: speaker diarization

The speaker diarization system is inspired by the system described in [2]. It was developed for the transcription and diarization tasks, with the goal of minimizing both word error rate and speaker error rate. It rests upon a segmentation and a hierarchical agglomerative clustering. Furthermore, this system uses MFCC features as audio descriptors [2, 7, 17].

The system is composed of a segmentation step followed with a clustering step. Speaker diarization needs to produce homogeneous speech segments. Errors such as having two distinct clusters (i.e., detected speakers) corresponding to the same real speaker could be easily corrected by merging both clusters. In this article, we focus the study on the clustering step and the segmentation step is based on a perfect manual segmentation (ground truth).

The clustering algorithm is based upon a hierarchical agglomerative clustering. The initial set of clusters is composed of one segment per cluster. Each cluster is modeled by a Gaussian with a full covariance matrix. The ΔBIC measure (cf equation 1) is employed to select the candidate clusters to group as well as to stop the merging process. The two closest clusters i and j are merged at each iteration until $\Delta BIC(i,j) > 0$.

Let $|\Sigma_i|$, $|\Sigma_j|$ and $|\Sigma|$ be the determinants of gaussians associated to the clusters i, j and $i+j$ and λ be a parameter to set up. The penalty factor P (eq. 2) depends on d, the dimension of the features, as well as on n_i and n_j, referring to the total length of cluster i and cluster j respectively. The $\Delta BIC(i,j)$ measure between the clusters i and j is then defined as follows:

$$\Delta BIC(i,j) = \frac{n_i+n_j}{2}\log|\Sigma| - \frac{n_i}{2}\log|\Sigma_i| - \frac{n_j}{2}\log|\Sigma_j| - \lambda P, \quad (1)$$

$$\text{with } P = \frac{1}{2}\left(d + \frac{d(d+1)}{2}\right) + log(n_i+n_j). \quad (2)$$

This speaker diarization system is the first stage of most state-of-the-art systems for TV or radio recording as the one based on GMM or i-vectors[1, 8]. GMM and i-vectors are both statistical models which represent audio data. The generated clusters have a high purity (i.e. each cluster contains mostly only one speaker) and the system is fast.

2.3 User input and Real-time reassignment

User input consists in validating, or correcting, the speaker labels estimated by the diarization system. The proposed active-learning strategy consists in associating each correction, defined as a mismatch between the speakers C_i and C_j, to the computation of new speaker models, trained on the validated speech segments. The resulting models are based on a single gaussian, which is fast to compute, and assumed to be more accurate than the models inferred during the initialization. These simple speaker models are then used to re-estimate the ΔBIC distance with the remaining speech segments involved to the last mismatch (segments attributed to C_i and C_j only).

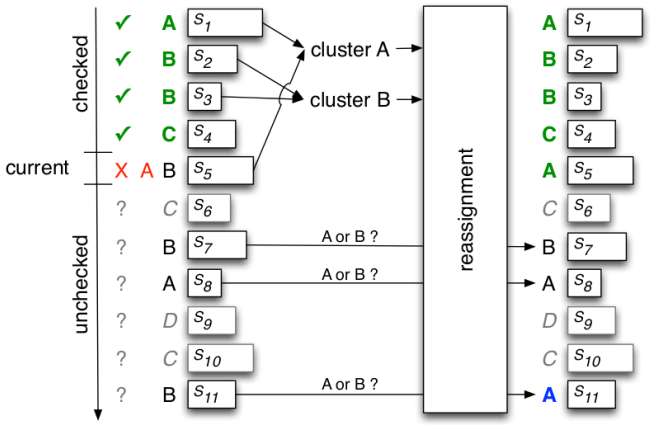

Figure 2: Example of user-input and reassignment

An illustration of these interactions is provided in figure 2. In this example, four speakers (A, B, C, D) have been inferred through the automatic initialization step. The user has manually validated the four first speech segments ($S_1 ... S_4$) before reporting a speaker label modification for segment S_5, tagged as speaker B instead of speaker A. The resulting action of the active-learning system, consists to create speaker models for the mismatching speakers only (A and B). These models are used to re-estimate the labels of the remaining segments tagged with A or B (segments S_7, S_8 and S_{11}), and may lead to a speaker label modification (segment S_{11}). Remaining speech segments tagged with other labels (C and D) are not re-estimated. The modified diarization is updated before the annotator moves to the next segment S_6. The process iterates until the last segments are reached.

3 Evaluation

3.1 Corpus

Experiments were performed on TV recordings drawn from the corpora of ANR-REPERE challenge[4] . The ANR-REPERE is a challenge organized by the LNE (French national laboratory of metrology and testing) and ELDA (Evaluations and Language resources Distribution Agency) in 2010-2014. This challenge is a project in the area of the multimedia recognition of people in television documents. The aim is to find the identities of people who speak along with the quoted and written names at each instant in a television show. The data comes from two French channels (BFM and LCP). Shows were recorded from two French digital terrestrial television channels.

The ANR-REPERE project has started since 2010 and evaluations are set up in 2013 and 2014. In this paper, we merge the 2013 evaluation corpus and the 2014 evaluation corpus to build the corpus called REPERE in the below sections. The table 1 give us some statistics about this corpus. The duration reported in table 1 shows that only a part of the data is annotated and evaluated.

Statistics	REPERE
Show number	15
Recording number	90
Recording time	34h30
Annotation time	13h11
Speaker number	571

Table 1: 2013-2014 news and debate TV recordings from REPERE corpus.

The current diarization systems are less efficient with spontaneous speech mainly present in debates than with prepared speech from news [6]. We have chosen this corpus because of the variety of the shows. The corpus is balanced between prepared and spontaneous speech and composed of street interviews, debates and news shows.

It is common to accept a ± 250 millisecond tolerance on segment boundaries for the recordings with prepared speech and far less for the recordings with spontaneous speech. Having and using a reference segmentation for the segmentation step, we do not normally have segmentation errors. Therefore, we do not use any tolerances on segment boundaries.

Most of the diarization systems are not able to detected overlap speech zones [4, 16, 24]. In the following described experiments, we remove overlap speech from the evaluation and consider it as a non-speech area. Figure 3 shows the segment duration after the superposed speech deletion.

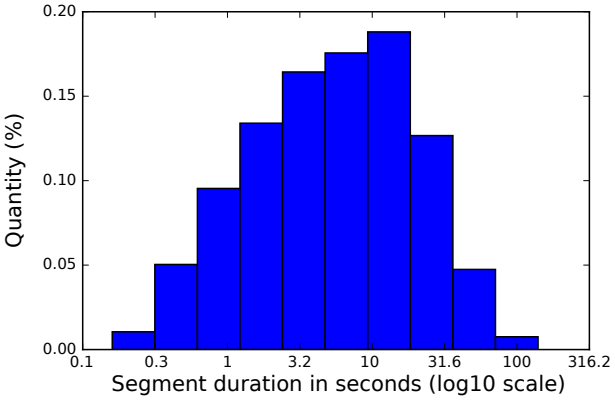

Figure 3: Segment duration of REPERE corpus

3.2 Metrics

3.2.1 Diarization

The metric used to measure performance in the speaker diarization task is the Diarization Error Rate (DER) [18]. DER was introduced by NIST as the fraction of speaking time which is not attributed to the correct speaker, using the best matching between references and hypothesis speaker labels. The scoring tool is available in the $sidekit/s4d$ toolkit[14].

In order to evaluate the impact of a reassignment, we use the percentage of pure clusters with respect to the total number of clusters. We also use the well-known purity as defined in [9] which is the ratio between the number of frames by the dominating speaker in a cluster and the total number of frames in this cluster. This measure is used in order to evaluate the purity of hypothesis clusters according to the assignment provided by reference clusters. To evaluate the action applied by a human, we simply use some counters. These counters will be in the form of percentages in this paper.

3.2.2 Keystroke Saving Rate

The DER and the purity measure the quality of a diarization. The evaluation of the user input is difficult, as the proposed metric needs to be as much as possible reproducible and objective [10]. In our case, the human interactions are simulated.

The proposed method is inspired from a previous work on computer assisted transcription [15]. In this paper the authors proposed to evaluate the human interactions with the Keystroke Saving Rate (KSR) [23].

The KSR method has been developed for AAC (Augmentative and Alternative Communication) systems, so that handicapped persons can use it. It is computed according to the number of keyboard strokes made by the user to write a message. In our case, the strokes corresponds to the number of actions made by the annotator to correct the diarization. To compute the KSR, we assume that the annotator will always choose the best strategy to minimize the number of actions.

[4] `http://www.defi-repere.fr/`

We suppose here that the annotator can make two kinds of actions for a current segment: the reassignment to another cluster (reassignment) or the assignment to a new cluster (creation). The annotator can create a new cluster when the first segment of a given speaker is checked. The number of creations in the whole document, denoted by n_c, is constant for any reassignment even if the threshold λ in equation 1 differs. Similarly the total number of segments reassigned by the user is denoted by n_r and the number of segments is n_s. We define the KSR as the ratio of the sum of the numbers of created clusters and the reassigned segments n_r given the number of segments in the initial diarization (equation 3):

$$KSR = \frac{n_c + n_r}{n_s} \times 100. \quad (3)$$

A KSR equal to 0% corresponds to a perfect speaker diarization in which each segment is assigned to the true corresponding speaker. In this case, the annotator does not reassign any segments. Conversely, a KSR equal to 100% corresponds to the worse speaker diarization in which each segment is assigned to the wrong speaker. Therefore, the annotator needs to change the assignment of all the segments, if the corrections are not gradually propagated in the rest of the document.

4 Results

4.1 Speaker diarization

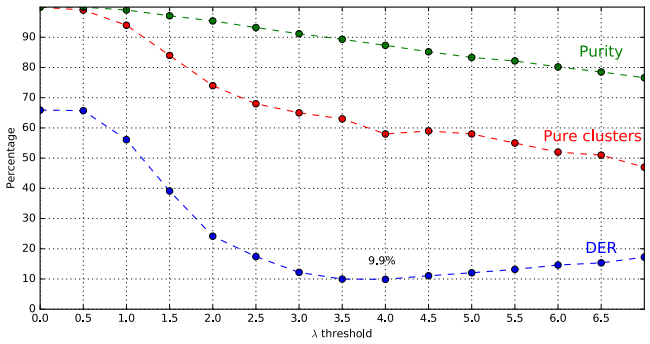

Figure 4: Initial diarization: DER, % of pure clusters and average cluster purity

The speaker diarization is based on hierarchical clustering where each speaker is modeled by a gaussian with a full covariance computed over acoustic features. The acoustic features are composed of 12 MFCCs with energy, and are not normalized (the background channel helps to segment and cluster the speaker) [2, 7, 17].

As mentioned previously, we use the ground truth segmentation as input of the clustering algorithm (corresponding to the stage 1 in figure 1) and the overlapping speaker segments are removed in the ground truth.

Figure 4 shows the DER of the speaker diarization for different λ thresholds (cf. equation 1). The lower DER is 9.9% for a λ threshold of 4.0. Compared to literature [8], this DER is rather low, which is mainly due to ground truth segmentation: the segments contain the voice of a single speaker, overlap segments are removed, as well as there are no missed speech and no false alarm speech segments.

4.2 Active-learning system

In our experiments, the human annotator is simulated with the ground truth speaker annotations. The main objective is to decrease the number of actions performed by an annotator to obtain a perfect diarization. To reach this goal, we compare the KSR obtained with or without the human corrections taken into consideration (i.e. with or without an active-learning reassignment) using various λ thresholds for the speaker diarization.

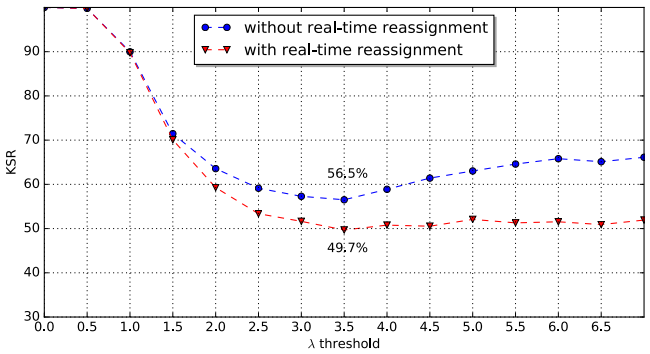

Figure 5: KSR

The real-time segment reassignment stage (stage 3 in figure 1) uses the same parameters as the initial diarization: 12MFCC+energy, full covariance gaussian and BIC metric to label the unchecked segments
.

Figure 5 gives the KSR of the system with real-time reassignment (including stages 2 & 3) and the system without real-time reassignment (including stage 2 only). The KSR decreases until $\lambda = 3.5$ in both systems and increases when λ is upper. The KSR is 56.5% and 49.7% respectively without reassignment and with reassignment when λ is equal to 3.5. About half segments are manually corrected (49.7%) and the 6.8% in absolute value are reassigned to the correct speaker automatically after a user correction.

In the most favorable case when λ is at 3.5, the DER is low, about 10% and the average cluster purity is equal to 90%. In the same time, only 60% of the clusters are 100% pure (cf. figure 4). The difference between these indicators can be explained by the fact that, unlike the DER, the KSR does not take into account the duration of the segments . Most of the errors come from the small segments, and these ones are numerous (cf. figure 3).

The KSR remains almost static when λ is greater than 3.5 in the system with reassignment, whereas the choice of the parameter λ is more critical to minimize the number of actions in the system without reassignment. Finally, one can notice that the system with reassignment always obtains a lower KSR whatever the λ value, except for $\lambda = 0$ where the KSR is equal to 100% in both cases.

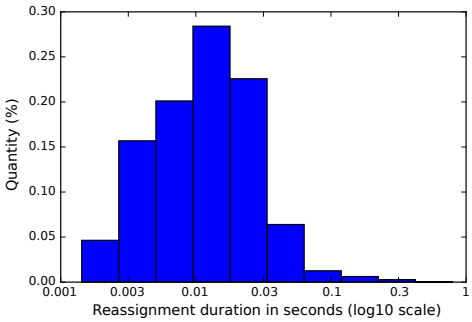

Figure 6: Time of reassignment after each user correction.

After each user correction, the unchecked segments are clustered again in the reassignment stage. The process is generally fast, since the duration takes less than 0.03 second in 95% of cases, so it is interesting to notice that this stage could be done in real time without any impact on the user interface (figure 6).

5 Conclusion & prospects

In this paper, we attempt to find a way to help a human to segment and cluster the speakers in an audio or audio-visual document. We propose a method that takes into consideration the annotator corrections by modifying the allocation of the unchecked segments. The proposed computer assisted method allows us to obtain a noticeable reduction in the number of required corrections. Not only is our method effective, but the corrections are also made quickly. Thanks to its fast treatment, this could be applied in a real application without impacting the reactivity of the interface and without increasing the work intensity of the annotator.

Some future improvements should be done on the base of this preliminary work. Firstly, we plan to minimize the number of user actions by applying a constrained clustering to reassign all unchecked segments and to create or delete clusters. Another improvement would be to integrate the automatic segmentation in the correction process.

6 Acknowledgments

This research was partially supported by the European Commission, as part of the Event Understanding through Multimodal Social Stream Interpretation (EUMSSI) project (contract number FP7-ICT-2013-10) in which the LIUM is involved.

References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, 'Speaker diarization: A review of recent research', **20**(2), 356–370, (Feb 2012).

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

[3] Thierry Bazillon, Yannick Estève, and Daniel Luzzati, 'Transcription manuelle vs assistée de la parole préparé et spontanée', *Revue TAL*, (2008).

[4] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, 'Overlapped speech detection for improved speaker diarization in multiparty meetings', in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4353–4356. IEEE, (2008).

[5] Mbarek Charhad, Daniel Moraru, Stéphane Ayache, and Georges Quénot, 'Speaker identity indexing in audio-visual documents', in *Content-Based Multimedia Indexing (CBMI2005)*, (2005).

[6] Ruchard Dufour, Vincent Jousse, Yannick Estève, Fréderic Béchet, and Georges Linarès, 'Spontaneous speech characterization and detection in large audio database', *SPECOM, St. Petersburg*, (2009).

[7] Grégor Dupuy, *Les collections volumineuses de documents audiovisuels: segmentation et regroupement en locuteurs*, Ph.D. dissertation, Université du Maine, 2015.

[8] Grégor Dupuy, Sylvain Meignier, Paul Deléglise, and Yannick Esteve, 'Recent improvements on ilp-based clustering for broadcast news speaker diarization', in *Proc. Odyssey Workshop*, (2014).

[9] Jean-Luc Gauvain, Lori Lamel, and Gilles Adda, 'Partitioning and transcription of broadcast news data.', in *ICSLP*, volume 98, pp. 1335–1338, (1998).

[10] Edouard Geoffrois, 'Evaluating interactive system adaptation', in *The International Conference on Language Resources and Evaluation*, (2016).

[11] Jerry Goldman, Steve Renals, Steven Bird, Franciska De Jong, Marcello Federico, Carl Fleischhauer, Mark Kornbluh, Lori Lamel, Douglas W Oard, Claire Stewart, et al., 'Accessing the spoken word', *International Journal on Digital Libraries*, **5**(4), 287–298, (2005).

[12] Nicolas Hervé, Pierre Letessier, Mathieu Derval, and Hakim Nabi, 'Amalia.js: An open-source metadata driven html5 multimedia player', in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, MM '15, pp. 709–712, New York, NY, USA, (2015). ACM.

[13] Juliette Kahn, *Parole de locuteur: performance et confiance en identification biométrique vocale*, Ph.D. dissertation, Avignon, 2011.

[14] Anthony Larcher, Kong Aik Lee, and Sylvain Meignier, 'An extensible speaker identification sidekit in python', in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5095–5099. IEEE, (2016).

[15] Antoine Laurent, Sylvain Meignier, Teva Merlin, and Paul Deléglise, 'Computer-assisted transcription of speech based on confusion network reordering', in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 4884–4887. IEEE, (2011).

[16] Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, 'Speaker diarization: A review of recent research', *Audio, Speech, and Language Processing, IEEE Transactions on*, **20**(2), 356–370, (2012).

[17] Lindasalwa Muda, Mumtaj Begam, and I Elamvazuthi, 'Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques', *arXiv preprint arXiv:1003.4083*, (2010).

[18] NIST. The rich transcription spring 2003 (RT-03S) evaluation plan. `http://www.itl.nist.gov/iad/mig/tests/rt/2003-spring/docs/rt03-spring-eval-plan-v4.pdf`, February 2003.

[19] Roeland Ordelman, Franciska De Jong, and Martha Larson, 'Enhanced multimedia content access and exploitation using semantic speech retrieval', in *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*, pp. 521–528. IEEE, (2009).

[20] Julien Pinquier and Régine André-Obrecht, 'Audio indexing: primary components retrieval', *Multimedia tools and applications*, **30**(3), 313–330, (2006).

[21] Sue E Tranter and Douglas A Reynolds, 'An overview of automatic speaker diarization systems', *IEEE Transactions on Audio, Speech, and Language Processing*, **14**(5), 1557–1565, (2006).

[22] Félicien Vallet, Jim Uro, Jérémy Andriamakaoly, Hakim Nabi, Mathieu Derval, and Jean Carrive, 'Speech trax: A bottom to the top approach for speaker tracking and indexing in an archiving context', in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), (2016).

[23] Matthew EJ Wood and Eric Lewis, 'Windmill-the use of a parsing algorithm to produce predictions for disabled persons', *PROCEEDINGS-INSTITUTE OF ACOUSTICS*, **18**, 315–322, (1996).

[24] Martin Zelenák and Javier Hernando, 'The detection of overlapping speech with prosodic features for speaker diarization.', in *INTERSPEECH*, pp. 1041–1044, (2011).