What Makes a Speaker Recognizable in TV Broadcast? Going Beyond Speaker Identification Error Rate

Delphine Charlet[1], *Johann Poignant*[2], *Hervé Bredin*[2], *Corinne Fredouille*[3], *Sylvain Meignier*[4]

[1] Orange Labs – Lannion, France
[2] LIMSI – CNRS – Orsay, France.
[3] CERI/LIA – University of Avignon – Avignon, France
[4] LIUM – Université du Mans – Le Mans, France
delphine.charlet@orange.com

Abstract

Speaker identification approaches for TV broadcast are usually evaluated and compared based on global error rates derived from the overall duration of missed detection, false alarm and confusion. Based on the analysis of the output of the systems submitted to the final round of the French evaluation campaign REPERE, this paper highlights the fact that these average metrics lead to the incorrect intuition that current state-of-the-art algorithms partially recognize all speakers. Setting aside incorrect diarization and adverse acoustic conditions, we show that their performance is in fact essentially bi-modal: in a given show, either all speech turns of a speaker are correctly identified or none of them are. We then proceed with trying to understand and explain this behavior, through perfomance prediction experiments. These experiments show that the most discriminant speaker characteristics are – first – their total speech duration in the current show and – then only – the amount of training data available to build their acoustic model.

Index Terms: speaker recognition, error analysis, TV broadcast

1. Introduction

For about five years, tremendous progress has been made in the speaker recognition field, especially for the speaker verification task in a phone environment. Supported by the evaluation campaigns organized by the National Institute of Standards and Technology (NIST)[1, 2], this progress mainly relies on the development of the i-vector paradigm, which has definitely overtaken classical UBM/GMM [3], or SVM [4] approaches. Inspired from the Joint Factor Analysis (JFA), which had already been applied with success in speaker detection, and which aims at estimating speaker and channel/session subspaces separately, the simpler and powerful i-vector-based modeling paradigm [5] makes no distinction between the two subspaces thanks to a single total variability space, which covers both the speaker and session/channel variability. A large amount of studies has been dedicated to the enhancement of this paradigm by coupling it with different channel compensation techniques [5, 6, 7], or by investigating various scoring approaches, which directly embed channel compensation [8, 5, 9, 10, 11].

With regards to performance analysis, studies also mainly focus on speaker verification task. In this framework, [12] have proposed a typology of speakers, using a menagerie lexicon, based on the observed properties of speakers to be more or less prone to miss detection or false alarm. Besides this typology, efforts have been made to identify and quantify the impact of the main factors that influence the performance of speaker verification. In [13], the authors report dramatic variations of performance, when varying the choice of the training session, for a similar amount of training data; similar trend is observed, on a lesser extent, for testing data. [14] investigate a range of variability factors, divided into "intrinsic" and "extrinsic" variations where "intrinsic" refers to internal speaker variability issues such as speech style and vocal efforts, and "extrinsic" refers to sources of variability external to the speaker, such as microphone, channels, noise. Their experiments show the strong impact of the variation of the speech style. In the field of speaker diarization for conference meetings, [15] have tried to identify the main factors contributing to errors, and to quantify their impact, through a set of oracle experiments. Their analysis showed that the speech detector, followed by the overlapped speech, were the main causes of errors.

Very few recent studies have concerned speaker identification in TV broadcast based on state-of-the-art speaker recognition systems. However, this context implies some non-trivial specificities such as the widely varying amount of training and testing data per speaker, the properties of speech segments - duration, number, acoustic quality, etc - implied in the identification decision while processing an entire TV show, generally issued from a preliminary speaker diarization step, and finally the coverage of speaker dictionary used by the system and its direct impact on an open-set identification task (closer to real life applications). In this paper, we propose to perform such a performance analysis, on the 3 systems submitted to the final round of the REPERE challenge [16], which has enabled the development of multimodal identification systems. Here, the analysis is restricted to the so-called "mono-modal" systems, which only use speaker voice to identify speakers. We are interested in analyzing the performance obtained individually for each speaker, and the influence of some of their characteristics (for instance in terms of speech turns duration, etc) on the obtained performance. In Section 2, the corpus, the systems and the evaluation metrics are introduced. In Section 3, the performance analysis is done, and the influence of some features on the performance is investigated in Section 4 through a performance prediction paradigm.

2. Experimental protocol

2.1. Evaluation metrics

We are interested in analysing speaker identification system performance, and particularly, the influence of some characteristics, in the training and testing data. Thus, one speaker in a given show is considered as the unit of analysis, the so-called

System	PERCOL	QCOMPERE	SODA
Diarization	two stage spk. diarization [17] + overlapping speech detection	multi-stage spk. diarization [18]	i-vector [19]
SID	ALIZE v3.0 toolkit [20]	Bob toolkit [21]	ALIZE v3.0 toolkit [20]
feature	19 LFCC + δ coef, + δ energy + 11 $\delta\delta$ coef	15 PLP-like cepstrum coef [22] 15 δ coef + δ energy	19 MFCC + δ coef, + δ energy + 11 $\delta\delta$ coef
UBM	gender independent 512 diagonal Gaussians	multi-lingual 256 diagonal Gaussians	gender independent 1024 diagonal Gaussians
i-vector	200 dim TVS estimated from 1200 spk. and 7500 sessions	400 dim TVS estimated from 39356 speech segments (around 15 seg./spk.)	300 dim TVS estimated from 680 spk. and 4150 sessions
Normalisation	cepstral mean subtraction and variance normalization	Feature warping normalization with a sliding window of 3 s. [23].	cepstral mean subtraction and variance normalization
Training data for i-vector	533 spk. id., min 30s, max 2mn30 (if higher, a set of i-vectors extracted) REPERE+ETAPE+French radio+web	706 spk. id., min 30s REPERE+ETAPE+French radio	680 spk. id., min 1mn , max 12 min REPERE+ETAPE+French radio+web
Decision	CDS joined with Within-Class covariance normalization for session/channel compensation	PLDA Eigen Factor Radial-based length normalization [24]	PLDA Eigen Factor Radial-based length normalization

Table 1: System comparison, TVS : total variability space, CDS: Cosine Distance Scoring

$SpkShow$. One speaker appearing in 2 different videos is considered as 2 distinct $SpkShow$.

In this analysis, we adopt the point of view of the references: for each $SpkShow_i$ in the reference, the performance metric of the biometric system is defined as the F-measure of the detection of $SpkShow_i$. More precisely, defining $T_i^{\text{reference}}$ the total duration of $SpkShow_i$ in the reference, $T_i^{\text{hypothesis}}$ the total duration where $SpkShow_i$ is the system response and T_i^{correct} the total duration of correct identification of $SpkShow_i$, Precision and Recall can be computed for each $SpkShow_i$:

- $\text{precision}_i = \frac{T_i^{\text{correct}}}{T_i^{\text{hypothesis}}} \qquad \text{recall}_i = \frac{T_i^{\text{correct}}}{T_i^{\text{reference}}}$
- $Fm_i = \frac{2 * \text{precision}_i * \text{recall}_i}{\text{precision}_i + \text{recall}_i}$

Thus, $Fm_i = 0$ means that $SpkShow_i$ was never correctly identified, whereas $Fm_i = 1$ means that $SpkShow_i$ is perfectly identified, without miss detection nor false alarm.

2.2. REPERE Corpus

The REPERE challenge [16] is an evaluation campaign on multimodal person recognition (phase 1 took place in January 2013 and phase 2 in January 2014). The systems evaluated in our experiments are the "mono-modal" systems (voice-based speaker identification) submitted to phase2, on `test2` data set, composed of 62 videos recorded from 8 different types of show (including news and talk shows) broadcasted by two French TV channels [25]. 10 hours of speech are annotated, and contain 477 non-anonymous $SpkShow$, which have on average 6.2 speech turns each, for a mean duration of speech turn equal to 12.1s.

2.3. System description

Monomodal speaker identification systems used in this paper for the error analysis were developed by the three research consortia involved in the REPERE challenge : PERCOL, QCOMPERE, and SODA. For all the systems, the speaker identification process relies on a typical i-vector framework, applied on speech segment clusters resulting from associated speaker diarization systems. Only SODA system [19] fully integrates the i-vector framework for both the speaker diarization and identification processes. Instead, PERCOL and QCOMPERE speaker diarization systems [18, 17] are based on a more standard multi-stage hierarchical clustering. Table 1 provides detailed information about system configuration individually. It is interesting to notice that QCOMPERE and SODA speaker identification systems follow very similar speaker modeling strategies. Indeed, they are based on about the same number of speaker models (706 and 680 respectively), for which only one i-vector is estimated if a minimum amount of training data is available (30s for QCOMPERE, 1mn for SODA). PERCOL system exhibits 533 speaker models only and proposes a different manner for extracting the corresponding i-vectors in order to take into account the possibly large amount of training data available for some speakers: (1) for a given speaker, an i-vector is extracted only if a minimum 30s long training data are available, (2) if the duration of training data for a given speaker is longer than 2mn30, a set of i-vectors is extracted, each of them on the basis of 2mn30 duration, the last one having to respect rule (1). Yet, for all the consortia, the initial training corpus made available for the REPERE challenge was enriched by additional audio sources (other French TV shows recorded earlier for the ETAPE evaluation campaign [26], French radio shows and data collected on the web). Finally, for the decision, QCOMPERE and SODA systems use a similar PLDA-based scoring coupled with a length-normalization approach (Eigen Factor Radial technique for QCOMPERE and for SODA). A basic Cosine Distance Scoring combined with a session/channel compensation technique (Within-Class covariance normalization) is used in PERCOL system.

3. Performance analysis

Table 2 shows the average Fm related to $SpkShow$ for the different systems, and for the oracle system which is made of the best system for each $SpkShow$. We can notice the large number of $SpkShow$ which are not in the dictionary of the three systems: about 40% for each system. As they do not have any model, they obviously cannot be identified, leading to a rather poor global Fm. More interestingly, the number of in-dictionary $SpkShow$ that are not recognised at all is not negligible: they represent between 23.5% and 31.5% of the in-dictionary $SpkShow$, depending on the system.

Figure 1 plots the distribution of all the $SpkShow$ in the

	PERCOL	QCOMPERE	SODA	Oracle
average Fm	36.1	38.1	35.1	46.2
average Fm for in dict. $SpkShow$	62.8	68.4	61.9	72.2
#$SpkShow$ out of dict.	200	209	204	172
#$SpkShow$ in dict.	277	268	273	305
#$SpkShow$ in dict. with $Fm = 0$	79	63	86	63

Table 2: Average system performance related to $SpkShow$

system dictionaries, according to their performance expressed in terms of Fm, for the different systems. We can see that the average performance (between 61.9% and 72.2%) is not at all representative of performance obtained for each $SpkShow$: speakers are either not recognized or well recognized. Indeed, if we compute the average performance for $SpkShow$ which have $Fm \neq 0$, the average Fm grows to 87.9% for PERCOL, 89.5% for QCompere and 90.3% for SODA.

To evaluate the impact of the automatic speaker diarization, the analysis of the speaker performance when systems rely on the reference speaker diarization is carried out. Results for PERCOL system is shown in Figure 2. The bi-modal distribution of performance (speaker either well recognized, or not at all recognized) is dramatically emphasized with the reference diarization. The comparison of the speaker identification performance between using the reference or automatic speaker diarization, carried out on PERCOL and SODA systems, shows that 38 $SpkShow$ (out of 277 in-dictionary speakers) for PERCOL and 14 $SpkShow$ (out of 273 in-dictionary speakers) for SODA present a null F-measure ($Fm_i = 0$) with the automatic speaker diarization and a F-measure above 90% with the reference one. For these particular $SpkShow$, the quality of the automatic speaker diarization is the main reason of the poor speaker identification performance. A fine-grained analysis of the speaker diarization outputs highlighted segment frontier errors, clustering confusion errors, or both of them. In addition, we considered the $SpkShow$ for which a null F-measure is obtained even with the reference diarization: 41 for PERCOL and 72 for SODA. For half of these $SpkShow$, the amount of testing data was less than 10s. For the other half, the amount of training data could not explain the poor performance as, on average, more than 600s were available for each of those $SpkShow$. Focusing on the 12 $SpkShow$ which have a null F-measure with reference diarization in both systems, the analysis revealed that these segments had very poor acoustic quality: a large amount of overlapped speech for 4 $SpkShow$ (from 20 to 90% of overlapped speech according to $SpkShow$), an entire interview made by phone for one $SpkShow$, poor sound quality with reverberation for another one, and large background noise (street, assembly background voices, applause, etc) or music for 8 of them.

If we analyze performance obtained per speaker, independently of the shows in which they appear, we distinguish the speakers occurring in only one show (single speakers), and the speakers occurring in several shows (recurrent speakers). The 305 $SpkShow$ which are in at least one system dictionary comes from 141 speakers, 88 being single speakers and 53 being recurrent speakers. These 53 recurrent speakers count for 217 $SpkShow$, among which 35 $SpkShow$ have a null F-measure in the oracle system. Among these 35 $SpkShow$ with null F-measure, 23 comes from 13 speakers which have a good

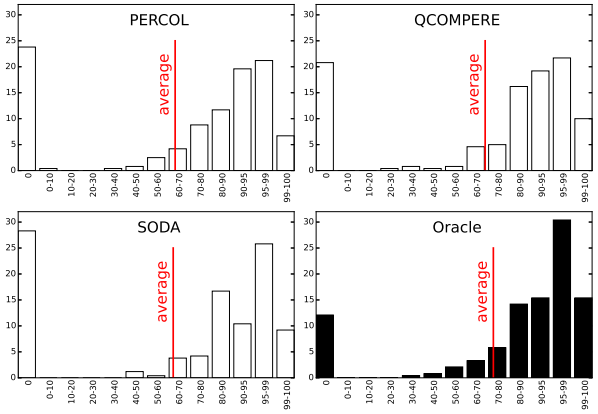

Figure 1: Distribution of $SpkShow$ according to system performance expressed in terms of Fm

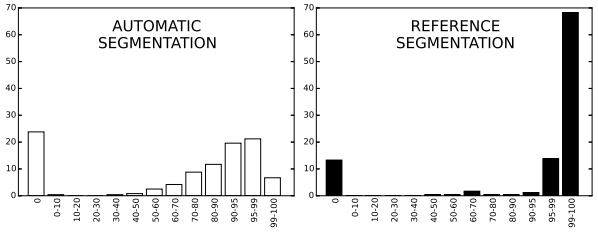

Figure 2: Effect of diarization errors on PERCOL system.

F-measure in other shows (with an average oracle non-null F-measure=91.7%) and 12 originate from 5 speakers which always have null F-measure. Thus, we can conclude that the influence of the show (testing data) is very strong: for a same given speaker model, we can observe null performance or very good performance depending on the show.

4. Predicting speaker recognizability

We have seen in the previous section that speaker identification performance is essentially bi-modal: either a speaker is not recognized at all or it is very well recognized. This section aims at uncovering the speaker characteristics explaining why some speakers are recognized (✓) and others are not (×)? To answer this question, we first try to automatically classify the speakers into those two classes. In case we succeed, by analyzing the speaker characteristics contributing the most to this prediction, we should be able to identify the characteristics that facilitate or hamper the identification.

4.1. $SpkShow$ characteristics

Numerous characteristics could explain why a $SpkShow$ is recognized or not, including linguistic or prosodic characteristics or the background noise. In this paper, we only study two families of characteristics – derived from the amount of training data used for speaker modeling, or related to the distribution of speech segments uttered by each speaker in the test set.

The first set of characteristics includes the duration of training data available for each reference speaker (from the REPERE corpus, other corpora, or both) and the corresponding number of training sessions. For each $SpkShow$, these characteristics are

	All characteristics			Optimal subset		
	P.	R.	F-measure	P.	R.	F-measure
✓	94.5	93.4	94.0	95.6	94.6	95.0
×	75.8	79.0	77.4	80.0	82.7	81.4

Table 3: Prediction performance

obtained from the oracle system (*i.e.* the system that performs the best for this particular $SpkShow$ among the three systems).

For each $SpkShow$, the second set of characteristics includes the number of speech turns, their total (or average) duration or the duration of the longest speech turn. It also includes characteristics related to the level of interactions of a $SpkShow$, such as the number and total duration of overlapped speech segments or the average pause duration before and after each speech turn of a $SpkShow$.

4.2. Prediction of oracle performance

Given a $SpkShow$ and its corresponding set of characteristics, we aim at predicting whether the oracle system is able to (at least partially) recognize ✓ it or not (at all) ×. Focusing on oracle system enables to draw conclusions which are not specific of a given system, but related to the best performance we can expect from state-of-the-art technology. As the corpus is limited (only 305 different $SpkShow$) and unbalanced (63 × vs. 242 ✓), we proceed using leave-one-out cross-validation and evaluate this classification experiment as two complementary detection tasks using precision, recall and F-measure.

Not all $SpkShow$ characteristics are meaningful features for this task, and some can even degrade the classification performance. Hence, feature selection is applied using the following heuristic. Starting with the whole set of characteristics, each iteration removes the characteristic whose removal leads to the best performance. The optimal subset of characteristics is selected as the one leading to the best overall performance.

As far as the classification algorithm is concerned, we chose to use a decision tree (rather than a more sophisticated black box classifier), with the Scikit-learn implementation [**?**], as the analysis of its internal structure allows for easy interpretation of the results and the importance of each characteristic. Table 3 contains the experimental results. It shows that it is possible to predict, with pretty good performance, whether a $SpkShow$ will be recognized (✓ F-measure $= 95.0\%$) or not (× F-measure $= 81.4\%$).

4.3. What makes a speaker recognizable?

Now that we showed that it is possible to predict whether a $SpkShow$ is recognizable or not, this section aims at providing more insight into why this is the case.

Figure 3 provides the distributions of *feature importance* for the six characteristics selected in the optimal subset, computed over the 305 leave-one-out cross-validation rotations. Here, in the case of decision trees, feature importance is defined as the Gini coefficient and is related to the number of times a characteristic is used in the tree and how discriminant it is on average. More information on this metric can be found in [27]

Interestingly, the two most important features are related to the duration of speech turns in the test set – characteristics related to the amount of training data only appear at rank #3 and #6. The presence at rank #4 of the characteristic defined as the average duration of the pause after each speech turn of a

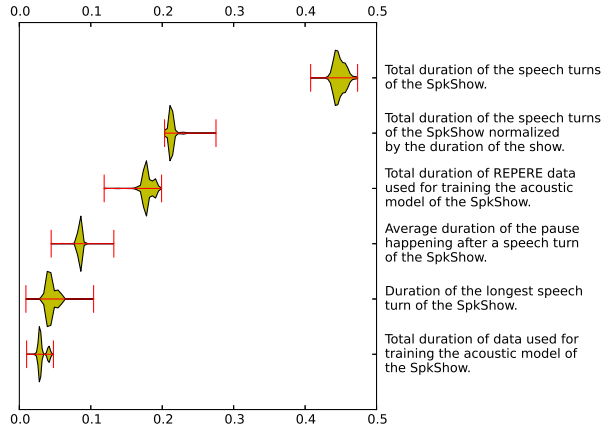

Figure 3: Distribution of feature importance.

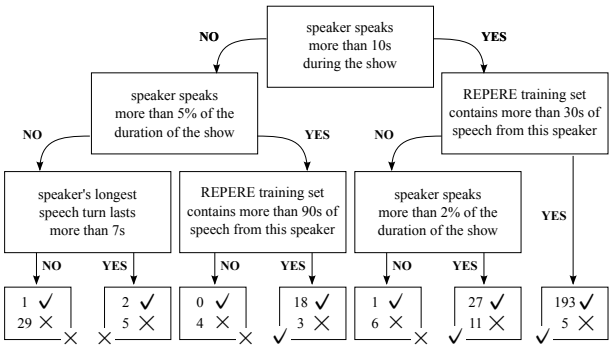

Figure 4: Not recognized (×) vs. (partially) recognized (✓)

speaker is somewhat suprising. This could be explained by the fact that long pauses between two speech turns of two different speakers may ease the segmentation process (whose influence is discussed in Section 3) and therefore increase speaker recognizability. Finally, Figure 4 is a graphical illustration of a decision tree (for sake of readability, the visualization is truncate at depth 3) trained on the whole set of $SpkShow$ and based on the optimal subset of characteristics discussed in previous section. The boxes at the bottom represent the leaves of the decision tree. For each leaf, the label given by the tree is in the bottom right corner, and the box details the actual composition of $SpkShow$ classified in this leaf. Noticeably, the rightmost leaf concentrates 195 out of the 242 $SpkShow$ which are recognized by the system (✓), with only 2 simple rules about the minimal amount of testing data (10s) and training data in REPERE (30s).

5. Conclusion

In this paper, performance analysis has been done on 3 state-of-the-art i-vectors based speaker identification systems submitted on the REPERE challenge. It is shown that the performance distribution is essentially bi-modal: a speaker in a given show is either not recognized at all, or well recognized. A performance prediction paradigm has been developed, in order to predict if a speaker, with given characteristics in training and testing data will be recognized or not. This framework yields interesting prediction results and enables to identify the characteristics which contribute the most to the correct recognition of speakers.

6. References

[1] C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, and J. Hernandez-Cordero, "The 2012 nist speaker recognition evaluation." in *INTERSPEECH*, 2013, pp. 1971–1975.

[2] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The nist 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.

[4] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in *NEURAL NETWORKS SIGNAL PROCESS PROC IEEE*, vol. 2, 2000, pp. 775–784.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[6] P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Odyssey: The Speaker and Language Recognition Workshop*, 2012, pp. 157–164.

[7] A. Kanagasundaram, D. Dean, S. Sridharan, M. McLaren, and R. Vogt, "I-vector based speaker recognition using advanced channel compensation techniques," *Computer Speech & Language*, vol. 28, no. 1, pp. 121–140, 2014.

[8] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[9] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[10] Y. Jiang, K.-A. Lee, Z. Tang, B. Ma, A. Larcher, and H. Li, "Plda modeling in i-vector and supervector space for speaker verification." in *INTERSPEECH*, 2012.

[11] P.-M. Bousquet, J.-F. Bonastre, and D. Matrouf, "Exploring some limits of gaussian plda modeling for i-vector distributions," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[12] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation," in *ICSLP*, 1998.

[13] J. Kahn, N. Audibert, S. Rossato, and J. F. Bonastre, "Intra-speaker variability effects on speaker verification performance," in *Odyssey*, 2010, pp. 21–25.

[14] C. Greenberg, A. Martin, and M. Przybocki, "The 2011 best speaker recognition interim assessment," in *Odyssey*, 2012, pp. 275–282.

[15] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *in Proceedings of the Interspeech*, 2007.

[16] J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, and P. Joly, "A presentation of the REPERE challenge," in *CBMI*, 2012, pp. 1–6.

[17] D. Charlet, C. Barras, and J.-S. Lienard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7707–7711.

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

[19] G. Dupuy, S. Meignier, P. Deléglise, and Y. Estève, "Recent improvements towards ILP-based clustering for broadcast news speaker diarization," 2014.

[20] A. Larcher, J.-F. Bonastre, B. G. Fauve, K.-A. Lee, C. Lévy, H. Li, J. S. Mason, and J.-Y. Parfait, "Alize 3.0-open source toolkit for state-of-the-art speaker recognition." in *INTERSPEECH*, 2013, pp. 2768–2772.

[21] A. Anjos, L. E. Shafey, R. Wallace, M. Günther, C. McCool, and S. Marcel, "Bob: a free signal processing and machine learning toolbox for researchers," in *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*. ACM Press, October 2012.

[22] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[23] J. Pelecanos and S. Sridharan, "Feature Warping for Robust Speaker Verification," in *Proceedings of Odyssey 2001 - The Speaker Recognition Workshop*, Crete, Greece, June 2001, pp. 213–218.

[24] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, 2011.

[25] A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert, and L. Quintard, "The REPERE Corpus: a Multimodal Corpus for Person Recognition," in *International Conference on Language Resources and Evaluation*, 2012.

[26] G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE Corpus for the Evaluation of Speech-based TV Content processing in the French language," in *International Conference on Language Resources, Evaluation and Corpora*, Turkey, 2012.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.