Segmentation et Regroupement en Locuteurs: comment évaluer les corrections humaines

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

Segmentation et Regroupement en Locuteurs: comment évaluer les corrections humaines

Broux Pierre-Alexandre[1, 2] Doukhan David[2] Petitrenaud Simon[1]
Meignier Sylvain[1] Carrive Jean[2]
(1) Laboratoire informatique de l'université du Maine (LIUM - EA 4023), Avenue Olivier Messiaen, F-72085 Le Mans, France
(2) Institut national de l'audiovisuel (Ina), 18 Avenue des Frères Lumière, 94360 Bry-sur-Marne, France
pabroux@ina.fr, ddoukhan@ina.fr, simon.petit-renaud@univ-lemans.fr, sylvain.meignier@univ-lemans.fr, jcarrive@ina.fr

RÉSUMÉ

Dans cet article, nous présentons un simulateur dédié à l'évaluation des corrections humaines sur la tâche de Segmentation et Regroupement en Locuteurs (SRL). Nous proposons quatre actions élémentaires afin de corriger une SRL et un automate pour simuler la séquence de corrections. Une mesure est proposée pour évaluer le coût de correction. Le simulateur est évalué en utilisant des émissions françaises d'information tirées du corpus REPERE.

ABSTRACT

Computer-assisted speaker diarization : how to evaluate human corrections

In this paper, we present a framework to evaluate the human correction of a speaker diarization. We propose four elementary actions to correct the diarization and an automaton to simulate the correction sequence. A metric is described to evaluate the correction cost. The framework is evaluated using French broadcast news drawn from the REPERE corpus.

MOTS-CLÉS : Segmentation et Regroupement en Locuteurs, annotation, Interactions Homme-Machine (IHM), évaluation.

KEYWORDS: Speaker diarization, annotation, Human-Computer Interaction (HCI), evaluation.

1 Introduction

Le travail présenté dans cet article a été réalisé pour répondre à certains objectifs de l'Institut national de l'audiovisuel (Ina). L'Ina est une institution publique en charge de la préservation et de la valorisation du patrimoine audiovisuel français. La valorisation repose en partie sur l'annotation de collections de documents audiovisuels. L'annotation consiste à enrichir les documents avec des titres, des résumés, des mots-clés ou les noms des participants pour répondre aux requêtes des clients et des usagers de l'Ina, qu'ils soient journalistes, producteurs, réalisateurs ou chercheurs. Cependant, en raison du nombre croissant de documents et du nombre limité de documentalistes, beaucoup de documents restent peu ou pas documentés. Les informations fournies par la documentation varient grandement selon le type d'archives : les émissions d'information (journaux télévisés et magazines d'actualité) sont habituellement finement documentées, tandis que d'autres programmes tels que les jeux, les documentaires, les émissions de variété ou de téléréalité le sont plus sommairement.

Une des solutions pour faciliter l'annotation et améliorer l'accès aux documents est d'utiliser les technologies de reconnaissance automatique de la parole et du locuteur, comme peuvent le proposer Charhad *et al.* (2005); Ordelman *et al.* (2009); Vallet *et al.* (2016). La tâche de SRL est une étape de pré-traitement nécessaire pour l'identification du locuteur (Bonastre *et al.*, 2000) ou la transcription de parole (Anguera *et al.*, 2012) dans des émissions télévisées. Les tâches de SRL et d'identification permettent de déterminer « qui parle quand ». Les systèmes de SRL sont généralement fondés sur des méthodes de segmentation et regroupement non supervisées, estimant le nombre de locuteurs et découpant le flux audio en segments de parole étiquetés par des labels anonymes. Cependant, les systèmes de SRL à l'état de l'art ne sont toujours pas suffisamment précis pour être employés tels quels dans les applications de l'Ina, principalement à cause de la large variété des collections. La variété porte sur la période temporelle qui va de la fin du dix-neuvième siècle à nos jours, le type d'émissions ou les conditions d'enregistrement. Les interventions humaines sont donc la plupart du temps requises pour obtenir des annotations robustes. De plus, l'annotation de parole entièrement manuelle ne peut pas être une solution raisonnable au regard du coût important du processus. En effet, neuf heures sont requises pour effectuer l'annotation manuelle d'une heure de parole spontanée (transcription de parole et identification des locuteurs) (Bazillon *et al.*, 2008).

Dans cet article, nous proposons un simulateur pour expérimenter des méthodes de SRL assistées par l'humain afin de réduire le coût de correction d'une segmentation en locuteurs. Plus précisément, les objectifs sont de construire un automate qui simule les corrections des annotateurs et de proposer une mesure qui évalue le coût de ces corrections. Dans cet article, nous présentons dans un premier temps l'état de l'art dans le domaine d'annotation des systèmes de reconnaissance de la parole et du locuteur. Ensuite, nous décrivons le système de SRL assisté par l'humain proposé ainsi qu'une nouvelle mesure pour évaluer de tels systèmes. Dans la partie suivante, nous définissons les actions utilisées pour corriger la SRL. Avant de conclure, nous mesurons la durée de chaque action pour construire la mesure proposée et nous proposons une évaluation reposant sur un système oracle.

2 Travaux précédents

L'annotation humaine d'un document audio est une tâche longue et souvent fastidieuse. Elle est généralement réalisée manuellement avec des logiciels d'annotation comme *Transcriber* (Barras *et al.*, 2001) ou *ELAN* (Wittenburg *et al.*, 2006). Dans Bazillon *et al.* (2008), les auteurs ont montré que la correction des sorties d'un système de transcription automatique de la parole permet de diminuer le temps d'annotation. Une méthode d'apprentissage actif (active learning en anglais), proposée dans Budnik *et al.* (2014), utilisée en conjonction avec des systèmes de reconnaissance automatique du locuteur et du visage, réduit davantage le nombre d'interactions homme-machine. Récemment, dans Broux *et al.* (2016), nous avons proposé un système qui assiste la SRL et réduit le nombre d'interventions humaines. Dans ce travail, l'annotateur corrige seulement les erreurs de regroupement en locuteurs et la segmentation est supposée être parfaite ainsi que sans erreurs. Les deux derniers articles cités se concentrent sur la correction des erreurs de regroupement en locuteurs et négligent les erreurs de segmentation. De plus, les auteurs supposent sans le justifier que toutes les corrections ont le même coût.

3 Système de SRL assisté par l'humain

3.1 Description du système

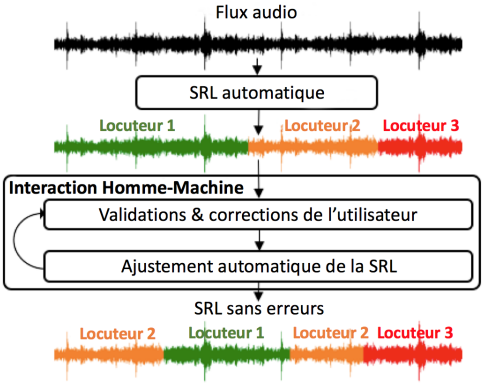

FIGURE 1 – Architecture du système de SRL assisté par l'humain

La figure 1 présente l'architecture du système de SRL assisté par un humain que nous proposons. Il est composé de deux parties principales. La première consiste à appliquer un système de SRL automatique sur un flux audio. Une segmentation initiale du flux est alors obtenue. La seconde consiste à demander à un humain de corriger la sortie de la première partie. Chaque correction humaine est à son tour prise en considération par un système qui améliore la SRL en réaffectant des segments à des locuteurs différents. Cet ajustement rend généralement plus faciles les actions restantes de l'annotateur. Selon l'objectif visé, l'annotateur peut effectuer des corrections sur le regroupement en locuteurs et/ou sur la segmentation. À la fin du processus, le taux d'erreurs de SRL, plus connu sous le nom de Diarization Error Rate (DER (NIST, 2003)), devrait avoir diminué et même être nul si toutes les erreurs de segmentation et de classification sont corrigées.

3.2 Simulateur expérimental

À partir de l'architecture présentée dans la précédente section, plusieurs règles ont été définies :

1. l'annotateur est simulé par un automate et ne fait aucune erreur ;
2. l'annotateur corrige l'émission du début à la fin dans l'ordre temporel afin de valider l'annotation automatique faite a posteriori ;
3. seul le tour de parole courant, qui vient d'être écouté, peut être corrigé par l'annotateur.

La première règle permet d'éviter la modélisation aléatoire et complexe des erreurs pouvant être commises par un humain. De plus, cette simplification permet d'avoir un document sans erreurs à la fin du processus de correction et ainsi un DER à 0%. Corriger à partir du début jusqu'à la fin est supposé aider la compréhension de l'annotateur et améliorer la correction. Cette règle et la dernière sont des conditions expérimentales choisies pour faciliter notre problème. Elles peuvent être remises en cause par la suite.

3.3 Mesure proposée : HCIQ

Le DER mesure la qualité de la segmentation et du regroupement en locuteurs (NIST, 2003). Cependant, il n'est pas pertinent pour évaluer le temps de création ou de correction d'un fichier de SRL. Une nouvelle mesure inspirée du Keystroke Saving Rate (KSR) utilisé dans la prédiction de mot pour les personnes ayant des difficultés de communication (Wood & Lewis, 1996) est proposée. Nous l'appelons Human-Computer Interaction Quantity (HCIQ). Cette mesure estime le coût des interventions d'un humain pour la correction d'une SRL. Elle peut être calculée à la fois pour les systèmes assistés et pour les systèmes où un humain corrige la SRL seul. En outre, ainsi que le DER, la mesure HCIQ peut être calculée pour chaque enregistrement ou pour un ensemble d'enregistrements audio/vidéo. Elle est définie par la formule suivante :

$$HCIQ = \sum_{i=1}^{K} w_i n_i,$$

où i correspond à un type d'action de correction dans l'interface, w_i est son coût associé, n_i le nombre de fois que l'annotateur a appliqué ce type d'action et K est le nombre de types d'action différents. Plus la mesure HCIQ est faible, plus le temps de correction de l'annotation sera faible. Au passage, elle permet de comparer différents systèmes de SRL assistés d'une manière objective pour un corpus donné.

La mesure HCIQ, en tant que telle, ne permet pas de comparer des corpus différents. Pour y remédier, nous proposons la formule suivante :

$$HCIQ_n = \frac{HCIQ}{d},$$

où $HCIQ$ est normalisé par d la durée du corpus sur lequel a été calculé le $HCIQ$. Cette normalisation permet de rester dépendant de la spécifité du corpus (nombre de locuteurs, parole spontanée, etc). La mesure HCIQ$_n$ est un ratio du nombre de corrections à faire pour une unité de temps. Plus la valeur est élevée pour un corpus donné, plus ce dernier requiert des corrections.

La mesure HCIQ se rapproche de celle proposée dans Guillaumin *et al.* (2009) où les auteurs proposent une mesure reflétant l'effort nécessaire pour un humain à la correction d'images mal labelisés.

4 Annotateur et outils de SRL assistés

4.1 Logiciel d'annotation : Transcriber

Afin de choisir les actions humaines nécessaires à la correction, nous nous reposons sur *Transcriber*, un logiciel de référence dans la transcription de la parole et l'annotation. Ce logiciel permet de couper un flux audio en segments. Chaque segment correspond à une zone de parole et est associé à un nom de locuteur. Ce nom, ou label, peut être enrichi par des informations tel que le genre ou la langue native du locuteur. Dans *Transcriber*, les actions de segmentation sont "*Créer une frontière*", "*Supprimer une frontière*" et "*Déplacer une frontière*". L'action "*Créer une frontière*" ajoute une frontière en coupant un segment en deux parties, l'action "*Supprimer une frontière*" fusionne deux segments consécutifs et l'action "*Déplacer une frontière*" déplace la frontière d'un segment incorrectement positionné.

Concernant les actions de regroupement en locuteurs, *Transcriber* offre les actions "*Créer un label locuteur*" et "*Changer le label locuteur*". La première permet de créer un nouveau label locuteur pour le segment sélectionné tandis que la dernière permet de changer le label locuteur en sélectionnant un autre label parmi la liste des labels déjà créés.

4.2 Actions de correction

Afin de faciliter la création d'un annotateur simulé par un automate, les séries d'actions seront déterministes. Nous voulons qu'aucune action ne puisse être substituée par un ensemble d'actions fournissant la même correction. Une des actions de *Transcriber* ne remplit pas ce critère. L'action "*Déplacer une frontière*" peut être remplacée par les deux actions suivantes : "*Créer une frontière*" et "*Supprimer une frontière*".

Pour résumer, nous avons gardé deux actions pour modifier les frontières des segments et deux actions pour modifier les labels affectant le regroupement en locuteurs. En combinant ces actions, nous pouvons décrire toutes les corrections d'une manière unique. Finalement, les quatres actions sélectionnées utilisées dans la mesure HCIQ sont :

- "*Créer une frontière*",
- "*Supprimer une frontière*",
- "*Créer un label locuteur*",
- "*Changer le label locuteur*".

5 Expériences

5.1 Corpus

Les expériences ont été appliquées sur des enregistrements télévisuels de la campagne d'évaluation de 2013 du défi ANR-REPERE [1]. Les émissions télévisuelles proviennent de deux chaînes françaises (BFM et LCP). Le tableau 1 décrit le corpus utilisé dans les expériences. Le corpus est équilibré : il

Nombre d'émissions	7
Nombre d'enregistrements	28
Temps d'enregistrement	14h17
Temps d'annotation	2h57
Nombre de locuteurs	212

TABLE 1 – Description de REPERE test 2013

contient de la parole spontanée et préparée. Il est constitué de micros-trottoirs, de débats et d'émissions d'informations mais seule une partie des données est annotée (Kahn *et al.*, 2012).

1. http://www.defi-repere.fr/

5.2 Mesure de la durée des actions

Dans la section 4.2, nous avons selectionné quatre actions de correction. Maintenant, nous proposons une méthode pour estimer la durée moyenne de chaque action. L'historique des clics de la souris et des frappes du clavier dans *Transcriber* permet de déterminer indirectement les actions successives et d'évaluer précisement la durée de chaque action. Pour enregistrer cette trace d'exécution, il est nécessaire de modifier le code source de *Transcriber*. Une entrée dans le fichier des traces, c'est-à-dire un clic de souris ou une frappe de clavier, contient trois types d'information : le temps du clic ou de la frappe, le nom du module actif et un commentaire (figure 2). Le module identifie un élément de l'interface utilisateur tandis que le commentaire donne des informations précises sur l'évènement en cours. Le fichier des traces lui-même n'est pas suffisant pour déterminer les

```
[1319494751] :: [Player] [Strategy: Play/Pause; Pause at 654.942]
[1319493381] :: [LabelWindow] [Open window; Edit an existing Turn]
[1319491287] :: [Label] [Edit an existing speaker thanks to LabelWindow]
[1319480451] :: [LabelWindow] [Validate; Close Window]
```

FIGURE 2 – Exemple d'un fichier des traces

actions d'une manière automatique. En effet, l'annotateur peut faire des erreurs ou prendre une pause durant la session d'annotation contrairement à notre automate simulant un annotateur idéal. Pour résoudre ce problème, l'enregistrement de l'écran utilisateur, conservé sous la forme d'un film, est manuellement segmenté en actions en interprétant conjointement la vidéo et le fichier des traces. Chaque segment créé correspond donc précisément à la durée mesurée des actions actuelles. Afin

Action	**Nombre d'occurrences**	**Moyenne (sec)**	**Écart type (sec)**
Créer un label locuteur	28	12,7	6,0
Changer le label locuteur	32	7,6	3,8
Créer une frontière	38	12	7,6
Supprimer une frontière	46	5,1	2,3

TABLE 2 – Durée des actions - 20 min des données REPERE test 2013

de faciliter l'annotation de chaque action et de minimiser les erreurs d'interprétation, seules les régions avec peu de paroles spontanées et sans paroles superposées ont été annotées. Le tableau 2 montre les résultats de la durée des actions. Les actions les plus chronophages sont "*Créer un label locuteur*" et "*Créer une frontière*", avec une moyenne comprise entre 12 et 13 secondes. La première action requiert d'entrer un label locuteur (et éventuellement d'autres méta-données du locuteur), alors que la seconde action requiert de regarder et écouter le signal pour détecter la frontière de locuteurs. On notera qu'il est généralement nécessaire d'écouter le signal plusieurs fois afin de placer une nouvelle frontière. L'action nommée "*Changer le label locuteur*" a une durée moyenne de 7,6 secondes. Affichée dans une fenêtre contextuelle, elle consiste à sélectionner le label locuteur correct dans une liste déroulante. Chercher un label dans une liste déroulante demande moins d'efforts cognitifs que créer une frontière. L'action la plus rapide est l'action "*Supprimer une frontière*". Elle requiert d'arrêter l'écoute lorsqu'une frontière erronée est détectée et de la supprimer par une simple combinaison de touches au clavier.

5.3 Évaluation d'un système oracle

La simulation de l'annotateur repose sur deux types d'information pour déterminer si une correction est requise au temps t :

1. une différence ou non entre le segment de référence (vérité terrain) et le segment de l'hypothèse ;
2. la correspondance ou non entre les labels locuteurs de la référence et l'hypothèse qui minimise le DER.

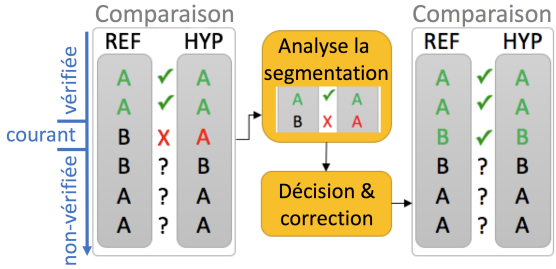

FIGURE 3 – Illustration d'un annotateur simulé

En cas de discordance au niveau de la segmentation ou du label au temps t entre la référence et l'hypothèse, une correction est nécessaire. L'annotateur simulé corrige en premier les erreurs de segmentation, puis les erreurs de regroupement en locuteurs (figure 3). Après chaque correction, le système peut lancer un système de SRL sur la partie non vérifiée (les segments avec un temps de début $> t$) en prenant en considération les segments déjà vérifiés (segments avec un temps de fin $\leq t$). Dans le cas d'un système initial sans erreur de segmentation, la correction du regroupement en locuteurs est facile à mettre en place (Broux *et al.*, 2016) car les segments de l'hypothèse sont identiques aux segments de la référence. La correction de la segmentation est plus difficile, l'annotateur simulé a besoin de prendre en considération la précision des frontières de la référence. Pour résoudre ce problème, une tolérance de plus ou moins 250 ms est généralement appliquée aux frontières des segments de référence pour le calcul du DER. Nous appliquons la même tolérance pour éviter les nombreuses micro-corrections, généralement inutiles. Ainsi, avant de procéder à l'évaluation d'une possible différence entre la zone de segmentation de la référence et celle de l'hypothèse, toute frontière de l'hypothèse appartenant à une zone de tolérance est déplacée à coût nul afin d'être alignée avec la frontière de la référence.

L'annotateur simulé devient un système oracle quand aucun ajustement automatique n'est effectué au fur et à mesure des corrections. L'évaluation du système oracle est reportée dans le tableau 3. Le HCIQ du corpus test est de 331,6 minutes et correspond à la somme de toutes les estimations de durée de correction (tableau 3). La SRL utilisée comme entrée de l'oracle est fournie par le système de SRL entièrement automatique décrit dans Meignier & Merlin (2010). Le DER de la SRL avant correction est de 13,85%. Le nombre d'occurrences des actions de segmentation est environ une fois et demie plus important que le nombre d'actions de regroupement en locuteurs (respectivement 1142 et 758). Les erreurs de segmentations représentent environ 65% du temps de correction total (210,5 minutes). "*Créer une frontière*" est l'action la plus coûteuse, car elle correspond à environ 52% des corrections globales. Pour un enregistrement audio de 2h57 (177 minutes), un annotateur passera 3h17 (197,2

Action	Nombre d'occurrences	Durée estimée (min)
Créer un label locuteur	295	62,4
Changer le label locuteur	463	58,7
Créer une frontière	986	197,2
Supprimer une frontière	156	13,3

TABLE 3 – Correction pour le système oracle - REPERE test 2013. Durée estimée (durée moyenne $\times$ nombre d'occ.)

minutes) à créer des frontières. Si l'annotateur simulé corrige seulement les erreurs de regroupements en locuteurs, le DER est de 5,59% à la fin du processus de correction. Ces 5,59% d'erreurs sont dus à la mauvaise segmentation. Ce résultat montre que les erreurs de segmentation et de regroupement en locuteurs contribuent globalement à 40% et 60% du DER respectivement. Comparativement, les erreurs de segmentation correspondent au coût de correction principal en termes de HCIQ.

Corpus	F	M	S	P	C	H	T	H_n
ESTER test 2003	0,77	0,56	17,53	9,76	20,80	477,2	592	0,81
ESTER test 2009	1,45	0,86	13,05	8,67	28,67	482,0	430	1,12
ETAPE test 2012	14,93	0,35	18,74	9,96	10,16	793,7	418	1,90
REPERE test 2013	0,76	3,60	9,49	11,47	8,68	331,6	177	1,87

TABLE 4 – Comparaison des HCIQ_n obtenus à partir des corrections du système oracle sur divers corpus. F : $\text{DER}_{false\,alarm}(\%)$; M : $\text{DER}_{missed\,speaker}(\%)$; S : $\text{DER}_{confusion}(\%)$; P : Pureté (%) ; C ; Couverture (%) ; H : HCIQ (min) ; T : Temps d'annotation (min) ; H_n : $\text{HCIQ}_n(min)$

Le tableau 4 permet de comparer le HCIQ_n de différents corpus. Il prouve que le corpus REPERE est un des corpus qui requiert le plus de corrections pour une unité de temps car il nécessite en moyenne 1,87 minutes de corrections humaines pour 1 minute de signal audio. En outre, il montre que les corpus ETAPE et REPERE, principalement plus composés de parole spontanée (faux départs, répétitions, parole superposée, interjections, etc (Bazillon *et al.*, 2008)) que les corpus ESTER, obtiennent des scores HCIQ_n élevés.

6 Conclusion

Dans cet article, nous avons proposé un simulateur pour évaluer des systèmes intéractifs de SRL prenant en compte les corrections humaines. La combinaison de quatre actions permet de décrire les étapes de correction d'une manière unique. Nous avons proposé une mesure pour déterminer précisément la durée de chaque action afin d'évaluer le coût des interactions homme-machine. L'évaluation du système oracle sur le corpus REPERE test 2013 montre que les corrections de segmentation prennent plus de temps que les corrections de regroupement en locuteurs. Les résultats de l'oracle montre également l'importance des erreurs de segmentation sur le HCIQ et le DER. La correction des erreurs de segmentation fait croître le HCIQ tandis qu'elle affecte de façon négligable le DER. Seule la correction des erreurs de classification fait diminuer directement le DER. Les prochains travaux se concentreront sur le développement d'un système de SRL intégré pour réduire le temps de correction.

Références

ANGUERA X., BOZONNET S., EVANS N., FREDOUILLE C., FRIEDLAND G. & VINYALS O. (2012). Speaker diarization : A review of recent research. *ieee-tsap*, **20**(2), 356–370.

BARRAS C., GEOFFROIS E., WU Z. & LIBERMAN M. (2001). Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, **33**(1), 5–22.

BAZILLON T., ESTÈVE Y. & LUZZATI D. (2008). Transcription manuelle vs assistée de la parole préparé et spontanée. *Revue TAL*.

BONASTRE J.-F., DELACOURT P., FREDOUILLE C., MERLIN T. & WELLEKENS C. (2000). A speaker tracking system based on speaker turn detection for nist evaluation. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 2, p. II1177–II1180 : IEEE.

BROUX P.-A., DOUKHAN D., PETITRENAUD S., MEIGNIER S. & CARRIVE J. (2016). An active learning method for speaker identity annotation in audio recordings. In *1st International Workshop on Multimodal Media Data Analytics (MMDA), In conjuction with the 22nd European Conference on Artificial Intelligence (ECAI)*.

BUDNIK M., POIGNANT J., BESACIER L. & QUÉNOT G. (2014). Automatic propagation of manual annotations for multimodal person identification in tv shows. In *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on*, p. 1–4 : IEEE.

CHARHAD M., MORARU D., AYACHE S. & QUÉNOT G. (2005). Speaker identity indexing in audio-visual documents. In *Content-Based Multimedia Indexing (CBMI2005)*.

GUILLAUMIN M., VERBEEK J. & SCHMID C. (2009). Is that you ? metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, p. 498–505 : IEEE.

KAHN J., GALIBERT O., QUINTARD L., CARRÉ M., GIRAUDEL A. & JOLY P. (2012). A presentation of the repere challenge. In *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, p. 1–6 : IEEE.

MEIGNIER S. & MERLIN T. (2010). Lium spkdiarization : an open source toolkit for diarization. In *CMU SPUD Workshop*, volume 2010.

NIST (2003). 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`.

ORDELMAN R., DE JONG F. & LARSON M. (2009). Enhanced multimedia content access and exploitation using semantic speech retrieval. In *Semantic Computing, 2009. ICSC'09. IEEE International Conference on*, p. 521–528 : IEEE.

VALLET F., URO J., ANDRIAMAKAOLY J., NABI H., DERVAL M. & CARRIVE J. (2016). 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).

WITTENBURG P., BRUGMAN H., RUSSEL A., KLASSMANN A. & SLOETJES H. (2006). Elan : a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006, p. 5th.

WOOD M. E. & LEWIS E. (1996). Windmill-the use of a parsing algorithm to produce predictions for disabled persons. *PROCEEDINGS-INSTITUTE OF ACOUSTICS*, **18**, 315–322.