Impact de la détection de la parole pour différentes tâches de traitement automatique de la parole

Florent Desnous, Anthony Larcher, Sylvain Meignier

Impact de la détection de la parole pour différentes tâches de traitement automatique de la parole

Florent Desnous[1, 2] Anthony Larcher[1] Sylvain Meignier[1]
(1) LIUM, Le Mans Université, France
{florent.desnous, anthony.larcher, sylvain.meignier}@univ-lemans.fr
(2) Institute for Infocomm Research - A*STAR, Singapore

RÉSUMÉ

Dans cet article, nous proposons de comparer plusieurs systèmes de détection de la parole et leurs impacts sur deux tâches du traitement de la parole : la Segmentation et le Regroupement de Locuteurs (SRL) et la Reconnaissance Automatique de la Parole (RAP). Des systèmes à base de mixtures de Gaussiennes (GMM), de réseaux de neurones profonds (DNN) et récurrents (RNN) sont comparés, ainsi que l'utilisation d'un système de RAP pour détecter la frontière des mots. Les expériences présentées ici ont été conduites sur les corpus issus des campagnes d'évaluation ESTER1 et 2, ETAPE et REPERE1, constitués d'émissions de radio et de télévision française.

ABSTRACT

Impact of speech activity detection systems for different automatic speech processing tasks

In this article, we compare several Speech Activity Detection (SAD) systems and how they interact with two other speech processing tasks : Speaker Diarization and Automatic Speech Recognition (ASR). We study systems based on Gaussian Mixture Models (HMM/GMM), Deep and Recurrent Neural Networks (DNN and RNN) as well as an ASR system used to detect word borders. Experiments were made on French television and radio corpora : ESTER 1 and 2, ETAPE and REPERE1.

MOTS-CLÉS : détection de la parole, segmentation et regroupement de locuteurs, transcription automatique de la parole, apprentissage profond.

KEYWORDS: speech activity detection, speaker diarization, automatic speech recognition, deep learning.

1 Introduction

La détection de la parole est une étape importante dans plusieurs tâches de traitement de la parole. Elle permet d'extraire les segments de parole du signal, tout en ignorant le bruit, la musique, le silence, etc. Le but est d'une part de ne garder que des informations pertinentes pour la modélisation de la parole ou du locuteur et d'autre part de réduire la quantité de calculs nécessaire pour les traitements suivants.

De nombreuses approches permettent la détection de la parole. Il peut s'agir d'un seuillage sur l'énergie du signal ou d'autres paramètres acoustiques (Renevey & Drygajlo, 2001), d'une segmentation en utilisant des modèles de Markov cachés (*Hidden Markov Models*, HMM)(Kingsbury *et al.*, 2002; Gauvain *et al.*, 2002), de réseaux de neurones profonds (*Deep Neural Networks*, DNN)(Ryant *et al.*, 2013) ou récurrents (RNN)(Hughes & Mierle, 2013). Des mixtures de Gaussiennes (GMM) associées à des HMM sont souvent utilisées grâce au compromis qu'elles offrent entre légèreté et

Statistiques	ESTER 1	ESTER 2	ETAPE	REPERE 1
Nature	radio	radio	radio+TV	TV
Parole spontanée	peu	peu	beaucoup	variable
# de stations	6	4	3	2
# d'émissions	9	8	11	7
# d'enregistrements	18	26	15	28
Durée totale	10h	7h	8h30	14h
Durée annotée	9h	6h	6h	3h
# de locuteurs uniques	342	250	148	158
Moyenne loc./émission	20,17	11,53	10,33	7,5

TABLE 1 – Description du contenu des corpus évalués

performances. Cependant, ces dernières années, les performances des modèles neuronaux se sont grandement améliorées (Shahsavari *et al.*, 2017; Kaur & Sohal, 2017; Gelly & Gauvain, 2018).

En général, les systèmes de SAD servent comme première étape pour à d'autres tâches de traitement de la parole mais sont rarement optimisés pour celles-ci. Ce papier à pour but de comparer plusieurs système de détection de la parole afin de déterminer si l'un d'entre eux serais optimal pour la segmentation et le regroupement de locuteurs et la reconnaissance automatique de la parole.

Nous proposons de comparer plusieurs approches de détection de la parole en les utilisant pour des tâches de Segmentation et Regroupement de Locuteurs (SRL) et de Reconnaissance Automatique de la Parole (RAP). Nous utiliserons une approche GMM-HMM, trois approches basées sur les réseaux de neurones (récurrent et profond) et une approche utilisant un système de RAP.

Les différents corpus utilisés seront détaillés en Section 2 et les différents systèmes de détection de la parole en Section 3. La section 4 décrit les conditions d'évaluation (systèmes et métriques), tandis que les les résultats de ces comparaisons sont exposés en Section 5.

2 Corpus

Les corpus utilisés sont issus d'émissions de radio et de télévision française, il s'agit des corpus utilisés lors des campagnes d'évaluation ESTER1 (Gravier *et al.*, 2004), ESTER 2 (Galliano *et al.*, 2009), ETAPE (Gravier *et al.*, 2012) et REPERE 1 (Giraudel *et al.*, 2012). Leurs caractéristiques sont résumées dans le tableau 1. L'évaluation des différents systèmes se fait sur 87 enregistrements de journaux, de débats politiques ou d'émissions culturelles contenant des conditions acoustiques très difficiles.

Le corpus d'entraînement ainsi que le corpus de développement d'ESTER1 sont utilisés pour l'apprentissage des systèmes de détection de la parole. Le corpus d'apprentissage est partiellement annoté en 8 classes comme décrit dans (Meignier & Merlin, 2010). Les 8 classes consistent en 2 classes pour les silences en studio et au téléphone, 4 classes pour la parole en studio (propre , bruitée, avec de la musique ou d'un autre type), une classe pour la parole téléphonique et une classe pour les jingles et la musique pure. L'annotation a été réalisée de manière semi-automatique à partir d'un alignement forcé de la transcription de référence. Le tableau 2 indique les durées de chaque classe.

Classe	Durée
Silence studio	80 min
Silence téléphone	35 min
Parole propre (F0, F1)	54 min
Parole téléphonique (F2)	62 min
Parole bruitée (F4)	10 min
Parole et musique (F3)	38 min
Parole autre (FX)	142min
Jingles	85min

TABLE 2 – Description du corpus d'apprentissage Parole/Silence/Musique

3 Détection de la parole

3.1 Définition de la tâche

Le but d'un système de détection de la parole est de différencier les zones de parole des zones de non-parole au sein d'un signal audio, afin d'utiliser uniquement les segments de parole dans un système de reconnaissance du locuteur, de reconnaissance de la parole ou d'autres applications similaires.

La notion de non-parole inclut généralement tout ce qui est silence, bruit et musique, mais sa définition peut être élargie ou réduite selon la tâche. La notion de parole est aussi dépendante de la tâche : on peut vouloir garder la parole bruitée, superposée à de la musique, ou conserver uniquement la parole sans nuisances. Supprimer les segments de non-parole du signal permet d'éviter d'influencer les modèles d'apprentissage avec de l'information non pertinente et de réduire la quantité de calcul nécessaire à la tâche visée. Modéliser les caractéristiques de la non-parole est difficile, les sources de nuisances étant potentiellement nombreuses et très variables. La parole a moins de diversité et de variabilité, ce qui la rend comparativement plus simple à modéliser.

3.2 GMM-HMM

Le premier système de détection de la parole utilise des mixtures de gaussiennes (*Gaussian Mixture Models*, GMM) comme décrit dans (Meignier & Merlin, 2010). Il est réalisé à l'aide de la plateforme SIDEKIT (Larcher *et al.*, 2016) et de son extension S4D (*SIDEKIT for Diarization* [1]).

Le système est composé d'un modèle de Markov caché à 8 états, chacun d'eux associé à un GMM à 16 composantes diagonales. Ces 8 états/GMM représentent une classe acoustique du corpus d'apprentissage, représentant des états de la parole.

Les GMMs associés à l'HMM sont entraînés par l'algorithme EM-ML. En test, un décodage de Viterbi est utilisé pour retirer les zones de non-parole et garder les autres segments. Les pénalités de transmission entre les états ont été déterminées expérimentalement à partir d'un corpus de développement (campagne ESTER1).

Un post-traitement est nécessaire à la suite du décodage de Viterbi : le début et la fin de chaque segment de parole sont étendus de 0,5s afin de minimiser les imprécisions du décodage. Les segments de non-parole d'une durée inférieure à 0,25s sont réaffectés en parole.

1. https ://projets-lium.univ-lemans.fr/sidekit/

3.3 DNN-HMM

De façon similaire au GMM-HMM, le système se basant sur un réseau de neurones profond, est entraîné à classifier un vecteur acoustique en l'une des 8 classes du corpus. Les probabilités a posteriori générées par le DNN sont utilisées dans un décodage en Viterbi pour obtenir les segments de parole. L'étape de post-traitement est effectuée de la même manière que pour le système GMM-HMM. À la différence du GMM-HMM où 8 états sont décodés, ce système n'en utilise que 3 : les distributions de probabilités sont normalisées avant le décodage pour ne garder que la parole, le silence et la musique.

Le DNN acoustique est développé avec SIDEKIT et Theano (Bergstra *et al.*, 2010). Il est composé de 4 couches de 1000 neurones activées par une *sigmoïde*, et d'une couche à 8 sorties activée par un *Softmax*.

3.4 DNN « mimic »

Inspiré des travaux de (Rohdin *et al.*, 2017), ce système consiste à entraîner un DNN à reproduire les sorties du GMM utilisé en 3.2. L'hypothèse étant que le DNN a une meilleure capacité de généralisation que le GMM. Le DNN est construit à partir de Keras (Chollet, 2016) et est composé de 2 couches de 600 neurones activées par une fonction $\tanh$, et d'une couche de 8 neurones à activation linéaire (les sorties du GMM étant des log-probabilités). L'erreur quadratique moyenne (*Mean Squared Error*, MSE) est utilisée en tant que fonction de coût pour l'apprentissage du réseau.

De la même façon que pour le GMM-HMM, un décodage de Viterbi et un post-traitement sont effectués afin de ne garder que les segments de parole.

3.5 Segmentation par transcription automatique

Un système de transcription automatique de la parole génère des mots ainsi que leurs frontières. L'idée est d'utiliser ces frontières comme segments de parole et de les évaluer tels quels.

Ce système est initialisé avec les segments du GMM-HMM mentionné précédemment. Le temps de calcul étant en fonction du nombre de locuteurs, un regroupement hiérarchique est effectué au préalable (décrit plus loin dans cet article). Les sorties du système de RAP sont traitées pour ne garder que l'information sur les frontières de mots, tout en supprimant les phonèmes trop longs considérer comme des erreurs de transcription.

3.6 LSTM bidirectionnel

Les LSTM (*Long Short-Term Memory* (Sepp Hochreiter & Jürgen Schmidhuber, 1997)) sont des réseaux de neurones récurrents possédant une mémoire interne à court et long terme. Le mode bidirectionnel permet la prise en compte d'un contexte acoustique plus large. Ce système permet de s'affranchir du décodage de Viterbi.

L'architecture utilisée s'inspire de (Yin *et al.*, 2017) mais adaptée à la détection de la parole et réalisée avec Keras. Il s'agit de deux couches de BLSTM à 64 et 40 neurones en sortie, deux couches de DNN à 40 et 10 neurones et d'une sortie à un neurone. Le réseau utilise un contexte long de 3,2s extrait toutes les 0.8s (avec un chevauchement de 75%). Le réseau produisant des segments du même format qu'en entrée, ceux-ci sont regroupés (moyenne) afin d'avoir un vecteur de probabilités par

fichier audio. Les segments de parole sont générés en utilisant deux seuils (activation/désactivation) sur les probabilités a posteriori du réseau.

L'apprentissage de ce système nécessite d'avoir une annotation continue du corpus utilisé pour l'apprentissage. Or le corpus d'apprentissage n'est que partiellement annoté. Pour ce système, les annotations utilisées sont les segments de parole pour la classe positive et les écarts entre ces segments pour la classe négative (non-parole).

4 Évaluation

4.1 Tâches évaluées

Les différents systèmes de détection de la parole ont été évalués dans deux tâches de traitement de la parole : la Segmentation et le Regroupement de Locuteurs (SRL), et la transcription automatique de la parole.

4.1.1 Segmentation et Regroupement de locuteurs

La tâche de segmentation et du regroupement de locuteurs consiste à répondre à la question « qui parle quand ? » dans un fichier audio. Typiquement, un système de SRL comporte quatre étapes : la paramétrisation, la détection de la parole, la segmentation en tour de parole, et le regroupement des tours par locuteurs uniques. L'étape de détection de la parole permet de supprimer les silences, bruits et musique qui dégraderaient la qualité des étapes suivantes. La segmentation en tour de parole a pour but de construire des segments de parole homogène et pure contenant un seul locuteur, pour enfin effectuer les regroupements (ou *clustering*) jusqu'à obtenir des classes représentant les locuteurs de l'enregistrement.

Le système utilisé ici est développé au LIUM, proche de celui décrit par (Dupuy *et al.*, 2014), mais réalisé à partir de l'extension S4D de SIDEKIT. Dans un premier temps, une segmentation acoustique est effectuée en utilisant le critère d'information bayésien (*Bayesian Information Criteria*, BIC), suivi par un regroupement hiérarchique utilisant la même métrique (*Hierarchical Agglomerative Clustering*, BIC-HAC) et un décodage de Viterbi, afin de réajuster les frontières des segments. À ce stade, généralement, chaque locuteur est représenté par plusieurs classes. Une dernière étape de regroupements est réalisée. Des *i*-vectors (Dehak *et al.*, 2011) sont extraits à partir de chaque classe, et une matrice de distance est calculée. Les regroupements sont affinés en traitant cette matrice comme un problème de Programmation Linéaire en Nombres Entiers (PLNE).

4.1.2 Tâche : Reconnaissance automatique de la parole

La tâche de Reconnaissance Automatique de la Parole (RAP) permet de générer le texte énoncé dans un fichier audio. Le système de transcription utilisé est similaire au système développé par le LIUM (Rousseau *et al.*, 2014) durant la campagne REPERE à partir de l'outil Kaldi (Povey *et al.*, 2011).

Le système de RAP prend en entrée chacune des segmentations générées par les systèmes de détection de la parole après l'étape de regroupement hiérarchique BIC décrite en 4.1.1.

4.2 Paramètres acoustiques pour les détecteurs de parole

Tous les systèmes évalués utilisent des MFCC (*Mel Frequency Cepstral Coefficients*). Les vecteurs sont extraits tous les 10ms sur une fenêtre de 25ms. Pour les systèmes GMM-HMM et MIMIC, 13 coefficients sans l'énergie sont extraits auxquels s'ajoute leurs dérivées (dimension 26). Pour le DNN-HMM, un contexte de ± 31 trames (31 + 1 + 31 trames $\times$ 26 = 1638 coefficients) est ajouté.

Le BLSTM utilise une paramétrisation différente : 11 coefficients MFCC sont extraits de la même façon, auxquels sont ajoutées leurs dérivées premières et secondes avec leur énergie respective (11 + 12 + 12 = 35 coefficients).

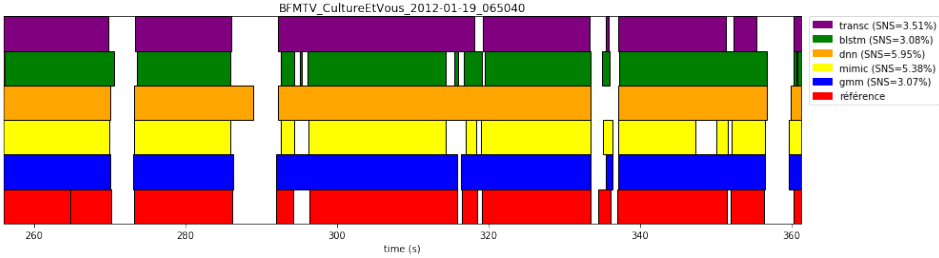
4.3 Métriques d'évaluation

Un système de détection de la parole est évalué à l'aide de deux scores : le taux de fausse alarme (FA), où le système détecte de la parole là où la référence n'en indique pas, et le taux de parole manquée (Miss), où le système échoue à détecter de la parole présente dans la référence. Ces deux taux peuvent être sommés pour mesurer le taux d'erreur parole vs non-parole (SNS).

Pour le regroupement de locuteurs, la métrique la plus utilisée est le DER (*Diarization Error Rate*), qui est composé de trois types d'erreurs : détection manquée, fausse détection et substitution de locuteurs par rapport à la référence. En complément au DER, les mesures de Pureté et de Couverture des segments sont utilisées comme décrites dans la documentation de l'outil utilisé *pyannote.metrics* (Bredin, 2017).

Enfin, pour la transcription, la métrique utilisée est le taux d'erreur de mots (*Word Error Rate*, WER), qui combine les erreurs de substitution, de suppression et d'insertion de mots.

FIGURE 1 – Exemple de segments générés



5 Résultats

La figure 1 montre les segments générés par les différents systèmes de SAD pour un fichier du corpus REPERE. On remarque que tous les systèmes réussissent à détecter les longs silences. Le DNN ne parvient pas à détecter les silences de courte durée et produit des segments de parole trop longs. Au contraire, le système utilisant la RAP et le système MIMIC génèrent des silences trop longs ou inexistants dans la référence, produisant un taux de détection manquée élevée (voir 3).

Le tableau 3 résume les résultats obtenus pour les 5 systèmes évalués. Au niveau de la tâche de SRL, le système GMM-HMM reste le meilleur système avec 0.24% de DER d'avance en absolu par rapport au deuxième meilleur système.

Le DNN-HMM produit le taux de parole manquée le plus faible, mais aussi le plus haut taux de fausse alarme : les segments générés sont plus longs qu'ils ne devraient l'être. En conséquence ce système est le moins performant en termes de DER.

Le système « MIMIC » joue bien son rôle avec des taux d'erreur proches de ceux du GMM-HMM. Les résultats proposés sont ceux de la cinquième itération d'apprentissage du réseau.

Logiquement, le système utilisant la segmentation du système de RAP produit le plus faible taux de fausse alarme du lot grâce aux alignements qu'il effectue. Cependant le taux de parole manquée est aussi le plus élevé : certains éléments de la parole ne sont pas détectés comme des mots donc supprimés. Le DER qui en résulte est néanmoins le deuxième meilleur du groupe, mais ce système est pénalisé par un temps de calcul élevé par rapport aux autres systèmes.

Enfin, le LSTM bidirectionnel offre des taux d'erreur similaires au GMM-HMM, mais une augmentation 0.1% absolue du taux d'erreur total se traduit par une augmentation nette de 0.55% DER, classant ce système en avant-dernière position.

Quant aux différents taux d'erreur de mots, ceux-ci ne varient que très peu. Nous pouvons tout de même constater qu'il est préférable d'avoir un faible taux de parole manquée pour avoir un meilleur WER.

Un SAD « parfait » produirait un taux DER de 7.93% et un WER de 14.00%. Il reste donc une marge d'amélioration pour ces systèmes, bien que celle-ci soit faible.

Le corpus d'évaluation, composé de 87 émissions très diverses, permet d'obtenir une performance moyenne des systèmes. Cependant le système GMM-HMM est devancé légèrement par la majorité des autres systèmes sur le corpus ESTER1 pour la tâche de SRL. Ce corpus est le corpus le plus facile en moyenne pour cette tâche. Il contient peu de parole spontanée et de parole superposée.

6 Conclusion

Cette étude conduite à partir d'un même corpus d'apprentissage nous a permis de comparer plusieurs systèmes de détection de la parole sur deux tâches de traitement de la parole : la segmentation et regroupement en locuteurs et la transcription. Le système classique GMM-HMM s'est avéré être le plus robuste grâce à son apprentissage de plusieurs modèles plutôt qu'un simple parole/non-parole binaire sur les deux tâches évaluées.

Sur les 87 enregistrements constituant un large corpus de test contenant des enregistrements nature divers, nous n'avons pas pu démontrer que les systèmes à base DNN surpassaient un système classique GMM-HMM. Il est à noter toutefois que le corpus d'apprentissage, construit au départ pour l'apprentissage du GMM-HMM développé durant la campagne ESTER1, n'a pas été remis en question.

Dans la mouvance actuelle qui consiste à développer des systèmes entièrement neuronaux (end-to-end), il faut noter que les performances des systèmes neuronaux, sans dépasser le système GMM-HMM, offrent des perspectives importantes. En effet, les systèmes présentés dans cette étude sont optimisés pour une tâche de détection de parole qui ne représente pas un cas d'usage mais seulement une première étape nécessaire aux systèmes finaux qui apportent une valeur ajoutée. L'intégration des systèmes de détection de parole neuronaux au sein d'une architecture complètement neuronale permettrait d'optimiser cette étape de sélection directement pour la tâche visée. Cette intégration constitue la prochaine étape de nos travaux.

Systèmes	Corpus	FA	MISS	SNS	DER	Pureté	Couv.	WER
GMM-HMM	ESTER1	0.59	0.35	0.94	6.50	94.18	96.90	11.07
	ESTER2	1.11	0.20	1.30	6.26	96.57	97.76	11.59
	ETAPE	3.19	0.24	3.43	15.69	84.30	85.29	25.96
	REPERE	0.53	0.95	1.48	9.30	88.74	91.25	14.89
	Moyenne	1.39	0.35	**1.74**	**8.95**	91.44	93.34	**15.02**
DNN-HMM	ESTER1	0.67	0.30	0.97	5.85	94.34	96.99	11.21
	ESTER2	1.93	0.47	2.40	7.91	96.83	97.76	11.99
	ETAPE	3.68	0.17	3.86	17.49	83.41	83.72	26.15
	REPERE	0.80	0.42	1.23	10.64	86.56	90.53	14.49
	Moyenne	1.80	**0.33**	2.13	9.71	90.70	92.85	15.07
MIMIC	ESTER1	0.45	0.50	0.95	6.41	94.25	97.00	11.12
	ESTER2	0.69	0.31	1.01	5.89	96.94	97.79	11.65
	ETAPE	2.86	0.46	3.33	17.10	83.46	83.69	26.16
	REPERE	0.41	2.04	2.46	10.50	87.76	90.76	16.11
	Moyenne	1.13	0.61	**1.74**	9.29	91.10	92.93	15.48
TRANSCR.	ESTER1	0.33	0.89	1.22	5.66	96.03	96.90	-
	ESTER2	0.47	0.53	1.01	6.19	97.05	97.46	-
	ETAPE	2.52	0.74	3.25	17.65	83.80	84.94	-
	REPERE	0.42	1.32	1.74	10.29	89.33	90.86	-
	Moyenne	**0.94**	0.80	**1.74**	9.19	**92.07**	93.06	-
BLSTM	ESTER1	0.59	0.32	0.91	6.55	94.43	96.34	11.08
	ESTER2	1.26	0.32	1.58	6.59	96.81	97.66	12.00
	ETAPE	3.15	0.31	3.46	17.20	82.68	85.62	26.21
	REPERE	0.49	1.29	1.78	10.09	87.62	91.90	15.80
	Moyenne	1.42	0.42	1.84	9.50	90.92	**93.46**	15.48

TABLE 3 – Résultats des systèmes de détection de parole appliqués en segmentation et regroupement et locuteur et en transcription automatique.
FA : détection de la parole erronée, MISS : détection manquée de parole, SNS : FA+MISS.
DER : erreur de segmentation et de regroupement en locuteur, Pureté et Couv. : pureté et couverture en locuteur des segments. WER : erreur de transcription automatique.

Remerciements

Nous tenons à remercier Kong Aik Lee et Rafael E. Banchs de l'*Institute for Infocomm Research* (I^2R) ainsi qu'Antoine Laurent du LIUM pour leur aide précieuse.

Références

BERGSTRA J., BREULEUX O., BASTIEN F., LAMBLIN P., PASCANU R., DESJARDINS G., TURIAN J., WARDE-FARLEY D. & BENGIO Y. (2010). Theano : A CPU and GPU math compiler in Python. In *Proc. 9th Python in Science Conf*, p. 1–7.

BREDIN H. (2017). pyannote. metrics : a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*.

CHOLLET F. (2016). keras : Deep Learning for humans. original-date : 2015-03-28T00 :35 :42Z.

DEHAK N., KENNY P. J., DEHAK R., DUMOUCHEL P. & OUELLET P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*.

DUPUY G., MEIGNIER S., DELÉGLISE P. & ESTEVE Y. (2014). Recent improvements on ilp-based clustering for broadcast news speaker diarization. In *Odyssey 2014*.

GALLIANO S., GRAVIER G. & CHAUBARD L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*.

GAUVAIN J.-L., LAMEL L. & ADDA G. (2002). The LIMSI Broadcast News transcription system. *Speech Communication*, **37**(1), 89–108.

GELLY G. & GAUVAIN J. L. (2018). Optimization of RNN-Based Speech Activity Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **26**(3), 646–656.

GIRAUDEL A., CARRÉ M., MAPELLI V., KAHN J., GALIBERT O. & QUINTARD L. (2012). The REPERE Corpus : a multimodal corpus for person recognition. In *LREC*, p. 1102–1107.

GRAVIER G., ADDA G., PAULSON N., CARRÉ M., GIRAUDEL A. & GALIBERT O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC-Eighth international conference on Language Resources and Evaluation*, p.ña.

GRAVIER G., BONASTRE J.-F., GEOFFROIS E., GALLIANO S., MCTAIT K. & CHOUKRI K. (2004). The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *LREC*.

HUGHES T. & MIERLE K. (2013). Recurrent neural networks for voice activity detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* : IEEE.

KAUR S. & SOHAL J. S. (2017). Speech Activity Detection and its Evaluation in Speaker Diarization System. *INTERNATIONAL JOURNAL*, **16**(1).

KINGSBURY B., JAIN P. & ADAMI A. (2002). A hybrid HMM/TRAPS model for robust voice activity detection. In *Seventh International Conference on Spoken Language Processing*.

LARCHER A., LEE K. A. & MEIGNIER S. (2016). An extensible speaker identification sidekit in Python. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, p. 5095–5099 : IEEE.

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

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y. & SCHWARZ P. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* : IEEE Signal Processing Society.

RENEVEY P. & DRYGAJLO A. (2001). Entropy based voice activity detection in very noisy conditions. p. 1887–1890.

ROHDIN J., SILNOVA A., DIEZ M., PLCHOT O., MATEJKA P. & BURGET L. (2017). End-to-end DNN Based Speaker Recognition Inspired by i-vector and PLDA. *arXiv :1710.02369 [cs, eess]*.

ROUSSEAU A., BOULIANNE G., DELÉGLISE P., ESTÈVE Y., GUPTA V. & MEIGNIER S. (2014). LIUM and CRIM ASR system combination for the REPERE evaluation campaign. In *International Conference on Text, Speech, and Dialogue*, p. 441–448 : Springer.

RYANT N., LIBERMAN M. & YUAN J. (2013). Speech activity detection on youtube using deep neural networks. In *INTERSPEECH*, p. 728–731.

SEPP HOCHREITER & JÜRGEN SCHMIDHUBER (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780.

SHAHSAVARI S., SAMETI H. & HADIAN H. (2017). Speech activity detection using deep neural networks. In *2017 Iranian Conference on Electrical Engineering (ICEE)*, p. 1564–1568.

YIN R., BREDIN H. & BARRAS C. (2017). Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks. p. 3827–3831 : ISCA.