Prédiction de performance des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs

Zied Elloumi, Benjamin Lecouteux, Olivier Galibert, Laurent Besacier

Prédiction de performance des systèmes de reconnaissance automatique de la parole à l'aide de réseaux de neurones convolutifs

Zied Elloumi[1,2], **Benjamin Lecouteux**[2], **Olivier Galibert**[1], **Laurent Besacier**[2]

[1] *Laboratoire national de métrologie et d'essais (LNE) , France*
prénom.nom@univ-grenoble-alpes.fr

[2] *Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France*
prénom.nom@lne.fr

RÉSUMÉ. *Dans ce travail, nous nous intéressons à la tâche de prédiction de performance des systèmes de transcription de la parole. Nous comparons deux approches de prédiction: une approche de l'état de l'art fondée sur l'extraction explicite de traits et une nouvelle approche fondée sur des caractéristiques entraînées implicitement à l'aide des réseaux neuronaux convolutifs (CNN). Nous essayons ensuite de comprendre quelles informations sont capturées par notre modèle neuronal et leurs liens avec différents facteurs. Pour tirer profit de cette analyse, nous proposons un système multitâche qui se montre légèrement plus efficace sur la tâche de prédiction de performance.*

ABSTRACT. *This paper focuses on the ASR performance prediction task. Two prediction approaches are compared: a state-of-the-art performance prediction based on engineered features and a new strategy based on learnt features using convolutional neural networks. We also try to better understand which information is captured by the deep model and its relation with different conditioning factors. To take advantage of this analysis, we then try to leverage these 3 types of information at training time through multi-task learning, which is slightly more efficient on ASR performance prediction task.*

MOTS-CLÉS : *prédiction de performance, reconnaissance de la parole continue à grand vocabulaire, réseau neuronal convolutif.*

KEYWORDS: *performance prediction, large vocabulary continuous speech recognition, convolutional neural networks.*

1. Introduction

Prédire la performance d'un système de reconnaissance automatique de la parole (SRAP) sur de nouveaux enregistrements (par exemple de nouveaux types de programmes TV ou radio jamais rencontrés auparavant) est un Graal important de la reconnaissance automatique de la parole. Résoudre une telle tâche permet de prévoir la difficulté de transcription d'une nouvelle collection de documents audio et d'avoir une idée de l'effort nécessaire qui sera demandé à des annotateurs humains pour produire des transcriptions (références) correctes à partir de transcriptions automatiques (hypothèses).

Un système de prédiction de performance prend en entrée des données textuelles (l'hypothèse du système de SRAP) et/ou acoustiques (le signal source ayant donné l'hypothèse) et prédit une performance (un taux d'erreurs de mots ou *word error rate*) associée sans disposer de la transcription de référence. Ceci peut être fait à différentes granularités : tour de parole, document, type de programme TV ou radio, etc. De plus, pour la tâche de prédiction de performance, le système de reconnaissance de la parole est généralement considéré comme une « boîte noire », c'est-à-dire un système pour lequel nous n'aurons accès au fonctionnement interne, ni aux N meilleures hypothèses.

La prédiction de performance de transcription automatique à partir d'une collection de signaux est légèrement différente de la détection d'erreurs ou de l'estimation de mesures de confiance sur une sortie de système de reconnaissance de la parole (SRAP) car elle a pour but de donner une estimation générale de la difficulté de la tâche de transcription sur un ensemble d'enregistrements. Par ailleurs, résoudre une telle tâche de prédiction peut aider à mieux analyser les difficultés rencontrées par les systèmes de SRAP sur différents types d'enregistrements et les facteurs qui affectent leurs performances.

1.1. *Contribution*

Dans ce travail, nous nous intéressons à la tâche de prédiction de performance des systèmes de reconnaissance de la parole (SRAP) sur des collections d'émissions TV ou radio. Cet article reprend des résultats déjà publiés en anglais (Elloumi *et al.*, 2018) où nous proposons un corpus ainsi qu'un protocole d'évaluation pour la prédiction de performance de SRAP associés à des expérimentations fondées sur deux approches de prédiction des performances fondées sur des traits explicites (*engineered features*) et sur des traits entraînés au cours de l'apprentissage de réseaux de neurones convolutifs (*learnt features*). Nous y ajoutons une contribution originale visant à analyser les représentations apprises et les informations capturées par le réseau de neurones.

Nous commençons par présenter un corpus en français large et hétérogène (multiples programmes TV ou radio, mélange de la parole non spontanée et spontanée, différents accents) dédié à cette tâche, ainsi que le protocole d'évaluation utilisé. Nous comparons deux approches de prédiction : une approche de l'état de l'art fon-

dée sur des traits explicites et une nouvelle approche fondée sur des réseaux neuronaux convolutifs (CNN). L'utilisation jointe de traits textuels et acoustiques n'apporte pas de gains dans l'approche de l'état de l'art, tandis qu'elle permet d'obtenir de meilleures prédictions en utilisant les CNN. Nous montrons également que les CNN prédisent clairement la distribution des taux d'erreurs sur une collection d'enregistrements, contrairement à l'approche de l'état de l'art qui génère une distribution éloignée de la réalité. Nous essayons ensuite de comprendre quelles informations sont capturées par notre modèle neuronal et leurs liens avec différents facteurs (style de parole, accent, etc.). Nos expériences montrent que les représentations intermédiaires dans le réseau encodent spontanément des informations sur le style de la parole, l'accent du locuteur ainsi que le type d'émission. Pour tirer profit de cette analyse, nous proposons une approche multitâche qui se montre légèrement plus efficace sur la tâche de prédiction de performance.

1.2. *Plan*

Ce document est organisé comme suit : dans un premier temps, nous présentons dans la section 2 les travaux existants sur la tâche de prédiction de performance ainsi que sur l'analyse des représentions intermédiaires apprises par les réseaux de neurones. Dans la section 3, nous présentons notre protocole d'évaluation. Nous comparons par la suite dans la section 4 deux approches de prédiction (TranscRater *vs* CNN). Dans la section 5, nous détaillons la méthodologie utilisée pour analyser et évaluer des représentations intermédiaires apprises par notre meilleur système neuronal profond. Ensuite, nous présentons, dans la section 6, les systèmes de prédiction multitâche qui exploitent, au cours de l'apprentissage, des informations telles que style de parole, accent du locuteur et type d'émission. Finalement, nous concluons notre travail dans la section 7.

2. Travaux liés

De nombreux travaux ont proposé d'estimer des mesures de confiance afin de détecter les erreurs dans les sorties des SRAP. La détection des erreurs consiste à étiqueter chaque mot en entrée comme « correct » ou « incorrect » (tâche de classification). Les mesures de confiance ont été introduites pour la tâche de détection des mots hors vocabulaire (OOV) par Asadi *et al.* (1990) et exploitées par Young (1994) qui a utilisé les probabilités *a posteriori* (WPP) pour la tâche de reconnaissance de la parole.

La tâche de prédiction des performances va au-delà de l'estimation de confiance puisqu'elle ne se concentre pas sur un système de reconnaissance automatique de la parole donné ni sur des treillis ou des N meilleures hypothèses. Elle a pour but de donner une estimation générale de la difficulté de la tâche de transcription. Tandis que la tâche de prédiction de performance est traitée à l'aide d'approches à base de régression, la tâche d'estimation de confiance est traitée avec des approches à base de classification. Plusieurs travaux se fondent essentiellement sur des traits acoustiques

pour prédire les performances, Hermansky *et al.* (2013) exploitent des caractéristiques temporelles du signal vocal (*Mean Temporal Distance* – calculées sur le signal et corrélées avec le rapport signal sur bruit) pour prédire la performance. Ferreira *et al.* (2018) proposent d'analyser le comportement de l'énergie à court terme du bruit et de la parole en tenant compte de divers facteurs tandis que le système RAP est considéré comme une boîte noire. Les auteurs comparent deux approches de régression (MLP et linéaire) en prenant en compte la variabilité des systèmes de RAP en fonction du volume et du type de bruit. Les performances obtenues montrent que la régression MLP est meilleure que la régression linéaire. Meyer *et al.* (2017) proposent une méthode de prédiction de performance de RAP apprise avec des données propres et évaluée sur 10 types de bruits inconnus et une large gamme de rapports signal/bruit des corpus DRE01 (Dreschler *et al.*, 2001), Noisex et BBC SOUND EFFECTS. Les résultats montrent que le bruit dans les données influence la qualité des systèmes de prédiction. Negri *et al.* (2014) proposent d'autres types de traits (autres que le signal) comme les informations internes d'un SRAP, des caractéristiques acoustiques, des caractéristiques hybrides, et des caractéristiques textuelles. Trois scénarios de prédiction ont été proposés afin d'étudier l'impact de présence/absence de caractéristiques particulières extraites des SRAP, ainsi que l'effet de l'homogénéité/non-homogénéité des données d'apprentissage et d'évaluation sur la qualité des systèmes de prédiction. Les performances obtenues montrent que la qualité des systèmes de prédiction dépend de l'homogénéité entre les données d'entraînement et les données d'évaluation. Jalalvand *et al.* (2016) ont proposé un outil *open source* nommé TranscRater qui se fonde essentiellement sur l'extraction de traits (caractéristiques phonétiques, syntaxiques, acoustiques et issues du modèle de langue) et utilise un algorithme fondé sur une régression pour prédire un taux d'erreurs. Le SRAP est considéré comme une « boîte noire », et l'évaluation a été effectuée sur les données de la campagne CHiME-3. [1] Dans ce travail, qui se rapproche le plus de notre contribution, les expérimentations montrent que les caractéristiques acoustiques (issues directement du signal) n'ont pas d'influence sur la qualité du système de prédiction.

Les travaux présentés précédemment s'appuient sur des traits (ou *features*) prédéfinis qui exigent des outils et des ressources spécifiques pour une langue afin de prédire la performance. À l'aide des CNN, nous visons à proposer une méthode flexible (ne dépendant pas de la langue) qui se fonde sur des représentations apprises au cours de l'apprentissage du système. Un autre apport de notre travail est d'encoder les informations du signal de parole dans un CNN pour la prédiction de performance des SRAP. L'encodage du signal pour un CNN peut être effectué à partir de la sortie d'un module de traitement du signal (*front end*) qui transforme le signal en une suite de vecteurs acoustiques (Piczak, 2015 ; Sainath *et al.*, 2015 ; Jin *et al.*, 2016). Cependant, certains travaux récents ont directement utilisé le signal brut en entrée d'un CNN pour la reconnaissance vocale (Sainath *et al.*, 2015 ; Palaz *et al.*, 2015) et pour la classification de signaux de parole (Dai *et al.*, 2017). Des travaux liés à cet objectif ont aussi

1. `http://spandh.dcs.shef.ac.uk/chime_challenge/chime2015/`

consisté à détecter des classes phonétiques à partir de signaux de parole analysés avec des réseaux neuronaux (Pellegrini et Mouysset, 2016 ; Nagamine *et al.*, 2015).

La contribution principale présentée ici est la prédiction de performance à l'aide de réseaux de neurones qui se fondent sur des traits multimodaux : acoustiques et textuels observés au cours de l'apprentissage. Toutefois, il est important d'interpréter les représentations intermédiaires apprises par le réseau afin de comprendre quelles informations ont été capturées. Des travaux récents sur la tâche de reconnaissance automatique de la parole ont proposé d'analyser les représentations capturées par les SRAP profonds. Mohamed *et al.* (2012) et Belinkov et Glass (2017) ont analysé les représentations intermédiaires apprises (d'un SRAP profond) en utilisant la visualisation t-SNE (Maaten et Hinton, 2008). Ils essaient aussi de comprendre quelles couches capturent mieux les informations phonétiques en entraînant un classifieur de phonèmes peu profond. Par ailleurs, Wu et King (2016) ont évalué les représentations de plusieurs variantes de LSTM pour une tâche de synthèse vocale. Wang *et al.* (2017) ont, quant à eux, proposé une étude sur trois types de représentations apprises pour une tâche de reconnaissance de locuteur : i-vecteur, d-vecteur et s-vecteur (fondé sur un réseau LSTM). Des tâches de classification annexes ont été conçues pour mieux comprendre comment sont encodées les informations sur les locuteurs. Un apprentissage multitâche est également proposé pour intégrer ces différents types de représentations, ce qui mène à une meilleure performance d'identification du locuteur. Nous trouvons aussi des travaux similaires dans d'autres applications du traitement automatique du langage naturel (TALN), comme en traduction automatique neuronale par exemple. Parmi ces travaux récents, nous pouvons citer les travaux de Shi *et al.* (2016) et Belinkov *et al.* (2017) qui ont essayé de comprendre les représentations apprises par un système de traduction neuronal. Ces représentations sont fournies à un classifieur peu profond afin de prédire des étiquettes syntaxiques (Shi *et al.*, 2016), grammaticales ou sémantiques (Belinkov *et al.*, 2017). L'analyse montre que les couches inférieures sont meilleures pour l'étiquetage grammatical.

3. Protocole d'évaluation

Nous nous intéressons à la prédiction des performances de systèmes de transcription de la parole sur des émissions non vues durant l'apprentissage. Notre objectif est de prédire la performance lorsque les informations internes d'un SRAP sont indisponibles. Notre cas d'étude se fonde sur un système de prédiction de performance n'utilisant que les transcriptions automatiques (fournies par un SRAP) et/ou le signal audio afin de prédire la performance de la transcription correspondante. Les transcriptions de référence (humaines) ne sont disponibles que pour évaluer le système de transcription de la parole et pour produire un rapport de performance. Un corpus nommé $\textit{Train}_{pred}$ est utilisé pour construire nos systèmes de prédiction. Il est constitué de triplets {signaux, transcription automatique, performance} pour 75 k tours de parole. Le corpus $\textit{Test}_{pred}$ est utilisé pour évaluer le système de prédiction, il contient aussi les triplets {signaux, transcription automatique, performance} (6,8 k tours de

parole) mais la performance est inconnue du système au moment de la prédiction et dévoilée seulement au moment de l'évaluation de la qualité de prédiction.

Afin d'implémenter ce protocole, nous avons besoin d'un système de reconnaissance de la parole pour produire les transcriptions automatiques et la performance associée pour l'intégralité des corpus Train$_{pred}$ et Test $_{pred}$, ce qui nous permettra d'entraîner et d'évaluer des systèmes de prédiction de performance.

3.1. *Corpus*

Les données utilisées dans notre protocole proviennent de différentes collections d'émissions en français :

1) un sous-ensemble du corpus Quaero[2] qui contient 41 heures de discours radiodiffusés de différents programmes de radio et de télévision français sur divers sujets ;

2) les données du projet ETAPE (Gravier *et al.*, 2012) qui comportent 37 heures d'émissions de radio et de télévision (principalement des discours spontanés avec des locuteurs qui se chevauchent) ;

3) des données des campagnes d'évaluation ESTER 1 & ESTER 2 (Galliano *et al.*, 2005) qui contiennent 111 heures d'enregistrement audio transcrit. Ce sont principalement des programmes de radio français et africains (mélange de discours préparés et plus spontanés : parole du présentateur, interviews, reportages) ;

4) les données de la campagne d'évaluation REPERE (Kahn *et al.*, 2012) :
54 heures d'émissions transcrites de parole spontanée (des débats TV) et de la parole préparée (journaux télévisés).

Comme décrit dans le tableau 1, nos données contiennent de la parole non spontanée (NS) et de la parole spontanée (S). Les données d'entraînement (Train$_{SRAP}$) de notre système de transcription de la parole automatique sont sélectionnées à partir des données non spontanées qui correspondent essentiellement à des journaux télévisés. Les données utilisées pour la tâche de prédiction (*Train*$_{pred}$ et *Test*$_{pred}$) sont un mélange des deux styles de parole (S et NS). Il est important de mentionner que les émissions du corpus *Test*$_{Pred}$ n'existent pas dans le *Train*$_{Pred}$ et *vice versa*. En outre, des émissions plus difficiles (ayant des taux d'erreurs plus élevés) ont été sélectionnées pour *Test*$_{Pred}$. La distribution détaillée des taux d'erreurs sur notre corpus Test$_{pred}$ est donnée plus loin dans la figure 3. Dans le tableau 2, nos émissions ayant un style de parole spontanée ont systématiquement un taux d'erreurs plus élevé (de 28,74 % à 45,15 % selon l'émission) par rapport aux émissions ayant un style de parole non spontanée (de 12,06 % à 25,41 % selon l'émission). Cette division S et NS nous permettra de comparer nos systèmes de prédiction de performance sur différents types de documents contenant du discours NS et S.

2. http ://www.quaero.org

	Train$_{SRAP}$	Train$_{Pred}$	Test$_{Pred}$
NS	100 h 51	30 h 27	4 h 17
S	-	59 h 25	4 h 42
Durée	100 h 51	89 h 52	8 h 59
TEM	-	**22,29**	**31,20**

Tableau 1. *Distribution de nos corpus entre les styles de parole non spontanée (NS) et spontanée (S) – TEM = taux d'erreurs mots (WER)*

Source	Émission	Mots	TEM
	Non spontanées (NS)		
Quaero	Arte News (AN)	3 726	12,06
ESTER 2	Tvme (T)	10 706	18,44
Quaero	France Culture TEMPS (FCT)	10 091	20,92
Quaero	Fab histoire (FH)	10 022	22,76
ESTER 2	Africa1 (A1)	15 257	25,41
	Spontanées (S)		
Quaero	Ce soir ou jamais (CSOJ)	10 992	28,74
REPERE	Planete showbiz (PS)	15 946	36,74
REPERE	Culture et vous (CV)	16 026	39,79
ETAPE	La place du village (PV)	20 396	45,15

Tableau 2. *Performance sur le corpus Test$_{Pred}$ en termes de TEM (taux d'erreurs de mots)*

3.2. *Métriques d'évaluation*

Nous avons utilisé la boîte à outils LNE-Tools (Galibert, 2013) afin d'évaluer la qualité de notre système de transcription et produire les rapports de performance en termes de taux d'erreurs de mots (TEM). La parole superposée et les tours de parole vides sont supprimés. Comme il est mentionné dans le tableau 1, nous avons obtenu 22,29 % de TEM sur le corpus Train$_{pred}$ et 31,20 % de TEM sur le corpus Test$_{pred}$.

Afin d'évaluer la tâche de prédiction de performance, nous utilisons la métrique *Mean Absolute Error* (MAE) définie comme suit :

$$MAE = \frac{\sum_{i=1}^{N} |TEM_{Ref}^{i} - TEM_{Pred}^{i}|}{N} \quad [1]$$

avec N le nombre d'unités (tours de parole ou document complet).

Nous utilisons également le coefficient de corrélation de rang Kendall entre le score de référence et la sortie du système de prédiction au niveau des tours de parole. Plus le score Kendall est proche de 1, plus les performances prédites sont proches des vraies performances mesurées.

3.3. *Système de reconnaissance automatique de la parole*

Afin de produire des transcriptions automatiques pour le système de prédiction de performance, nous avons construit un système de transcription automatique de la parole fondé sur la boîte à outils KALDI (Povey *et al.*, 2011), en suivant la recette standard. Un système hybride HMM-DNN a été appris en utilisant le corpus Train_{SRAP} (100 heures de journaux diffusées par ESTER, REPERE, ETAPE et Quaero). Nous avons entraîné un modèle de langue 5-grammes à partir de plusieurs corpus français (3 milliards de mots au total : EUbookshop, TED2013, Wit3, GlobalVoices, Gigaword, Europarl-v7, MultiUN, OpenSubtitles2016, DGT, News Commentary, News WMT, *Le Monde*, Trames, Wikipédia) et les transcriptions de notre jeu de données $Train_{SRAP}$ en utilisant l'outil SRILM (Stolcke *et al.*, 2002). Pour le modèle de prononciation, nous avons utilisé la ressource lexicale BDLEX (De Calmès et Pérennou, 1998) ainsi que l'outil de conversion automatique de graphèmes à phonèmes LIA_Phon[3] afin de trouver les variantes de prononciation de notre vocabulaire (limité à 80 k mots).

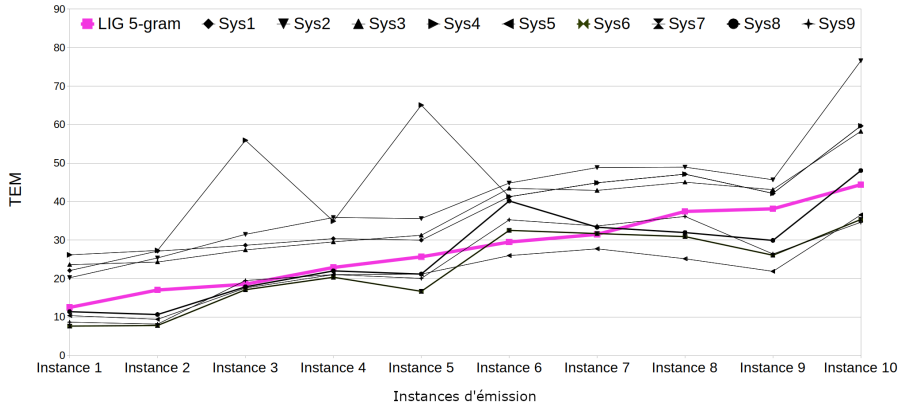


Figure 1. *Performance (TEM) de notre SRAP comparé aux performances d'autres systèmes sur des données identiques. Chaque instance en abscisse représente une instance d'un type d'émission.*

Dans la figure 1, nous comparons les performances de notre système LIG 5-gram (colorées en rose) à celles obtenues au cours des différentes campagnes d'évaluation (colorées en noir) sur les mêmes instances d'émissions (10 au total). Le système que nous avons développé est situé au milieu des systèmes proposés. Cela signifie que notre système produit des transcriptions correctes et que les performances sont corrélées avec celles des autres systèmes.

3. `http://lia.univ-avignon.fr/chercheurs/bechet/download/lia_phon.v1.2.jul06.tar.gz`

4. Prédiction de performance

Dans cette section, nous présentons une étude comparative entre une approche de prédiction fondée sur des caractéristiques explicites (système de base) et une nouvelle approche de prédiction fondée sur des caractéristiques entraînées en utilisant un réseau de neurones convolutif. Les transcriptions automatiques des corpus Train$_{pred}$ et Test$_{pred}$ ont été obtenues par notre système SRAP présenté dans la section 3.3. Les rapports de performance ont été générés par l'outil LNE-Tool (un taux d'erreurs de mots par tour de parole).

4.1. *Prédiction fondée sur des traits explicites (baseline)*

Afin d'avoir un système de prédiction de l'état de l'art (extraction des traits prédéfinis), nous avons adapté l'outil TranscRater (Jalalvand *et al.*, 2016) de l'anglais vers le français. Ce dernier s'appuie sur l'extraction de traits explicites (*engineered features*) pour prédire la performance de chaque entrée en termes de TEM. En exploitant des résultats empiriques antérieurs dans (Negri *et al.*, 2014 ; de Souza *et al.*, 2013 ; Jalalvand *et al.*, 2015b ; de Souza *et al.*, 2015), TranscRater exploite l'algorithme Extremely Randomized Trees (Geurts *et al.*, 2006) pour l'apprentissage du système. La sélection des traits est effectuée avec l'algorithme Randomized Lasso (Meinshausen et Bühlmann, 2010). Les principaux hyperparamètres du modèle sont optimisés à l'aide d'une grille de recherche avec une validation croisée sur l'ensemble des données d'apprentissage, afin de minimiser l'erreur absolue moyenne (MAE) entre les vrais TEM et les TEM prédits.

TranscRater est capable d'extraire 63 traits de quatre types :

– **9 traits morphosyntaxiques (POS)** : permet de capturer la plausibilité de la transcription d'un point de vue syntaxique en utilisant l'outil Treetagger (Schmid, 1995). Pour chaque mot compris dans un tour de parole transcrit, un score de prédiction d'étiquette POS est attribué au niveau du mot lui-même ainsi qu'au précédent et au suivant. Cette fenêtre glissante de 3 mots est utilisée pour calculer la valeur moyenne de l'ensemble du tour de parole transcrit. De plus, le vecteur de traits comporte également le nombre et le pourcentage de classes de *tokens* (nombres, noms, verbes, adjectifs et adverbes). Ces traits ont été testés dans diverses conditions (données propres ou bruitées, microphones simples ou multiples) (Jalalvand *et al.*, 2015a ; Jalalvand *et al.*, 2015c).

– **3 traits issus du modèle de langue (LM)** : permet de capturer la plausibilité de la transcription selon un modèle n-gramme. Ils comprennent la moyenne des probabilités des mots, la somme des log-probabilités et le score de perplexité pour chaque transcription. Un modèle 5-grammes est entraîné en utilisant l'outil SRILM (Stolcke *et al.*, 2002) sur l'ensemble des corpus textes de 3 milliards de mots mentionné dans la section 3.3 ;

– **7 traits lexicaux (LEX)** : les traits sont extraits à partir du lexique de notre système de transcription : un vecteur de traits contenant la fréquence des catégories de phonèmes liées à la prononciation de chaque mot ;

– **44 traits acoustiques (SIG)** : ils capturent des informations sur le signal d'entrée (conditions générales d'enregistrement, accents spécifiques au locuteur). Pour l'extraction des traits, TranscRater calcule 13 paramètres de type MFCC (en utilisant openSMILE (Eyben *et al.*, 2010)), leurs dérivées, accélération et log-énergie, fréquence fondamentale (F0), probabilité de voisement, contours d'intensité et le *pitch* pour chaque trame de parole. Pour l'ensemble du signal d'entrée, le vecteur de traits SIG est obtenu en calculant la moyenne des valeurs de chaque trame.

4.2. *Prédiction par les réseaux neuronaux convolutifs (CNN)*

Afin de prédire le TEM, nous proposons une nouvelle approche de régression supervisée fondée sur des réseaux de neurones convolutifs. Notre réseau prend en entrée des données textuelles et/ou des données acoustiques (signal brut, des MFCC ou des spectrogrammes). Suivant notre protocole expérimental, le système de RAP est considéré comme une boîte noire, et seuls les signaux et/ou les transcriptions automatiques sont fournis pour créer et évaluer des systèmes de prédiction. Nous avons construit notre modèle en utilisant à la fois Keras (Chollet *et al.*, 2015) et Tensorflow[4].

Pour l'entrée textuelle, nous proposons une architecture inspirée de Kim (2014) (verte dans figure 2). L'entrée est un tour de parole complété à N mots (N est défini comme la longueur de la plus longue phrase dans notre corpus complet) présenté sous forme d'une matrice EMBED de taille N $\times$ M (M = la dimension des représentations vectorielles de mots). Ainsi, chaque ligne de la matrice EMBED correspond à une représentation vectorielle (appelée aussi *embeddings*) d'un mot. Ces *embeddings* ont été initialisés à l'aide d'un modèle pré-entraîné sur l'ensemble des corpus textuels (3 milliards de mots mentionnés dans la section 3.3) en utilisant l'outil Word2Vec (Mikolov *et al.*, 2013) puis, mis à jour automatiquement au moment de l'apprentissage du réseau.

Chaque opération de convolution implique un filtre w qui est appliqué à un segment de h mots pour produire une nouvelle caractéristique. Par exemple, la caractéristique c_i est générée à partir des mots $x_{i:i+h-1}$ comme :

$$c_i = f(w \cdot x_{i:i+h-1} + b) \quad [2]$$

où b est un terme de biais et f est une fonction non linéaire. Ce filtre est appliqué à chaque segment du mot dans le tour de parole pour produire un vecteur (*feature map*) $c = [c_1, c_2 ... c_{n-h+1}]$. Ensuite, une opération de type *max-pooling* (Collobert *et al.*, 2011) prend les 4 plus grandes valeurs de c, qui sont ensuite moyennées. L'ensemble des filtres W fournit une entrée de taille fixe (nombre de filtres de W * nombre de

4. `https://www.tensorflow.org`

régions de convolution) aux deux couches cachées entièrement connectées (256 et 128 unités) suivies respectivement d'une régularisation de type $dropout$ (0,2 et 0,6) avant la prédiction de la performance (TEM).

Pour l'entrée du signal, nous utilisons la meilleure architecture proposée dans (Dai *et al.*, 2017) (colorée en rouge dans la figure 2). Il s'agit d'un CNN profond avec 17 couches convolution + *max-pooling* suivies d'une opération d'agrégation (*global average pooling)* et de trois couches cachées complètement connectées (512, 256 et 128 unités). Nous avons ajouté une régularisation (*dropout*) de 0,2 entre les deux dernières couches (256 et 128). Nous proposons plusieurs méthodes pour encoder le signal avec le CNN en utilisant Librosa (McFee *et al.*, 2015) : les échantillons du signal brut (RAW - SIG), le spectrogramme (MEL-SPEC) ou des coefficients MFCC.

Afin de prédire un taux d'erreurs (TEM) à l'aide des réseaux CNN, nous proposons deux approches différentes :

– **CNN**$_{Softmax}$: nous utilisons les probabilités Softmax et un vecteur fixe externe nommé TEM$_{Vector}$ pour calculer le TEM prédit (TEM$_{Pred}$). TEM$_{Vector}$ et les probabilités Softmax doivent avoir la même dimension. TEM$_{Pred}$ est alors défini comme suit :

$$TEM_{Pred} = \sum_{C=1}^{NC} P_{Softmax}(C) * TEM_{Vector}(C) \quad [3]$$

NC est la dimension du vecteur TEM$_{Vector}$. Dans nos expériences, NC est égale à 6 et TEM$_{Vector}$ = [0 %, 25 %, 50 %, 75 %, 100 %, 150 %] ;

– **CNN**$_{ReLU}$: nous appliquons la fonction ReLU (la taille de sortie est égale à 1) à la dernière couche cachée du réseau. Cette fonction permettra d'estimer directement le TEM en retournant une valeur de type réel entre 0 et $+\infty$.

Pour l'utilisation jointe des données textuelles et acoustiques, nous fusionnons les deux dernières couches cachées de CNN EMBED et CNN RAW - SIG (ou MEL-SPEC ou MFCC) en les concaténant et en les faisant passer à une nouvelle couche cachée (de taille 128) avant la prédiction de TEM avec le CNN$_{Softmax}$ ou le CNN$_{ReLU}$ (représentées par des lignes en pointillé dans la figure 2). Nous entraînons par la suite le réseau de la même manière.

Contrairement aux traits de l'approche de base (extraction qui nécessite d'avoir défini les traits au préalable, on parle dans ce cas d'*engineered features*), les traits CNN textuels sont extraits et entraînés à partir des représentations vectorielles des mots (on parle alors de *learnt features*). Ces traits sont appris par le réseau neuronal jusqu'à ce que le comportement désiré soit obtenu.

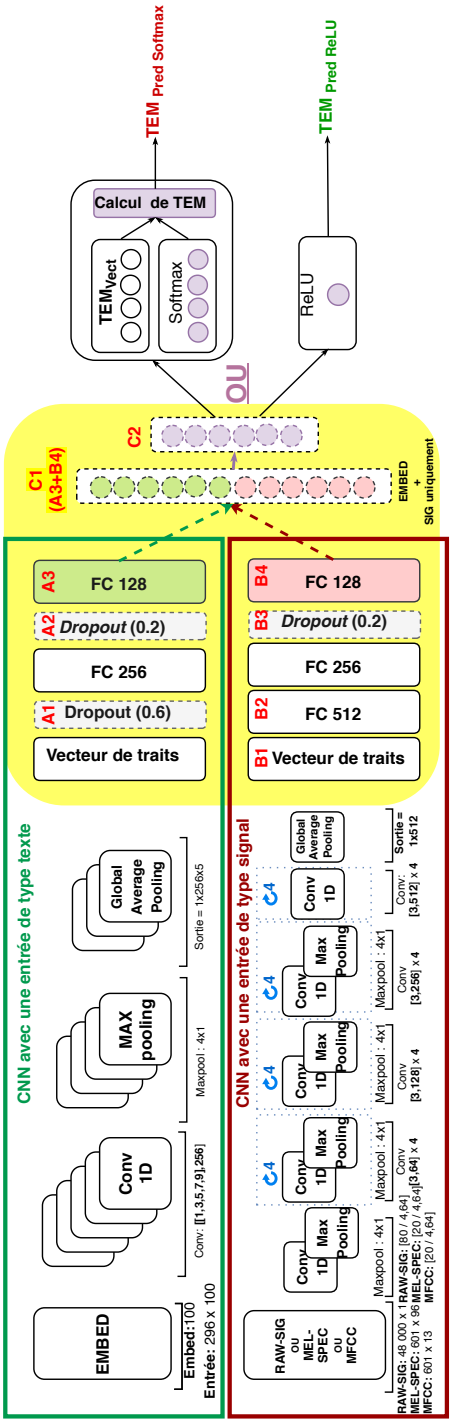


Figure 2. *Architecture de nos CNN à partir d'entrées texte (vert) et signal (rouge). Les couches avec des lignes en pointillé correspondent à l'utilisation conjointe texte + signal*

4.3. *Expériences et résultats*

Dans cette section, nous comparons les deux approches de prédiction de performance des SRAP : prédiction fondée sur des caractéristiques explicites et prédiction fondée sur des caractéristiques entraînées en utilisant les CNN. La prédiction par l'outil TranscRater s'appuie sur les traits issus de la sortie du SRAP et du signal (POS, LEX, LM et SIG), tandis que le CNN est fondé sur la sortie du SRAP et le signal brut. Pour le CNN, nous sélectionnons aléatoirement 10 % des données Train$_{Pred}$ comme corpus de développement DEV. Le reste est considéré comme un corpus d'apprentissage du réseau (TRAIN). Dix modèles de prédiction sont entraînés selon 10 sélections aléatoires de la partition TRAIN et DEV.

L'entraînement est effectué à l'aide de l'algorithme Adadelta (Zeiler, 2012) sur des mini-batchs de taille 32 avec 50 époques d'apprentissage.[5] La métrique MAE est utilisée à la fois comme fonction de perte (coût) et comme mesure d'évaluation. Après la phase d'apprentissage, nous prenons le meilleur modèle (parmi les 10 sélections aléatoires de la partition TRAIN et DEV) obtenu en termes de MAE sur le corpus DEV et nous l'évaluons sur le corpus Test$_{Pred}$.

Nous étudions plusieurs entrées pour le CNN :

1) entrées textuelles (transcription automatique) uniquement (**EMBED**) : l'entrée du réseau est une matrice de dimension 296 $\times$ 100 (296 est la longueur de l'hypothèse SRAP la plus longue dans notre corpus ; 100 est la dimension des *embeddings* de mots pré-entraînés sur le grand corpus de texte de 3,3 milliards de mots). Nous utilisons des tailles différentes de fenêtres de filtre *h* de [1, 3, 5, 7, 9] avec des filtres de taille 256 ;

2) signal brut uniquement (**RAW - SIG**) : les modèles sont entraînés sur des tours de parole de 6 secondes et échantillonnés à 8 kHz seulement (pour éviter les problèmes de surcharge de mémoire au cours de l'apprentissage du CNN). Les tours de parole courts ($< 6s$) sont complétés par des zéros (silence). Notre entrée est un vecteur de dimension 48 000 $\times$ 1. Les paramètres des filtres sont détaillés dans la figure 2 ;

3) spectrogramme seulement (**MEL-SPEC**) : nous utilisons la même configuration que pour le signal brut ; nous avons des vecteurs en entrée de dimension 96 (chaque dimension correspond à une plage de fréquence particulière) extraits toutes les 10 ms (la fenêtre d'analyse est de 25 ms). Notre entrée a donc une dimension 601 $\times$ 96,[6] ;

4) paramètres **MFCC** seulement : nous calculons 13 MFCC[7] toutes les 10 ms pour fournir au réseau CNN une entrée de dimension 601 $\times$ 13 ;

5) entrées conjointes (texte et signal) (EMBED + RAW - SIG ou EMBED + MEL - SPEC ou EMBED + MFCC) : dans ce cas, nous concaténons les dernières couches

5. L'algorithme d'optimisation et le nombre d'époques sont définis suivant les performances obtenues sur le corpus de DEV.

6. Les paramètres détaillés des filtres sont représentés dans la figure 2

7. Par souci de comparaison avec Transrater, les dérivées premières et secondes ne sont pas prises en compte.

cachées des réseaux CNN texte et signal (lignes en pointillé dans la figure 2).

4.3.1. *Résultats*

Les lignes TranscRater du tableau 3 présentent les résultats obtenus avec le système de base fondé sur la régression. Nous pouvons observer que la meilleure performance est obtenue avec les caractéristiques textuelles POS + LEX + LM (MAE de 22,01 %) alors que l'ajout du SIG n'améliore pas le modèle (MAE de 21,99 %). Cette impossibilité à intégrer correctement les caractéristiques issues du signal, dans les modèles de TranscRater , a également été observée par Jalalvand *et al.* (2016).

Modèle	Input	MAE	Kendall
Caractéristiques textuelles (TXT)			
TranscRater	POS + LEX + LM	22,01	**44,16**
CNN$_{Softmax}$	EMBED	**21,48**	38,91
CNN$_{ReLU}$	EMBED	22,30	38,13
Caractéristiques acoustiques (SIG)			
TranscRater	SIG	25,86	23,36
CNN$_{Softmax}$	RAW - SIG	25,97	23,61
CNN$_{ReLU}$	RAW - SIG	26,90	21,26
CNN$_{Softmax}$	MEL - SPEC	29,11	19,76
CNN$_{ReLU}$	MEL - SPEC	26,07	24,29
CNN$_{Softmax}$	MFCC	**25,52**	**26,63**
CNN$_{ReLU}$	MFCC	26,17	25,41
Caractéristiques textuelles et acoustiques (TXT + SIG)			
TranscRater	POS + LEX + LM + SIG	21,99	45,82
CNN$_{Softmax}$	EMBED + RAW - SIG	**19,24**	**46,83**
CNN$_{ReLU}$	EMBED + RAW - SIG	20,56	45,01
CNN$_{Softmax}$	EMBED + MEL-SPEC	20,93	40,96
CNN$_{ReLU}$	EMBED + MEL-SPEC	20,93	44,38
CNN$_{Softmax}$	EMBED + MFCC	19,97	44,71
CNN$_{ReLU}$	EMBED + MFCC	20,32	45,52

Tableau 3. *TranscRater vs CNN*$_{Softmax}$ *vs CNN*$_{ReLU}$ *évalués au niveau de la phrase avec une métrique MAE ou Kendall sur le corpus Test*$_{pred}$

En utilisant des caractéristiques textuelles uniquement, nous constatons que CNN$_{Softmax}$ et CNN$_{ReLU}$ ont des performances équivalentes (meilleurs en termes de MAE mais moins performantes en termes de Kendall) par rapport au modèle de TranscRater . CNN$_{Softmax}$ montre une meilleure performance que CNN$_{ReLU}$ en termes de MAE et de coefficient de corrélation.

Toutefois, il faut noter aussi que la tâche de prédiction de performance est difficile en s'appuyant essentiellement sur les caractéristiques acoustiques (MAE supé-

rieur à 25 %). Cependant, parmi les différentes entrées du signal testées, de simples MFCC conduisent à une meilleure performance en termes de MAE et Kendall. Bien que l'utilisation conjointe de caractéristiques textuelles et acoustiques n'ait pas donné des bons résultats pour la prédiction par TranscRater , elle mène à de meilleures performances en utilisant les CNN. La meilleure performance est obtenue avec le système CNN$_{Softmax}$ (EMBED + RAW - SIG) [8] qui dépasse l'approche de régression (le MAE est réduit de 21,99 % à 19,24 %, et la corrélation entre les vrais TEM et les TEM prédits est améliorée de 45,82 % à 46,83 % en termes de Kendall). Un *test de Wilcoxon* [9] permet de confirmer que la différence entre les TEM prédits par ces deux systèmes est significative ($p < 0,001$).

4.3.2. *Analyse des taux d'erreurs de mots (TEM) prédits*

Le tableau 4 présente les TEM prédits sur le corpus TEST en utilisant les deux approches de prédiction (TranscRater et CNN) pour les différents styles de parole (spontanée et non spontanée). Les performances montrent que notre approche (à − 3,83 % du TEM références) est meilleure que l'approche de régression (− 5,38 %) sur l'ensemble du corpus. Les performances montrent que le système CNN a bien prédit le TEM sur la parole non spontanée et spontanée. Le TEM$_{Pred}$ est à − 2,54 % sur la parole non spontanée et à − 4,84 % sur la parole spontanée. En revanche, la méthode de régression n'arrive pas à bien prédire la performance sur la parole spontanée (− 10,11 %).

	NS	S	NS + S
TEM$_{REF}$	21,47	38,83	31,20
TEM$_{Pred}$ TranscRater	**22,08**	28,72	25,82
TEM$_{Pred}$ CNN$_{Softmax}$	18,93	**33,99**	**27,37**
#Tours Parole	3,1 k	3,7 k	6,8 k
#Mots$_{REF}$	49,8 k	63,3 k	113,1 k

Tableau 4. *TranscRater (POS + LEX + LM + SIG) vs CNN$_{Softmax}$ (EMBED + RAW - SIG) des TEM prédits (moyennés sur toutes les phrases) par type de parole (NS ou S) sur le corpus Test$_{pred}$*

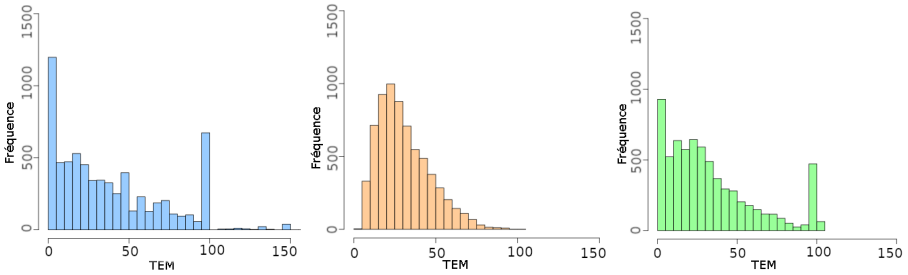
La figure 3 présente l'analyse de prédiction de TEM au niveau des tours de parole.[10] Elle montre la distribution des tours de parole en fonction de leur TEM réel ou prédit. Il est clair que la prédiction CNN permet d'approximer la vraie distribution de TEM sur le corpus Test$_{pred}$. La distribution produite par TranscRater ressemble à une

8. Les MAE obtenus sur les 10 modèles atteignent entre 15,24 % et 15,96 % sur les corpus de développement, tandis que les MAE obtenus sur le corpus Test$_{Pred}$ sont entre 19,24 % et 20,70 %

9. `http://www.r-tutor.com/elementary-statistics/non-parametric-methods/wilcoxon-signed-rank-test`

10. Les sorties des modèles sont disponibles sur `http://www.lne.fr/LNE-LIG-WER-Prediction-Corpus`

distribution gaussienne autour de la moyenne TEM observée sur les données d'apprentissage. Il est également intéressant de relever que les deux pics de TEM = 0 % et TEM = 100 % sont prédits correctement par notre système CNN.



(a) Référence (b) TranscRater (c) CNN

Figure 3. *Distribution des tours de parole en fonction de leurs TEM : (a) référence (b) prédit par le meilleur système TranscRater, (c) prédit par le meilleur système CNN*

Dans la figure 4, nous comparons les meilleurs systèmes de prédiction CNN et TR en termes de MAE sur le corpus Test_{pred} au niveau du type d'émission afin de comprendre l'effet du style de parole sur la tâche de prédiction de performance. Comme décrit, nous avons classé les émissions de nos corpus en deux groupes : parole non spontanée (NS) et parole spontanée (S). Les performances obtenues confirment que la performance sur la parole spontanée est plus difficile à prédire que sur la parole non spontanée. Dans la partie spontanée, nous remarquons que l'écart entre la courbe CNN et la courbe TR est plus large que pour la parole non spontanée. Cela signifie que le système CNN est capable de prédire un TEM élevé, alors que le système TR prédit une performance autour du TEM moyen observé sur les données d'entraînement Train_{pred}.

5. Évaluation des représentations apprises

5.1. *Méthodologie*

Dans cette section, nous essayons de comprendre ce que notre meilleur système de prédiction de performance (EMBED + RAW - SIG) a appris. Nous analysons les représentations textuelles et acoustiques obtenues par notre architecture. Nous nous inspirons de travaux de Belinkov et Glass (2017) : le modèle pré-entraîné (EMBED + RAW - SIG) est utilisé pour générer des représentations au niveau des tours de parole. Nous nous intéressons à l'analyse des représentations qui correspondent à différentes couches supérieures de notre réseau (colorées en jaune dans la figure 2). Ces représentations sont utilisées par la suite pour entraîner un classifieur peu profond et résoudre des tâches de classification annexes telles que :

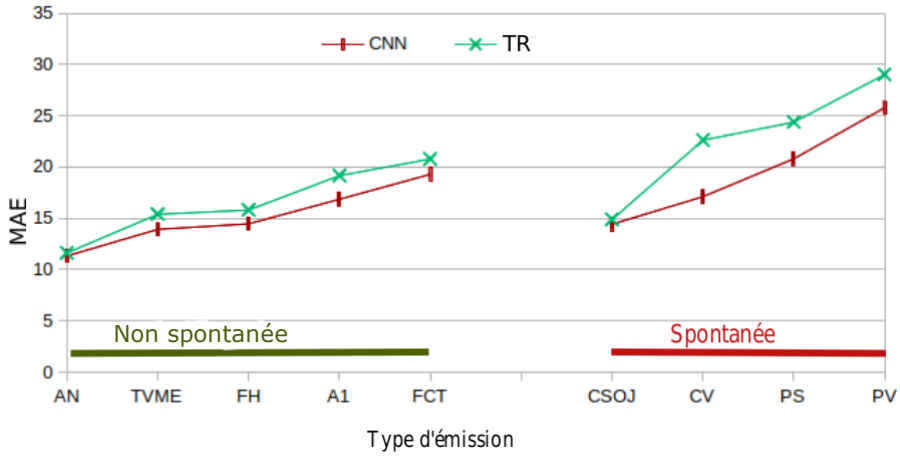


Figure 4. *Évaluation des systèmes de prédiction sur le corpus $Test_{pred}$ en termes de MAE au niveau du type d'émission*

– **STYLE :** classer les tours de parole entre les styles de parole (S et NS) (voir le tableau 5) ;

– **ACCENT :** classer les tours de parole entre locuteurs natif et non natif (comme il est indiqué dans le tableau 5), nous avons utilisé les annotations des locuteurs fournies avec nos données afin d'étiqueter nos tours de parole entre natifs et non natifs ;

– **ÉMISSION :** classer les tours de parole suivant les émissions. Comme cela est décrit dans le tableau 6, chaque tour de parole de notre corpus est étiqueté avec le nom de l'émission.

Les performances de ces classifieurs peu profonds nous permettront de savoir quelles informations (style, accent, émission) sont le mieux capturées par quelles couches du réseau ; c'est-à-dire ce que modélise un réseau CNN qui prédit les performances d'un SRAP.

Comme analyse visuelle, nous projetons également un exemple des représentations dans un espace à deux dimensions à l'aide de l'algorithme *t-sne* (t-distributed stochastic neighbor embedding).[11]

5.2. *Classifieur peu profond pour l'analyse*

Nous avons construit trois classifieurs peu profonds (ÉMISSION, STYLE, ACCENT) avec une architecture similaire. Le classifieur est un réseau neuronal supervisé

11. `https://lvdmaaten.github.io/tsne/code/tsne_python.zip`

avec une seule couche cachée (la taille de la couche cachée est fixée à 128) suivie d'un *dropout* (taux de 0,5) et d'une non-linéarité *ReLU*. Enfin, une couche Softmax est utilisée afin de convertir la sortie du réseau en une catégorie prédite. Nous avons choisi un classifieur simple et peu profond car nous nous intéressons à l'évaluation de la qualité des représentations apprises par notre modèle de prédiction SRAP, plutôt qu'à l'optimisation des tâches de classification secondaires. La taille de l'entrée du réseau dépend de la couche à analyser (voir figure 2).

L'apprentissage est effectué en utilisant l'algorithme Adam (Kingma et Ba, 2014) (en utilisant les paramètres par défaut) sur des mini-batchs de taille 16. La fonction de coût est l'entropie croisée. Les modèles sont entraînés avec 30 époques. Après l'apprentissage, nous conservons le modèle ayant les meilleures performances sur l'ensemble DEV et nous l'évaluons sur le corpus TEST (voir section suivante pour les détails sur DEV et TEST). Les sorties du classifieur sont évaluées en termes de taux de bonne classification (*accuracy*).

5.3. *Données*

Nous avons utilisé les mêmes données que celles proposées dans la section 3.1. Nous récupérons tout d'abord les corpus d'apprentissage (TRAIN) et de développement (DEV) du meilleur modèle obtenu (EMBED + RAW - SIG), tout en gardant le même corpus de TEST (Test$_{Pred}$), sachant que les émissions du corpus Test$_{Pred}$ n'existent ni dans le corpus TRAIN ni dans le corpus DEV.

Catégorie	**TRAIN**	**DEV**	**TEST**
Non spontanée	54 250	6 101	**3 109**
Spontanée	**13 277**	**1 403**	3 728
Native	44 487	4 945	5 298
Non native	**23 040**	**2 559**	**1 539**

Tableau 5. *Distribution de nos tours de parole entre des styles non spontanés et spontanés, accents natifs et non natifs*

Émission	**TRAIN**	**DEV**	**TEST**
FINTER-DEBATE	7 632	833	-
FRANCE3-DEBATE	928	77	-
LCP-PileEtFace	**4 487**	525	-
RFI	25 565	2 831	-
RTM	24 198	2 745	-
TELSONNE	4 717	**493**	-
Total	67 527	7 504	-

Tableau 6. *Nombre des tours de parole pour chaque émission*

Les tableaux 5 et 6 décrivent l'ensemble des données disponibles en termes de tours de parole pour chaque tâche de classification. Nous constatons clairement que les données sont déséquilibrées pour les trois catégories (STYLE, ACCENT, ÉMISSION). Étant donné que nous nous intéressons à évaluer le pouvoir discriminant de nos représentations apprises pour ces trois tâches, nous avons extrait une version équilibrée de nos données TRAIN, DEV, TEST en filtrant les étiquettes surreprésentées (le nombre final de tours de parole conservés correspond aux nombres en gras dans les tableaux 5 et 6). Le corpus TEST ne contient aucun type d'émission présent dans le tableau 6, car selon notre protocole expérimental, les émissions du corpus TEST (voir tableau 2) n'existent pas dans les corpus TRAIN et DEV et *vice versa*.

5.4. *Résultats*

Pour chaque tâche de classification, nous avons construit un classifieur peu profond en utilisant les représentations cachées des caractéristiques EMBED (texte), RAW - SIG (signal) et EMBED + RAW - SIG en entrée. Le tableau 7 présente les résultats expérimentaux obtenus sur les corpus DEV et TEST séparés par deux barres verticales (||). Les performances des systèmes de classification sont toutes supérieures à un taux de bonne classification correspondant à une décision aléatoire (> 50 % pour les tâches STYLE et ACCENT et > 20 % pour la tâche ÉMISSION). Cela montre que l'apprentissage d'un système de prédiction de TEM profond produit des représentations (au niveau des couches) qui contiennent une quantité significative d'informations sur le style de parole, l'accent du locuteur ainsi que sur le type d'émission. La prédiction du style des tours de parole (spontanée et non spontanée) est légèrement plus facile que la prédiction de l'accent (natifs et non-natifs), en particulier à partir de l'entrée de type texte (EMBED). Cela pourrait être lié à la durée courte (< 6 s) des tours de parole, étant donné que l'identification de l'accent a probablement besoin de séquences plus longues. Nous observons également que l'utilisation du texte et de la parole améliore les représentations apprises pour la tâche STYLE alors que cela est moins clair pour la tâche ACCENT (étant donné que l'amélioration observée sur DEV n'est pas confirmée sur TEST).

Enfin, l'entrée textuelle est significativement meilleure que l'entrée acoustique pour toutes les tâches de classification, alors que nous anticipions de meilleures performances sur l'entrée acoustique pour la tâche ÉMISSION (le signal transmet des informations sur les caractéristiques acoustiques d'un programme diffusé). Parmi les représentations analysées, les sorties des CNN (A1, B1) conduisent aux meilleurs résultats de classification, ceci est cohérent avec les résultats de la littérature qui présentent les convolutions comme de bons extracteurs de traits. En utilisant les couches supérieures (entièrement connectées), nous remarquons que la performance se dégrade. Cela signifie que l'information sur le style de parole, l'accent du locuteur ou l'émission et plutôt capturée dans les couches moins hautes de notre architecture neuronale de prédiction de performance de SRAP.

Couche	Dim.	ÉMISSION	STYLE	ACCENT
EMBED				
A1	1 280	**57,12**||-	**80,72**||68,99	**70,75**||66,54
A2	256	54,89||-	80,01||**69,56**	69,30||69,43
A3	128	51,04||-	79,23||68,27	68,25||**70,89**
RAW - SIG				
B1	512	**42,35**||-	**72,92**||**58,64**	64,60||**55,85**
B2	512	41,22||-	72,20||58,41	64,44||54,84
B3	256	41,22||-	72,38||58,44	64,50||54,65
B4	128	40,77||-	72,38||58,52	**64,74**||54,87
EMBED + RAW - SIG				
C1 (A3+B4)	256	**57,04**||-	**81,29**||70,36	**71,41**||**65,98**
C2	128	53,06||-	79,62||**70,55**	70,01||65,20
Aléatoire	-	**20,00**	**50,00**	**50,00**

Tableau 7. *Performances des systèmes de classification émission (sur le corpus DEV uniquement), style et accent en termes de taux de bonne classification en utilisant les représentions apprises durant l'apprentissage de notre système de prédiction*

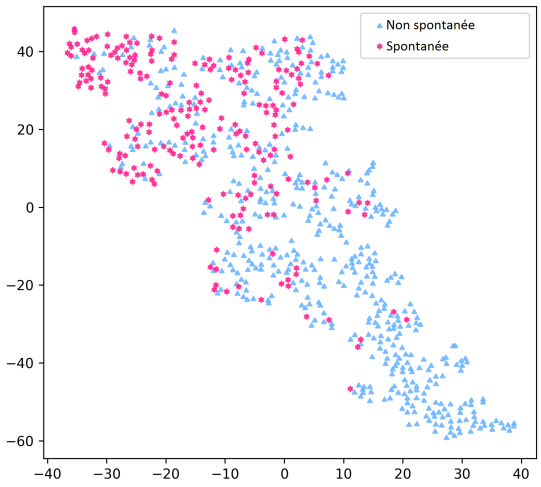
Dans la figure 5, nous visualisons un exemple de représentations des tours de parole de la couche C2 (EMBED + RAW - SIG) en utilisant t-SNE. Pour une durée fixe[12] de 4 à 5 s (716 tours de parole) et de 5 à 6 s (489 tours de parole), les tours de parole non spontanée sont colorés en bleu tandis que les tours de parole spontanée sont en rose. La couche C2 produit des *clusters* qui montrent que les tours de parole spontanée se trouvent dans la partie supérieure gauche de l'espace 2D. Cela suggère que la représentation cachée C2 véhicule une information (signal faible) sur le style de parole.

Enfin, la figure 6 présente la matrice de confusion produite à l'aide de la couche C2 (EMBED + RAW - SIG). Les classifieurs ont très bien prédit la catégorie *TELSONNE* (taux de bonne classification de 82 %), qui contient de nombreux appels téléphoniques des auditeurs de la radio. Cette émission est assez différente des quatre autres émissions de DEV (débats et actualités).

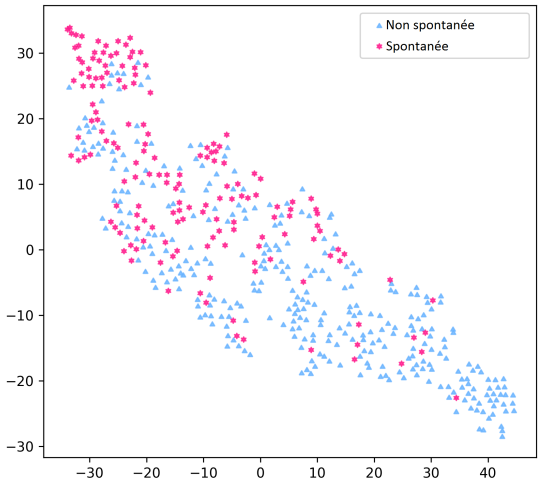
6. Apprentissage multitâche

Dans la section précédente, nous avons montré que les couches cachées de notre système de prédiction capturaient une information sur le style de la parole, l'accent et le type d'émission. Cela suggère que ces trois types d'informations pourraient être utiles pour structurer l'apprentissage des modèles neuronaux de prédiction de perfor-

12. D'après nos expériences, les représentations des tours de parole très courts (ayant une durée inférieure à 2 s) capturent plus difficilement l'information sur le style de parole.



(a) de 4 à 5 secondes



(b) de 5 à 6 secondes

Figure 5. *Visualisation des représentations des tours de parole de la couche C2 pour les différents styles de parole (spontanée, non spontanée) : (a) des tours de parole ayant une durée de 4 à 5 s et (b) de 5 à 6 s*

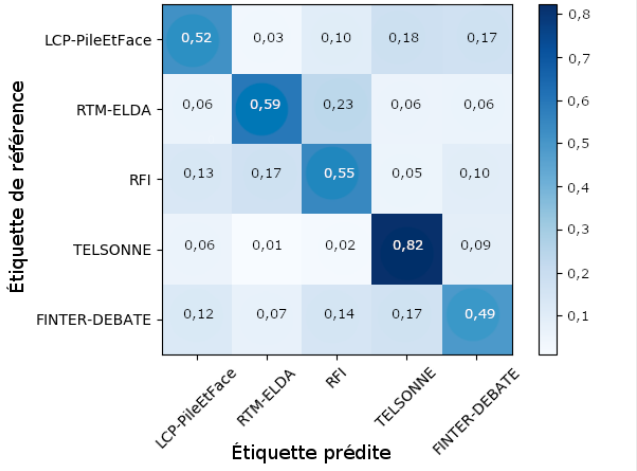


Figure 6. *Matrice de confusion de la classification ÉMISSION en utilisant les représentations de la couche C2 (EMBED + RAW - SIG) comme entrée - évaluée sur le corpus DEV*

mance. Dans cette section, nous examinons l'impact de la connaissance de ces étiquettes (style, accent, émission) au moment de l'apprentissage sur les performances des systèmes de prédiction. Pour cela, nous effectuons un apprentissage multitâche en fournissant des informations supplémentaires sur le type d'émission, le style de parole ainsi que l'accent du locuteur pendant l'apprentissage. L'architecture du modèle multitâche est similaire au modèle de prédiction de TEM (monotâche) présenté dans la figure 2 en ajoutant des sorties supplémentaires : une fonction Softmax est ajoutée pour chaque nouvelle tâche de classification après la dernière couche entièrement connectée (C2). La dimension de sortie dépend essentiellement de la tâche visée : 6 pour les tâches ÉMISSION et 2 pour les tâches STYLE et ACCENT.

Nous utilisons la totalité des données (non équilibrées) décrites dans les tableaux 5 et 6. L'entraînement du modèle multitâche utilise $Adadelta$. Les modèles sont appris pendant 50 époques avec une taille de mini-batch de 32. La métrique MAE est utilisée comme fonction de coût pour la tâche de prédiction, tandis que l'entropie croisée est utilisée pour les tâches de classification secondaire. Nous définissons aussi une fonction de coût composite dans le cas de l'apprentissage multitâche : nous attribuons une pondération de 1 pour le coût MAE (tâche principale) et une pondération plus petite de 0,3 pour le ou les coûts d'entropie croisée (tâche de classification secondaire).

Après la phase d'apprentissage, nous prenons le modèle qui donne le meilleur MAE sur le corpus DEV et nous l'évaluons sur le corpus TEST. Nous expérimentons plusieurs modèles qui traitent simultanément les 1, 2, 3 et 4 tâches. Les modèles sont

évalués avec une métrique spécifique pour chaque tâche : MAE et Kendall[13] pour la tâche de prédiction TEM et le taux de bonne classification (*accuracy*) pour les tâches de classification.

Les tableaux 8 et 9 résument les résultats expérimentaux sur les corpus DEV et TEST séparés par deux barres verticales (||). Nous avons considéré le modèle monotâche décrit dans la section 4.2 comme un système de référence.

Nous rappelons que nous avons évalué la tâche de classification ÉMISSION uniquement sur l'ensemble DEV (les émissions du corpus TEST n'existent pas dans notre TRAIN).

Tout d'abord, nous constatons que la performance des tâches de classification dans les scénarios multitâches est très bonne : nous sommes capables de former des systèmes efficaces de prédiction de performance SRAP qui annotent simultanément les tours de parole analysés en fonction de leur style de parole, de leur accent et de l'origine du programme de diffusion. De tels systèmes multitâches pourraient être utilisés comme outils de diagnostic pour analyser et prédire les TEM sur de grandes collections acoustiques.

De plus, nos meilleurs systèmes multitâches montrent une meilleure performance (MAE, Kendall) par rapport au système de base. Cela signifie que le fait de donner implicitement les informations sur le style, l'accent et le type d'émission peut être utile pour structurer l'apprentissage du système de prédiction.

Par exemple, pour les systèmes à deux tâches, le meilleur modèle est obtenu sur les tâches TEM + ÉMISSION avec une différence respective de + 0,41 % et + 2,25 % en termes de MAE et Kendall (sur le corpus DEV) par rapport au système de base sur la tâche de prédiction TEM.

Il faut cependant noter que l'impact de l'apprentissage multitâche sur la tâche principale (prédiction de la performance) est limité : de légères améliorations sur le corpus TEST sont observées en termes de MAE et Kendall. Néanmoins, les systèmes appris semblent complémentaires étant donné que leur combinaison (moyennage, sur l'ensemble des systèmes multitâches, du TEM prédit au niveau des tours de parole) conduit à une amélioration significative des performances (voir dernière ligne du tableau 8 pour MAE et Kendall).

13. Corrélation entre les vraies valeurs TEM (références) et les valeurs TEM prédites.

Tâche de prédiction de performance		
Modèles	**MAE**	**Kendall**
Baseline : monotâche		
TEM	15,24||19,24	45,00||46,83
2 tâches		
TEM ÉMISSION	**14,83**||19,15	**47,25**||47,05
TEM STYLE	15,07||19,66	45,92||45,49
TEM ACCENT	15,05||19,60	46,17||45,60
3 tâches		
TEM STYLE ACCENT	15,12||20,23	45,75||44,09
TEM ÉMISSION ACCENT	14,94||19,76	46,19||43,61
TEM ÉMISSION STYLE	14,90||**19,14**	45,87||**47,28**
4 tâches		
TEM ÉMISSION STYLE ACCENT	15,15||19,64	45,59||45,42
COMBINAISON TOUTES SORTIES	**14,50** ||**18,87**	**48,16**||**48,63**

Tableau 8. *Évaluation de la prédiction de performance du SRAP avec des modèles multitâches ($DEV||TEST$) en termes de MAE et Kendall*

Tâche de classification			
Modèles	**ÉMISSION**	**STYLE**	**ACCENT**
2 tâches			
TEM ÉMISSION	**99,29**||-	-	-
TEM STYLE	-	99,01||65,24	-
TEM ACCENT -	-	91,72|| 75,30	
3 tâches			
TEM STYLE ACCENT	-	98,63||69,07	88,99|| **77,46**
TEM ÉMISSION ACCENT	98,38||-	-	89,87||71,44
TEM ÉMISSION STYLE	99,12||-	**99,47**||**81,98**	-
4 tâches			
TEM ÉMISSION STYLE ACCENT	99,04||-	99,29||81,55	**91,92**||73,60

Tableau 9. *Évaluation des tâches de classification secondaires des modèles multitâches ($DEV||TEST$) en termes de taux de bonne classification*

7. Conclusion et perspectives

Dans ce travail, nous avons abordé la tâche de prédiction de performance des systèmes de transcription automatique de la parole. Dans un premier temps, nous avons proposé un corpus hétérogène en français spécifique pour cette tâche. Nous avons proposé par la suite de comparer deux différentes approches de prédiction de perfor-

mance : une approche fondée sur des traits prédéfinis (*engineered features*) en utilisant l'outil TranscRater et notre nouvelle approche fondée sur des traits estimés au cours de l'apprentissage d'un système neuronal de type CNN (*learnt features*). Nos expérimentations montrent que l'approche de prédiction par les CNN est meilleure que l'approche de prédiction de base (par TranscRater) en termes de scores MAE et Kendall. Plus précisément, l'utilisation conjointe en entrée des textes et signaux ne donne pas de résultats positifs pour les systèmes TranscRater , tandis qu'elle permet de meilleures performances en utilisant des CNN. Nous montrons également que les CNN prédisent correctement la distribution des taux d'erreurs de mots (TEM) sur une collection d'enregistrements, contrairement à TranscRater qui prédit une distribution proche d'une distribution gaussienne autour du TEM moyen observé dans le corpus d'apprentissage.

Dans un second temps, nous avons essayé de comprendre ce qu'apprend le système CNN en analysant les représentations intermédiaires produites par notre meilleur système de prédiction (CNN$_{Softmax}$ EMBED + RAW - SIG). Afin de comprendre quelles sont les informations capturées par le modèle au cours de l'entraînement, nous avons suivi une méthode d'analyse inspirée d'un article récent de Belinkov et Glass (2017) publié l'an dernier à la conférence NIPS. L'idée est d'utiliser les représentations apprises pour des tâches de classification annexes (ou de les visualiser). Nos expérimentations montrent que notre modèle capture des informations sur le style de parole, l'accent du locuteur et le type d'émission durant l'apprentissage du système. Enfin, nous avons étudié le potentiel d'un apprentissage structuré consistant à donner implicitement ces trois informations au moment de l'entraînement du système de prédiction via un apprentissage multitâche. Les performances obtenues montrent que la création d'un système multitâche améliore légèrement la prédiction de TEM tout en générant une prédiction correcte d'informations additionnelles telles que le style de parole, l'accent du locuteur et le type d'émission qui peuvent être des informations complémentaires utiles.

À partir de nos expérimentations, plusieurs perspectives de recherche peuvent être envisagées. Tout d'abord, nous souhaitons améliorer notre approche proposée (CNN) en exploitant des nouveaux types de traits à l'entrée de notre réseau tels que des traits de type : POS, LEX et LM. De plus, nous souhaitons expérimenter des architectures de type réseaux siamois pour apprendre des représentations prenant en compte explicitement des informations sur le style de parole, l'accent du locuteur, et le type d'émission. Ces représentations intermédiaires pourraient être intégrées dans notre système de prédiction pendant la phase d'entraînement afin d'améliorer la qualité du système. Enfin, nous souhaitons aussi étudier l'effet de la quantité des données d'apprentissage Train$_{Pred}$ sur la qualité des systèmes de prédiction, et étudier la robustesse des deux méthodes de prédiction proposées (TR et CNN) lorsqu'elles sont entraînées et/ou évaluées sur des transcriptions automatiques issues d'un autre SRAP.

8. Bibliographie

Asadi A., Schwartz R., Makhoul J., « Automatic detection of new words in a large vocabulary continuous speech recognition system », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1990.

Belinkov Y., Glass J., « Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems », *Advances in Neural Information Processing Systems*, p. 2438-2448, 2017.

Belinkov Y., Màrquez L., Sajjad H., Durrani N., Dalvi F., Glass J., « Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks », *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, vol. 1, p. 1-10, 2017.

Chollet F. *et al.*, « Keras », , https ://github.com/fchollet/keras, 2015.

Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., Kuksa P., « Natural language processing (almost) from scratch », *Journal of Machine Learning Research*, vol. 12, n° 8, p. 2493-2537, 2011.

Dai W., Dai C., Qu S., Li J., Das S., « Very deep convolutional neural networks for raw waveforms », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 421-425, 2017.

De Calmès M., Pérennou G., « BDLEX : a lexicon for spoken and written French », *Proceedings of 1st International Conference on Langage Resources & Evaluation*, p. 1129-1136, 1998.

de Souza J. G. C., Buck C., Turchi M., Negri M., « FBK-UEdin participation to the WMT13 quality estimation shared task », *Proceedings of the Eighth Workshop on Statistical Machine Translation*, p. 352-358, 2013.

de Souza J. G., Zamani H., Negri M., Turchi M., Daniele F., « Multitask learning for adaptive quality estimation of automatically transcribed utterances », *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 714-724, 2015.

Dreschler W. A., Verschuure H., Ludvigsen C., Westermann S., « ICRA noises : artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment : Ruidos ICRA : Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos », *Audiology*, vol. 40, n° 3, p. 148-157, 2001.

Elloumi Z., Besacier L., Galibert O., Kahn J., Lecouteux B., « ASR PERFORMANCE PREDICTION ON UNSEEN BROADCAST PROGRAMS USING CONVOLUTIONAL NEURAL NETWORKS », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Eyben F., Wöllmer M., Schuller B., « Opensmile : The Munich Versatile and Fast Open-source Audio Feature Extractor », *International Conference on Multimedia*, MM '10, ACM, New York, NY, USA, p. 1459-1462, 2010.

Ferreira S., Farinas J., Pinquier J., Rabant S., « Prédiction a priori de la qualité de la transcription automatique de la parole bruitée », *Proc. XXXIIe Journées d'Études sur la Parole*, p. 249-257, 2018.

Galibert O., « Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech. », *in* F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, P. Perrier (eds), *Interspeech*, ISCA, p. 1131-1134, 2013.

Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J.-F., Gravier G., « The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. », *Interspeech*, p. 1149-1152, 2005.

Geurts P., Ernst D., Wehenkel L., « Extremely randomized trees », *Machine learning*, vol. 63, n° 1, p. 3-42, 2006.

Gravier G., Adda G., Paulson N., Carré M., Giraudel A., Galibert O., « The ETAPE corpus for the evaluation of speech-based TV content processing in the French language », *LREC-Eighth international conference on Language Resources and Evaluation*, p. na, 2012.

Hermansky H., Variani E., Peddinti V., « Mean temporal distance : Predicting ASR error from temporal properties of speech signal », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 7423-7426, 2013.

Jalalvand S., Falavigna D., Matassoni M., Svaizer P., Omologo M., « Boosted acoustic model learning and hypotheses rescoring on the CHiME-3 task », *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, IEEE, p. 409-415, 2015a.

Jalalvand S., Negri M., Daniele F., Turchi M., « Driving rover with segment-based asr quality estimation », *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, vol. 1, p. 1095-1105, 2015b.

Jalalvand S., Negri M., Falavigna D., Turchi M., « Driving ROVER with Segment-based ASR Quality Estimation », 01, 2015c.

Jalalvand S., Negri M., Turchi M., de Souza J. G., Falavigna D., Qwaider M. R., « Transcrater : a tool for automatic speech recognition quality estimation », *Proceedings of ACL-2016 System Demonstrations. Berlin, Germany : Association for Computational Linguistics*p. 43-48, 2016.

Jin M., Song Y., Mcloughlin I., Dai L.-R., Ye Z.-F., « LID-senone extraction via deep neural networks for end-to-end language identification », *Proc. of Odyssey*, 2016.

Kahn J., Galibert O., Quintard L., Carré M., Giraudel A., Joly P., « A presentation of the REPERE challenge », *Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on*, IEEE, p. 1-6, 2012.

Kim Y., « Convolutional neural networks for sentence classification », *arXiv preprint arXiv :1408.5882*, 2014.

Kingma D. P., Ba J., « Adam : A Method for Stochastic Optimization », *CoRR*, 2014.

Maaten L. v. d., Hinton G., « Visualizing data using t-SNE », *Journal of machine learning research*, vol. 9, n° 11, p. 2579-2605, 2008.

McFee B., Raffel C., Liang D., Ellis D. P., McVicar M., Battenberg E., Nieto O., « librosa : Audio and music signal analysis in python », 2015.

Meinshausen N., Bühlmann P., « Stability selection », *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, vol. 72, n° 4, p. 417-473, 2010.

Meyer B. T., Mallidi S. H., Kayser H., Hermansky H., « Predicting error rates for unknown data in automatic speech recognition », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 5330-5334, 2017.

Mikolov T., Sutskever I., Chen K., Corrado G., Dean J., « Distributed Representations of Words and Phrases and their Compositionality », *NIPS*, 2013.

Mohamed A.-r., Hinton G., Penn G., « Understanding how deep belief networks perform acoustic modelling », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 4273-4276, 2012.

Nagamine T., Seltzer M. L., Mesgarani N., « Exploring how deep neural networks form phonemic categories », *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Negri M., Turchi M., de Souza J. G., Falavigna D., « Quality Estimation for Automatic Speech Recognition. », *COLING*, p. 1813-1823, 2014.

Palaz D., Doss M. M., Collobert R., « Convolutional neural networks-based continuous speech recognition using raw speech signal », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 4295-4299, 2015.

Pellegrini T., Mouysset S., « Inferring phonemic classes from CNN activation maps using clustering techniques », *Interspeech*, p. pp-1290, 2016.

Piczak K. J., « Environmental sound classification with convolutional neural networks », *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, IEEE, p. 1-6, 2015.

Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P. *et al.*, « The Kaldi speech recognition toolkit », *IEEE 2011 workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

Sainath T. N., Weiss R. J., Senior A., Wilson K. W., Vinyals O., « Learning the speech front-end with raw waveform CLDNNs », *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Schmid H., « Treetagger| a language independent part-of-speech tagger », *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, vol. 43, p. 28, 1995.

Shi X., Padhi I., Knight K., « Does string-based neural MT learn source syntax ? », *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 1526-1534, 2016.

Stolcke A. *et al.*, « SRILM-an extensible language modeling toolkit. », *Interspeech*, vol. 2002, p. 2002, 2002.

Wang S., Qian Y., Yu K., « What Does the Speaker Embedding Encode ? », *Interspeech*, vol. 2017, p. 1497-1501, 2017.

Wu Z., King S., « Investigating gated recurrent neural networks for speech synthesis », *CoRR*, 2016.

Young S. R., « Recognition Confidence Measures : Detection of Misrecognitions and Out-Of-Vocabulary Words », *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, p. 21-24, 1994.

Zeiler M. D., « ADADELTA : An Adaptive Learning Rate Method », *CoRR*, 2012.