XXXIIe Journées d'Études sur la Parole
4-8 June 2018, Aix-en-Provence, France

Prédiction *a priori* de la qualité de la transcription automatique de la parole bruitée

Sébastien Ferreira[1, 2] Jérôme Farinas[1] Julien Pinquier[1] Stéphane Rabant[2]
(1) IRIT, Université de Toulouse, CNRS, Toulouse, France
(2) Authôt, 52 Avenue Pierre Semard, 94200, Ivry-sur-Seine, France
prenom.nom@irit.fr[1],sferreira@authot.com,srabant@authot.com

RÉSUMÉ

De nombreuses sources de variabilité dégradent les performances d'un système de Reconnaissance Automatique de la Parole (RAP). Dans cette étude, les dégradations provoquées par le type et le niveau de bruit sont explorées afin de prédire *a priori* la qualité de la RAP, i.e. avant même le décodage. Notre méthode se fonde sur une séparation spectrale de la parole et du bruit afin de produire un modèle de régression. L'expérimentation a été réalisée sur le corpus Wall Street Journal, bruité avec le corpus NOISEX-92 (17 types de bruit) que nous appliquons à 9 niveaux de rapport signal à bruit. La méthode de régression proposée obtient moins de 8% d'erreur moyenne entre le Word Error Rate (WER) prédit et le WER réellement obtenu par le système de transcription automatique de la parole.

ABSTRACT

A priori prediction of the quality of automatic speech to text conversion for noised speech.

Many sources of variability degrade the performance of Automatic Speech Recognition (ASR) system. In this study, the degradations caused by the type and level of noise are explored in order to predict the *a priori* quality of ASR, i.e. even before decoding. Our method is based on a spectral separation of speech and noise to produce a regression model. The experiment was carried out on the Wall Street Journal corpus, noised with the NOISEX-92 corpus (17 types of noise) that we apply at 9 levels of signal-to-noise ratio. The proposed regression method obtains less than 8% of mean error between the predicted Word Error Rate (WER) and the actual WER obtained by automatic speech to text conversion system.

MOTS-CLÉS : prédiction d'erreur, reconnaissance automatique de la parole, analyse du bruit, séparation de la parole et du bruit.

KEYWORDS: error prediction, automatic speech recognition, noise analysis, speech/noise discrimination.

1 Introduction

Les progrès effectués dans le domaine de la Reconnaissance Automatique de la Parole (RAP) permettent de transcrire un fichier audio en texte dans des situations de plus en plus complexe. Les performances obtenues par ces systèmes sont fortements liées aux méthodes et aux données utilisées pour l'apprentissage des modèles acoustiques et linguistiques. Actuellement, il n'existe pas de système de RAP qui serait efficace pour toutes les situations, car il existe de nombreuses sources de variabilité dans un signal de parole : l'environnement acoustique, la voix du locuteur, le manière

10.21437/JEP.2018-29

de parler, l'interaction entre les locuteurs, la thématique du discours... Pour réduire l'impact de ces différentes sources de variabilité, il est courant d'utiliser des systèmes spécifiques pour chaque cas d'utilisation : la dictée vocale, la commande vocale, les enregistrements télévisuels et radiophoniques, les réunions, l'enseignement, l'automobile...

Pour sélectionner le meilleur système de RAP, sans information préalable sur le fichier traité, il serait intéressant de pouvoir prédire *a priori* la qualité des transcriptions. Le Word Error Rate (WER) est une métrique couramment utilisée pour évaluer la qualité de la transcription. Il existe de nombreuses méthodes pour prédire le WER, mais, ces méthodes utilisent des informations qui dépendent des résultats d'un système de RAP : mesures de confiance (CM) (Jiang, 2005; Ghannay *et al.*, 2015), probabilités *a posteriori*, paramètres lexicaux et syntaxiques, posteriogram sur les phonèmes (Meyer *et al.*, 2017) ou diverses statistiques calculées sur les données d'entraînement (Hermansky *et al.*, 2013).

Dans cet étude, nous cherchons une méthode qui n'exige pas de score interne du système de RAP et qui puisse être calculée avant d'utiliser le système lui-même. En effet, la prédiction du WER obtenue par un système de RAP sur un fichier audio devra être déterminée *a priori*. Il existe de nombreuses sources d'erreur de reconnaissance possibles. Ces erreurs sont causées en particulier par l'environnement sonore, les dialectes sous-représentés, les abréviations, les noms propres, les voix atypiques, une mauvaise maîtrise du langage, une thématique trop spécifique... Afin de ne pas mélanger toutes les sources d'erreur possibles, cette première étude se concentre sur les dégradations de la parole uniquement causées par le bruit. Pour représenter un grand nombre de bruits ambiants, nous avons artificiellement bruité les données pour différents types de bruits et à différents niveaux de RSB (Rapport Signal sur Bruit), d'une manière similaire à l'expérience de Haitian Xu (Xu *et al.*, 2007) pour évaluer la prédiction *a priori* obtenue par notre méthode.

Tout d'abord, le système de prédiction du WER est présenté dans la section 2. Ensuite, le cadre expérimental est décrit dans la section 3. Puis, les résultats obtenus sont présentés dans la section 4.

2 Système d'estimation du WER

Les modèles acoustiques utilisés par les systèmes de RAP modélisent généralement des paramètres plus ou moins complexes liés à l'énergie à court terme du signal : par exemple des paramètres issus d'un spectrogramme. Or, comme le bruit provoque une perturbation de l'énergie à court terme, les paramètres utilisés par les modèles acoustiques, même s'il sont plus robustes, se retrouvent dégradés. De plus, lorsque le bruit devient important, i.e. losrque le RSB est très faible (voir négatif), le bruit recouvre la parole. Il est devient alors très difficile de distinguer l'énergie provenant de la parole de celle provenant du bruit. Une analyse de ce recouvrement a été faite dans de nombreux domaines comme l'estimation du RSB (nis, 1994), la détection d'activité vocale (Voice Activity Detection - VAD) (Yiming & Rui, 2015) et l'amélioration de la parole (Speech Enhancement - SE) (Ruwei *et al.*, 2016). Afin de quantifier l'impact du bruit sur un signal, il est courant d'estimer le RSB. Cependant, le RSB ne permet pas, à lui seul, de quantifier l'impact sur la qualité de la transcription des systèmes de RAP. Le type de bruit, la localisation du signal de parole en temps et en fréquence et la robustesse au bruit du système de RAP utilisé sont aussi à prendre en compte (figure 1). Nous proposons d'étudier le comportement de l'énergie à court terme du signal tout en tenant compte des différents facteurs cités précédement.

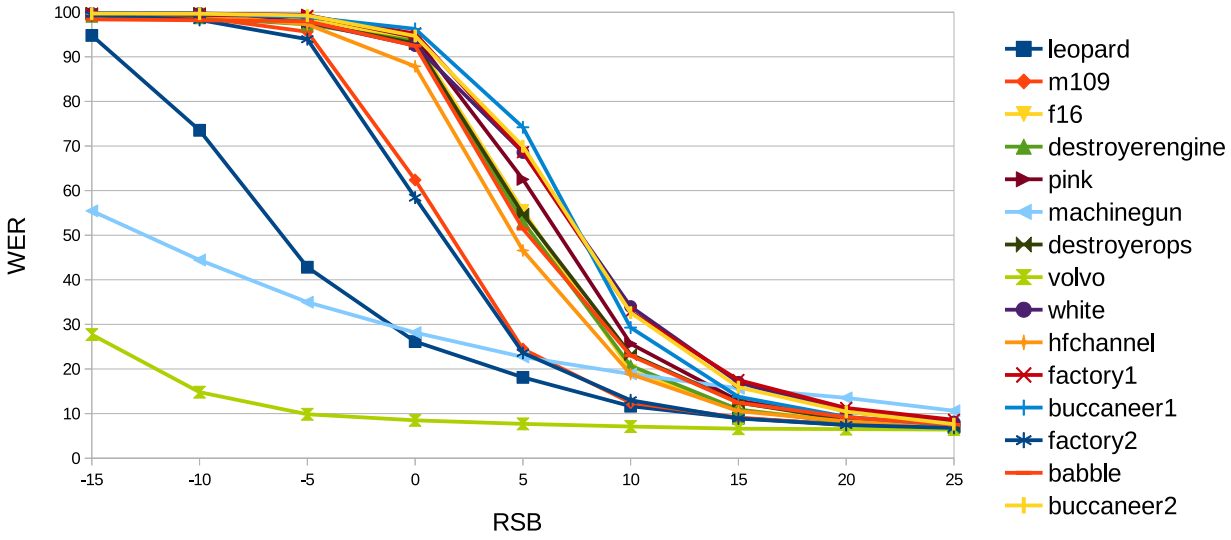

FIGURE 1 – Erreurs obtenues par un système entraîné dans des conditions propres en fonction du type de bruit et du RSB.

2.1 Architecture globale

Le système est composé de 4 étapes (voir figure 2) :

1. Calcul de la Transformée de Fourier Discrète (TFD) avec un fenêtrage de 512 points et un recouvrement de moitié puis normalisée sous l'échelle [0 ;1].
2. Détermination d'un masque binaire pour séparer la parole du bruit (voir section 2.2).
3. Extraction de paramètres pour chaque bande (voir section 2.3).
4. Régression entre les paramètres et les WER issus de l'apprentissage (voir section 2.4).

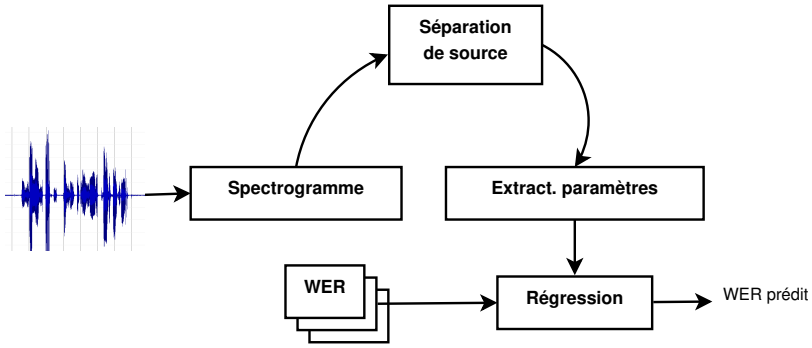

FIGURE 2 – Architecture du système de prédiction du WER.

2.2 Séparation de la parole et du bruit

Le bruit, en recouvrant le signal de parole, provoque trois types d'erreurs de reconnaissance :

— suppression : un phonème n'est pas reconnu, alors qu'il existe.

— insertion : un phonème est reconnu, alors qu'il n'est pas présent.
— substitution : un phonème erroné est reconnu.

Afin d'étudier la superposition de la parole et du bruit, une séparation est effectuée. Parmi les multiples méthodes de séparation de sources, il existe la méthode des masques. Le calcul d'un masque (binaire ou pondéré) permet de séléctionner les energies à court terme provenant de la parole. Dans ce domaine, le masque binaire idéal (Ideal Binary Mask - IBM) correspond au masque qui permet de séléctionner uniquement la parole. Il est généralement calculé pour un signal enregistré dans des conditions non bruitées. Le calcul automatique d'un masque binaire (Binary Mask - BM), le plus proche possible de l'IBM, est un enjeu majeur pour le domaine de l'analyse de scènes acoustiques (Wang, 2005). Actuellement, la détermination du BM qui minimise l'écart avec l'IBM se fonde sur une mesure de RSB local.

Dans notre méthode nous calculons un BM pour chaque bande Bark (Zwicker, 1961) du spectrogramme afin d'étudier le comportement de l'énergie à court terme de la parole et du bruit séparément :

$$BM(f,t) = \begin{cases} 1, & \text{si } E(f,t) \geq \omega * \overline{E(f)} \\ 0, & \text{sinon} \end{cases}$$

avec :

$$\overline{E(f)} = \frac{1}{t_{max}} * \sum_{t=1}^{t_{max}} E(f,t)$$

et f la fréquence, t la trame, t_{max} le nombre total de trames, E l'énergie à court terme et ω la pondération.

Il est important de noter que le seuil $T = \omega * \overline{E(f)}$ peut varier grandement en fonction de la bande Bark. Pour déterminer le BM optimal, on cherche le premier $\omega > \frac{max(\Delta densities)}{C}$ avec C une constante et $\omega_{optimal} > \omega_{maximum}$ (voir figure 3). Suite à la détermination du BM (deuxième image de la figure 4), nous éliminons les artefacts causés par les bruits résiduels : nous appliquons un masque local sur le BM afin de sélectionner que les zones ayant une densité supérieure à θ. Le résultat de cette amélioration est visible dans la troisième partie de la figure 4. Pour plus de détails, vous pouvez écouter quelques résultats ici [1].

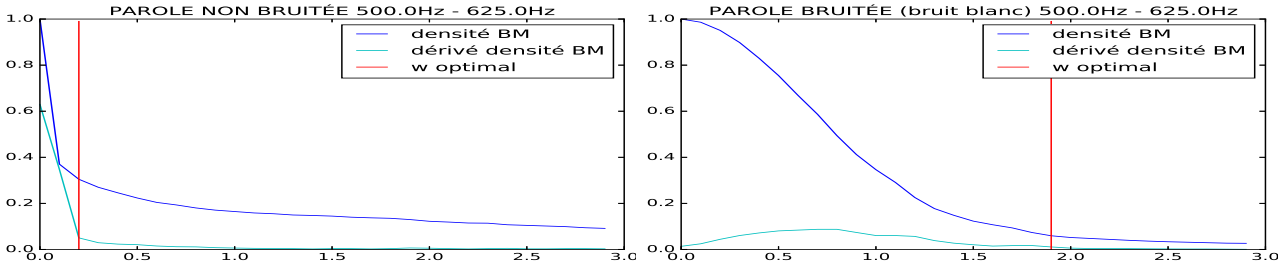

FIGURE 3 – Évolution de la densité du BM en fonction de ω. À gauche le fichier non bruité, à droite le même fichier bruités avec du bruit blanc.

1. https://goo.gl/MAiXb1

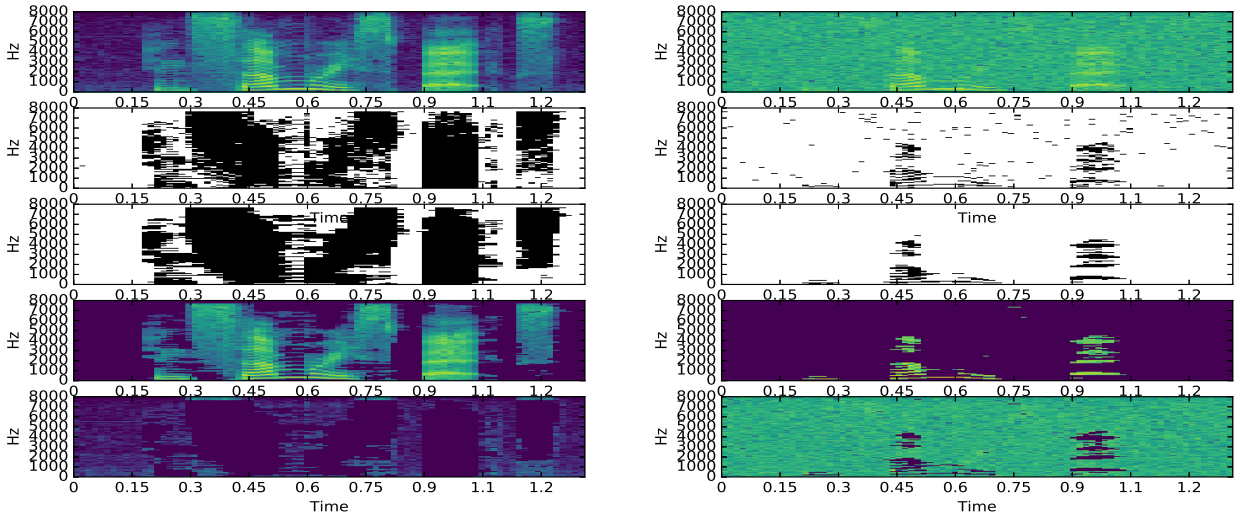

FIGURE 4 – Séparation de la parole et du bruit. À droite le fichier propre et à gauche le même fichier bruité avec du bruit blanc. De haut en bas : spectrogramme, BM, BM filtré, parole isolé, bruit isolé.

2.3 Extraction de paramètres

L'objectif est d'analyser le comportement de l'énergie à court terme du bruit et de la parole. Pour chaque bande Bark de 0 à 4400 Hertz le 5ème percentile, le 95e percentile et les 4 premiers moments (moyenne, variance, skewness et kurtosis) sont extraits. Une dernière mesure, similaire au Mean Crossing Rate (MCR) (Eyben, 2015), permet de quantifier le nombre de variations significatives de l'énergie à court terme. Ce paramètre est en quelque sorte un Zero Crossing Rate (ZCR) calculé sur la dérivée de l'énergie. Si les variations sont trop faibles (en dessous d'un seuil fixé à 0,2 dans notre cas) alors elles ne sont pas compatibilisées.

2.4 Régression

Un modèle de régression est calculé entre les paramètres extraits précédemment et le WER obtenu par différents systèmes de RAP. Cette regression permet de prendre en compte la variabilité des systèmes de RAP en fonction du volume et du type de bruit (Xu *et al.*, 2007) :

— les sytèmes multi-conditions sont plus performants sur des données bruitées,
— les sytèmes appris sur des données propres sont plus efficaces sur les données propres,
— les systèmes appris sur un type de bruit sont plus efficaces sur ce type de bruit.

Ces différences de performance proviennent de l'adéquation entre les données d'entrainement et les données de test. De plus, certains systèmes de RAP appliquent des algorithmes de SE en prétraitement, qui sont eux aussi plus ou moins efficaces selon le type de bruit. Deux types de modèle de régression ont été utilisés : la régression linéaire et la régression par Perceptron Multi-Couches (PMC). Scikit-learn (Pedregosa *et al.*, 2011) a été utilisé pour calculer les régressions. Notre MLP utilise 3 couches cachées et 18 (bandes Bark) * 6 neurones par couche.

3 Expériences

3.1 Corpus

Le corpus Wall Street Journal (WSJ) (John, DVD Philadelphia Linguistic Data Consortium 1993; wsj, Philadelphia Linguistic Data Consortium 1994), est fréquemment utilisé en RAP. Ce corpus a été choisie pour limiter les erreurs induites par le modèle de langage, les accents et les disfluences. Les sous-ensembles train_si284, dev93 et eval92 ont été séléctionnés lors de cette expérience. The details concerning the data size are indicated in the table 1.

Nous utilisons également le corpus NOISEX-92 (Varga & Steeneken, 1993) pour bruiter artificiellement nos données. Ce corpus est composé de 15 types de bruit d'une durée de 3min 56s chacun. La parole et le bruit ont été mixés pour neuf niveaux de RSB (de -15dB à 25dB par pas de 5dB) pour les 15 types de bruit (voir tableau 2). La fonction *v_addnoise* de la boîte à outils VoiceBox (Brookes, 2006) a été utilisée pour faire le mixage. La séléction de x secondes de bruit pour bruiter un tour de parole est aléatoire. Suite au mixage, nous disposons d'un corpus final composé de 83h pour 136 conditions différentes : 15 (types de bruit) * 9 (niveaux de SNR) * x (phrases sélectionnées dans la table 1) + x (pour la condition parole propre).

Les données train_si284 sont utilisées pour entraîner les différents modèles acoustiques des systèmes de RAP. Les données dev93 et eval92 ont été séparées en deux nouveaux sous-ensembles pour entraîner et pour tester la régression.

TABLE 1 – Corpus WSJ utilisé.

nom	nb locuteurs	nb phrases	temps
train_si284	284	37318	81h 15min
dev93	10	503	1h 5min
eval92	8	333	42min
Total	302	38154	83h

TABLE 2 – Types de bruit utilisés.

nom		
pink	factory1	destroyerengine
f16	babble	machinegun
white	leopard	destroyerops
m109	factory2	buccaneer1
volvo	hfchannel	buccaneer2

3.2 Vérité terrain

Afin de réaliser une vérité terrain pour notre système de prédiction, différents systèmes de RAP ont été entrainés via Kaldi (Povey *et al.*, 2011) :

— un sytème pour des conditions propres,
— un système multi-conditions, entraîné avec les 15 types de bruit en fixant le RSB à 10dB,
— 15 systèmes mono-condition, en fixant le type de bruit et le RSB à 10dB.

Les systèmes de RAP ont été réalisés grâce à la recette de Karel Vesely [2] : DNN-HMM sur des triphones, vecteur de 40 dimensions (MFCC-LDA-MLLT-fMLLR), vocabulaire de 20k et le modèle de langage est un n-gram. Pour information, dans les conditions propres, notre système de RAP appris avec train_si284 obtient 5,84% de WER sur dev93 et 3,42% sur eval92.

Pour entraîner et tester la régression entre les paramètres extraits par notre méthode et le WER, un découpage des sous ensembles dev93 et d'eval92 a été effectué. Pour chaque locuteur, 60% des phrases provenant de dev93 et eval92 sont séléctionnées pour l'entraînement et 40% pour le test :

2. `http://kaldi-asr.org/doc/dnn1.html`

soit 1h 4min pour l'entraînement et 43min pour le test. Ce découpage a été effectué afin de limiter l'impact du locuteur.

4 Résultats

Pour déterminer le coefficient ω optimal pour le masque binaire (voir section 2.2), 30 masques binaires ont été testés en faisant varier ω de 0 à 3 par pas de 0,1. Ces 30 masques permettent de calculer l'évolution de la densités des BM. La constante C a été fixée à 15 de manière empirique. Pour éliminer les valeurs aberrantes et ainsi améliorer le BM, la valeur de θ a été fixée à 0,4.

Afin d'évaluer la performance de la prédiction du WER (voir tableau 3), nous analysons l'erreur de prédiction absolu (PE pour Prediction Error) et l'écart type (SD pour Standard Deviation). La PE est calculée en moyennant les différences entre la prédiction et le WER réel pour chaque tour de parole. Le SD permet d'observer la dispersion de la prédiction à l'échelle d'un tour de parole. Dans la tableau 3, la PE et le SD sont affichés pour les deux régressions testées (linéaire et MLP). Le WER est prédit en utilisant différents systèmes de RAP pour évaluer l'indépendance de la méthode en fonction du système de RAP utilisée. Les résultats obtenus par le système de RAP ayant montré 5 différentes évolutions du WER en fonction du type de bruit, nous avons choisi d'afficher les résultats des 5 systèmes mono-condition correspondants. Les scores de PE obtenus sont généralement inférieurs à 8 sauf pour les systèmes mono-condition babble et factory. Par contre, la SD de la mesure reste importante (entre 10 et 11). Cette valeur indique que le WER ne peut pas être prédit efficacement pour une seule phrase : une fenêtre temporelle plus importante doit être utilisée. Cette même conclusion a été faite lors de l'expérience réalisé par Meyer (Meyer *et al.*, 2017), qui expliquait qu'un certain nombre de tours de parole devait être utilisé pour obtenir une prédiction sufisamment stable. Nous pouvons constater que la regression MLP obtient de meilleurs résultats que la régression linéaire.

TABLE 3 – Résultats des prédictions de WER pour différents systèmes de RAP.

		clean	multi-cond.	babble	fact.2	leopard	volvo	machinegun
Linéaire	PE	8.57	9.76	9.70	9.63	9.44	8.99	8.63
	SD	10.24	12.34	11.22	11.44	11.66	11.29	10.51
MLP	PE	6.89	7.88	8.55	8.07	7.92	7.27	7.13
	SD	10.09	11.60	10.72	11.00	10.78	10.96	10.36

Pour explorer les résultats, l'indépendance de la prédiction en fonction du type du bruit et du RSB, la PE a aussi été calculée pour chaque type de bruit et de RSB. De plus, afin d'évaluer l'indépendance de la prédiction en fonction du locuteur et du RSB, la PE a été calculée pour chaque locuteur et RSB.

Sur la figure 5, nous remarquons que la prédiction est liée au WER ciblé : les résultats sont plus précis lorsque le système obtient un WER faible ou important. Cette variabilité se constate car l'impact du modèle de langage sur le WER n'est pas ici quantifié. Par exemple, combien de phonèmes corrects sur un mot et ses voisins permettent d'identifier une suite de mots ? De plus, nous pouvons constater que le WER prédit est généralement sous estimé.

Nous remarquons aussi sur la figure 5 que le type de bruit n'influe pas sur la qualité de la prédiction : la méthode semble indépendante aux types de bruit.

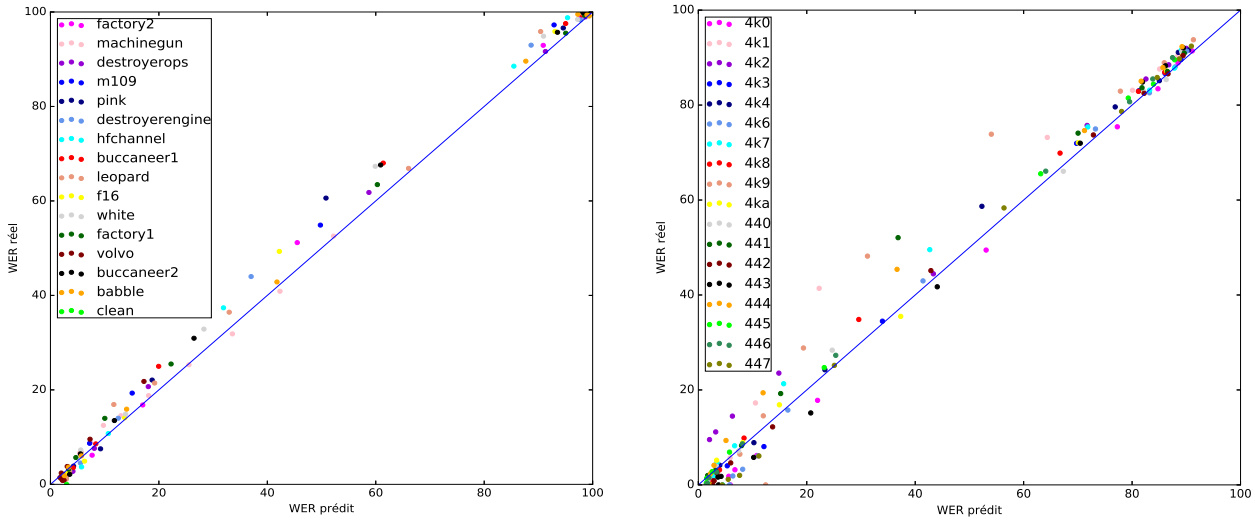

FIGURE 5 – À gauche, la différence entre la moyenne réelle et la prédiction (moyenne calculé pour le même type de bruit et de RSB). À droite, la différence entre la moyenne réelle et la prédiction (moyenne calculé pour le même locuteur et RSB).

Les locuteurs ont un impact sur les performances de la prédiction (voir figure 5). Par exemple, la prédiction est plus précise pour le locuteur 445 que pour le locuteur 4k9. Cependant, comme les textes dictés ne sont pas identiques entre les locuteurs, nous ne pouvons pas savoir si cet impact provient de la voix du locuteur, de la vitesse d'élocution ou du contenu des phrases lues.

5 Conclusions

Quel que soit la parole produite, le bruit va la dégrader. Connaître précisement l'impact du bruit sur la qualité de la transcription automatique de la parole, permet d'estimer un maximum ateignable par le système de RAP. Notre étude prouve qu'il est possible d'estimer *a priori* le WER obtenu par un système de RAP sur un fichier audio. En effet, les résultats montrent qu'avec une fenêtre temporelle suffisamment large, la prédiction est proche du WER réel (majoritairement inférieure à 8% d'erreur). Notre corpus étant composé de tours de parole indépendants, actuellement seul le calcul de la moyenne des WER prédits pour chaque tour de parole est proposé. Cependant, il est tout a fait possible d'imaginer une autre combinaison de scores. Par exemple, le filtrage des prédictions aberrantes ou l'utilisation d'un autre opérateur que la moyenne pour combiner les différents scores. Ce travail s'est focalisé uniquement sur les dégradations provoquées par le bruit afin d'analyser plus finement son impact sur le WER. Les paramètres extraits du bruit et de la parole séparée sont donc suffisamment corrélés au WER pour permettre une prédiction efficace pour de la parole lue.

La méthode de prédiction était, pour le moment, dépendante du locuteur afin d'analyser précisément l'impact du bruit. Une analyse de l'influence du locuteur sur le WER, comme la vitesse d'élocution ou le genre semble également intéressante. Cette seconde analyse permettrait d'obtenir idéalement une prédiction *a priori* indépendante du locuteur...

Références

(1994). Nist speech quality assurance (spqa) package v2.3. [Online]. Available : https ://www.nist.gov/itl/iad/mig/tools.

(Philadelphia : Linguistic Data Consortium, 1994). CSR-II (WSJ1) complete LDC94S13A.

BROOKES M. (2006). Voicebox : A speech processing toolbox for matlab. [Online]. Available : https ://goo.gl/hVRjXZ.

EYBEN F. (2015). Real-time speech and music classification by large audio feature space extraction. p. 20–21. Springer.

GHANNAY S., ESTÈVE Y. & CAMELIN N. (2015). Word embeddings combination and neural networks for robustness in asr error detection. In *European Signal Processing Conference*.

HERMANSKY H., VARIANI E. & PEDDINTI V. (2013). Mean temporal distance : Predicting asr error from temporal properties of speech signal. In *Int. Conf. Acoust. Speech Signal Process* : IEEE.

JIANG H. (2005). Confidence measures for speech recognition : A survey. *Speech Communication*, p. 45 (4) 455–470.

JOHN, ET AL. G. (DVD. Philadelphia : Linguistic Data Consortium, 1993). CSR-I (WSJ0) complete LDC93S6A.

MEYER B., MALLIDI S., KAYSER H. & HERMANSKY H. (2017). Predicting error rates for unknown data in automatic speech recognition. In *Int. Conf. Acoust. Speech Signal Process* : IEEE.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P., SILOVSKY J., STEMMER G. & VESELY K. (2011). The kaldi speech recognitiion toolkit. In *Workshop on Automatic Speech Recognition and Understanding*, p. 1–4 : IEEE.

RUWEI L., YANAN L., YONGQIANG, L.AND LIANG L. & WEILI C. (2016). Ilmsaf based speech enhancement with dnn and noise classification. *Speech Communication*, **85**, 53–70.

VARGA A. & STEENEKEN H. (1993). Assessment for automatic speech recognition : II NOISEX-92 : A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, **12**, 247–251.

WANG D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. *Speech Separation by Humans and Machines*, p. 181–197.

XU H., DALSGAARD P., TAN Z.-H. & LINDBERG B. (2007). Noise condition-dependent training based on noise classification and snr estimation. *IEEE transactions on audio, speech, and language processing*, **15**(8), 2431–2443.

YIMING S. & RUI W. (2015). Voice activity detection based on the improved dual-threshold method. In *Int. Con. on Intelligent Transportation, Big Data and Smart City*, p. 996–999.

ZWICKER E. (1961). Subdivision of the audible frequency range into critical bands. *The Journal of the Acoustical Society of America*, **33**, 248–248.