Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Une nouvelle mesure de la réverbération pour prédire les performances a priori de la transcription de la parole

Sébastien Ferreira[1,2], Jérôme Farinas[1], Julien Pinquier[1], Julie Mauclair[1] et 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

Dans cette étude, nous explorons la prédiction *a priori* de la qualité de la transcription automatique de la parole dans le cas de la parole réverbérée enregistrée avec un seul microphone. Cette prédiction est faite avant le décodage pour informer les utilisateurs de la qualité de la transcription attendue. Dans cette étude, nous nous concentrons uniquement sur les pertes de performance liées à la réverbération. Une nouvelle mesure de réverbération appelée « Excitation Behavior » est introduite. Cette mesure exploite le résidus de la prédiction linéaire sur les fenêtres voisées du signal de parole. L'expérience a été menée sur le corpus Wall Street Journal, réverbéré par des réponses impulsionnelles provenant du REVERB Challenge. Par rapport aux autres mesures de réverbération testées, notre mesure obtient une amélioration relative de 20% de la prédiction du taux d'erreur (aussi bien au niveau des phonèmes que des mots).

ABSTRACT

A new reverberation measure to predict *a priori* ASR performance

In this study, we explore the *a priori* prediction of the quality of automatic speech transcription in the case of reverberant speech recorded with a single microphone. This prediction is made before decoding to inform users of the expected transcription quality. We studied only the performance losses related to reverberation. A new reverberation measure called "Excitation Behavior" is introduced. This measure exploits the residuals of the linear prediction on the voiced windows of the speech signal. The experiment was conducted on the Wall Street Journal corpus, reverberated with impulse responses from the REVERB Challenge. Compared to the other reverberation measurements tested, our measurement obtains a relative prediction improvement of more than 20% (both at phone and word level).

MOTS-CLÉS : prédiction de performance, reconnaissance automatique de la parole, réverbération.

KEYWORDS: performance prediction, automatic speech recognition, reverberation.

1 Introduction

Au cours de la dernière décennie, les systèmes de Reconnaissance Automatique de la Parole (RAP) ont atteint de bonnes performances sur de la parole « propre ». L'amélioration de la robustesse des systèmes de RAP par rapport aux différents environnements acoustiques est un problème de recherche toujours d'actualité. Même si actuellement les systèmes de RAP sont de plus en plus

Volume 1 : Journées d'Études sur la Parole, pages 226–234.
HAL : hal-02798541.

robustes, la qualité de la transcription reste fortement dépendante de la qualité du signal de parole. Dans les systèmes commerciaux de RAP, les enregistrements soumis à une tâche de transcription ont généralement (voir même quasi-exclusivement) un seul canal audio. Ce constat limite les méthodes pouvant être employées afin d'améliorer la robustesse des systèmes car les algorithmes exploitant plusieurs microphones, comme les traitements par faisceaux (« beamforming » en anglais), ne peuvent pas être utilisés (Kinoshita *et al.*, 2016).

Comme nous l'avions fait précédemment lors d'une étude sur la parole bruitée (Ferreira *et al.*, 2018), l'objectif de cet article est de prédire *a priori* la qualité de la transcription des systèmes de RAP pour la parole réverbérée. L'analyse est *a priori* car elle est entièrement effectuée avant le décodage de la parole. L'avantage de cette prédiction est d'informer au plus tôt un utilisateur de la qualité de la transcription attendue. Nous avons choisi de traiter uniquement le cas de la réverbération afin d'isoler les sources de dégradation de la qualité de la parole. Ces travaux ont vocation à être utilisés par des systèmes commerciaux de transcription automatique de la parole (comme par exemple dans le modèle économique de la société Authôt), donc certaines contraintes ont été fixées :

- la mesure doit être non-intrusive,
- la réponse impulsionnelle de la pièce d'enregistrement est inconnue,
- la prédiction doit être réalisée avant le décodage : aucun score de confiance ou hypothèse de transcription ne peuvent être utilisés,
- le signal est mono-canal,
- le système de RAP doit être considéré comme une boîte noire.

Afin de créer cette prédiction, nous cherchons des mesures qui quantifient l'impact de la réverbération sur la qualité de la transcription automatique de la parole. Nous cherchons un score d'« intelligibilité de la parole » pour les systèmes de RAP lorsque le signal de parole est réverbéré.

Les méthodes d'estimation de la réverbération qui satisfont l'ensemble de nos contraintes se répartissent en deux catégories : l'estimation aveugle du T60 (temps de réverbération) et les scores d'intelligibilité de la parole non-intrusive. Le temps de réverbération ou T60 est défini comme le temps nécessaire pour que le niveau de pression acoustique diminue de 60 dB après extinction de la source d'excitation sonore (ISO 3382, 1997). Dans de nombreuses situations, le T60 est estimé (et non calculé) car la réponse impulsionnelle de la pièce (RIR pour Room Impulse Response) est inconnue. Il existe deux méthodes pour estimer le T60 : la distribution de décroissance spectrale (SDD pour Spectral Decay Distribution) qui estiment la distribution de la décroissance de l'enveloppe de puissance du signal dans le temps afin d'estimer le temps de réverbération (Dumortier & Vincent, 2014) et l'analyse des résidus de la Prédiction Linéaire (PL) (Keshavarz *et al.*, 2012). Les méthodes orientées vers l'intelligibilité de la parole reposent sur la quantification des déformations du signal de parole. Par exemple, dans cet article (Falk *et al.*, 2010), ce sont les caractéristiques spectrales de modulation qui sont exploitées. Les méthodes qui estiment le T60 n'ont pas été conçu pour informer de la distance entre le locuteur et le microphone : il s'agit pourtant d'une variable importante de la qualité de la parole. Les méthodes orientées vers l'intelligibilité de la parole ont été conçu pour prédire l'intelligibilité humaine. Les tests que nous avons effectués sur la prédiction *a priori* de la qualité de la transcription avec ces mesures n'ont pas été satisfaisants. Pour ces raisons, nous avons décidé de créer une nouvelle mesure, que nous appelons Excitation Behaviour (EB), qui quantifie l'impact de la réverbération sur les performances des systèmes de RAP.

Le calcul de cette mesure est détaillée en section 2. Dans la section 3, le protocole d'expérimentation est décrit afin d'évaluer l'EB : le taux d'erreur mots (WER pour Word Error Rate) et le taux d'erreur phonèmes (PER pour Phone Error Rate) sont prédits sur différentes conditions de réverbération et discutés en section 4.

2 La mesure Excitation Behaviour

2.1 Architecture

L'Excitation Behaviour (EB) est un paramètre qui appartient aux méthodes d'analyse des résidus de la Prédiction Linéaire (PL). Les résidus de la PL ou le signal d'erreur de prédiction $e(n)$ est la différence entre la parole d'entrée et la parole estimée (Makhoul, 1975).

$$e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

où a_k correspond aux coefficients de la prédiction linéaire (LPC pour Linear Prediction Coefficient), p l'ordre du filtre et $s(n)$ l'échantillon de parole.

Sur les fenêtres voisées de la parole, le résidu de la PL contient des informations sur l'instant de fermeture glottale et sur la source d'excitation (Ananthapadmanabha & Yegnanarayana, 1979). Lorsque la parole est réverbérée, la différence entre les impulsions glottales et la source d'excitation est plus faible (voir figure 1). C'est cette distorsion des résidus de la PL qui sera exploitée par l'EB. L'architecture de l'extraction de l'EB se fait en trois étapes : la sélection automatique des fenêtres de parole voisées, l'extraction du résidu de la PL et le calcul d'une valeur statistique basée sur des ratio entre percentiles de la distribution de ce résidu.

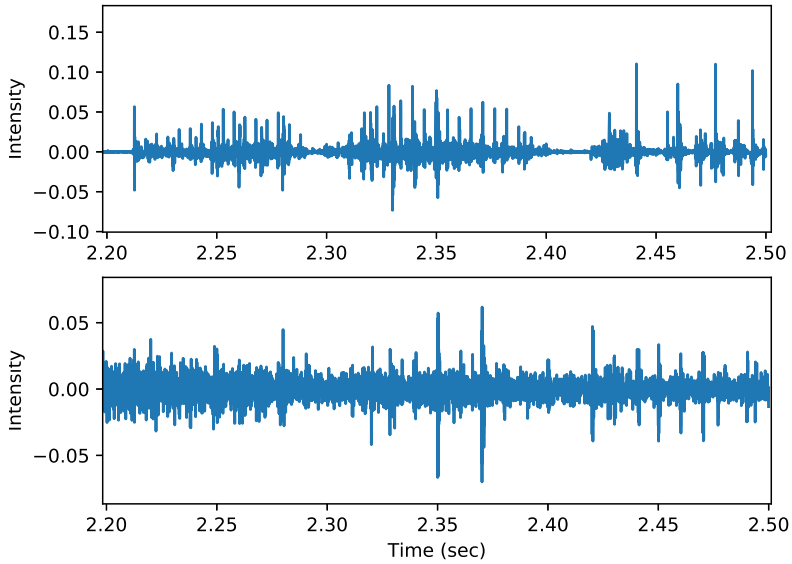

FIGURE 1 – Résidus de la PL d'un signal de parole. En haut signal propre. En bas version réverbérée

2.2 Sélection des fenêtres voisées

Pour sélectionner les fenêtres voisées de la parole, nous analysons la différence entre la racine quadratique moyenne (RMS pour Root Mean Square) et le taux de passage par zéro (ZCR pour Zero Crossing Rate) du signal. Une fenêtre de 16 ms est sélectionnée comme parole prononcée lorsque :

$$\frac{RMS_{trame}}{\max RMS_{global}} - \frac{ZCR}{2} > 0 \tag{2}$$

Un exemple de sélection de fenêtres est présenté dans la figure 2. Pour filtrer d'avantage cette sélection, seules les fenêtres voisées suivies de deux autres fenêtres voisées sont considérées comme voisées.

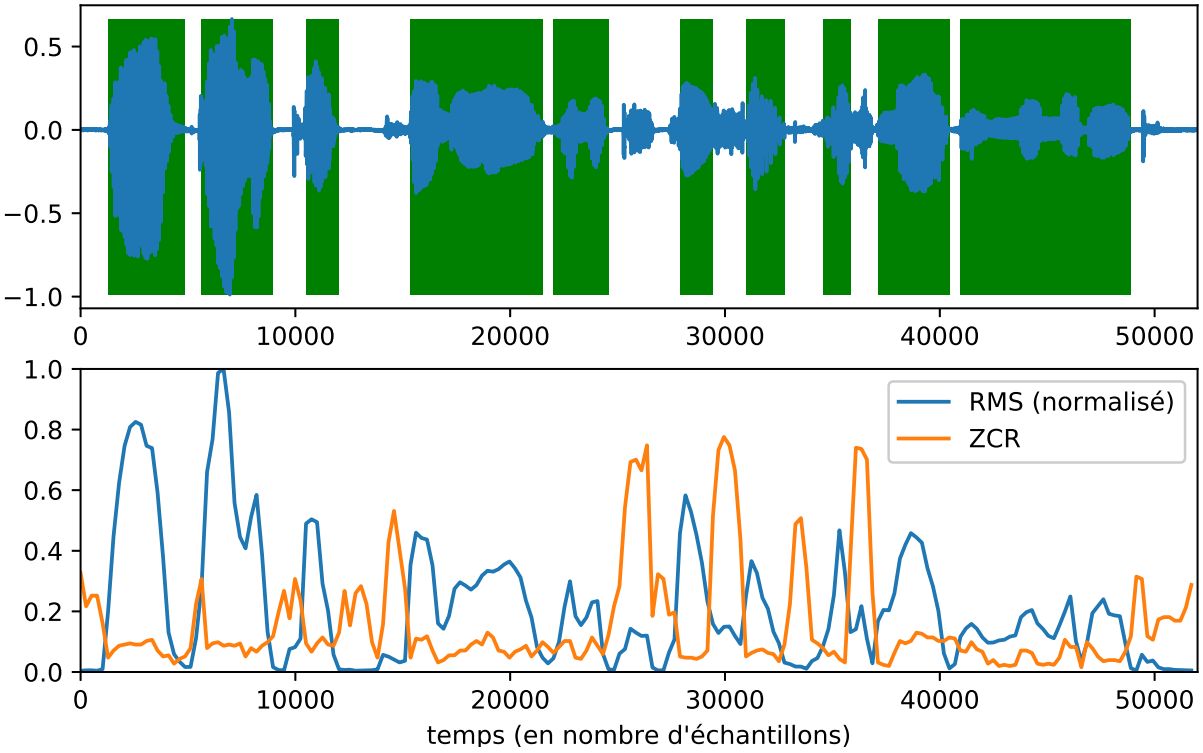

FIGURE 2 – En haut, le signal avec en vert les zones voisées sélectionnées automatiquement (équation 2). En bas, les valeurs du RMS normalisés et du ZCR correspondantes.

2.3 Calcul de la mesure

Après avoir sélectionné les fenêtres voisées, le signal est préaccentué. Pour calculer les LPC du signal, un modèle autorégressif est estimé avec l'algorithme de Levinson-Durbin pour minimiser l'erreur de prédiction (résidus de la PL). Les résidus de la PL sont calculés avec une fenêtre de 10ms et un ordre de 21. Pour chaque succession de 4 trames voisées, la transformée de Hilbert est utilisée sur les résidus de la PL et est normalisée par la valeur maximale : le maximum correspond à l'impulsion glottal la plus grande. La distribution de la transformée de Hilbert normalisée des résidus de la PL est affectée par la réverbération. Si nous observons les valeurs p_{90}, p_{50} ou p_{10} seules, la réverbération augmente ces valeurs. Cependant, ces valeurs de percentile sont très variables, selon le locuteur. C'est pourquoi le rapport défini dans l'équation 4 est calculé pour définir la mesure de l'EB, tel que :

$$p_{90}p_{50} = p_{90} - p_{50} \quad \text{et} \quad p_{50}p_{10} = p_{50} - p_{10} \tag{3}$$

où p_x définit les percentiles x^{th}. Avec la liste des $p_{90}p_{50}$ et des $p_{50}p_{10}$ obtenues pour la phrase, nous calculons le score EB avec :

$$EB = \frac{\overline{p_{90}p_{50}}}{\overline{p_{50}p_{10}}} \tag{4}$$

L'EB est proche de 1,8 lorsque la réverbération provoque beaucoup d'erreur par les système de RAP, et supérieur à 2,8 lorsque la réverbération est négligeable par les systèmes de RAP.

3 Expériences

3.1 Corpus

Le corpus Wall Street Journal est largement utilisé en RAP : WSJ0 (Garofalo *et al.*, 1993) et WSJ1 (Garofolo *et al.*, 1994). Ce corpus a été choisi pour limiter les erreurs induites par les modèles de langage, les accents et les disfluences. Les données *train_si284* (sans ajout de réverbération) sont utilisées pour entraîner le système de RAP. Pour travailler avec la parole réverbérée, les sous-ensembles *dev93* et *eval92* ont été convolués avec des réponses impulsionnelles des pièces (RIR pour Room Impulse Responses) mesurées dans différentes conditions. Les RIR proviennent du REVERB challenge (Kinoshita *et al.*, 2013). Les RIR enregistrées permettent de simuler 12 conditions de réverbération différentes : 6 pièces de tailles différentes (2 petites, 2 moyennes et 2 grandes) avec 2 types de distances entre un haut-parleur et un réseau de microphones (proche = 50 cm et loin = 200 cm). Les temps de réverbération des petites, moyennes et grandes pièces sont respectivement d'environ 0,25, 0,5 et 0,7s. Les sous-ensembles *Dev93* et *eval92* ont donc 13 conditions de réverbération différentes (le +1 vient du cas sans réverbération). Ces deux sous-ensembles sont respectivement utilisés pour entraîner et tester le modèle de régression.

3.2 Architecture du système de prédiction

Les systèmes de prédiction des performances des systèmes de RAP sont souvent appris avec une méthode de régression supervisée (voir Figure 3). L'objectif de cette régression est de modéliser le lien entre les caractéristiques extraites et la performances du système RAP : les systèmes de RAP sont plus ou moins robustes à certaines distorsions. Cette régression nous permet d'évaluer à quel point les différentes mesures de réverbération testées permettent de prédire l'impact de la réverbération sur les performances du système de RAP.

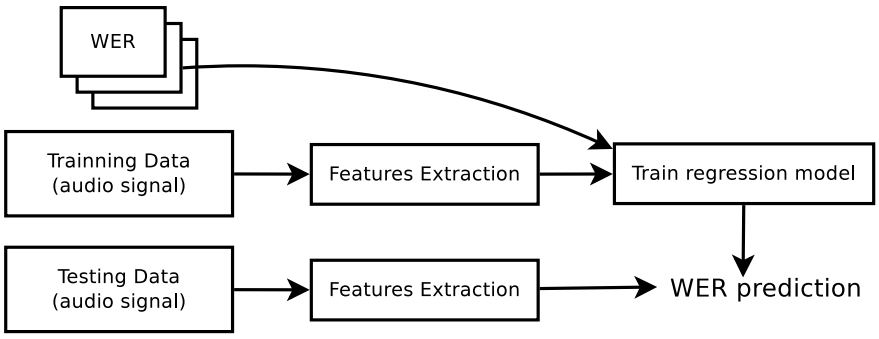

FIGURE 3 – Architecture du systèmes de prédiction de la qualité de la transcription.

3.3 Vérité terrain

Afin d'établir une vérité terrain pour notre système de prédiction, un système de RAP a été entraîné. Nous avons utilisé la recette de Karel Vesely avec Kaldi (Veselỳ *et al.*, 2013). Le système est hybride avec un réseau de neurones profond et des modèles de Markov cachés (DNN-HMM) entraîné par entropie croisée. Dans des conditions propres (train et test), le système de RAP entraîné avec *train_si284* obtient 5,84% de WER sur *dev93* et 3,42% sur *eval92*.

Pour construire le décodeur acoustico-phonétique utilisé pour prédire le PER, nous phonétisons au préalable les corpus d'entraînement et de test. Le dictionnaire de prononciation est modifié pour être composé uniquement des phonèmes possibles. Le modèle de langage est remplacé par un 1-gram appris sur un corpus de texte phonétisé. Ces changements permettent d'obtenir au décodage, une séquence de phonèmes sans aucun impact du modèle de langage. Une fois ces modifications apportées, le décodeur acoustico-phonétique est entraîné de la même manière que celle décrite au paragraphe précédent pour le système de RAP.

Pour calculer les modèles de régression, nous utilisons un perceptron multi-couche (MLP pour Multi-Layer Perceptron) que nous avons entraîné avec Scikit-learn.La MLP est assez simple et utilise une seul couche composée 3 neurones. Le but de la regression MLP est seulement d'obtenir une régression non linéaire.

3.4 Mesures de réverbération testées

Pour comparer la mesure de l'EB décrite dans la section 2, nous avons utilisé plusieurs mesures de l'état de l'art (voir le tableau 1) :

- SRMR+ : SRMR et SRMR normalisés (Falk *et al.*, 2010),
- Slope : ici c'est la valeur de "floored ratio of spectral substraction" (Tachioka *et al.*, 2013),
- Neg-side : une méthode de SDD qui utilise la variance négative et le skewness (Dumortier & Vincent, 2014),
- LP-kurto : moyenne des kurtosis des résidus de la PL d'ordre 10 (trame de 32ms) (Gillespie *et al.*, 2001).

4 Résultats & discussion

Afin d'évaluer la performance de la prédiction du WER, nous avons calculé l'erreur de prédiction absolue moyenne (MAE pour Mean Average Error) et l'écart-type (SD pour Standard Deviation). Dans (Willmott & Matsuura, 2005), les auteurs indiquent que le MAE est une mesure plus naturelle que le RMSE, et qu'elle est non ambiguë. Le MAE et le SD sont indiqués pour toutes les conditions de réverbération testées. Les résultats de la prédiction sont présentés dans le tableau 1.

En ce qui concerne la prédiction *a priori* du WER, l'EB obtient une meilleure précision de prédiction que les autres mesures de réverbération testées. Une amélioration relative de 23% de la prédiction de l'erreur moyenne de WER est obtenue avec EB par rapport à SRMR+. De plus, l'erreur de prédiction est moins dispersée.

Sachant que notre méthode de prédiction n'utilise pas d'informations linguistiques, nous avons voulu

TABLE 1 – Resultats de prédiction (WER et PER) avec une régression MLP.

	SRMR+	Slope	Neg-side+	LP-kurto	EB
WER (%)					
MAE	17,75	18,44	17,45	18,33	**13,66**
SD	14,26	14,95	13,75	15,69	**12,63**
PER (%)					
MAE	10,76	12,59	10,82	11,43	**7,86**
SD	8,14	9,04	7,75	9,15	**6,25**

observer les performances de prédiction du PER d'un système de décodage acoustico-phonétique : cela permet de retirer l'influence du modèle de langage. Le but est de savoir si la prédiction caractérise bien les distorsions acoustiques de la parole réverbérée. Nous pouvons voir, comme prévu, que la prédiction du PER est plus précise que la prédiction du WER. La précision de la prédiction de PER par notre méthode est meilleure que celle des autres mesures de réverbération testées. Une amélioration relative de 27% de la prédiction de l'erreur moyenne de PER est obtenue avec EB par rapport à SRMR+.

La mesure EB permet d'obtenir une prédiction des performance des système de RAP plus précise. Cependant, la précision reste insatisfaisante à l'échelle des phrases. La durée d'énonciation des phrases du corpus utilisé varie de 3 à 16 secondes, avec une moyenne de 7 secondes. Pour une tâche aussi difficile que la prédiction *a priori* des performances d'un système de RAP, une fenêtre temporelle plus longue est nécessaire. Dans la grande majorité des cas, les enregistrements soumis à une tâche de transcription automatique par les utilisateurs des services de RAP commerciaux dépassent 3 minutes. Nous avons donc analysé l'évolution de la prédiction en cumulant le nombre de phrases dans les même condition pour le même locuteur.

Dans la figure 4, la précision de la prédiction du WER devient plus précise, en fonction du nombre de tours de parole utilisés. Ainsi, la prédiction du WER atteint 7,13 d'erreur absolue moyenne avec 20 énoncés utilisés (durée d'environ 140 s). La prédiction de PER atteint 5,04 d'erreur absolue moyenne avec 20 énoncés utilisés (voir figure 5).

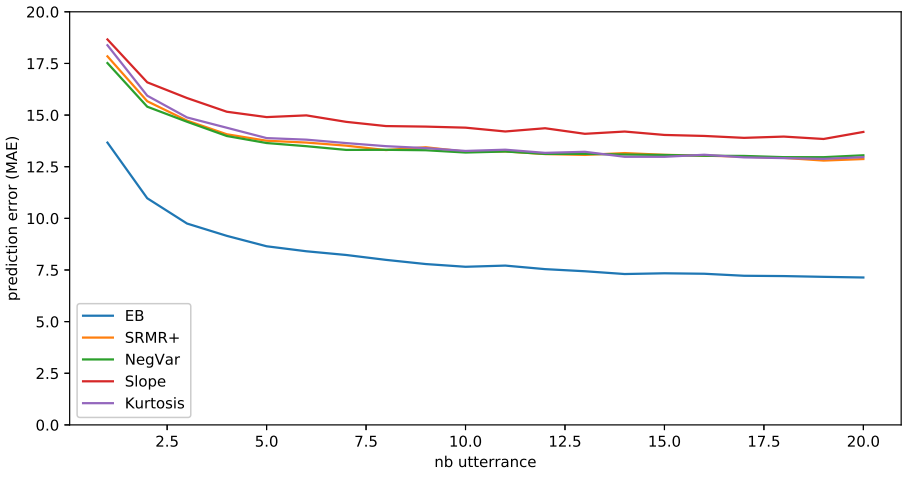

FIGURE 4 – Moyenne des erreurs de prédiction du WER en fonction du nombre de phrases utilisées.

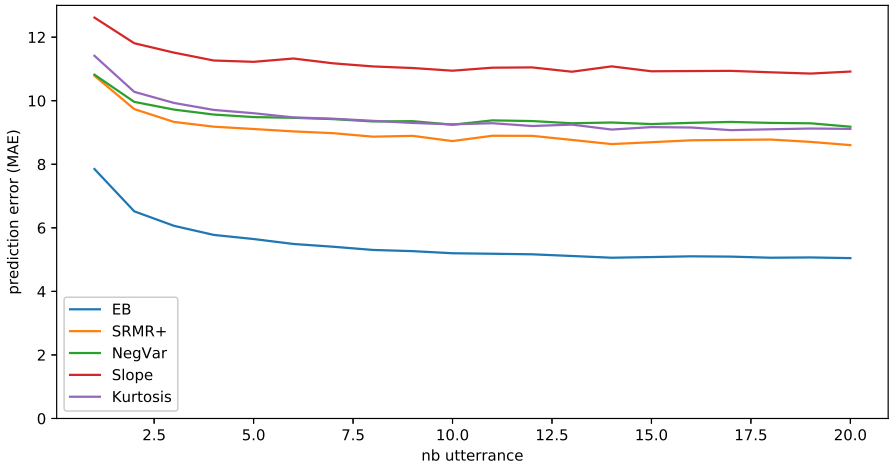

FIGURE 5 – Moyenne des erreurs de prédiction du PER en fonction du nombre de phrases utilisées.

5 Conclusions

Lorsqu'un signal de parole réverbéré est transcrit par un système de RAP, un grand nombre d'erreurs de transcription est possible. Dans les systèmes de RAP commerciaux, les fichiers audio sont principalement enregistrés avec un seul microphone : les méthodes exploitant plusieurs microphones ne peuvent pas être utilisées pour améliorer la robustesse des systèmes. Il est difficile de savoir dans quelle mesure la réverbération affecte les systèmes de RAP car le comportement de ces systèmes est très différent de celui des humains (lorsque la parole est réverbérée). Le principal objectif de ce papier est de prédire la qualité de la transcription *a priori* de la parole réverbérée, tout en respectant les contraintes des systèmes commerciaux de transcription automatique de la parole.

Nous avons contribué à l'élaboration de la mesure que nous appelons EB, pour quantifier l'impact de la réverbération sur les systèmes de RAP. L'EB extrait une mesure statistique sur le résidu de la PL, qui est bien corrélé à la performance des systèmes de RAP. Pour évaluer l'EB, nous avons prédit le WER et le PER obtenus avec les différentes mesures de réverbération testées. Au niveau d'une phrase, la mesure EB fournit une amélioration relative de la précision de prédiction du WER et du PER qui est de 20% supérieure aux autres mesures testées. La mesure EB obtient une erreur moyenne de prédiction de WER de 7,13 lorsque 140 secondes sont analysées, et une erreur moyenne de prédiction de PER de seulement 5,04. Dans tout les cas, la mesure de l'EB fournit une prédiction plus précise que les autres mesures testées.

Pour prédire *a priori* plus précisément la qualité de transcription des systèmes de RAP, d'autres analyses acoustiques du signal de parole pourraient être réalisées. Il serait intéressant d'observer l'influence du bruit et de la musique superposée. D'autre part, certaines variabilités au niveau des locuteurs, comme le débit de parole, le sexe ou l'âge, pourraient également améliorer la précision de la prédiction *a priori* de la qualité de la transcription automatique de la parole.

Références

ANANTHAPADMANABHA T. & YEGNANARAYANA B. (1979). Epoch extraction from linear

prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27**(4), 309–319.

DUMORTIER B. & VINCENT E. (2014). Blind rt60 estimation robust across room sizes and source distances. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5187–5191 : IEEE.

FALK T. H., ZHENG C. & CHAN W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, **18**(7), 1766–1774.

FERREIRA S., FARINAS J., PINQUIER J. & RABANT S. (2018). Prédiction a priori de la qualité de la transcription automatique de la parole bruitée. *JEP*.

GAROFALO J. *et al.* (1993). CSR-I (WSJ0) complete LDC93S6A. Linguistic Data Consortium, Philadelphia, USA.

GAROFOLO J., GRAFF D., PAUL D. & PALLETT D. (1994). CSR-II (WSJ1) Complete. Linguistic Data Consortium, Philadelphia, USA.

GILLESPIE B. W., MALVAR H. S. & FLORÊNCIO D. A. (2001). Speech dereverberation via maximum-kurtosis subband adaptive filtering. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, volume 6, p. 3701–3704 : IEEE.

ISO 3382 (1997). *Acoustics : measurement of the reverberation time of rooms with reference to other acoustical parameters*. Standard, International Organization for Standardization, Genève, Suisse.

KESHAVARZ A., MOSAYYEBPOUR S., BIGUESH M., GULLIVER T. A. & ESMAEILI M. (2012). Speech-model based accurate blind reverberation time estimation using an lpc filter. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(6), 1884–1893.

KINOSHITA K., DELCROIX M., GANNOT S., P. HABETS E. A., HAEB-UMBACH R., KELLERMANN W., LEUTNANT V., MAAS R., NAKATANI T., RAJ B., SEHR A. & YOSHIOKA T. (2016). A summary of the REVERB challenge : state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, **2016**(1). DOI : 10.1186/s13634-016-0306-6.

KINOSHITA K., DELCROIX M., YOSHIOKA T., NAKATANI T., SEHR A., KELLERMANN W. & MAAS R. (2013). The reverb challenge : A common evaluation framework for dereverberation and recognition of reverberant speech. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, p. 1–4 : IEEE.

MAKHOUL J. (1975). Linear prediction : A tutorial review. *Proceedings of the IEEE*, **63**(4), 561–580.

TACHIOKA Y., HANAZAWA T. & IWASAKI T. (2013). Dereverberation method with reverberation time estimation using floored ratio of spectral subtraction. *Acoustical Science and Technology*, **34**(3), 212–215.

VESELỲ K., GHOSHAL A., BURGET L. & POVEY D. (2013). Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, p. 2345–2349.

WILLMOTT C. J. & MATSUURA K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, **30**(1), 79–82.