# Analyse du bruit pour la prédiction de la qualité de la transcription automatique de la parole

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

[1]prenom.nom@irit.fr,

[2]sferreira@authot.com,srabant@authot.com

**Résumé –** De nombreuses sources de variabilité peuvent dégrader les performances des systèmes de Reconnaissance Automatique de la Parole (RAP). Dans cette étude, les dégradations provoquées par le bruit sont analysées afin de prédire *a priori* la qualité de la RAP, i.e. avant décodage. Notre méthode d'extraction de paramètre, nommée Sub-band Statistical Feature (S-SF), se base sur une séparation de la parole et du bruit. Une fois séparée, des statistiques sont extraites par bande fréquentielle. Pour relier ces paramètres à un système de RAP, un modèle de régression est calculé. L'expérimentation a été réalisée sur le corpus Wall Street Journal, bruité avec le corpus NOISEX-92 (15 types de bruit) que nous appliquons à 9 niveaux de rapport signal sur bruit. La méthode de régression proposée obtient 8,75 d'erreur de prédiction de WER sur un système de RAP entraîné avec des données non-bruitées. Lorsque 20 tours de parole sont utilisés (durée d'environ 140s), l'erreur de prédiction décroît à 5,82. Notre extraction de S-SF permet une amélioration relative de 20% par rapport à l'extraction des Sub-band Signal-to-Noise Ratio (S-SNR). Cette prédiction peut être utilisée pour ignorer des portions d'audio dont la transcription automatique de la parole est de mauvaise qualité et pour informer l'utilisateur, au plus tôt, de la qualité de la transcription pouvant être obtenue.

**Abstract –** Many sources of variability can degrade the performance of Automatic Speech Recognition (ASR) systems. In this study, noise-induced impairments are analyzed to predict the *a priori* quality of an ASR, i.e. before decoding. Our parameter extraction method called Sub-band Statistical Feature (S-SF) is based on a separation of speech and noise. Once separated, statistics are extracted by frequency bands. To link these parameters to an ASR system, a regression model is calculated. The experiment was conducted on the Wall Street Journal corpus, noised with NOISEX-92 corpus (15 types of noise) that we apply at 9 levels of Signal-To-Noise Ratio (SNR). The proposed regression method obtains a WER prediction error of 8.75 on a ASR system trained with clean data. When multiple utterances are used, mean error achieved 5.82 with 20 utterances (duration about 140s). The statistical feature extracted by our method, compared to classical sub-band SNR extraction, obtained a relative amelioration of 20% of mean error prediction. This prediction can be used to ignore portions of audio whose automatic transcription of speech is of poor quality and to inform the user as soon as possible of the quality of the transcription that can be obtained.

## 1 Introduction

Les progrès effectués dans le domaine de la Reconnaissance Automatique de la Parole (RAP) permettent d'utiliser cette technologie dans des situations de plus en plus diverses. Cependant, la qualité de la transcription de la parole dépend grandement du cas d'utilisation : par exemple, la qualité audio et le degré de spontanéité peuvent fortement impacter celle-ci. La qualité de la transcription dépend également de la différence entre les données d'apprentissage et de test pour les modèles acoustique et linguistique. Ainsi, il est courant d'utiliser des systèmes spécifiques pour chaque cas et type d'utilisation : la dictée vocale, les réunions, la télévision, la radio, les conférences...

Avec la démocratisation des technologies vocales, les services de transcription automatique doivent être capables de traiter une grande hétérogénéité de documents, notamment au niveau acoustique avec la présence potentielle de bruit, de réverbération, de saturation... Afin que l'expérience utilisateur soit réussie, il est important d'informer au plus tôt de la qualité possible de la transcription automatique [1].

Nous cherchons donc ici à estimer la qualité de la transcription automatique avant le Décodage Acoustico-Phonétique (DAP). Nous nous sommes fixés certaines contraintes. D'une part, le système de transcription doit être utilisé comme une boîte noire : nous ne voulons pas extraire de scores internes ou connaître les données d'entraînement. Cela permet d'utiliser aisément n'importe quel système de transcription. D'autre part, l'estimation de la qualité doit être effectuée avant le décodage. Cela permet notamment de ne pas transcrire inutilement certains fichiers. Nous souhaitons que cette mesure soit indépendante du locuteur. Et idéalement, la méthode doit être peu coûteuse en temps de calcul.

Le taux d'erreur sur les mots, en anglais Word Error Rate

(WER), est la métrique de référence permettant d'évaluer la performance des systèmes de transcription. Il existe de nombreuses méthodes pour prédire ce WER. Ces méthodes peuvent exploiter les scores internes (et mesures de confiances) du système de RAP [2, 3], les probabilités *a posteriori* des phonèmes [4], le texte transcrit [5], les statistiques des données d'entraînement [6]. Cependant ces méthodes ne respectent pas les contraintes que nous nous sommes fixées précédemment.

Nous avons choisi de nous concentrer dans cette étude sur l'impact du bruit sur le WER. Pour estimer l'impact du bruit sur la parole, le Signal to Noise Ratio (SNR) est une métrique couramment utilisée. Mais le SNR seul n'explique pas les chutes de performance des systèmes RAP, le type de bruit est aussi à prendre en compte (voir figure 1). Ce papier explore une nouvelle méthode d'extraction de paramètres, le Sub-band Statistical Feature (S-SF), pour estimer l'impact du bruit sur le WER. Pour tester le S-SF nous avons artificiellement bruité de la parole à différents niveaux de SNR, avec différents type de bruit. Nous comparerons cette méthode avec l'utilisation du Sub-band SNR (S-SNR). Pour évaluer la performance de la prédiction, nous évaluerons l'erreur moyenne absolue entre la prédiction et le WER sur différents systèmes de RAP. Nous avons utilisé le même processus pour évaluer la performance de prédiction des phonèmes, en anglais Phone Error Rate (PER), d'un DAP afin d'évaluer l'extraction de paramètres sans influence du modèle de langage. En effet cette mesure sera plus proche de la performance acoustique de la RAP.

Le système de prédiction du WER est présenté en section 2, le cadre expérimental en section 3, puis, les résultats obtenus sont exposés en section 4.

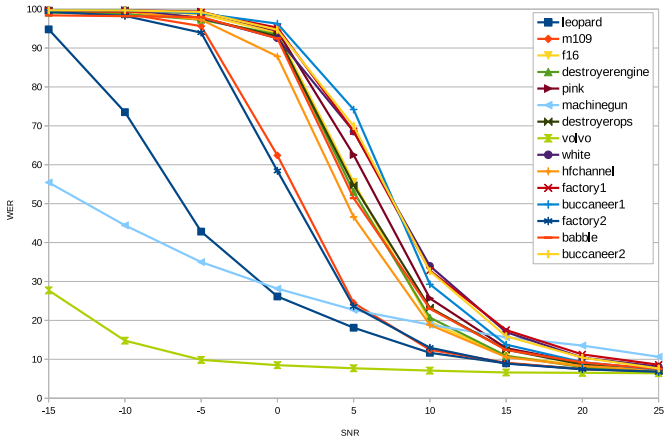


Figure 1 – WER obtenu en fonction du type de bruit et du SNR. Les types de bruit proviennent du corpus NOISEX-92 [7].

## 2 Système de prédiction du WER

Il est composé de 4 étapes (voir figure 2). Mise à part le calcul du masque binaire, les autres étapes sont assez classiques et n'influencent que légèrement les résultats.

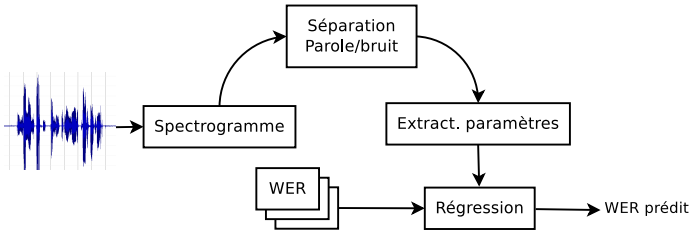


Figure 2 – Architecture du système de prédiction.

**1. Spectrogramme.** L'analyse du signal est classiquement fondée sur le calcul d'un spectrogramme avec une taille de fenêtres de 512 points (32 ms) et un recouvrement de moitié.

**2. Masque binaire.** Pour séparer la parole et le bruit, nous utilisons un Masque Binaire (MB) [8]. Le principe est d'identifier pour tous les couples (temps, fréquences) du spectrogramme, si l'amplitude provient de la parole ou du bruit.

Pour déterminer notre MB, nous calculons pour chaque bande fréquentielle Bark [9] :

$$MB(f,t) = \begin{cases} 1, & \text{si } E(f,t) \geq \omega_{f_{bin}} * \overline{E(f_{bin})} \\ 0, & \text{sinon} \end{cases}$$

avec $f$ la fréquence, $t$ la trame d'analyse, $\overline{E(f)}$ la moyenne des amplitudes sur la bande Bark ciblée, et $\omega_{f_{bin}}$ une pondération calculée pour chaque bande Bark. Pour déterminer $\omega_{f_{bin}}$, nous utilisons l'algorithme 1. La constante $Cnst$ a été fixée à 15 de manière empirique.

**algorithm 1** Détermination de $\omega_{f_{bin}}$ (voir figure 3)

1: Calcul de HC, l'histogramme cumulé divisé par $\overline{E(f_{bin})}$
2: $dHC \leftarrow |\Delta(1 - HC)|$
3: $Cnst \leftarrow 15$
4: $\omega_{f_{bin}} \leftarrow min(x)$ tq $\begin{cases} dHC(x) < max(dHC)/Cnst \\ x > argmax(dHC) \end{cases}$

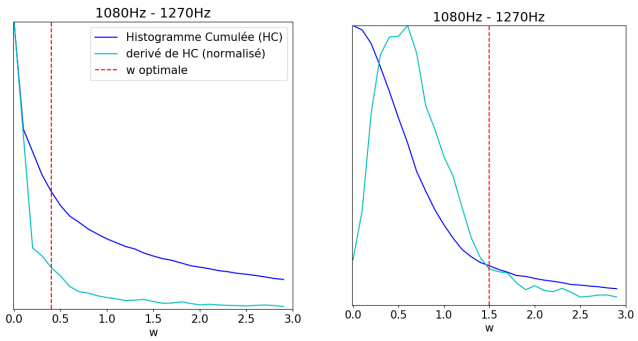


Figure 3 – Évolution de HC et dHC en fonction de $\omega$ (bande fréquentielle 1080-1270). À gauche sur un extrait propre. À droite sur le même extrait avec un bruit blanc à 0 dB.

Un filtrage est effectué sur MB pour éliminer les artefacts résiduels. Nous appliquons pour toutes les valeurs du MB un masque local carré de taille 9 qui modifie à 1 la valeur centrale si le masque local est $> 4$ sinon la valeur est de 0. Sur la

figure 4, nous pouvons observer le spectrogramme d'un extrait audio, le MB, le bruit et la parole séparée [1].

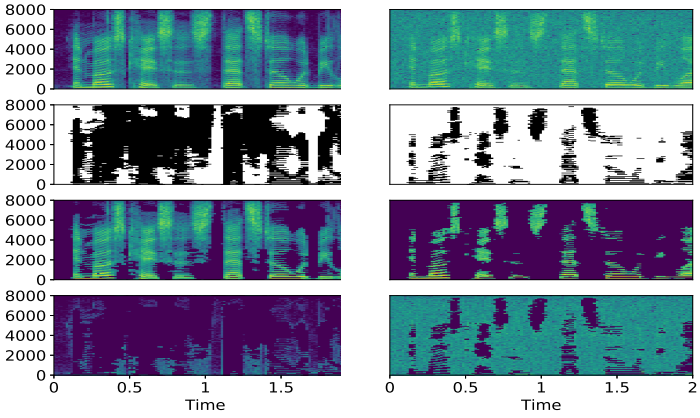


Figure 4 – Séparation de la parole et du bruit. À droite le fichier propre et à gauche le même fichier bruité. De haut en bas : spectrogramme, MB, parole isolée, bruit isolé.

**3. Paramètres statistiques.** Pour chaque bande Bark allant de 0 à 4400 Hz, nous calculons, différents paramètres statistiques : la somme des amplitudes et différents moments (moyenne, variance, kurtosis, etc.). Ces paramètres sont extraits pour les spectrogrammes de parole et de bruit précédemment séparés grâce au MB, ainsi que pour le spectrogramme initial. L'intérêt de cette partie est d'avoir plusieurs descripteurs discriminants.

**4. Modèle de régression.** Un modèle de régression est calculé entre les paramètres extraits et le WER obtenu par un système de RAP. Le corpus d'apprentissage utilisé ici (pour la régression) est différent de celui ayant servi à apprendre le système de RAP. La régression utilisée est un simple Multi-Layer Perceptron (MLP) avec 1 couche cachée de 6 neurones. Un MLP a été choisi car la régression est non-linéaire.

# 3 Expériences

**Données.** Le corpus de parole utilisé est le Wall-Street Journal (WSJ0 [10] et WSJ1 [11]), et plus précisément les sous-ensembles *train_si284*, *dev93* and *eval92*. Le corpus NOISEX-92 [7] permet de bruiter artificiellement la parole. La parole a été mixée avec 15 types de bruit pour 9 niveaux de SNR (allant de -15dB à 25dB tous les 5dB). Nous obtenons ainsi 136 conditions de bruits différentes (15 types de bruit * 9 niveaux de SNR + 1 condition non-bruité). Les données *train_si284* sont utilisées pour entraîner le systèmes de RAP. Les données *dev93* et *eval92* sont utilisées pour entraîner et tester le modèle de régression, respectivement.

1. Des exemples de fichier audio filtrés par le MB ainsi calculés sont disponibles ici `https://frama.link/bSykWP6q`

Pour construire le système de RAP, nous avons utilisé une recette du WSJ de Kaldi : la version de Karel Vesely's [12]. Le système est un Deep Neural Network - Hidden Markov Model (DNN-HMM). Le système obtient 5,84% de WER sur *dev93* et 3,42% sur *eval92* sur des données propres (non bruitées). Pour construire le DAP, la recette est similaire. Le dictionnaire et le modèle de langage ont été retirés du modèle afin de ne pas influencer la reconnaissance de phonèmes.

**SNR par bande (S-SNR)** Afin de comparer et d'évaluer l'efficacité de notre méthode, nous utilisons l'outil du NIST, le Speech Quality Assurance Package (SPQA) [2] pour extraire le SNR global et le SNR par bande (18 bandes Bark). Les paramètres ainsi extraits sont utilisés pour calculer un modèle de régression sur les mêmes données que notre méthode. Nous avons choisi d'extraire le SNR par bande afin de mieux quantifier la forme spectrale des différents types de bruit.

# 4 Résultats

Pour évaluer la performance de prédiction, nous calculons la moyenne des Erreurs de Prédiction (PE) et l'écart-type de ces erreurs (SD). Les résultats sont affichés dans le tableau 1 pour les deux méthodes d'extraction de paramètres: le S-SF décrit dans ce papier et la méthode S-SNR.

Table 1 – Prédiction avec le S-SF et le S-SNR

| Prédiction du WER | | |
|---|---|---|
| | S-SNR | S-SF |
| PE | 11,02 | 8,75 |
| SD | 12,28 | 11,02 |

| Prédiction du PER | | |
|---|---|---|
| | S-SNR | S-SF |
| PE | 9.59 | 5,82 |
| SD | 11.09 | 5,60 |

Le S-SF obtient de meilleures performances que le S-SNR pour la prédiction du WER. L'avantage du S-SNR est le faible nombre de paramètres et le temps de calcul réduit. Cependant, pour un tour de parole, la prédiction du S-SF n'est pas assez précise avec un score de PE de 8,75.

Concernant la prédiction du PER, le S-SF obtient là-aussi de meilleures performances que le S-SNR. La prédiction du PER est plus précise que la prédiction du WER, avec une amélioration relative de 33,4%. L'écart de performance provient de l'influence du modèle de langage qui peut rattraper des erreurs sur les enchaînements de mots. La durée des tours de parole varie de 3s à 16s avec une moyenne de 7s. En général les enregistrements soumis par les utilisateurs de services de transcription dépassent les 3 min, donc nous avons regardé l'influence du nombre de tours de parole pour voir a partir de quel moment la prédiction est la plus qualitative. Les regroupements ne mélangent pas les locuteurs.

Dans la figure 5 nous pouvons voir la performance de prédiction du WER et du PER de la méthode d'extraction S-SF en fonction du nombre de tours de parole utilisés (20 tours

2. `https://www.nist.gov/itl/iad/mig/tools`

de parole correspond à une durée d'environ 140 secondes). La prédiction est plus précise lorsque la fenêtre temporelle est plus large et l'écart de performance entre la prédiction du WER et du PER reste constant quelque soit le nombre de tours de parole utilisé. Lorsque 20 tours de parole sont utilisés, la performance de prédiction du WER atteint 5,82 de PE et la performance de prédiction du PER atteint 2,98 de PE.

Nous avons également observé la précision de la prédiction pour des bruits inconnus. Nous avons utilisé le sous-ensemble *eval93* comme corpus de parole et DEMAND [13] comme corpus de bruit. Trois types de bruit (restaurant, station et car) ont été mixés à la parole pour les même niveaux de SNR (de -15dB à 25dB). Pour le S-SNR, nous obtenons un score PE de 13,03 et un SD de 13,02 pour prédire le WER. Pour le S-SF, les scores de PE et de SD sont respectivement de 9,24 et 10,89. Lorsque les bruits sont inconnus, le S-SF reste plus performant que le S-SNR.

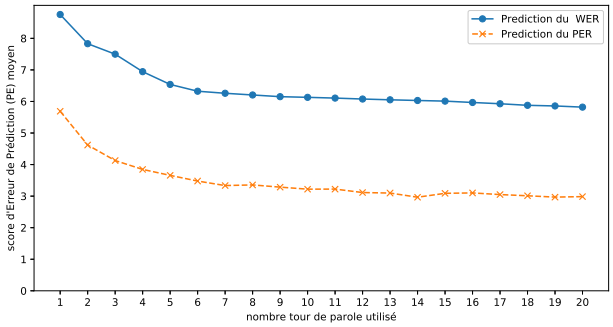


Figure 5 – Moyenne des erreurs de prédiction en fonction du nombre de tours de parole utilisé.

## 5 Conclusion

Notre étude cherche à évaluer précisément l'impact du bruit sur les système de RAP pour estimer *a priori* la qualité de la transcription. Cette estimation est calculée avant le décodage et sans connaître les extraits sonores utilisés pour entraîner le système de RAP. Nos travaux se sont concentrés uniquement sur les distorsions induites par le bruit afin d'analyser son impact sur la qualité de la transcription. Nous avons proposé une méthode d'extraction de paramètre, le S-SF, qui permet d'extraire des paramètres de la parole et du bruit séparé grâce a un masque binaire qui sont fortement corrélé à la qualité de la transcription. Lorsque la fenêtre temporelle est suffisamment large, l'erreur moyenne de prédiction du WER est de 5,82. Les meilleures performances que nous obtenons avec le PER (erreur moyenne de prédiction de 2,98) montrent bien que notre mesure est bien corrélé avec la réalité acoustique du signal traité. Notre système permet d'informer au plus tôt un utilisateur de la qualité de la transcription qu'il pourra possiblement obtenir. De plus nous avons observé que la prédiction du WER n'est que légèrement impacté (de 8,75 à 9,24) par des types de bruit inconnus, en testant notre prédiction avec 3 nouveaux types de bruits. Ces expériences sont à poursuivre en analysant d'autres dégradations acoustiques classiques telles que la réverbération et la superposition de parole.

## References


[1] J. Barcenilla and J. M. C. Bastien, "L'acceptabilité des nouvelles technologies: quelles relations avec l'ergonomie, l'utilisabilité et l'expérience utilisateur?," *Le travail humain*, vol. 72, no. 4, pp. 311–331, 2009.

[2] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech communication*, vol. 45, no. 4, pp. 455–470, 2005.

[3] S. Ghannay, Y. Esteve, and N. Camelin, "Word embeddings combination and neural networks for robustness in asr error detection," in *IEEE 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1671–1675, 2015.

[4] B. T. Meyer, S. H. Mallidi, H. Kayser, and H. Hermansky, "Predicting error rates for unknown data in automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5330–5334, 2017.

[5] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-wer," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 20–24, 2018.

[6] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: Predicting asr error from temporal properties of speech signal," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7423–7426, 2013.

[7] A. Varga and H. J. Steeneken, "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*, vol. 12, no. 3, pp. 247–251, 1993.

[8] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197, Springer, 2005.

[9] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer Series in Information Sciences, Springer-Verlag, 1990.

[10] G. John, et al., "CSR-I (WSJ0) complete LDC93S6A," DVD. Philadelphia: Linguistic Data Consortium, 1993.

[11] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-II (WSJ1) Complete," 1994.

[12] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks.," in *Interspeech*, vol. 2013, pp. 2345–2349, 2013.

[13] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust.*, 2013.