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

Analyse de l'effet de la réverbération sur la reconnaissance automatique 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

La Reconnaissance Automatique de la Parole (RAP) est moins performante lorsque le signal de parole est de mauvaise qualité. Dans cette étude, nous analysons les erreurs commises par les systèmes de RAP lorsque la parole transcrite est réverbérée afin de mieux comprendre les raisons de ces erreurs. Notre analyse permet de mettre en valeur les erreurs dues notamment à un mauvais alignement phonétique. Nous avons pu constater que les phonèmes de courte durée sont majoritairement supprimés lors du décodage phonétique. De plus, les phonèmes détectés, qu'ils soient corrects ou pas, ont tendance à avoir la même durée, ce qui est anormal pour certaines classes phonétiques comme les voyelles courtes ou les plosives. Nous avons aussi analysé les principales confusions entre les différentes classes phonétiques. Finalement, nous avons pu montrer que les erreurs lors de l'alignement phonétique des systèmes de transcription automatique entraînent beaucoup d'erreurs de détection.

ABSTRACT

Analyzing how reverberation affects Automatic Speech Recognition

Automatic Speech Recognition (ASR) is less effective when the speech signal is of poor quality. In this study, we analyze the errors made by ASR systems when the transcribed speech is reverberated in order to better understand the reasons for these errors. Our analysis allows us to highlight errors due to phonetic misalignment. We have found that short duration phonemes are mostly suppressed during phonetic decoding. Moreover, the detected phonemes, whether they are correct or not, tend to have the same duration. This is abnormal for certain phonetic classes such as short vowels or plosives. We also analyzed the main confusions between the different phonetic classes. We were able to show that errors in the phonetic alignment of automatic transcription systems lead to many detection errors.

MOTS-CLÉS : reconnaissance automatique de la parole, réverbération, analyse d'erreur.

KEYWORDS: automatic speech recognition, reverberation, error analysis.

1 Introduction

Au cours de la dernière décennie, les systèmes de Reconnaissance Automatique de la Parole (RAP) ont atteint de bonnes performances. Néanmoins, la robustesse de ces systèmes reste insatisfaisante par rapport aux humains (Kinoshita *et al.*, 2016), notamment lorsque le signal de parole est réverbéré (surtout pour les fichiers enregistrés avec un seul microphone). Pour la parole réverbérée, le comportement des systèmes de RAP semble être différent de celui des humains (Lippmann, 1996).

Volume 1 : Journées d'Études sur la Parole, pages 235–243.
HAL : hal-02798542.

Dans (Sehr *et al.*, 2010), nous pouvons apprendre que les 50 premières millisecondes de la réponse impulsionelle de la salle d'enregistrement (RIR pour Room Impulse Response) affectent peu les performances des systèmes de RAP, contrairement aux millisecondes suivantes (réverbération tardive). Dans (Junqua, 1997), les auteurs présentent une tentative de caractériser la sensibilité d'un dispositif de reconnaissance de phonèmes en fonction de la source de distorsion du canal. On peut voir que les grandes classes phonétiques ne sont pas affectées de la même manière par la réverbération. Dans (Parada *et al.*, 2014), les auteurs montrent la robustesse relative à la réverbération de chaque phonème, et proposent un modèle pour estimer la confusion de chaque phonème. La méthode utilise l'indice de clarté C50, qui est bien corrélé avec les performances de la RAP (Parada *et al.*, 2016).

Nous proposons d'analyser les erreurs des systèmes de RAP dues à la réverbération. Les résultats détaillés de la substitution des phonèmes et de la durée des phonèmes détectés (ou supprimés) permettront d'observer les raisons des mauvaises performances de la RAP pour la parole réverbérée. Plutôt que TIMIT (Fisher, 1986), nous avons choisi de travailler sur un autre corpus, le "Wall Street Journal" (WSJ) (Paul & Baker, 2003), avec une recette plus récente que les précédentes études sur le sujet pour créer notre système de RAP : utilisation de DNN (Deep Neural Network) et adaptation fMLLR (feature space Maximum Likelihood Linear Regression).

Dans la section 2, nous présentons le contexte de cette étude et nous rappelons le mécanisme de réverbération et ses différentes formes. Dans la section 3, nous exposons notre plan d'expérimentation, le matériel et les systèmes utilisés dans cette étude. Dans la section 4, nous présentons les résultats que nous discuterons dans la section 5.

2 Contexte

Supposons qu'un signal s qui est traité dans un environnement acoustique réaliste soit modélisé par :

$$y(t) = s(t) \circledast h(t) + n(t)$$

où t représente le temps, h la RIR et n le bruit de fond. Le RIR correspond à l'enregistrement d'un bruit impulsif (un clap) afin d'enregistrer les résonances. Dans ce modèle, nous pouvons remarquer que le bruit de fond est indépendant de la parole (distorsion additive), et la réverbération est fortement corrélée à la parole (distorsion convolutionnelle). La RIR décrit de manière précise la propriété de réverbération d'une pièce. La figure 1 présente un exemple schématique tiré de l'article (Valimaki *et al.*, 2012).

Le RIR est composé de trois parties :

- **Trajet direct** (« Direct Path ») : l'onde sonore est directement capturée par le microphone.
- **Réverbération précoce** (« Early Reverberation ») : les ondes sonores sont réfléchies une fois. La coloration spectrale du signal de parole est due à cette réverbération précoce. Cela n'affecte que très peu les performances de la RAP.
- **Réverbération tardive** (« Late Reverberation ») : les ondes sonores sont réfléchies plusieurs fois. Le flou temporel du signal de parole est dû à la réverbération tardive. Ceci affecte grandement les performances de la RAP.

La distorsion principalement gênante, provoquée par la réverbération, est le flou temporel. Le flou temporel provoque un chevauchement du phonème précédent sur le phonème actuel. Sur la figure 2, tirée de l'article (Petrick *et al.*, 2008), nous pouvons voir l'énergie des phonèmes qui se chevauchent. Dans cette illustration, **v** correspond au phonèmes voisés, **u** au phonèmes non-voisés et **r** a la

réverbération due au flou temporel. Maintenant la question est de savoir comment se comporte les systèmes de RAP losque les phonèmes se chevauchent ?

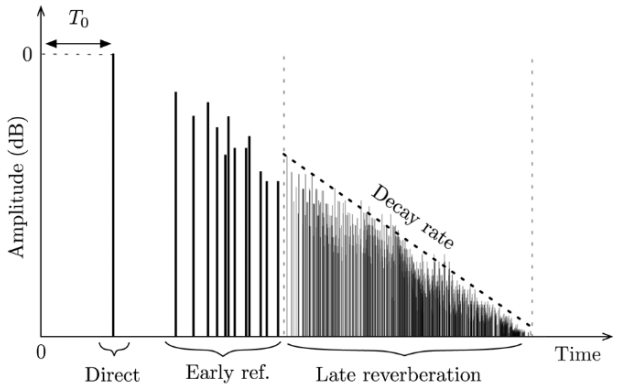

FIGURE 1 – Schéma d'une RIR générique, extrait de (Valimaki *et al.*, 2012).

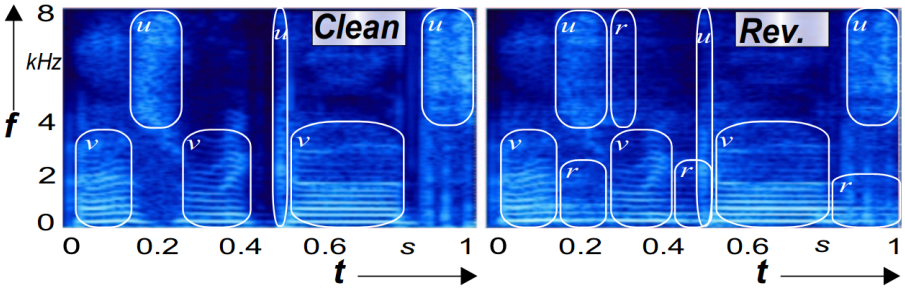

FIGURE 2 – Illustration des perturbations causées par la réverbération, extrait de (Petrick *et al.*, 2008).

3 Corpus et vérité terrain

Le corpus de parole utilisé vient du WSJ (Paul & Baker, 2003), et plus précisément le WSJ0 (Garofalo *et al.*, 1993) et le WSJ1 (Garofolo *et al.*, 1994). Les RIR provenant du REVERB Challenge (Kinoshita *et al.*, 2013) sont utilisées afin de réverbérer la parole artificiellement. Elles permettent de simuler 12 conditions différentes de réverbération : 6 pièces de tailles différentes (2 petites, 2 moyennes et 2 grandes) avec 2 types de distances au microphone (proche : 50 cm et lointaine : 200 cm). Les temps de réverbération T60 des petites, moyennes et grandes pièces sont respectivement d'environ 250, 500 et 700 millisecondes. Les salles de taille similaire ont été combinées, pour au final, obtenir 6 conditions réverbérées et une condition sans ajout de réverbération.

Pour créer le système de RAP, nous avons utilisé la recette Kaldi de Karel Vesely (Veselỳ *et al.*, 2013). Le système est hybride avec un réseau de neurones profond et des modèles de Markov caché (DNN-HMM) entraînés avec la cross-entropie sur le sous ensemble *train_si284* du WSJ (sans réverbération ajouté). Les données utilisées pour notre analyse sont composées des sous-ensembles *dev93* et *eval92* du WSJ qui ont été convolués par les différentes RIR pour obtenir 7 conditions de réverbération.

Afin d'analyser plus finement les erreurs commises par les systèmes de RAP, nous avons voulu nous détacher de l'influence du modèle de langage. Pour cela, nous utilisons un décodeur acoustico-phonétique pour prédire une suite de phonèmes (plutôt qu'une suite de mots). Pour concevoir ce

système, 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.

4 Résultats

4.1 Décodage par le système de RAP

Nous avons décodé les sous-ensembles *dev93* et *eval92* pour les 7 conditions décrites dans la section 3. Les WER (Word Error Rate) et les PER (Phone Error Rate), liés à la taille de la pièce et à la distance au microphone, sont indiqués dans le tableau 1. Le terme « propre » correspond au fichier originel (sans convolution avec une réponse impulsionnelle).

TABLE 1 – Résultats de WER et PER en fonction des différentes conditions de réverbération : moyenne, écart-type et pourcentages de substitution, insertion, délétion de phonèmes.

Taille salle	Propre	Petite		Moyenne		Grande	
Distance		proche	loin	proche	loin	proche	loin
WER en %							
Moyenne	4,9	6,9	12,9	16,5	52,1	20,1	78,4
Écart-type	8,1	9,2	13,7	15,8	24,4	16,8	17,5
PER en %							
Moyenne	9,8	14,1	24,2	28,5	50,6	31,7	63,8
Écart-type	5,6	6,7	8,9	9,2	8,9	9,5	7,4
Ratio des erreurs des phonèmes en %							
Substitution	61,2	59,8	60,8	60,7	56,2	61,1	53,8
Insertion	15,3	15,5	11,4	9,4	4,0	9,2	2,1
Délétion	23,5	24,9	27,8	29,9	39,8	29,7	44,1
C50							
Moyenne		42,20	20,29	16,43	6,84	13,34	5,91

Dans les même conditions, nous pouvons constater que l'écart-type du WER est plus importante que l'écart-type du PER. Nous remarquons aussi, que passé un certain seuil de PER (environ 50%), le modèle de langage a plus de difficultés à retrouver les mots correct : une fois les 50% de PER atteint le WER est supérieur au PER.

Nous voyons que deux facteurs impactent fortement les performances des systèmes de transcriptions :

— la **taille de la pièce**. Plus le T60 est important et plus l'énergie provenant de la réverbération est importante.

— la **distance au microphone**. Plus la distance au microphone augmente et plus l'énergie de parole provenant du trajet direct est atténué.

Ces deux facteurs ont une influence directe sur la mesure du C50, qui est fortement corrélée avec les performances des systèmes de RAP (Parada *et al.*, 2016).

Comme la réverbération est une distorsion acoustique, nous allons dans la suite de cet article observer uniquement les erreurs du décodage phonétique car il est difficile de comprendre les erreurs commises

par les systèmes de RAP sans dissocier le modèle de langage.

Nous avons aussi observé, parmi les erreurs commises par le décodeur acoustico-phonétique, les ratios d'erreur provenant des substitutions, des insertions et des délétions des phonèmes dont les résultats sont visibles dans le tableau 1. Nous pouvons observer que la contribution des insertions est plus faible lorsque la réverbération augmente. Par contre, le nombre de délétions augmente. Les substitutions restent relativement stables, sauf dans les conditions les plus réverbérées où elles diminuent. Dans tous les cas testés, les substitutions restent la cause principale d'erreurs.

4.2 Durée des phonèmes

Comme la réverbération provoque un flou temporel du signal, nous avons décidé d'observer la durée des phonèmes transcrit par le décodeur acoustico-phonétique. Pour obtenir la durée de chaque phonème, nous avons utilisé les résultats issus de Kaldi. Cela implique que la durée des phonèmes est obtenue automatiquement (sans annotation manuelle) : les durées des phonèmes de référence seront les durées obtenues sur le décodage des tours de parole n'ayant pas été réverbérés artificiellement (condition propre) [1]. Dans un souci de lisibilité, nous avons choisi de regrouper les phonèmes par classe phonétique pour l'affichage des résultats.

Nous avons observé la durée des phonèmes corrects et incorrects (substitution et délétion) dans des conditions réverbérées que nous comparons avec la durée des phonèmes dans des conditions propres (non-réverbérées) (voir figure 3).

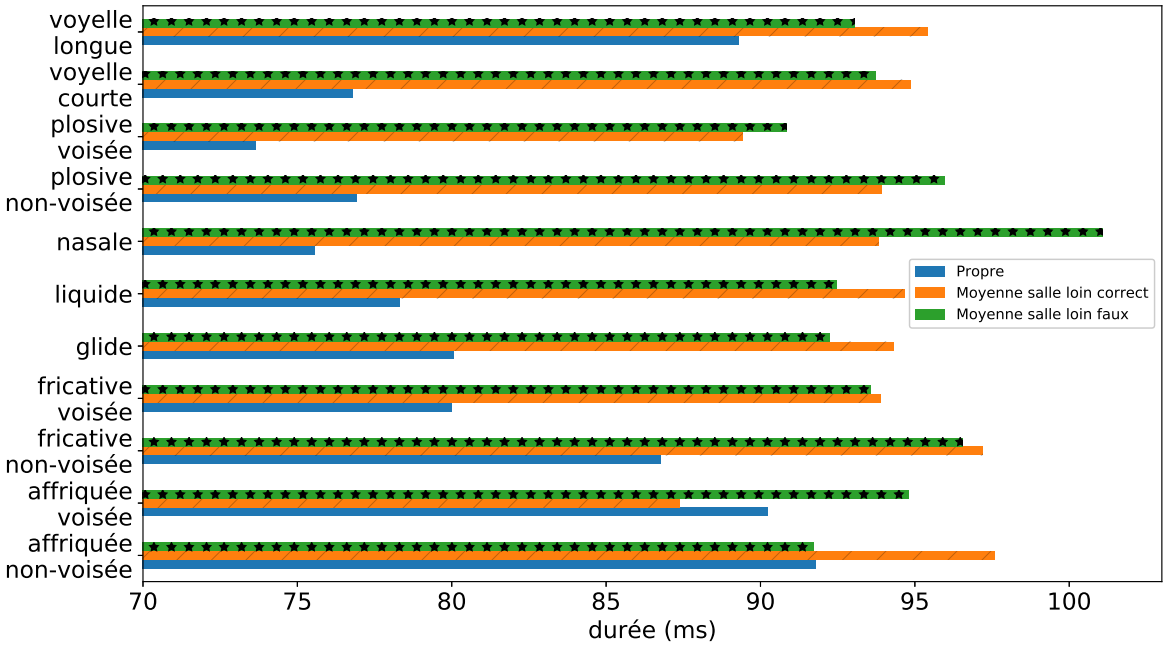

FIGURE 3 – Durée des phonèmes dans des conditions propre et réverbéré. Dans le cas réverbéré, les phonèmes correctement reconnus et les phonèmes erronés sont affichés séparément.

Sur la condition de réverbération salle moyenne et longue distance au microphone, la durée des

1. À noter que seuls les phonèmes correctement détectés dans des conditions propres sont pris en compte !

phonèmes réverbérés est globalement accrue pour attendre environ 95 ms en moyenne. Pour information, pour la condition *propre* la moyenne est de 82 ms, pour la condition *petite pièce lointaine* c'est 85 ms, et pour la condition *grande pièce lointaine* cela atteint 120 ms. Les phonèmes corrects ont globalement la même durée que les phonèmes substitués et insérés : hormis pour les consonnes affriquées et les nasales. En plus d'une augmentation de durée, nous avons aussi observé que les phonèmes tendent à avoir des durées similaires lorsque la réverbération s'accroît. Par exemple, la différence de durée entre voyelle courte et voyelle longue est clairement visible lorsque le fichier est non-réverbéré, mais lorsque le fichier est suffisamment réverbéré, les durées sont similaires.

Nous souhaitons maintenant analyser les phonèmes qui ont subit une délétion. Nous avons fait le lien entre les phonèmes supprimés dans des conditions réverbérées et leurs durées de référence (dans notre cas, cela correspond aux durées obtenues dans des conditions propres). Les résultats sont présentés dans le tableau 2.

TABLE 2 – Moyenne de la durée des phonèmes (condition propre) qui sont supprimés lorsque le fichier est réverbéré.

Taille salle	Petite		Moyenne		Grande	
Distance	proche	loin	proche	loin	proche	loin
Durée des délétion (ms)	58	61	66	68	67	71

Nous pouvons voir que les phonèmes qui sont supprimés ont en moyenne une durée moins importante. Ainsi, les phonèmes de faible durée ont plus de chance d'être supprimés lorsque la parole est réverbérée. Nous remarquons aussi que plus la réverbération augmente, et plus la durée moyenne des phonèmes supprimés augmente.

4.3 Résultats par classe phonétique

Commençons par observer le pourcentage de phonèmes correctement détectés en fonction de leur classe phonétique, dont les résultats sont affichés sur la figure 4a.

Nous avons choisi de montrer uniquement la condition de réverbération (moyenne salle, loin du microphone). Le comportement des résultats est similaire pour d'autres conditions de réverbération (plus la réverbération est importante et plus les résultats sont marqués). Sans réverbération, le système de transcription phonétique obtient des résultats similaires pour chaque phonème (environ 90% correct). Par contre, ce n'est pas le cas dans des conditions réverbérées. La classe phonétique des liquides est la moins impactée par la réverbération. Par contre, les consonnes affriquées voisées et les plosives sont les plus impactées. Les autres catégories sont moyennement impactées et obtiennent des résultats similaires. Nos résultats sont similaires à ceux trouvés dans cette étude (Junqua, 1997), ce qui montre que l'utilisation d'une recette de systèmes de RAP plus récente (DNN) ne modifie pas ces constats.

Afin d'identifier les erreurs, nous avons ensuite calculé une matrice de confusion entre classes phonétiques (figure 4b). Nous pouvons remarquer que les résultats sont très similaires [2].

Sur la figure 4a, l'impact des insertions et des substitutions entre même classe phonétique est pris

2. Hormis pour le cas des voyelles courtes (57% à 70%) qui s'explique par les insertions (25% des erreurs).

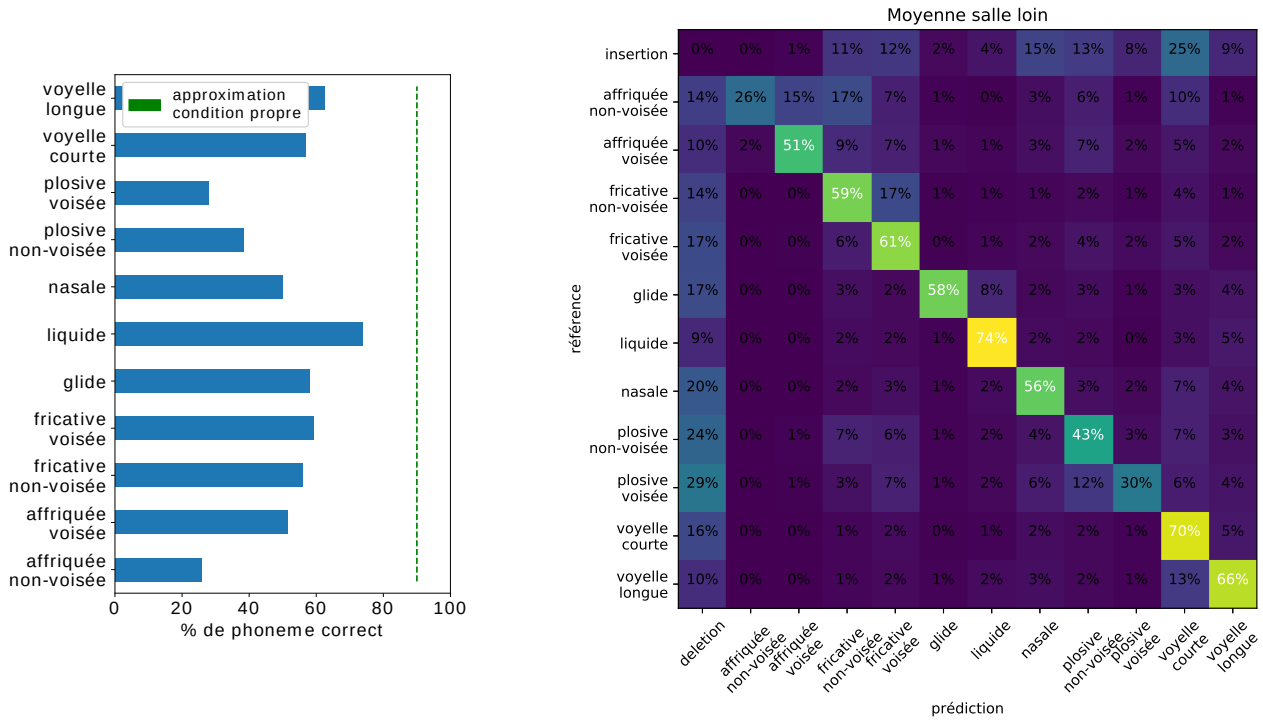

FIGURE 4 – (a) (partie à gauche) Pourcentage de phonèmes correctement détectés par classe phonétique (b) (partie de droite) et matrice de confusion dans une condition de réverbération forte (salle de taille moyenne, distance lointaine au micro).

en compte, ce qui n'est pas le cas lorsque nous observons la diagonale de la matrice de confusion (figure 4b). En effet, si nous observons les confusions entre phonèmes (et non au niveau des classes phonétiques), les substitutions sont majoritairement effectuées entre des classes phonétiques différentes (sauf pour les nasales).

Nous pouvons voir que les erreurs de détection proviennent principalement des suppressions, à l'exception des affriquées non-voisées, des fricatives non-voisées et des voyelles longues. Le nombre de délétions est toujours moins important que le nombre total de substitutions. Certaines classes phonétiques sont plus souvent supprimées que d'autres : les plosives et les nasales en particulier. Concernant les substitutions. Les fricatives non-voisées sont dans 17% des cas substituées par des fricatives voisées. La superposition avec le phonème précédent, provoquée par la réverbération, peut ajouter du voisement aux fricatives non-voisées. Le même effet se constate sur les affriquées, ou les non-voisées sont dans 17% des cas substituées par des voisées. Les affriquées non-voisées sont dans 17% des cas substituées par des fricatives non-voisées. Les affriquées partagent des caractéristiques communes aux plosives et aux fricatives. Le flou temporel provoqué par la réverbération rend plus similaire les affriquées aux fricatives car la composante plosive est fortement détériorée (étirée). Nous remarquons aussi que les plosives sont supprimées dans plus de 25% des cas. Les plosives se caractérisent par une phase de silence suivie par une impulsion. Lorsque que l'énergie provenant de la réverbération d'un précédent phonème se superpose aux plosives, il est plus bien plus difficile de détecté ce court silence. Concernant la substitution de 12% de plosives voisées en plosives non-voisées, l'effet inverse était plus attendu (de non-voisé à voisé).

Les autres substitutions remarquables, comme les plosives voisées qui sont substituées par des plosives non-voisées et les voyelles longues qui sont substituées par des voyelles courtes, sont plus difficilement explicables.

5 Discussion

La reconnaissance automatique de phonèmes dans la chaîne de traitement Kaldi est composée de deux étapes : l'alignement puis la classification. Le niveau de réverbération a un impact sur la précision de l'alignement des phonèmes. Nous pouvons le constater grâce à ces différentes observations :

— un allongement de la durée moyenne des phonèmes (82 ms pour la condition propre à 120 ms pour la condition la plus réverbérée),
— une uniformisation de la durées des phonèmes (voir figure 2),
— le ratio du nombre de délétions augmente (23,5% en condition propre et 44.1% dans la condition la plus réverbérée),
— la moyenne des durées des délétions de phonème augmente (de 58 ms pour une réverbération très faible à 71 ms pour la réverbération la plus forte),

Comme les phonèmes détectés sont de plus en plus long, de moins en moins de phonèmes peuvent être détectés. De plus, la réverbération cause de nombreuse erreurs de substitution, suite au chevauchement des phonèmes. Cela s'observe principalement par les nombreuses substitutions entre les phonèmes non-voisés et ceux voisés.

La distorsion du signal de parole provoquée par la réverbération est avant tout une superposition. L'effet d'un point de vue acoustique est similaire à de la conversation superposée. Les systèmes de transcription automatique de la parole sont conçus pour n'attribuer qu'un seul label phonétique par unités de temps. Or, dans des conditions réverbérées, le systèmes de RAP doit souvent choisir entre deux labels. Pour l'alignement, cela rend la détection de la transition entre deux phonèmes beaucoup plus difficile. Nous pensons que les systèmes automatiques favorisent l'allongement du phonème précédent plutôt que la transition vers le phonème suivant. Le début de certains phonèmes disparaît à cause de la superposition avec la réverbération du phonème précédent. Il est ainsi plus difficile de reconnaître ces phonèmes. Enfin, les phonèmes ayant une durée plus courte sont en général supprimés lorsqu'ils sont intégralement recouvert par la réverbération du phonème précédent : ce qui explique de nombreuses observations effectuées dans ce papier.

6 Conclusions

Dans cet article, nous avons analysé les erreurs de détection commises par les systèmes de RAP dues à la réverbération. Concernant la confusion entre phonèmes, nous avons obtenu des résultats similaires aux études précédentes, même avec une recette de RAP plus récente, utilisant des réseaux de neurones profond. De plus, nous avons montré les limites de l'alignement phonétique des systèmes de RAP actuels dans le cas de la parole réverbéré. Enfin, nous avons étudié les confusions en classes phonétiques, en expliquant lorsque cela a été possible les erreurs de substitution.

Parmi les perspectives de nos travaux, nous pensons que la détection d'anomalie de la durée des phonèmes (l'uniformisation de la durée par exemple) pourrait servir aux travaux de prédiction de la performance des systèmes de RAP, par exemple en utilisant les postériogrammes des phonèmes comme (Meyer *et al.*, 2017). Il serait également intéressant de comparer les erreurs phonétiques entre un système avec un alignement supervisé et un autre non supervisé. Il serait également pertinent de tester d'autres méthodes d'alignement comme les CTC (Connectionist Temporal Classification), afin de savoir si les erreurs d'alignement sont similaires à ce que nous avons observé. Enfin, nous pourrions observer la robustesse des systèmes de RAP appris sur des données réverbérées.

Références

FISHER W. M. (1986). The DARPA speech recognition research database : specifications and status. In *Proc. DARPA Workshop on Speech Recognition, Feb. 1986*, p. 93–99.

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.

JUNQUA J.-C. (1997). Impact of the unknown communication channel on automatic speech recognition : A review. In *Fifth European Conference on Speech Communication and Technology*.

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.

LIPPMANN R. (1996). Speech perception by humans and machines. In *Proceedings of the Workshop on the Auditory Basis of Speech Perception*, p. 309–316.

MEYER B. T., MALLIDI S. H., KAYSER H. & HERMANSKY H. (2017). Predicting error rates for unknown data in automatic speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5330–5334 : IEEE.

PARADA P. P., SHARMA D., LAINEZ J., BARREDA D., VAN WATERSCHOOT T. & NAYLOR P. A. (2016). A single-channel non-intrusive C50 estimator correlated with speech recognition performance. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, **24**(4), 719–732.

PARADA P. P., SHARMA D., NAYLOR P. A. & VAN WATERSCHOOT T. (2014). Reverberant speech recognition : A phoneme analysis. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, p. 567–571 : IEEE.

PAUL D. B. & BAKER J. M. (2003). The design for the Wall Street Journal-based CSR corpus. In *Proceedings Title*, volume II, p. 803–806 : IEEE.

PETRICK R., LOHDE K., LORENZ M. & HOFFMANN R. (2008). A new feature analysis method for robust asr in reverberant environments based on the harmonic structure of speech. In *Signal Processing Conference, 2008 16th European*, p. 1–5 : IEEE.

SEHR A., HABETS E. A., MAAS R. & KELLERMANN W. (2010). Towards a better understanding of the effect of reverberation on speech recognition performance. In *Proc. IWAENC*.

VALIMAKI V., PARKER J. D., SAVIOJA L., SMITH J. O. & ABEL J. S. (2012). Fifty years of artificial reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(5), 1421–1448.

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