Anonymisation de parole par quantification vectorielle

Pierre Champion, Denis Jouvet, Anthony Larcher

Anonymisation de parole par quantification vectorielle

Pierre Champion[1,2] Denis Jouvet[1] Anthony Larcher[2]
(1) Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
(2) LIUM, Le Mans Université, Avenue Olivier Messiaen, 72085 LE MANS CEDEX 9, France
{pierre.champion, denis.jouvet}@inria.fr, anthony.larcher@univ-lemans.fr

RÉSUMÉ

L'utilisation de la reconnaissance de parole se répand de plus en plus dans les assistants virtuels. Cependant, les signaux de parole contiennent de nombreuses informations sensibles telles que l'identité du locuteur, ce qui soulève des préoccupations quant à la protection des données personnelles. Les expériences présentées montrent que les représentations extraites par les couches profondes des réseaux de reconnaissance de la parole contiennent cette information. Dans cet article, nous cherchons à produire une représentation anonyme tout en préservant les performances de reconnaissance de la parole. Dans ce but, nous proposons d'utiliser la quantification vectorielle pour contraindre l'espace de représentation, et inciter le réseau à supprimer l'identité du locuteur. Le choix de la taille du dictionnaire de quantification permet d'ajuster le compromis entre l'utilité (reconnaissance de la parole) et le respect de la vie privée (masquage de l'identité du locuteur).

ABSTRACT

Privacy-Preserving Speech Representation Learning using Vector Quantization

With the popularity of virtual assistants (e.g., Siri, Alexa), the use of speech recognition is now becoming more and more widespread. However, speech signals contain a lot of sensitive information, such as the speaker's identity, which raises privacy concerns. The presented experiments show that the representations extracted by the deep layers of speech recognition networks contain speaker information. This paper aims to produce an anonymous representation while preserving speech recognition performance. To this end, we propose to use vector quantization to constrain the representation space and induce the network to suppress the speaker identity. The choice of the quantization dictionary size allows to configure the trade-off between utility (speech recognition) and privacy (speaker identity concealment).

MOTS-CLÉS : Anonymisation de la parole, Assistants vocaux, Reconnaissance du locuteur, Reconnaissance de parole.

KEYWORDS: Speech Anonymization, Voice Assistants, Speaker Recognition, Speech Recognition.

1 Introduction

Avec l'essor des assistants vocaux, de plus en plus d'objets connectés sont déployés chez les consommateurs. Ces assistants ont besoin d'une connexion internet et de serveurs centralisés pour fonctionner. Les signaux de parole de l'utilisateur sont envoyés à ces serveurs pour bénéficier d'une expérience confortable et accessible en permanence. Sur les serveurs, les fournisseurs de service font appel à des systèmes de reconnaissance automatique de parole et de compréhension du langage naturel afin de répondre à la demande de l'utilisateur. Cependant, les signaux de parole contiennent de nombreuses informations relatives au locuteur, on y retrouve des attributs sensibles comme le genre du locuteur,

son l'identité, son âge, ses sentiments, ses émotions, etc. Ces attributs sensibles peuvent être extraits et utilisés à des fins malveillantes. Cette collecte excessive, et sans précédent, de signaux de parole sert à établir des profils d'utilisateurs complets et à construire de très grands jeux de données, nécessaires pour enrichir et améliorer les modèles de reconnaissance et de compréhension. Ce transfert global des données vers les fournisseurs de services soulève de sérieuses questions à propos de la protection de la vie privée. Récemment, des systèmes de reconnaissance de parole embarqués ont été proposés affin de résoudre cette problématique. Cependant, les performances de ces systèmes sont encore restreintes dans les environnements peu favorables (c'est-à-dire, environnements bruyants, parole réverbérée, forts accents, etc.). La collecte de grands corpus de parole représentatifs des utilisateurs réels et des diverses conditions d'utilisation est nécessaire pour améliorer les performances. Mais cela doit être effectué tout en préservant la vie privée des utilisateurs, ce qui signifie au moins garder l'identité du locuteur privée.

Dans l'approche proposée, un encodeur réside sur chaque objet connecté et effectue des calculs locaux pour créer une représentation anonymisée de la parole. Ce processus de calcul défini par (Osia *et al.*, 2020) est adapté pour les assistants vocaux. Jusqu'à présent, les travaux suivants s'y sont inscrits : dans (Srivastava *et al.*, 2019), les auteurs emploient une méthode d'apprentissage antagoniste pour supprimer l'identité du locuteur dans un réseau de reconnaissance de la parole. Cependant, leur approche a eu un fable impact, le système de vérification locuteur n'a pas vu ses performances significativement dégradées. Dans (Koppelmann *et al.*, 2021), les auteurs cherchent à créer une représentation capable de détecter des mots de réveil sans être capable de décoder le contenu linguistique. Dans (Aloufi *et al.*, 2021), les auteurs ont étudié la discrétisation de la parole dans de multiples systèmes de reconnaissance de la parole afin de minimiser l'inférence de plusieurs attributs sensibles (comme le locuteur, l'émotion, le genre). Finalement, dans le Challenge Voice Privacy (VPC) 2020 (Tomashenko *et al.*, 2020), un protocole dédié et des métriques ont été proposés afin d'évaluer différentes méthodes d'anonymisation du locateur.

Dans cet article, notre travail est similaire à ceux de (Srivastava *et al.*, 2019; Aloufi *et al.*, 2021) ou nous nous concentrons sur la création d'une représentation anonymisée, où l'objectif est d'envoyer au fournisseur de services uniquement les informations qui lui sont nécessaires pour un bon fonctionnement du service. Dans le cas considéré des assistants vocaux, l'information relative au contenu linguistique doit être gardée alors que celle relative aux locuteurs doit être supprimée. L'encodeur effectuant l'anonymisation étudiée dans cet article est basé sur un système de reconnaissance de la parole. La représentation est extraite au niveau de la couche *bottleneck* du réseau. Ce type de représentation a pour but de compresser l'information afin qu'elle soit efficiente. Dans le cas d'un système de reconnaissance de la parole, les *bottlenecks* sont supposés encoder l'information du contenu linguistique, et ce, en étant invariants aux locuteurs. En utilisant le protocole d'évaluation du VPC, nous avons observé que les *bottlenecks* n'encodent pas uniquement l'information linguistique, le locuteur peut être lui aussi identifié à un degré élevé. Afin de mieux supprimer l'information relative au locuteur (donc améliorer l'anonymisation), nous avons introduit l'utilisation de la quantification vectorielle au niveau de la couche *bottleneck* du réseau de reconnaissance de la parole. La quantification vectorielle consiste en l'approximation d'un vecteur continu par un autre vecteur de même dimension, mais ce dernier appartenant à un ensemble fini de vecteurs (Gersho & Gray, 1992). La quantification vectorielle est fréquemment utilisée dans la compression de données avec pertes. Dans notre cadre d'utilisation, la quantification vectorielle permet d'imposer une contrainte sur le la couche *bottleneck*. Cette contrainte incite le réseau de reconnaissance de parole à encoder l'information du contenu linguistique dans un ensemble fini de vecteurs. De ce fait, les autres informations relatives au locuteur se retrouvent moins encodées par manque de capacité d'encodage. Nos contributions sont les

suivantes. Premièrement, nous évaluons à quelle hauteur l'information du locuteur est présente dans un *bottleneck* d'un système de reconnaissance de la parole. Deuxièmement, nous étudions l'impact que la quantification vectorielle a sur les performances de reconnaissance de parole et du locuteur. Troisièmement, nous montrons que les *bottlenecks* peuvent être utilisés pour générer un signal de parole audible permettant une potentielle annotation et réapprentissage du modèle de reconnaissance de la parole.

La structure du reste du document est la suivante. Dans la section 2, nous décrivons le cadre de travail et le modèle proposé pour anonymiser la parole. La section 3 explique le dispositif expérimental et présente nos résultats. Enfin, nous concluons et discutons des travaux futurs dans la section 4.

2 Processus de calcul hybride avec des calculs locaux et mutualisés

Dans cette section, nous présentons le cadre de travail hybride proposé par (Osia *et al.*, 2020) permettant d'effectuer des calculs locaux et d'autres mutualisés tout en respectant la vie privée des utilisateurs. L'objectif de ce processus de calcul est de partager une représentation de parole avec un fournisseur de service, mais ce, en anonymisant les données de parole au niveau du périphérique avant de les partager. Dans le contexte des assistants vocaux, la représentation anonymisée doit être riche en information relative au contenu linguistique tout en empêchant l'exposition d'informations sensibles qui pourrait potentiellement révéler des informations privées de l'utilisateur. Dans nos expériences, nous nous focalisons sur l'identité du locuteur, et considérons que cette information doit être supprimée. Dans ce processus de calcul hybride, la tâche compliquée est de concevoir l'encodeur qui extrait la représentation anonymisée, car le codage ou la modification du signal de parole peut nuire au bon fonctionnement de la tâche de reconnaissance de la parole. Dans la section suivante, nous décrivons l'architecture de l'encodeur utilisé pour anonymiser la parole.

2.1 Présentation du modèle

De par leurs fonctions d'apprentissage, les modèles acoustiques utilisés dans les systèmes de reconnaissance de parole cherchent à encoder l'information du contenu linguistique (par exemple via la classification temporelle des phonèmes). Ces modèles sont souvent conçus pour être invariants au locuteur dans le but de proposer les mêmes performances de reconnaissance à tout utilisateur. C'est pour ces raisons que nous avons choisi d'utiliser comme encodeur un modèle acoustique.

Nous utilisons une architecture *time delay neural network factorized* (*TDNN-F*) introduite par (Povey *et al.*, 2018). Elle est utilisée dans le cadre d'un système de reconnaissance de la parole hybride *Hidden Markov Model - Deep Neural Network (HMM-DNN)* (Povey *et al.*, 2011). Cette architecture a été reconnue comme l'une des plus efficientes dans un récent classement comparant les performances des modèles par rapport aux exigences matérielles (Georgescu *et al.*, 2021). L'architecture *TDNN-F* est donc appropriée pour l'utilisation embarquée nécessaire au fonctionnement du processus hybride avec des calculs locaux et d'autres effectuées par un serveur centraliser.

La fonction d'objectif *Lattice-Free Maximum Mutual Information (LF-MMI)* (Hadian *et al.*, 2018) est utilisée afin de réaliser un entraînement discriminatif des séquences. La fonction *MMI* traditionnelle vise à maximiser la probabilité postérieure :

$$\mathcal{L}_{mmi}(\lambda) = \sum_{r=1}^{R} \log P_\lambda\left(S_r \mid O_r\right) = \sum_{r=1}^{R} \log \frac{P_\lambda\left(O_r \mid S_r\right) P\left(S_r\right)}{\sum_S P_\lambda\left(O_r \mid S\right) P\left(S\right)} \tag{1}$$

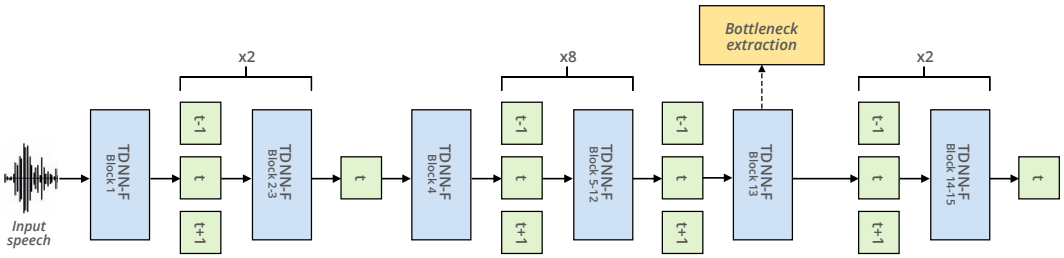

FIGURE 1 – Architecture du modèle *TDNN-F*, totalisant 15 couches. Les *bottleneck* sont extraits à partir de la 13e couche.

où λ est l'ensemble des paramètres du réseau de neurones, R est le nombre total de segments d'apprentissage, S_r est la transcription correcte du r^{eme} segment de parole O_r, $P(S)$ est la probabilité du modèle de langage pour la phrase S. La distribution $P(S)$ est considérée comme fixe, et est estimée avec un modèle de langage à partir des transcriptions d'entraînement. Le numérateur indique la vraisemblance de la prédiction pour une séquence de mots de référence, tandis que le dénominateur indique la vraisemblance totale de la prédiction pour toutes les séquences de mots possibles, ce qui équivaut à la somme sur toutes les séquences de mots possibles estimées par le modèle acoustique et le modèle de langage. Le numérateur encode les caractéristiques de supervision et il est spécifique à chaque segment, tandis que le dénominateur encode toutes les séquences de mots possibles et il est identique pour tous les segments. Cette fonction de coût est optimisée en maximisant le numérateur et en minimisant le dénominateur. *MMI* maximise la log-vraisemblance conditionnelle des probabilités globalement normalisées des transcriptions correctes.

Dans l'objectif d'obtenir une représentation anonymisée de la parole, nous extrayons des *bottlenecks* de faibles dimensions (D = 256 dimensions) depuis une couche profonde du réseau (la 13ème couche sur les 15 du réseau, cf. : figure 1). Il a été observé par (Adi *et al.*, 2019; Srivastava *et al.*, 2019; Li *et al.*, 2020) que ce type de représentation encode principalement l'information relative au contenu linguistique et supprime une partie de l'information de l'identité locuteur.

2.2 Introduction à la quantification vectorielle pour l'anonymisation

Afin d'améliorer l'anonymisation, nous proposons de contraindre la couche du réseau de neurones produisant les *bottlenecks* en ajoutant une couche de quantification vectorielle. La quantification vectorielle consiste en l'approximation d'un vecteur continu par un autre vecteur de même dimension, mais ce dernier appartenant à un ensemble fini de vecteurs (Gersho & Gray, 1992), ces vecteurs sont dénommés vecteurs prototypes. Dans la tâche d'apprentissage non supervisé de représentation discriminante via l'utilisation d'auto-encodeur, il a été observé que les vecteurs prototypes appris suite à une quantification vectorielle représentent principalement l'information relative aux phonèmes (van den Oord *et al.*, 2017; Chorowski *et al.*, 2019; Wu & Lee, 2020).

L'application de la quantification vectorielle dans un modèle acoustique a pour but d'inciter le modèle à supprimer l'information du locuteur, car la quantification vectorielle réduit la capacité d'encodage du réseau. Comparée aux tâches non supervisées, la fonction de coût d'un modèle acoustique impose explicitement que l'information phonétique soit encodée dans le *bottleneck*, de ce fait, nous pouvons appliquer une contrainte élevée en réduisant le nombre de vecteurs prototypes.

Étant donné la séquence audio d'entrée $s = (s_1, s_2, \ldots, s_T)$ de longueur T, l'encodeur *TDNN-F* produit les *bottlenecks* $h(s) = (h_1, h_2, \ldots, h_j)$ de longueur J ($J < T$ dû au sous-échantillonnage effectué par l'encodeur) où $h_j \in \mathbb{R}^D$ pour chaque pas temporel t, et D est la dimensionnalité de la représentation latente.

La couche de quantification vectorielle prend en entrée la séquence de vecteurs continus $h(s)$ et remplace chaque $h_j \in h(s)$ par un prototype du dictionnaire apprenable $E = \{e_1, e_2, \ldots, e_V\}$ de taille V, chaque $e_i \in \mathbb{R}^D$.

$$q(s) = \arg\min_{e_i} \|h(s) - e_i\|_2^2 \tag{2}$$

Le vecteur h_j est remplacé par son vecteur prototype e_v le plus proche en termes de distance euclidienne. Puisque la quantification est non différentiable (à cause de l'opération $\arg\min$), sa dérivée doit être approximée. Pour ce faire, nous utilisons un *straight-through estimator* (Bengio *et al.*, 2013) i.e., $\frac{\partial \mathcal{L}}{\partial h(s)} \approx \frac{\partial \mathcal{L}}{\partial q(s)}$. Les vecteurs prototype sont contraints de se rapprocher des vecteurs *bottlenecks* qu'ils remplacent par l'ajout d'une fonction de coût auxiliaire :

$$\mathcal{L}_{vq} = \|\mathrm{sg}\,[h(s)] - q(s)\|_2^2 \tag{3}$$

où $\mathrm{sg}[\cdot]$ désigne l'opération bloquant la rétropropagation du gradient, donc la mise à jour des poids. Cette opération est semblable à un k-means, mais appliquée à chaque minibatch pendant l'apprentissage, les prototypes du dictionnaire correspondant aux centroïdes d'un k-means. Étant donné que les *bottleneck* peuvent prendre n'importe quelle valeur, l'ajout d'une fonction de coût régularise l'encodeur a produire des *bottlenecks* proches des prototypes afin que l'apprentissage de l'encodeur ne diverge pas de l'apprentissage du dictionnaire :

$$\mathcal{L}_{vq\_reg} = \|h(x) - \mathrm{sg}[q(x)]\|_2^2 \tag{4}$$

La fonction de coût du modèle acoustique peut être alors exprimée par la somme des fonctions *mmi*, de quantification et de régularisation :

$$\mathcal{L} = \mathcal{L}_{mmi} + \mathcal{L}_{vq} + \lambda \mathcal{L}_{vq\_reg} \tag{5}$$

où λ désigne le coefficient du facteur de régularisation (nous avons utilisé $\lambda = 0.25$). Pour mettre à jour les prototypes du dictionnaire, nous utilisons une moyenne mobile exponentielle (EMA) (Łukasz Kaiser *et al.*, 2018). EMA met à jour le dictionnaire E indépendamment de l'optimiseur, l'apprentissage est donc plus robuste face aux différents choix d'optimiseurs et d'hyperparamètres (par example : le taux d'apprentissage, momentum).

3 Expériences

3.1 Jeux de données

Nous avons utilisé le corpus LibriSpeech (Panayotov *et al.*, 2015) pour toutes nos expériences. Les statistiques des jeux de données sont disponibles dans le tableau 1.

Le sous-ensemble LibriSpeech train-clean-100 a été utilisé pour apprendre le modèle acoustique. Les jeux de données utilisées pour évaluer les performances de reconnaissance de parole sont LibriSpeech test-clean et LibriSpeech test-other.

Le challenge Voice Privacy définit LibriSpeech train-clean-360 comme jeux d'apprentissage pour apprendre le système de vérification du locuteur. Il est important de remarquer que ce jeu d'apprentissage ne propose pas une grande variabilité intralocuteur du aux longues sessions d'enregistrement

TABLE 1 – Statistiques des jeux de données d'entraînement et de test.

	Taille	Nombre de locuteurs			Nombre d'utterances
		Femme	Homme	Total	
LibriSpeech : train-clean-100	100h	125	126	251	28539
LibriSpeech : train-clean-360	364h	439	482	921	104014
LibriSpeech : test-clean	5.4h	20	20	40	2620
LibriSpeech : test-other	5.1h	17	16	33	2939

des chapitres des livres audio. Entraîner le système de vérification du locuteur sur les *bottlenecks* d'un modèle acoustique n'est pas une tâche facile, car toute erreur de représentation effectuée par le modèle acoustique est propagée dans celui de vérification du locuteur. Pour atténuer cet effet, nous avons appris le système de reconnaissance du locuteur sur la combinaison de train-clean-100 et train-clean-360. Le modèle acoustique produit une très bonne représentation pour le sous-ensemble train-clean-100 (vu lors de l'entraînement du modèle acoustique), ce qui aide l'apprentissage du modèle de vérification du locuteur.

Conformément au challenge Voice Privacy, les performances en reconnaissance du locuteur ont été évaluées avec le jeu de donnée LibriSpeech test-clean. Parmi les 40 locuteurs de LibriSpeech test-clean, 29 d'entre eux sont sélectionnés, pour chaque locuteur un sous-ensemble totalisant 1 min de parole (après détection d'activité vocale) a été sélectionné pour l'ensemble d'enrôlement et le reste a été utilisé pour l'ensemble de test. Les nombres de test cible et imposteur sont détaillés dans le tableau 2.

TABLE 2 – Nombre de test de vérification dans l'ensemble de donnée d'évaluation.

	Type	Femme	Homme	Total
Librispeech : test-clean	Cible	548	449	997
	Imposteur	11196	9457	20653

3.2 Métriques et évaluation

Pour évaluer les performances du système en matière d'anonymisation (*capacité de dissimulation de l'identité du locuteur*) et d'utilité (*capacité à reconnaître le contenu linguistique*), deux systèmes et métriques sont utilisés.

Pour évaluer quantitativement la qualité de l'anonymisation, une architecture de vérification automatique du locuteur implémentée dans SideKit (Larcher *et al.*, 2016) est utilisée. Il s'agit d'un système x-vecteur composé de cinq couches TDNN suivit d'une couche de *statistics pooling* (Snyder *et al.*, 2018). La fonction de coût utilisée pour l'apprentissage est la *large margin softmax loss* (Liu *et al.*, 2019). La métrique d'évaluation est le taux d'égale erreur ($\text{EER}_{\%}$), plus l'$\text{EER}_{\%}$ est élevé, mieux les locuteurs sont anonymisés.

Pour l'utilité, le système de reconnaissance de parole transcrit la parole depuis la représentation *bottleneck*. La mesure du taux d'erreurs mots ($\text{WER}_{\%}$) est utilisée pour évaluer dans quelle mesure les *bottlenecks* encodent correctement l'information linguistique. Plus le $\text{WER}_{\%}$ est faible, mieux le contenu linguistique est encodé.

TABLE 3 – Résultats de la reconnaissance vocale et de la vérification du locuteur en fonction du nombre de vecteurs prototypes dans le dictionnaire de quantification. La mesure de l'intervalle de confiance pour l'EER et le WER est effectuée avec un ré-échantillonnage *bootstrap*.

Nb vecteurs prototypes	EER$_{\%}$		WER$_{\%}$	
	F	H	test-clean	test-other
(No VQ)	9.3 ±0.5	4.2 ±1.0	5.8 ±0.3	19.5 ±0.6
16	30.0 ±2.1	32.4 ±2.1	15.9 ±0.5	42.5 ±0.8
32	25.6 ±2.1	27.3 ±1.9	9.8 ±0.4	31.4 ±0.8
48	22.0 ±1.7	22.6 ±2.1	8.7 ±0.4	28.8 ±0.8
128	22.0 ±1.8	22.8 ±2.0	8.5 ±0.4	28.5 ±0.8
256	19.2 ±1.6	19.6 ±2.0	7.6 ±0.3	26.1 ±0.7
512	19.6 ±1.6	19.2 ±2.0	7.6 ±0.3	25.4 ±0.7
1024	17.9 ±1.6	18.3 ±1.8	7.2 ±0.3	24.7 ±0.7

3.3 Résultats et discussions

Le tableau 3 présente les résultats expérimentaux. La première ligne présente les scores de vérification du locuteur et de reconnaissance de parole pour les jeux de données test-clean et test-other, sans introduction de quantification vectorielle. Ces scores sont cohérents avec ceux reportés dans la littérature (Tomashenko *et al.*, 2020; Madikeri *et al.*, 2020). Les résultats de vérification du locuteur montrent que la représentation *bottleneck* d'un système de reconnaissance de parole de référence (No VQ) est capable de correctement discriminer les locuteurs. Il est à noter que pour cet exemple les femmes sont plus difficiles à différencier que les hommes, c'est-à-dire : 9,3 EER$_{\%}$ pour les femmes et 4,2 EER$_{\%}$ pour les hommes. Le WER$_{\%}$ sur LibriSpeech test-clean est de 5,8, valeur utilisée comme référence pour les expériences suivantes.

En contraignant la représentation *bottleneck* du réseau avec l'utilisation de quantification vectorielle, les performances de vérification du locuteur sont drastiquement réduites. Le nombre V de prototypes dans le dictionnaire de quantification contraint plus ou moins le modèle acoustique, avec V vecteurs prototypes l'information linguistique du signal est compressée dans un espace vectoriel discret de V vecteurs prototypes. Plus le dictionnaire est petit, plus le réseau doit trouver une transformation efficace pour représenter l'information linguistique, ce qui laisse moins de place pour encoder l'information relative au locuteur. Ainsi avec V = 16 les scores d'EER$_{\%}$ de vérification du locuteur sont de 30,0 pour les femmes et 32,4 pour les hommes valeurs plus élevées qu'avec V = 1024 ou le réseau obtient 17,9 pour les femmes et 18,3 pour les hommes. En comparaison avec le système de référence (No VQ), l'utilisation de la quantification vectorielle permet d'anonymiser les *bottlenecks*. Plus la valeur de V est petite, moins les *bottlenecks* sont représentatifs du locuteur, permettant une meilleure anonymisation.

Cependant, les performances de reconnaissance de parole, mesurées en termes de WER$_{\%}$, sont elles aussi impactées par la taille du dictionnaire de quantification. Avec V = 16 le WER$_{\%}$ est de 15,9 sur LibriSpeech test-clean, dégradation très importante par rapport à la valeur de référence de 5,8. En augmentant le nombre de vecteurs prototype, le WER$_{\%}$ redescend. Pour V = 1024 le WER$_{\%}$ est de 7,2. Le tableau 3 présente aussi les scores de reconnaissance de parole sur le jeu de données LibriSpeech test-other.

Le compromis entre de bonnes performances en reconnaissance de parole et une bonne anonymisation

est inhérent au problème de partage de données anonymisées (problème connu sous le nom de "*privacy-utility tradeoff*" (Li & Li, 2009)). Dans notre cadre de travail, ce compromis est paramétrable et peut être ajusté au souhait de l'utilisateur ou du fournisseur de service via la taille V du dictionnaire de quantification. De manière générale, le tableau 3 montre que plus V est faible, meilleure est l'anonymisation, mais cela est au prix d'une dégradation des performances en reconnaissance de parole. Et, inversement, plus V est grand, meilleure est la reconnaissance de parole au détriment d'une moins bonne anonymisation.

4 Conclusion

Dans cet article, nous avons appliqué le processus de calcul hybride respectueux de la vie privée, qui décompose un modèle neuronal en deux parties, un encodeur qui génère une représentation anonyme sur l'appareil de l'utilisateur, et un décodeur qui utilise cette représentation anonyme pour effectuer des calculs mutualisés. Nous avons étudié ce système dans le contexte des assistants vocaux. Comme encodeur, un modèle acoustique *TDNN-F* a été considéré, et nous avons montré ses limitations. En utilisant le jeu de donnée du challenge Voice Privacy, nous avons mesuré que le locuteur peut être vérifié à la hauteur de 9,3% d'EER pour les femmes et 4,2% d'ERR pour les hommes dans un modèle *TDNN-F* classique. Nous avons proposé d'utiliser un algorithme de quantification vectorielle afin de contraindre l'espace de représentation, forçant ainsi le modèle acoustique à uniquement encoder l'information phonétique. Cet algorithme est configurable en fonction de la taille du dictionnaire de quantification, ce qui permet d'ajuster le compromis entre de bonnes performances en reconnaissance de parole et une bonne anonymisation. Par exemple, avec un dictionnaire de 128 vecteurs, le locuteur est dramatiquement moins vérifiable, 22,0% d'EER pour les femmes et 22,8% d'ERR pour les hommes ce qui correspond a un gain 232%. Mais ce gain en anonymisation impacte les performances de reconnaissance de parole, le WER augmente de 47% (augmentation de 5,8% à 8,5% de WER). Dans les prochains travaux, nous prévoyons de générer de la parole à partir de ces représentations anonymes et d'évaluer les performances en reconnaissance de parole et masquage d'identité du locuteur à partir de la parole générée.

Remerciements

Ce travail a été réalisé avec le soutien de l'Agence nationale de la recherche française, dans le cadre du projet ANR DEEP-PRIVACY (18-CE23-0018) et de la Région Grand Est.

Références

ADI Y., ZEGHIDOUR N., COLLOBERT R., USUNIER N., LIPTCHINSKY V. & SYNNAEVE G. (2019). To reverse the gradient or not : an empirical comparison of adversarial and multi-task learning in speech recognition. In *IEEE ICASSP*.

ALOUFI R., HADDADI H. & BOYLE D. (2021). Configurable Privacy-Preserving Automatic Speech Recognition. In *Proc. Interspeech*.

BENGIO Y., LÉONARD N. & COURVILLE A. C. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv*.

CHOROWSKI J., WEISS R. J., BENGIO S. & VAN DEN OORD A. (2019). Unsupervised speech representation learning using wavenet autoencoders. *In IEEE TASLP*.

GEORGESCU A.-L., PAPPALARDO A., CUCU H. & BLOTT M. (2021). Performance vs. hardware requirements in state-of-the-art automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*.

GERSHO A. & GRAY R. M. (1992). Vector quantization and signal compression. In *The Kluwer international series in engineering and computer science*.

HADIAN H., SAMETI H., POVEY D. & KHUDANPUR S. (2018). End-to-end speech recognition using lattice-free mmi. In *Proc. Interspeech*.

KOPPELMANN T., NELUS A., SCHÖNHERR L., KOLOSSA D. & MARTIN R. (2021). Privacy-Preserving Feature Extraction for Cloud-Based Wake Word Verification. In *Proc. Interspeech*.

LARCHER A., LEE K.-A. & MEIGNIER S. (2016). An extensible speaker identification sidekit in python. *In IEEE ICASSP*.

LI C.-Y., YUAN P.-C. & YI LEE H. (2020). What does a network layer hear ? analyzing hidden representations of end-to-end asr through speech synthesis. *In IEEE ICASSP*.

LI T. & LI N. (2009). On the tradeoff between privacy and utility in data publishing. In *The 15th ACM SIGKDD*.

LIU Y., HE L. & LIU J. (2019). Large Margin Softmax Loss for Speaker Verification. In *Proc. Interspeech*.

MADIKERI S., TONG S., ZULUAGA-GOMEZ J., VYAS A., MOTLICEK P. & BOURLARD H. (2020). Pkwrap : a pytorch package for lf-mmi training of acoustic models. *ArXiv*.

OSIA S. A., SHAHIN SHAMSABADI A., SAJADMANESH S., TAHERI A., KATEVAS K., RABIEE H. R., LANE N. D. & HADDADI H. (2020). A hybrid deep learning architecture for privacy-preserving mobile analytics. *IEEE Internet of Things Journal*.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An asr corpus based on public domain audio books. In *IEEE ICASSP*.

POVEY D., CHENG G., WANG Y., LI K., XU H., YARMOHAMMADI M. & KHUDANPUR S. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Proc. Interspeech*.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLÍČEK P., QIAN Y., SCHWARZ P., SILOVSKÝ J., STEMMER G. & VESEL K. (2011). The Kaldi Speech Recognition Toolkit. *In IEEE ASRU*.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust dnn embeddings for speaker recognition. In *IEEE ICASSP*.

SRIVASTAVA B. M. L., BELLET A., TOMMASI M. & VINCENT E. (2019). Privacy-Preserving Adversarial Representation Learning in ASR : Reality or Illusion ? In *Proc. Interspeech*.

TOMASHENKO N., SRIVASTAVA B. M. L., WANG X., VINCENT E., NAUTSCH A., YAMAGISHI J., EVANS N., PATINO J., BONASTRE J.-F., NOÉ P.-G. & TODISCO M. (2020). Introducing the VoicePrivacy Initiative. *Proc. Interspeech*.

VAN DEN OORD A., VINYALS O. & KAVUKCUOGLU K. (2017). Neural discrete representation learning. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Eds., *Advances in Neural Information Processing Systems*.

WU D.-Y. & LEE H.-Y. (2020). One-shot voice conversion by vector quantization. In *IEEE ICASSP*.

ŁUKASZ KAISER, ROY A., VASWANI A., PARMAR N., BENGIO S., USZKOREIT J. & SHAZEER N. M. (2018). Fast decoding in sequence models using discrete latent variables. In *ICML*.