Analyse de l'anonymisation du locuteur sur de la parole émotionnelle

Hubert Nourtel, Pierre Champion, Denis Jouvet, Anthony Larcher, Marie Tahon

Analyse de l'anonymisation du locuteur sur de la parole émotionnelle

Hubert Nourtel[1⋆] Pierre Champion[1,2⋆] Denis Jouvet[1] Anthony Larcher[2]
Marie Tahon[2]
[1] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[2] LIUM - EA4023, Le Mans Université, Avenue Olivier Messiaen,72085 LE MANS CEDEX 9, France
⋆ Contributions équivalentes des auteurs
hubert.nourtel@inria.fr

RÉSUMÉ

Les données vocales contiennent des informations personnelles, telles que l'identité du locuteur ou son état émotionnel, pouvant être utilisées à des fins malveillantes. Parmi les études ayant abordé le sujet de la préservation de la confidentialité de la parole, l'initiative VoicePrivacy vise à promouvoir le développement d'outils de préservation de la vie privée pour les technologies vocales. L'objectif du VoicePrivacy Challenge 2020 (VPC) était de cacher l'identité du locuteur source tout en préservant les informations linguistiques. Cet article étudie l'impact du système d'anonymisation du locuteur du VPC sur les informations émotionnelles de la parole. Des modifications du système ont aussi été ajoutées pour essayer de masquer les émotions tout en limitant la dégradation du signal. Nos résultats montrent que la baseline du VPC masque légèrement les émotions des locuteurs mais certaines modifications permettent de masquer plus efficacement les émotions tout en conservant une bonne intelligibilité.

ABSTRACT

Evaluation of Speaker Anonymization on Emotional Speech

Speech data carries personal information, such as the speaker's identity and emotional state, which can be used for malicious purposes. Among the studies that have addressed the topic of speech privacy, the VoicePrivacy initiative aims to promote the development of privacy tools for speech technologies. The goal of the VoicePrivacy Challenge 2020 (VPC) was to hide the source speaker's identity while preserving the linguistic information. This paper studies the impact of the VPC speaker anonymization system on the emotional information of speech. Modifications to the system were also added in an attempt to mask the emotions while limiting signal degradation. Our results show that VPC baseline slightly masks the speakers' emotions, but some modifications can mask the emotions more effectively while maintaining good intelligibility.

MOTS-CLÉS : Anonymisation du locuteur, Privacité, Reconnaissance d'émotion.

KEYWORDS: Speaker Anonymization, Voice Privacy, Emotion Recognition.

1 Introduction

Les applications à commande vocale, telles que les enceintes intelligentes, sont devenues très populaires. Une grande quantité de données est nécessaire pour entrainer ces applications. Cela incite les

fournisseurs de services à collecter, traiter et stocker les données personnelles des utilisateurs dans des serveurs centralisés. La voix est l'une des modalités les plus sensibles car elle englobe de nombreux attributs discernables d'un locuteur tels que l'âge, le sexe, la santé, les traits de personnalité, le statut socio-économique, l'origine géographique, l'identité biométrique, les humeurs et les émotions (Kröger *et al.*, 2019). Étant donné que les données vocales entrent dans la catégorie des données personnelles (Nautsch *et al.*, 2019), les solutions de préservation de la confidentialité des données vocales deviennent de plus en plus importantes. En outre, de récentes réglementations, par exemple le Règlement Général sur la Protection des Données (RGPD) (European Parliament and Council, 2016) dans l'Union Européenne, mettent l'accent sur la préservation de la vie privée et la protection des données personnelles. Le système d'anonymisation utilisé dans cet article a été développé dans le cadre du Voice Privacy Challenge (VPC) (Tomashenko *et al.*, 2020) qui est l'une des premières tentatives de la communauté parole pour évaluer la recherche sur ce sujet en produisant des protocoles, des mesures, des corpus ainsi que des baselines dédiées.

L'objectif du système VPC est d'anonymiser le locuteur. Cette tâche est effectuée pour supprimer les informations permettant de retrouver l'identité du locuteur dans le signal de parole, tout en conservant les autres caractéristiques. La baseline du VPC utilise une approche d'anonymisation du locuteur (Fang *et al.*, 2019) basée sur les x-vecteurs et la conversion de la voix. La qualité de l'anonymisation dans le VPC est mesurée à l'aide d'un système de vérification du locuteur, qui évalue la capacité de dissimulation de l'identité du locuteur (mesure de confidentialité) et à l'aide d'un système de reconnaissance automatique de la parole pour évaluer la préservation et l'intelligibilité du contenu linguistique (mesure d'utilité) (Tomashenko *et al.*, 2020). Dans ce travail, nous étudions dans quelle mesure le contenu émotionnel d'un énoncé peut être récupéré après anonymisation du signal de parole.

La reconnaissance du locuteur et la confidentialité de la voix se concentrent généralement sur la parole dite "neutre". Cependant, dans le cas d'un discours expressif spontané, le signal audio contient des informations sur le locuteur, le contenu linguistique et des indices émotionnels. Le processus d'anonymisation peut être altéré par un discours émotionnel, pour lequel le signal vocal diffère fortement du discours "neutre".

Très peu de travaux ont traité le problème de la préservation de la vie privée dans le contexte des discours émotionnels. Dans (Dias *et al.*, 2018), les auteurs ont proposé des techniques de hachage préservant la distance et un chiffrement homomorphique pour protéger les données sensibles telles que les émotions. Les Generative Adversarial Networks (GAN) ont été utilisés comme couche intermédiaire entre les utilisateurs et les services sur le cloud pour modifier le discours d'entrée (Ranya *et al.*, 2019). Notre objectif est d'étudier l'impact de l'application d'un processus d'anonymisation sur les données émotionnelles, qui est censé cacher l'identité du locuteur, en mesurant les performances d'un système de reconnaissance des émotions (SER) sur ces données (avant et après anonymisation). Des travaux préliminaires sur ce sujet ont été entrepris précédemment (Nourtel *et al.*, 2021). Cet article apporte l'analyse d'une nouvelle technique d'anonymisation, la Differential Privacy (Shamsabadi *et al.*, 2022), ainsi qu'une modification du périmètre émotionnel étudié, la frustration n'étant plus incluse comme dans (Pappagari *et al.*, 2020; Cho *et al.*, 2018).

2 Système d'anonymisation

2.1 Système d'anonymisation du locuteur du VPC

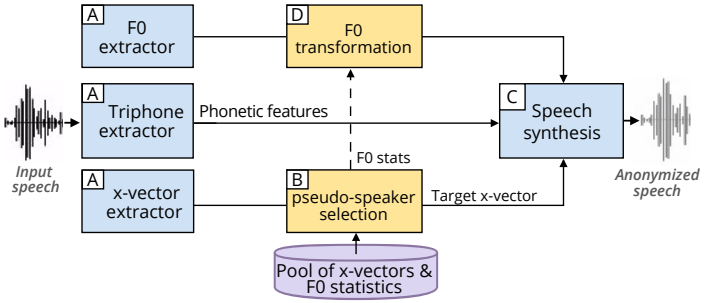

FIGURE 1 – Système d'anonymisation du locuteur de Voice Privacy. Les modules A, B et C font partie de la baseline. Le module D correspond à la modification apportée pour transformer les valeurs de F0. Figure provenant de l'article (Champion *et al.*, 2021).

La baseline du VPC [1] repose sur la séparation de l'identité du locuteur et du contenu linguistique d'un énoncé vocal en entrée. En supposant que ces caractéristiques sont effectivement séparées, une forme d'onde vocale anonymisée est générée en modifiant uniquement les caractéristiques qui encodent l'identité du locuteur. Sur la figure 1, les modules du *groupe A* extraient différentes caractéristiques du signal source : la fréquence fondamentale (F0), les caractéristiques phonétiques codant l'articulation des sons de la parole et une représentation du locuteur à l'aide de x-vecteur (Snyder *et al.*, 2018).

Le module B produit un x-vecteur cible correspondant à un pseudo-locuteur. Le x-vecteur du locuteur source est comparé à un ensemble de x-vecteurs externes pour sélectionner les 200 x-vecteurs les plus éloignés ; 100 d'entre eux sont sélectionnés de manière aléatoire et leur moyenne est calculée pour créer le x-vecteur d'un pseudo-locuteur anonyme. Enfin, *le module C* synthétise une forme d'onde vocale à partir du x-vecteur cible avec les caractéristiques phonétiques et la F0. L'anonymisation du locuteur est réalisée par la sélection du x-vecteur caractérisant le pseudo-locuteur cible.

2.2 Transformations de la F0

Dans le système original d'anonymisation du VPC, les valeurs de F0 extraites de la parole source sont utilisées inchangées par le synthétiseur vocal, quel que soit le pseudo-locuteur sélectionné. Des études ont examiné la conversion vocale conditionnée par la valeur de la F0 (Bahmaninezhad *et al.*, 2018; Qian *et al.*, 2020; Chappell & Hansen, 1998). Ils concluent que la modification de la F0 améliore la qualité de la voix convertie. Motivés par ces résultats, ainsi que par le fait que les émotions affectent indubitablement l'intonation, nous proposons de modifier les valeurs de F0 d'un énoncé source d'un locuteur donné (cf. *module D* dans la figure 1) en utilisant deux techniques distinctes décrites ci-dessous.

Transformation linéaire avec décalage aléatoire Dans cette méthode, pour alimenter le synthétiseur avec des valeurs de F0 proches du pseudo-locuteur sélectionné, les caractéristiques de la F0

1. Détails d'implémentation : https ://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2020_Eval_Plan_v1_4.pdf

du locuteur source sont transformées à l'aide d'une transformation linéaire (Champion *et al.*, 2021). Avant d'appliquer cette dernière, le contour des valeurs de F0 est modifié de manière aléatoire pour augmenter ou diminuer la plage de variation de F0 à l'aide d'un facteur d'échelle :

$$\log \hat{x}_t = \mu_y + \frac{\sigma_y}{\sigma_x} (\log x'_t - \mu_x) \text{ avec } x'_t = \mu_s + (x_t - \mu_s) \times \alpha \quad (1)$$

Ici, x_t représente la F0 du locuteur source à l'instant t, μ_x et σ_x représentent la moyenne et l'écart-type de la F0 du locuteur source avec une échelle logarithmique calculée sur tous ses énoncés. Les valeurs μ_y et σ_y représentent la moyenne et l'écart-type de la F0 du pseudo-locuteur avec une échelle logarithmique. x'_t correspond à la F0 à laquelle a été appliquée le facteur d'échelle α, échantillonné de façon aléatoire pour chaque énoncé à partir d'une distribution uniforme entre 0,8 et 1,2. μ_s représente la moyenne de la F0 sur les parties voisées du segment de parole en cours de transformation. Le calcul n'est effectué que sur les trames voisées. La moyenne et l'écart-type pour le pseudo-locuteur cible sont calculés en faisant la moyenne des F0 des trames voisées des 100 locuteurs sélectionnés pour obtenir le x-vecteur du pseudo-locuteur.

Differential Privacy Dans cette méthode (Shamsabadi *et al.*, 2022), la F0 est bruitée en respectant la propriété de Differential Privacy (DP). La F0 est d'abord extraite via un estimateur, puis est envoyée dans un autoencodeur convolutionnel entrainé pour reconstruire la F0. Le bruit est ajouté entre l'encodeur et le décodeur, selon une distribution Laplacienne centrée, avec un coefficient de bruitage ϵ, fixé à 10 pour la totalité de cet article. Le bruit n'est pas directement ajouté à la F0 extraite mais dans l'espace latent de l'autoencodeur associé pour ne pas détruire les intonations et les informations linguistiques du signal d'entrée. Cette architecture ajoute des perturbations locales qui sont plus spécifiques au locuteur. De cette manière, les intonations et informations linguistiques globales portées par la F0 sont conservées tout en réduisant la corrélation avec le locuteur d'origine.

Exemple de transformations de la F0 La figure 2 montre la comparaison des valeurs de la F0 entre un segment audio original et les segments audios avec les différentes transformations appliquées. Nous pouvons constater que la transformation linéaire ne modifie que très peu le profil de la F0 alors que la transformation avec Differential Privacy la modifie de façon plus prononcée.

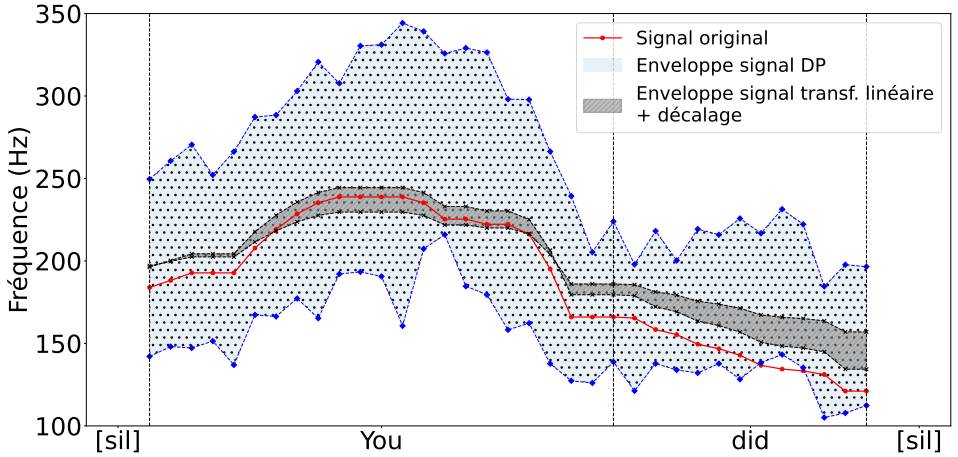

FIGURE 2 – Comparaison des valeurs de la F0 entre un signal original et ce même signal avec différentes transformations de la F0. Les enveloppes indiquent les plages de valeurs de la F0 anonymisée sur différentes expériences.

2.3 Les scénarios d'attaques du VPC

Dans le cadre du Voice Privacy Challenge, plusieurs séries de tests ont été réalisées en fonction de la connaissance de l'algorithme d'anonymisation par l'attaquant. Dans ce travail, nous nous concentrons sur les scénarios *Ignorant* et *Informé* de l'attaquant (Srivastava *et al.*, 2020, 2021). Dans le scénario *Ignorant*, l'attaquant ne sait pas que la parole est transformée. Ainsi, la mesure de la confidentialité est évaluée à l'aide de modèles entrainés sur des données originales non anonymisées, alors que l'évaluation est réalisée à l'aide de données anonymisées. Ce décalage conduit à la mesure d'une performance d'anonymisation plutôt bonne. À l'inverse, le scénario de l'attaquant *Informé* est entièrement conscient de l'algorithme d'anonymisation. Ces attaquants sont capables d'anonymiser un ensemble de données d'entraînement de la même manière que le fournisseur de services. Cet ensemble de données anonymes est ensuite utilisé pour entraîner le modèle d'évaluation.

3 Expérimentations

3.1 Dataset

Le corpus IEMOCAP (Busso *et al.*, 2008) est un corpus émotionnel utilisé à des fins d'expérimentation telles que la reconnaissance des émotions ou la reconnaissance de la parole. Il est composé de 12h de données audio-visuelles. Des dialogues improvisés et scénarisés entre 10 acteurs féminins et masculins en langue anglaise ont été enregistrés. Des microphones directionnels ont été utilisés pour capturer la parole de chaque locuteur. Cela implique que dans les fichiers audio, les deux locuteurs peuvent apparaître simultanément (chevauchement). En cas de chevauchement de la parole, le locuteur le plus proche du microphone est considéré comme dominant, et seulement sa parole sera transcrite dans les transcriptions de référence. En raison des microphones directionnels utilisés, le niveau de la voix qui se chevauche est beaucoup plus faible que le niveau de la voix du locuteur "dominant".

Les données sont segmentées par tour de parole du locuteur (dominant). Chaque tour a été annoté avec des catégories d'émotions par six annotateurs humains. Seuls les enregistrements sur lesquels la majorité des annotateurs étaient en accord ont été utilisés. À la suite de travaux antérieurs (Pappagari *et al.*, 2020; Cho *et al.*, 2018), nous ne considérons que quatre émotions : neutre, tristesse, colère et joie. La joie combine les annotations originales de la joie et de l'excitation pour équilibrer le nombre d'énoncés dans chaque classe d'émotion. Après sélection de l'ensemble d'émotions, il reste 5500 segments de parole pour une durée totale de sept heures équilibrés entre les quatre émotions.

3.2 Protocole d'évaluation

Dans cet article, nous nous concentrons sur l'évaluation des informations émotionnelles présentes dans le signal vocal, à la fois dans les énoncés originaux et dans les énoncés anonymisés correspondants. Ces évaluations sont réalisées à l'aide d'un système de reconnaissance des émotions de la parole (SER).

Le système de reconnaissance des émotions utilisé est basé sur un modèle de Support Vector Machine (SVM) entrainé sur le corpus IEMOCAP. Dans la littérature sur la SER, le SVM avec le noyau non linéaire de la fonction de base radiale est largement utilisé (Lee *et al.*, 2011; Tahon & Devillers, 2016) comme référence. Comme caractéristiques d'entrée, les caractéristiques eGeMAPS sont utilisées car elles fournissent une représentation minimaliste mais efficace des émotions (Eyben *et al.*, 2015).

Nous avons également utilisé des x-vecteurs comme caractéristiques d'entrée car des résultats récents (Pappagari *et al.*, 2020) ont montré qu'ils donnent de bons résultats pour la SER. Ces x-vecteurs sont extraits à l'aide de l'outil *Sidekit* (Larcher *et al.*, 2016). Le modèle utilisé est un half-Resnet, entrainé sur les données *VoxCeleb1,2* pour une tâche de vérification du locuteur. La reconnaissance des émotions est évaluée en utilisant les deux scénarios d'attaque du VPC *Ignorant*, où le modèle est entrainé sur les données originales, et *Informé*, où le modèle est entrainé sur les données anonymisées.

La métrique Unweighted Average Recall (UAR), définie comme étant la moyenne non pondérée du rappel par classe d'émotion, est utilisée pour mesurer les performances de reconnaissance des émotions. Un UAR élevé correspond à une bonne reconnaissance des émotions. Pour tenir compte de la petite taille de l'ensemble de données, l'entrainement et l'évaluation sont effectués en utilisant une validation croisée "leave-one-session-out", les sessions étant de tailles équivalentes. L'UAR est alors calculé globalement sur les cinq tests de la validation croisée. La littérature (Lee *et al.*, 2011) montre des résultats de 51% d'UAR dans des conditions similaires.

L'évaluation de l'utilité, qui porte sur la préservation et l'intelligibilité du contenu linguistique, est réalisée à l'aide de deux systèmes de reconnaissance automatique de la parole (ASR) fournis par les organisateurs du VPC et déjà entrainé sur des données LibriSpeech *train-clean-360* originales pour l'un et anonymisées pour l'autre. Les résultats sont rapportés avec le Word Error Rate (WER). Plus ce WER est faible, plus la parole est intelligible. Pour référence, ce système obtient un WER de 4.15% sur le corpus *Librispeech-test-clean*.

3.3 Résultats

3.3.1 Reconnaissance d'émotion

TABLE 1 – Résultats en terme d'UAR selon différentes configurations d'anonymisation sur IEMOCAP. Les entrées sont des features eGeMAPS (1ère colonne) ou des x-vecteurs (2ème colonne).

Attaquant	Signal	$UAR_{\%}$ eGeMAPS	$UAR_{\%}$ x-vecteur
-	Original (baseline)	56,23 ±1,31	57,07 ±1,31
Informé	VPC	48,00 ±1,32	50,75 ±1,32
	VPC + Linéaire	48,72 ±1,32	50,45 ±1,32
	VPC + DP	43,11 ±1,31	47,03 ±1,32
Ignorant	VPC	27,24 ±1,18	30,64 ±1,22
	VPC + Linéaire	26,42 ±1,17	29,02 ±1,20
	VPC + DP	25,52 ±1,15	30,49 ±1,22

En utilisant les scénarios d'attaque définis par le VPC (voir section 2.3), les scores de reconnaissance des émotions des attaquants *Ignorant* et *Informé* sont résumés dans le tableau 1 pour différentes configurations d'anonymisation et de features d'entrée.

Pour un attaquant *Ignorant*, l'UAR est entre 25% et 30%, ce qui est proche d'une estimation aléatoire pour la classification de quatre émotions. Dans ce cas, l'information émotionnelle semble donc fortement dégradée pour l'attaquant.

Pour un attaquant *Informé*, le taux de reconnaissance des émotions est plus élevé. On observe également, en comparant aux performances sur le signal original, que l'anonymisation du signal réduit l'UAR de façon plus importante avec des features eGeMAPS qu'avec des x-vecteurs en entrée. De plus, la technique de Differential Privacy semble avoir un impact plus important en terme de

masquage des émotions que les autres techniques d'anonymisation. Les matrices de confusions nous montrent que la classification de la colère est impactée de 16% alors que l'impact est plus limité et homogène sur les autres émotions. Un attaquant *Informé* peut donc toujours détecter des informations émotionnelles mais en quantité réduite par rapport au signal original.

TABLE 2 – Résultats en terme de WER et d'UAR sur IEMOCAP. La première ligne montre les résultats lorsqu'aucune anonymisation n'est effectuée. Les lignes suivantes montrent les résultats lorsque la parole est anonymisée selon diverses techniques. L'évaluation de la parole anonymisée est effectuée dans un contexte d'attaquant *Informé*.

Signal	$WER_{\%}$	Dégradation WER / Signal orignal	$UAR_{\%}$	Dégradation UAR / Signal original
Original (baseline)	37,88 ±0,77	-	57,07 ±1,31	-
VPC	41,86 ±0,80	10,51 %	50,75 ±1,32	11,07 %
VPC + Linéaire	41,62 ±0,80	9,87 %	50,45 ±1,32	11,60 %
VPC + DP	41,65 ±0,82	9,95 %	47,03 ±1,32	17,60 %

3.3.2 Paramètres prosodiques

Si l'on souhaite masquer des informations émotionnelles dans la parole anonymisée, il faut modifier les paramètres prosodiques (c'est-à-dire la fréquence fondamentale, l'intensité ou le rythme). La baseline d'anonymisation fournissant la F0, nous avons réalisé des expériences impliquant des modifications de la F0 (voir section 2.2).

Les résultats sont présentés dans le tableau 2 avec l'intervalle de confiance à 95 %. Sont affichés le WER et l'UAR sur le corpus IEMOCAP, à la fois pour les données originales et pour les données anonymisées, dans le contexte d'un attaquant informé (signal orignal $\Rightarrow$ modèle entrainé sur données originales, signal anonymisé $\Rightarrow$ modèle entrainé sur données anonymisées). Concernant l'UAR, le modèle utilisé est basé sur les x-vecteurs car donnant de meilleurs résultats sur la parole originale.

Nous constatons une dégradation d'environ 10% en terme de WER quel que soit le type d'anonymisation effectué. La dégradation étant constante, nous pouvons en déduire que les modifications de la F0 apportées dans cette expérimentation ne dégradent pas l'utilité du signal par rapport à l'anonymisation effectuée dans la baseline de VPC.

Nous voyons aussi que l'anonymisation via la baseline ou avec une transformation linéaire dégrade l'UAR d'environ 11%, ce qui est du même ordre de grandeur que la dégradation en WER. En revanche, l'anonymisation avec Differential Privacy dégrade l'UAR de 17%, ce qui est significativement supérieur à la dégradation en WER pour ce contexte. Les performances de reconnaissance d'émotion sont donc bel et bien impactées par certaines modifications de la F0.

4 Conclusion

Dans cet article, nous avons évalué l'application du système d'anonymisation du Voice Privacy Challenge sur de la parole émotionnelle. Concernant la métrique d'utilité, basée sur les performances de la reconnaissance automatique de la parole, nous avons observé une dégradation de 10 % du Word Error Rate (WER) sur les données anonymisées IEMOCAP par rapport au WER mesuré sur la parole originale. Cette dégradation est constante quelle que soit la technique d'anonymisation utilisée (avec ou sans transformation de la F0).

Concernant l'information sur les émotions véhiculée par le signal vocal, nous l'avons mesurée à l'aide de la métrique Unweighted Average Recall (UAR). Par rapport à la mesure sur les données originales, nous avons observé une dégradation de 11% de l'UAR sur les données anonymisées sans transformation ou avec une transformation linéaire de la F0. Cette dégradation passe à 17% sur les données anonymisées avec une transformation Differential Privacy de la F0.

Si l'on considère l'émotion comme une information précieuse à conserver dans le signal vocal anonymisé, il faut conserver une anonymisation sans transformation ou avec une transformation linéaire de la F0. En effet, la dégradation observée pour la reconnaissance des émotions est similaire à la dégradation observée sur le Word Error Rate (qui mesure l'utilité) et est réduite par rapport à la modification avec Differential Privacy.

Cependant, nous pouvons aussi considérer l'émotion comme une information personnelle que le système d'anonymisation devrait supprimer. Dans ce cas, il faut considérer une modification de la F0 via la technique de Differential Privacy, qui dégrade plus fortement la métrique de reconnaissance d'émotion que la métrique d'utilité. Pour aller plus loin dans le masquage des émotions, des recherches ultérieures pourront porter sur des modifications de la durée et de l'énergie, qui sont d'autres paramètres prosodiques porteurs d'informations sur les émotions.

Remerciements

Ce travail a été effectué avec le support de l'Agence Nationale de la Recherche dans le cadre du projet ANR DEEP-PRIVACY (18-CE23-0018) et de la Région Grand Est. Les expériences présentées dans ce document ont été réalisées à l'aide du banc d'essai Grid'5000, soutenu par un groupement d'intérêt scientifique hébergé par Inria et comprenant le CNRS, RENATER et plusieurs Universités ainsi que d'autres organismes (voir https ://www.grid5000.fr). Nous remercions Brij Mohan Lal Srivastava pour l'accès au code source d'implémentation de la technique Differential Privacy.

Références

BAHMANINEZHAD F., ZHANG C. & HANSEN J. (2018). Convolutional neural network based speaker de-identification. In *Odyssey 2018*, p. 255–260.

BUSSO C., BULUT M., LEE C.-C., KAZEMZADEH A., MOWER E., KIM S., CHANG J. N., LEE S. & NARAYANAN S. S. (2008). Iemocap : Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, **42**(4), 335–359.

CHAMPION P., JOUVET D. & LARCHER A. (2021). A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender. In *PPAI 2021 - The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence*.

CHAPPELL D. & HANSEN J. (1998). Speaker-specific pitch contour modeling and modification. In *ICASSP 1998*, volume 2, p. 885–888.

CHO J., PAPPAGARI R., KULKARNI P., VILLALBA J., CARMIEL Y. & DEHAK N. (2018). Deep neural networks for emotion recognition combining audio and transcripts. In *Interspeech 2018*, p. 247–251.

DIAS M., ABAD A. & TRANCOSO I. (2018). Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition. In *ICASSP 2018*, p. 2057–2061.

EUROPEAN PARLIAMENT AND COUNCIL (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard

to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

EYBEN F., SCHERER K. R., SCHULLER B. W., SUNDBERG J., ANDRÉ E., BUSSO C., DEVILLERS L. Y., EPPS J., LAUKKA P., NARAYANAN S. S. *et al.* (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, **7**(2), 190–202.

FANG F., WANG X., YAMAGISHI J., ECHIZEN I., TODISCO M., EVANS N. & BONASTRE J.-F. (2019). Speaker Anonymization Using X-vector and Neural Waveform Models. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, p. 155–160.

KRÖGER J. L., LUTZ O. H.-M. & RASCHKE P. (2019). *Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference*, In *Privacy and Identity Management*, volume 576, p. 242–258. Springer International Publishing.

LARCHER A., LEE K. A. & MEIGNIER S. (2016). An extensible speaker identification sidekit in python. In *ICASSP 2016*, p. 5095–5099.

LEE C.-C., MOWER E., BUSSO C., LEE S. & NARAYANAN S. S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, **53**(9), 1162–1171.

NAUTSCH A., JASSERAND C., KINDT E., TODISCO M., TRANCOSO I. & EVANS N. (2019). The gdpr & speech data : Reflections of legal and technology communities, first steps towards a common understanding. In *Interspeech 2019*, p. 3695–3699.

NOURTEL H., CHAMPION P., JOUVET D., LARCHER A. & TAHON M. (2021). Evaluation of Speaker Anonymization on Emotional Speech. In *SPSC 2021 - 1st ISCA Symposium on Security and Privacy in Speech Communication*, Virtual, Germany.

PAPPAGARI R., WANG T., VILLALBA J., CHEN N. & DEHAK N. (2020). X-vectors meet emotions : A study on dependencies between emotion and speaker recognition. In *ICASSP 2020*, p. 7169–7173.

QIAN K., JIN Z., HASEGAWA-JOHNSON M. & MYSORE G. J. (2020). F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder. In *ICASSP 2020*, p. 6284–6288.

RANYA A., HAMED H. & DAVID B. (2019). Emotionless : Privacy-preserving speech analysis for voice assistants. In *Privacy Preserving in Machine Learning (CCS19) Workshop*, London, UK.

SHAMSABADI A. S., SRIVASTAVA B. M. L., BELLET A., VAUQUIER N., VINCENT E., MAOUCHE M., TOMMASI M. & PAPERNOT N. (2022). Differentially private speaker anonymization.

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

SRIVASTAVA B. M., VAUQUIER N., SAHIDULLAH M., BELLET A., TOMMASI M. & VINCENT E. (2020). Evaluating voice conversion-based privacy protection against informed attackers. In *ICASSP 2020*, p. 2802–2806.

SRIVASTAVA B. M. L., MAOUCHE M., SAHIDULLAH M., VINCENT E. & BELLET A. E. A. (2021). Privacy and utility of x-vector based speaker anonymization. *Transactions on Audio, Speech and Language Processing*.

TAHON M. & DEVILLERS L. (2016). Towards a small set of robust acoustic features for emotion recognition : Challenges. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **24**(1), 16–28.

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. In *Interspeech 2020*, p. 1693–1697.