Extraction d'informations liées au locuteur depuis un modèle acoustique personnalisé

Salima Mdhaffar[1], Jean-François Bonastre[1], Marc Tommasi[2], Natalia Tomashenko[1], Yannick Estève[1]

(1) LIA, Avignon Université, France
(2) Université de Lille, CNRS, Inria, Centrale Lille, UMR 9189 - CRIStAL, Lille, France
[1]{prénom.nom}@univ-avignon.fr,[2]{prénom.nom}@inria.fr

RÉSUMÉ

Plusieurs services intégrés dans notre vie quotidienne utilisent la reconnaissance automatique de la parole (Apple-Siri, Amazon-Alexa...). Ces services s'appuient sur des modèles entraînés sur une grande quantité de données pour assurer leur efficacité. Les données utilisées sont collectées via les applications, à partir des interactions des utilisateurs. Elles contiennent souvent des informations sensibles, ce qui peut créer d'importants problèmes de confidentialité. Dans ce contexte, de nouveaux paradigmes d'apprentissage automatique ont été proposés, comme l'apprentissage fédéré et distribué. Tous deux permettent de créer des modèles personnalisés à partir de données privées. Seuls les modèles sont alors partagés, sans exposer les données elles-mêmes. Une question cruciale est de savoir si la diffusion de ces modèles acoustiques (MA) personnalisés peut aussi entraîner une fuite d'informations personnelles. Les résultats montrent qu'il est possible de retrouver des informations liées au locuteur en s'appuyant sur les modifications de poids d'un MA induites par l'adaptation locale sur ce locuteur.

ABSTRACT

Speaker Information Extraction from Personalized Acoustic Model

Several services integrated in our daily life use automatic speech recognition (Apple-Siri, Amazon-Alexa...). These services are based on models trained on a large amount of data to ensure their efficiency. The used data is collected via applications, from the users' interactions. It often contains sensitive information, which can create significant privacy issues. In this context, new machine learning paradigms have been proposed, such as federated and distributed learning. Both of them allow the creation of personalized models from private data. Only the models are then shared, without exposing the data itself. An important question is whether the dissemination of these personalized acoustic models (AM) can also lead to the leakage of personal information. The results show that it is possible to recover speaker-related information based on the changes in weight of an AM induced by local adaptation on that speaker.

MOTS-CLÉS : Modèles acoustiques personnalisés, apprentissage fédéré, vie privée.

KEYWORDS: Personalized acoustic model, federated learning, privacy.

1 Introduction

Plusieurs services présents dans notre vie quotidienne utilisent la reconnaissance automatique de la parole (Apple Siri, Amazon Alexa, Google Home...). Pour être performants, ces services ont besoin de modèles entraînés sur une grande quantité de données et utilisent des données collectées en mode production, venant de leurs utilisateurs, créant un risque sérieux en termes de confidentialité des données des utilisateurs. Les nouvelles réglementations sur les données, comme le Règlement général sur la protection des données (RGPD) dans l'Union européenne, changent les règles afin de protéger la vie privée des citoyens (Nautsch *et al.*, 2019). Afin d'améliorer les performances des modèles de la reconnaissance de la parole en exploitant l'expérience des utilisateurs sans accéder à leurs données, des solutions telles que l'apprentissage fédéré et l'apprentissage distribué ont été proposées. Elles consistent à échanger des modèles personnalisés, ou leurs gradients, au lieu des données brutes (Leroy *et al.*, 2019; Hard *et al.*, 2020; Guliani *et al.*, 2021; Yu *et al.*, 2021; Cui *et al.*, 2021) pour préserver la vie privée des utilisateurs. Dans le cadre de l'apprentissage distribué, un modèle personnalisé est un modèle qui a été adapté localement à un utilisateur (Mansour *et al.*, 2020). Dans un travail récent (Mdhaffar *et al.*, 2021), nous avons étudié une approche pour personnaliser un modèle acoustique hybride HMM/TDNN (Peddinti *et al.*, 2015) dans un contexte d'apprentissage collaboratif.

Dans cet article, nous étudions les informations contenues dans les modèles acoustiques personnalisés. En particulier, nous nous intéressons aux informations liées à l'identité et au sexe du locuteur qui peuvent être extraites à partir des modèles acoustiques personnalisés. Des travaux antérieurs ont étudié les représentations intermédiaires de la parole calculées dans des modèles neuronaux de bout en bout pour la reconnaissance de la parole. Ils ont illustré la manière dont ces modèles construisent des représentations phonétiques et graphémiques (Belinkov & Glass, 2017; Belinkov *et al.*, 2019), et ils ont montré comment la variabilité du locuteur et le bruit sont progressivement éliminés à mesure qu'on s'éloigne de la couche d'entrée d'un modèle neuronal profond (Li *et al.*, 2020). À notre connaissance, il n'existe aucune étude dans la littérature sur l'information contenue dans les changements des poids des réseaux neuronaux dus à la personnalisation d'un modèle acoustique.

L'article est organisé comme suit : la section 2 introduit un modèle acoustique personnalisé. La section 3 présente la méthode utilisée dans cette étude pour retrouver l'information contenue dans les changements des poids des réseaux neuronaux d'un modèle acoustique personnalisé. La section 4 présente notre protocole expérimental. La section 5 présente les résultats. Enfin, nos conclusions et nos perspectives futures sont énoncées dans la section 6.

2 Personnalisation d'un modèle acoustique

Dans notre scénario, un modèle acoustique générique est entraîné avec un jeu de données public. Il est ensuite déployé sur chaque terminal client où ce dernier est adapté par fine-tuning localement à partir des données utilisateur . Le fine-tuning consiste à poursuivre le processus d'apprentissage du modèle acoustique générique sur un petit jeu de données du locuteur, en veillant à éviter le sur-apprentissage. Le modèle résultant du fine-tuning est considéré comme un modèle personnalisé pour le locuteur local. La figure 1 illustre le processus de personnalisation d'un modèle acoustique.

Dans le cadre de l'apprentissage fédéré, les modèles personnalisés sont partagés pour être agrégé et améliorer un modèle global sans partager les données de l'utilisateur.

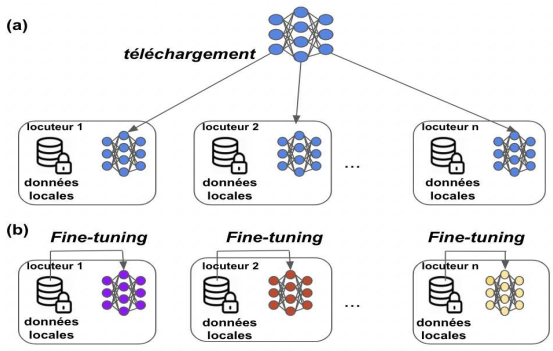

FIGURE 1 – Processus de personnalisation d'un modèle acoustique

3 Méthodologie proposée

Pendant la personnalisation des modèles acoustiques, les poids du modèle générique sont mis à jour. Nous supposons que ces mises à jour dépendent de certaines caractéristiques du locuteur. Nous émettons l'hypothèse qu'il est possible d'extraire des informations du locuteur en étudiant uniquement ces changements de poids. Dans cet article, nous analysons les informations liées au sexe et à l'identité des locuteurs. De plus, nous cherchons à savoir dans quelles couches cachées les modifications de poids sont particulièrement informatives.

3.1 Sexe du locuteur

Dans la première partie de cette étude, nous souhaitons étudier le niveau d'information liée au sexe capturé dans les modèles acoustiques neuronaux personnalisés, ou plus exactement dans les matrices de poids neuronaux correspondantes. Nous considérons que si cette information est encore présente dans les modèles personnalisés, les différences liées au sexe seront prédominantes et peuvent émerger lors d'un processus de regroupement automatique en deux classes : une classe dédiée aux femmes et l'autre classe est dédiée aux hommes. Pour cela, nous appliquons un regroupement agglomératif sur les matrices de poids de l'ensemble des modèles acoustiques personnalisés afin de créer une structure hiérarchique, jusqu'à avoir deux classes. Cette approche est basée sur un algorithme de classification non supervisé qui construit une arbre en partant des "feuilles" (poids d'une couche d'un modèle acoustique personnalisé) et procède par fusions successives des plus proches regroupements jusqu'à obtenir un regroupement unique "racine". La distance entre deux modèles acoustiques neuronaux est calculée avec la distance euclidienne appliquée aux matrices de poids de la même couche cachée.

3.2 Identité du locuteur

Dans la deuxième partie de cette étude, nous voulons évaluer la capacité d'identifier les locuteurs, toujours en ne considérant que les modifications appliquées aux matrices de poids lors de la personnalisation d'un modèle acoustique. Cependant, ces matrices de poids et même leurs couches cachées,

sont de très grande dimension. Les approches de réduction de la dimensionnalité comme l'Analyse en Composantes Principales (ACP) sont une solution potentielle mais le grand facteur de réduction visé, combiné à un nombre limité d'échantillons (un modèle par locuteur) pourrait résulter dans ce cas à une grande perte d'information discriminante.

Afin de résoudre ce problème, nous proposons d'appliquer une méthode inspirée de (Snyder *et al.*, 2018), qui consiste à apprendre un extracteur des représentations du locuteur. Nous proposons de construire une extracteur à l'aide d'un réseau de neurones appris sur les matrices de poids d'une couche cachée donnée des modèles neuronaux personnalisés de reconnaissance automatique de la parole. L'objectif de l'apprentissage est une tâche de reconnaissance du locuteur. Vu le nombre de données limité (notre jeu de données d'apprentissage est très petit), nous proposons de modifier la tâche de discrimination du locuteur en utilisant des classes de locuteurs comme labels de classification, au lieu des locuteurs. Ceci permet d'augmenter le nombre d'exemples par classe pendant la phase d'apprentissage, et donc de réduire le risque de sur-apprentissage. Les classes de locuteurs utilisées pour l'entraînement de l'extracteur ont été construites à partir d'un regroupement hiérarchique des i-vecteurs présents dans les données d'entraînement du modèle générique.

Afin de réduire le problème de mémoire de l'extracteur (les matrices d'entrée sont très larges) et de surmonter cette difficulté, nous avons conçu une structure spécifique pour notre extracteur. En partant d'un classificateur classique de réseau neuronal profond (DNN), nous appliquons une approche d'entrée multi-blocs. La matrice de poids est divisée en petits blocs qui sont liés séparément à une couche cachée dédiée. Un petit bloc de la matrice de poids d'entrée est composé de tous les poids liés à quelques unités neuronale dans la couche cachée ciblée dans le modèle acoustique. Par exemple, si la couche cachée ciblée H_t de l'architecture du modèle acoustique du système de reconnaissance de la parole contient n unités, la matrice des poids utilisée comme entrée de notre extracteur d'intégration du locuteur sera divisée en multiple de n blocs différents. Ensuite, les sorties de la couche cachée dédiée à chaque bloc sont concaténées pour créer la couche cachée suivante de l'extracteur basé sur un DNN, composé de couches entièrement connectées suivies de la couche finale softmax.

La structure de l'extracteur est illustrée dans la Figure 2. La couche cachée juste en dessous de la couche softmax représente la couche de la représentation vectorielle du locuteur. Le modèle neuronal résultant est capable d'extraire des représentations de locuteurs à partir des modèles acoustiques, y compris pour des locuteurs qui n'étaient pas présents dans les données d'apprentissage.

4 Protocole expérimental

4.1 Système de reconnaissance de la parole

Notre système de reconnaissance est fondé sur la boîte à outils de reconnaissance vocale Kaldi (Povey *et al.*, 2011). Nous utilisons les modèles acoustiques hybrides de type HMM/TDNN (Time Delay Neural Network). Notre modèle acoustique neuronal est composé de 13 couches avec une dimension de 512. Il prend comme entrée des paramètres MFCC de dimension 40 et des i-vecteurs de dimension 100. Nous avons appliqué deux stratégies d'augmentation des données sur les données d'apprentissage : perturbation de la vitesse (avec des facteurs 0,9 ; 1,0 et 1,1), et perturbation du volume. Le nombre total de paramètres est 13,8M.

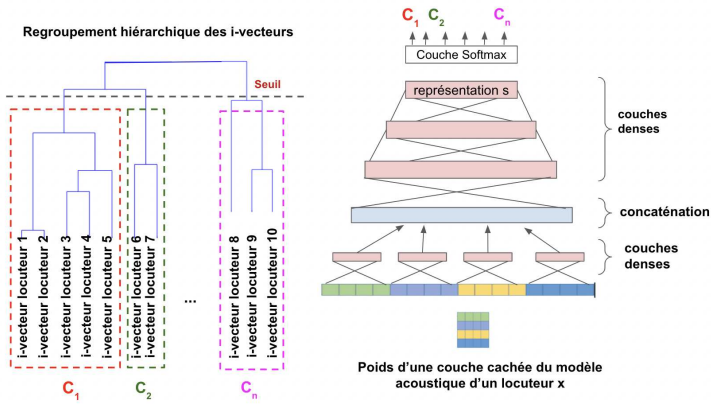

FIGURE 2 – Approche utilisée pour entraîner l'extracteur des représentations du locuteur à partir des poids d'une couche d'un modèle acoustique personnalisé

4.2 Données expérimentales

Les expériences ont été menées sur le corpus TED-LIUM 3 (Hernandez *et al.*, 2018). Ce jeu de données [1] contient 1495 présentations de conférences, correspondant à 420 heures de parole pour 2295 locuteurs. Comme dans les travaux (Mdhaffar *et al.*, 2021), nous avons utilisé une nouvelle partition de ce corpus. Le corpus a été divisé en 3 parties : *generic*, *p1* et *p2*. Les caractéristiques générales des ensembles de données obtenus sont présentées dans le tableau 1.

	generic	$p1$	$p2$
Duration (hours)	200	150	170
Duration of speech (hours)	170	125	150
# speakers	880	650	765
# speakers (duration>10 min)	-	463	581
# men	-	-	553
# women	-	-	212

TABLEAU 1 – TED-LIUM 3 dataset

Le corpus TED-LIUM3 ne comporte pas une annotation au sujet du sexe des locuteurs. Cette annotation a été ajouté manuellement pour l'ensemble de données $p2$ durant cette étude.

4.3 Modèles personnalisés

Un modèle acoustique générique est entrainé avec la partie *generic*. Les modèles personnalisés sont obtenus en *fine-tunant* le modèle générique sur les données du locuteur provenant de *p1* et *p2* : pour chaque locuteur, nous créons deux modèles personnalisés en utilisant séparément ses deux sessions

1. https ://lium.univ-lemans.fr/ted-lium3/

de cinq minutes. Alors, pour la plupart des locuteurs (locuteurs avec une durée > 10 minutes), deux modèles personnalisés différents sont obtenus.

Pour la reproductibilité des différentes expériences, les modèles acoustiques personnalisés ainsi que le modèle générique sont disponibles en ligne [2].

5 Résultats

Dans cette section, nous présentons les résultats de deux analyses (Mdhaffar *et al.*, 2022) : celle concernant l'information sur le sexe du locuteur et celle concernant l'identité du locuteur.

5.1 Sexe du locuteur

Il existe plusieurs méthodes utilisées pour évaluer la performance du regroupement agglomératif. Dans notre étude, nous utilisons la pureté. La pureté se concentre uniquement sur la maximisation du nombre total de vraies réponses positives par regroupement. Les valeurs de pureté sont comprises entre 0 et 1 (regroupement parfait). La pureté est définie comme

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |c_i \cap t_j| \tag{1}$$

où N est le nombre de locuteurs, k est le nombre de regroupements, c_i est un regroupement et t_j est le nombre de classifications pour le regroupement c_i.

La figure 3 montre les résultats pour les différentes couches cachées du réseau neuronal des modèles acoustiques pour les données dans *p2*. Nous observons qu'il est possible d'obtenir deux regroupements basés sur le sexe avec une valeur de pureté de 0,96 pour la couche 5. Nous observons que l'information sur le genre peut être identifiée dans les poids du réseau neuronal des modèles acoustiques. Les résultats montrent que l'information de genre peut être plus facilement identifiée dans les cinq premières couches.

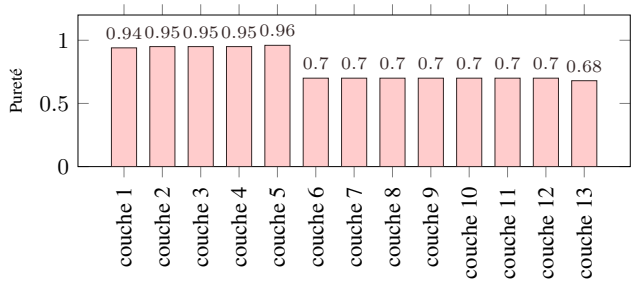

FIGURE 3 – Pureté de regroupement agglomératif des poids des couches cachées des modèles acoustiques personnalisés

2. https ://github.com/mdhaffar/Acoustic_model_personalisation

5.2 Vérification du locuteur

Tout d'abord, un extracteur de représentations des locuteurs est entraîné en utilisant chaque couche de nos modèles acoustiques comme entrée. Pour entraîner l'extracteur des représentations des locuteurs, nous utilisons 926 modèles de locuteurs personnalisés correspondant à 463+463 sous-ensembles uniques de *p1*. L'extracteur est entraîné avec les données *p2* pour extraire les représentations pour les locuteurs *p1* (sachant qu'il n'y a pas de chevauchement entre les locuteurs *p1* et *p2*). Respectivement, une deuxième expérience est menée avec *p1* comme ensemble de test et *p2* comme ensemble d'entraînement pour l'extracteur.

Le nombre de classes cibles (issue du regroupement hiérarchique des i-vecteurs des locuteurs présents dans les données d'entraînement) utilisées pour entraîner notre extracteur est fixé à 20 et la dimension des vecteurs de sortie (les représentations de locuteurs) est fixée à 100.

Au moment du test, l'extracteur entraîné est utilisé pour extraire les représentations du locuteur à partir de chaque instance d'entrée. Les instances d'entrée sont composées de certains poids de modèles acoustiques de réseaux de neurones personnalisés en fonction du locuteur. Nous utilisons une tâche de vérification des locuteurs pour évaluer la capacité à reconnaître les locuteurs à partir d'un poids de couche donné. Une simple distance cosinus est utilisée pour calculer le score de vérification pour une paire (enrollment, test).

Les données de chaque locuteur (voir section 4.2) sont divisées en deux sessions, notées $s1$ et $s2$. On obtient une paire "target", (x_i^{s1}, x_i^{s2}), par locuteur x_i. Les paires "non target", (x_i^{s1}, x_j^{s2}), sont formées en croisant la première session d'un locuteur donné avec toutes les secondes sessions des autres locuteurs.

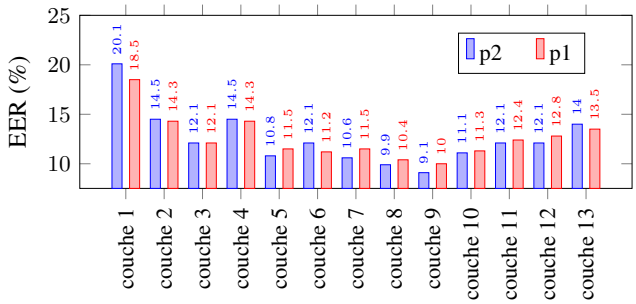

FIGURE 4 – Résultats en EER pour la vérification du locuteur

Nous présentons dans le tableau 4 l'ensemble des résultats obtenus. Les résultats sont exprimés en termes de EER (Equal Error Rate). La meilleure performance est obtenue en utilisant la couche 9 (9,07% EER pour *p2* et 10% EER pour *p1)*, montrant clairement que des informations spécifiques au locuteur peuvent être extraites des poids d'un modèle acoustique personnalisé. Afin de comparer les performances, nous avons également calculé la performance lorsque les vecteurs de poids sont utilisés directement pour calculer la distance cosinus, sans l'extracteur d'incorporation. L'EER est d'environ 48% dans ce cas pour *p2* (proche de la performance aléatoire). Ceci prouve l'efficacité de l'approche proposée pour extraire un encastrement du locuteur à partir des poids du modèle acoustique personnalisé.

6 Conclusion

Dans cette étude, nous avons montré qu'il est possible de retrouver le sexe et l'identité d'un locuteur à partir de l'analyse des modifications appliquées aux poids de son modèle acoustique personnalisé. Les expériences menées sur le jeu de données TED-LIUM3 montrent que l'information concernant le sexe du locuteur est apportée principalement par les mises à jour impactant les cinq premières couches d'un modèle acoustique HMM/TDNN composé de 13 couches cachées, alors que l'identité du locuteur est principalement intégrée dans les couches cachées intermédiaires (5 à 9). Afin d'analyser l'information liée à l'identité, nous avons également proposé une méthode originale pour construire un extracteur de représentations du locuteur à partir de matrices de poids personnalisées. Nous avons obtenu une pureté de 0,96 pour la reconnaissance du sexe sur les cinq premières couches et un EER de vérification du locuteur de 9% pour la couche 9. Ces résultats sont particulièrement intéressants pour de futurs travaux portant sur l'apprentissage distribué pour la préservation de la vie privée. Dans cette direction, nous avons proposé dans une étude en parallèle, deux modèles d'attaques pour les modèles personnalisés (Tomashenko *et al.*, 2022). L'approche utilisée pour ces modèles consiste à construire des empreintes de ces modèles personnalisés à partir des traces de leur application sur un jeu de données fixe et indépendant. Nous envisageons dans nos futurs travaux d'approfondir le travail présenté dans cet article en étudiant l'information de l'identité du locuteur sur des modèles acoustiques personnalisés entraînés sans i-vecteurs. Nous souhaitons également effectuer une analyse similaire en étudiant plus de caractéristiques du locuteur (accent, nationalité, âge, émotion, etc).

7 Remerciements

Ce travail a été financé par les projets ANR DEEP-PRIVACY (ANR18-CE23-0018) et VoicePersonae (ANR-18-JSTS-0001). Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2021-AD011012551 attribuée par GENCI.

Références

BELINKOV Y., ALI A. & GLASS J. (2019). Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition. In *Proc. Interspeech 2019*, p. 81–85.

BELINKOV Y. & GLASS J. (2017). Analyzing hidden representations in end-to-end automatic speech recognition systems. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 2438–2448.

CUI X., LU S. & KINGSBURY B. (2021). Federated acoustic modeling for automatic speech recognition. In *ICASSP*, p. 6748–6752.

GULIANI D., BEAUFAYS F. & MOTTA G. (2021). Training speech recognition models with federated learning : A quality/cost framework. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 3080–3084 : IEEE.

HARD A., PARTRIDGE K., NGUYEN C., SUBRAHMANYA N., SHAH A., ZHU P., MORENO I. L. & MATHEWS R. (2020). Training keyword spotting models on non-iid data with federated learning. In *Interspeech 2020*.

HERNANDEZ F., NGUYEN V., GHANNAY S., TOMASHENKO N. & ESTÈVE Y. (2018). TED-LIUM 3 : twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, p. 198–208 : Springer International Publishing.

LEROY D., COUCKE A., LAVRIL T., GISSELBRECHT T. & DUREAU J. (2019). Federated learning for keyword spotting. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6341–6345 : IEEE.

LI C.-Y., YUAN P.-C. & LEE H.-Y. (2020). What does a network layer hear ? analyzing hidden representations of end-to-end asr through speech synthesis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6434–6438 : IEEE.

MANSOUR Y., MOHRI M., RO J. & SURESH A. T. (2020). Three approaches for personalization with applications to federated learning. *CoRR*.

MDHAFFAR S., BONASTRE J.-F., TOMMASI M., TOMASHENKO N. & ESTÈVE Y. (2022). Retrieving speaker information from personalized acoustic models for speech recognition.

MDHAFFAR S., TOMMASI M. & ESTÈVE Y. (2021). Study on acoustic model personalization in a context of collaborative learning constrained by privacy preservation. *SPECOM*.

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* : ISCA.

PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, p. 3214–3218 : ISCA.

POVEY D., GHOSHAL A., BOULIANNE G., BURGET L., GLEMBEK O., GOEL N., HANNEMANN M., MOTLICEK P., QIAN Y., SCHWARZ P. *et al.* (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* : IEEE Signal Processing Society.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5329–5333 : IEEE.

TOMASHENKO N., MDHAFFAR S., TOMMASI M., ESTÈVE Y. & BONASTRE J.-F. (2022). Privacy attacks for automatic speech recognition acoustic models in a federated learning framework. In *ICASSP 2022*.

YU W., FREIWALD J., TEWES S., HUENNEMEYER F. & KOLOSSA D. (2021). Federated learning in ASR : Not as easy as you think. In *ITG Conference on Speech Communication*.