Scattering transform et réseaux convolutionels pour l'identification du locuteur

Wajdi Ghezaiel, Luc Brun, Olivier Lézoray, Myriam Mokhtari

Scattering transform et réseaux convolutionels pour l'identification du locuteur*

Wajdi Ghezaiel[1] Luc Brun[1] Olivier Lezoray[2] Myriam Mokhtari[1]

[1] Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, Caen France
[2] Normandie Université, UNICAEN, ENSICAEN, CNRS, GREYC Caen, France

{wajdi.ghezaiel, luc.brun, myriam.brun}@ensicaen.fr, olivier.lezoray@unicaen.fr

Résumé

Les assistants vocaux sont devenus très populaires ces dernières années. Les utilisateurs peuvent contrôler ces appareils intelligents par la voix et obtenir divers services. Combinés à la biométrie, ces dispositifs peuvent permettre de distinguer des profils utilisateurs et sécuriser l'usage de l'appareil. Dans ce scénario, quelques segments de discours de courte durée (2-4 sec.) sont utilisés pour l'authentification. Afin de limiter le nombre de paramètres utilisés pour l'apprentissage, nous proposons de combiner une Wavelet Scattering Transform (ST) et un réseau convolutif (CNN). Nos expérimentations montrent que la combinaison ST/CNN extrait efficacement les caractéristiques de l'identité du locuteur sur des discours de courte durée.

Mots Clef

Assistant vocal, identification du locuteur, réseau de neurones convolutifs, réseau hybride.

Abstract

Voice assistants have became very popular in recent years. Users can control these devices by voice and obtain various services. Combination of biometric technology not only allows to differentiate different profiles of users but also allows to secure the terminal. In this scenario, few short speech segments (2-4 sec.) are used for authentication. In order to limit the number of parameters used for learning, we propose to combine a Wavelet Scattering Transform (ST) and a convolutional network (CNN). The ST/CNN combination efficiently extracts the speaker's identity over short speeches

Keywords

Voice assistant, speaker identification, scattering transform, convolutional neural network, hybrid network.

1 Introduction

Les appareils Google Home et Amazon Alexa deviennent nos compagnons de tous les jours et de nombreuses tâches peuvent à présent être effectuées à l'aide de commandes vocales. Il s'agit notamment de jouer de la musique, de prendre un rendez-vous, de consulter la météo, de faire des achats en ligne ou de piloter des objets connectés. Certaines commandes vocales, telles que le paiement bancaire ou l'accès à certains services sensibles, nécessitent une authentification de l'utilisateur. Pour un tel scénario, il se peut qu'il n'y ait aucune contrainte sur les phrases utilisées. Ces phrases sont généralement courtes et pas nombreuses. Ceci constitue une contrainte pour un système d'authentification vocal, mais facilite l'utilisation du système d'authentification. Différentes études [1, 2, 3] ont montré que l'utilisation de discours vocaux courts peut induire une baisse des performances des systèmes d'authentification. Cette baisse de performance est principalement due à la faible quantité d'informations pertinentes sur l'identité du locuteur extraite à partir des courts discours utilisés en phase d'apprentissage ou en phase de test. L'identification du locuteur avec seulement quelques courts discours vocaux est donc un problème difficile.

Traditionnellement, les systèmes d'identification du locuteur sont basés sur l'extraction de caractéristiques vocales reposant sur la production et la perception de la parole, telles que les coefficients cepstraux (MFCC). Pendant la phase d'apprentissage, ces caractéristiques sont extraites pour la totalité des locuteurs. Un modèle de mélange Gaussien (GMM) est ensuite défini pour construire un modèle universel (UBM) [4, 5]. Chaque locuteur est représenté par une adaptation du vecteur GMM. Cependant, il a été montré [6, 7] qu'il est avantageux de traiter ces vecteurs GMM, en estimant la matrice de variabilité totale à partir du modèle UBM et utiliser l'analyse conjointe en facteurs pour extraire des vecteurs d'identité, appelés i-vecteurs. Pendant la phase d'authentification, le vecteur d'identité correspondant au discours test est comparé aux autres vecteurs de référence. La décision est faite soit par une simple distance cosinus entre les deux vecteurs, soit avec des techniques plus complexes telles que l'analyse discriminante linéaire probabiliste (PLDA) [8]. Les performances de ces deux méthodes de référence peuvent diminuer en cas de discours courts [9, 10].

Ces dernières années, l'apprentissage profond a fait son ap-

*Financé par BPI France, projet HomeKeeper.

parition dans de nombreux domaines liés à la reconnaissance de formes. Il a connu un succès remarquable pour la reconnaissance d'images [11] et le traitement du langage naturel [12]. Une tendance similaire a été observée dans l'identification du locuteur. Les réseaux neuronaux profonds (DNN) ont été utilisés avec les i-vecteurs pour calculer les statistiques de Baum-Welch [13], ou pour l'extraction de caractéristiques au niveau de chaque trame [14]. Des DNN ont également été proposés pour la discrimination des locuteurs [15, 16]. Récemment, un nombre croissant d'études ont tenté d'utiliser un réseau neuronal convolutif (CNN) [17] dans de nombreuses tâches en lien avec la parole [18, 19]. Certains travaux ont proposé d'apprendre directement les réseaux avec des spectrogrammes [20, 21] ou même avec un signal de parole [22, 23]. Les CNN ont une architecture adaptée au traitement direct des échantillons de parole, car le partage de poids, les filtres locaux et le pooling constituent des outils précieux pour découvrir des représentations robustes et invariantes. Cependant, afin de parvenir à un apprentissage efficace les réseaux CNN nécessitent de nombreux exemples étiquetés pour l'apprentissage ainsi que des ressources de calcul qui peuvent être considérables. Dans un contexte où seules quelques données étiquetées de courte durée sont disponibles, l'entraînement devient difficile et nécessite beaucoup de régularisation.

L'extraction de caractéristiques pertinentes est un point critique dans les systèmes d'authentification vocaux. Elle permet de réduire la taille des données requises pour l'apprentissage. La wavelet scattering transform (ST), ou transformée diffusion par ondelette, a connu un succès significatif dans diverses tâches de classification de signaux audio [24] et biomédicaux [25]. Sa structure est celle d'un réseau neuronal convolutif [26], mais avec des filtres fixes et non appris. Ce dernier point est important lorsque seuls quelques échantillons d'entraînement sont disponibles. Plus précisément, cette transformation opère des convolutions avec des ondelettes et des transformations non-linéaires pour assurer une invariance aux décalages temporels et une stabilité aux distorsions temporelles [27].

Les représentations par la transformée diffusion par ondelette peuvent être insérées dans n'importe quel système de classification ou de régression, qu'il soit profond ou peu profond. Les premières études d'Andén et Mallat [24] utilisent les ST avec des machines à supports vecteurs (SVM) avec des noyaux linéaires ou Gaussiens. Pour la reconnaissance vocale avec des bases de données importantes, le ST est utilisé avec cinq couches de réseaux neuronaux profonds (DNN) ou des réseaux convolutionnels profonds (ConvNets). Cette architecture n'a apporté que des améliorations marginales de la précision [28, 29]. Cependant, dans le cadre du challenge de reconnaissance vocale avec Zéros ressources [30], dont l'objectif est de découvrir de manière non supervisée des sous-mots et des mots à partir d'un discours continu, l'association du ST à un réseau siamois a permis un gain substantiel dans le compromis entre la discrimination inter-classes et la robustesse aux changements de locuteur [31].
Les caractéristiques de la transformée diffusion par ondelette ont fourni des résultats prometteurs sur l'ensemble de données TIMIT [32] pour la classification des phonèmes [23] et la reconnaissance de parole [29]. Dans cet article, nous explorons l'utilisation de la transformée diffusion par ondelette (ST) pour l'extraction des caractéristiques vocales tout en le couplant avec un réseau convolutionel CNN pour l'identification du locuteur. Le système proposé doit nécessiter un nombre limité de discours vocaux afin de fonctionner efficacement sur des assistants vocaux. Nous remplaçons alors la première couche convolutionnelle du réseau CNN par la transformée diffusion par ondelette. Dans ce réseau hybride, les coefficients de ST générés dans les premières couches capturent l'énergie dominante contenue dans les discours d'entrée.
Le reste de cet article est organisé comme suit. La section 2 traite de la transformée diffusion par ondelette. La section 3 décrit l'architecture hybride proposée, qui est une combinaison d'une transformée diffusion par ondelette et d'un réseau convolutionnel. La section 4 présente les résultats obtenus par le système proposé ainsi que ceux fournis par les systèmes de la littérature.

2 Wavelet Scattering Transform : Transformée diffusion par ondelette

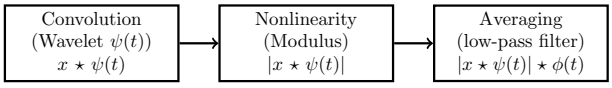

FIGURE 1 – Processus de la transformée diffusion par ondelette, où x est la donnée d'entrée, ψ une fonction d'ondelette et ϕ un filtre passe-bas.

La transformée diffusion par ondelette (ST) [24, 33], est une représentation profonde, obtenue par application itérative du module de la transformée en ondelettes. Elle a été définie de manière à être invariante aux translations temporelles du signal d'entrée et stable aux petites déformations. Les auteurs [24, 33] ont démontré que cette transformation à base d'ondelettes permet d'extraire des caractéristiques significatives du signal aux différentes échelles de décomposition. La ST a été appliquée avec succès à différentes tâches de classification, de textures [33, 34], de chiffres [33], de sons [24] ou à des ensembles complexes d'images [25]. De plus, il a été prouvé [35] que les coefficients de diffusion des ondelettes sont plus informatifs qu'une transformée de Fourier lorsqu'il s'agit de signaux courts pouvant subir de petites déformations et rotations.
La ST consiste en une cascade de transformations d'ondelettes et de transformations non-linéaires sur le module. Pour produire la ST d'un signal d'entrée x, trois opérations

successives sont nécessaires : convolution, non-linéarité et calcul de la moyenne (figure 1). Les coefficients de la ST sont obtenus en faisant la moyenne des coefficients du module d'ondelettes par un filtre passe-bas ϕ. Soit une onde-

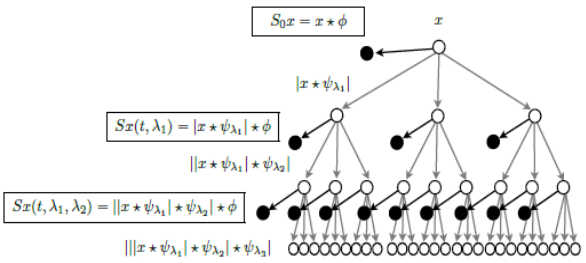

FIGURE 2 – Représentation hiérarchique des coefficients de diffusion sur plusieurs couches [24].

lette $\psi(t)$ correspondant à un filtre passe-bande avec une fréquence centrale normalisé à 1, et $\psi_\lambda(t)$ une base de filtres construite par dilatation de l'ondelette :

$$\psi_\lambda(t) = \lambda\psi(\lambda t) \tag{1}$$

où $\lambda = 2^{\frac{j}{Q}}, \forall j \in \mathbb{Z}$ avec Q représentant le nombre d'ondelettes par octave.
La largeur de bande de l'ondelette $\psi(t)$ est de l'ordre de $\frac{1}{Q}$, et par conséquent, la banque de filtres est composée de filtres passe-bande qui sont centrés dans le domaine fréquentiel en λ et ont une largeur de bande de fréquence $\frac{\lambda}{Q}$.
À l'ordre zéro, nous avons un seul coefficient donné par $S_0x(t) = x \star \phi(t)$ où $\star$ représente la convolution. Ce coefficient est proche de zéro pour les signaux audio. Au premier ordre, nous avons :

$$S_1x(t, \lambda_1) = |x \star \psi_{\lambda_1}| \star \phi(t) \tag{2}$$

Les coefficients de second ordre capturent les modulations d'amplitude haute fréquence sur chaque bande de fréquence de la première couche et sont obtenus par

$$S_2x(t, \lambda_1, \lambda_2) = ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \tag{3}$$

La figure 2 montre la hiérarchie des coefficients de ST. Cela ressemble un peu à la structure des réseaux neuronaux profonds, bien que dans la ST, chaque couche fournit une sortie, alors que la seule sortie de la plupart des réseaux neuronaux profonds est celle de la dernière couche. Cette décomposition en coefficients de diffusion de premier et de second ordre est appliquée au signal. Les coefficients de second ordre sont normalisés par les coefficients de premier ordre, afin de s'assurer qu'un ordre supérieur de décomposition dépend de la modulation d'amplitude du signal vocal. Les premier et deuxième ordres de la ST sont concaténés pour former un vecteur de caractéristiques pour une trame. Les caractéristiques des coefficients de la ST à la première échelle, correspondent aux coefficients MFCC

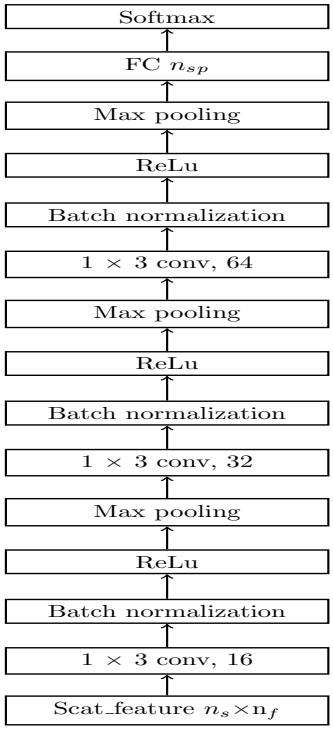

FIGURE 3 – Le réseau hybride proposé.

[24]. Les coefficients ST de la seconde échelle préservent plus de détails dans le signal vocal [24]. Cette représentation est invariante aux décalages temporels et stable aux déformations. Ainsi, pour assurer une invariance à des translations de fréquence, la ST est effectuée sur une échelle log de fréquence. Le logarithme est appliqué à chacun des coefficients du ST. Il est donc localement invariant aux décalages temporels et de fréquence et stable aux déformations temporelles et fréquentielles.

3 Architecture hybride de réseau

Le réseau hybride proposé est composé d'une ST pour l'extraction des informations relatives à l'identité du locuteur, suivi d'un réseau convolutionnel CNN pour la classification. Nous proposons donc d'initialiser la première couche de notre CNN avec les vecteurs issus de la ST. Nous utilisons deux familles d'ondelettes, l'ondelette de Morlet et l'ondelette de Gabor, pour obtenir les coefficients caractéristiques ST. Nous utilisons 8 filtres par octave pour le premier niveau et 1 filtre par octave au second niveau. Cette configuration a été choisie pour que la première décomposition ST corresponde à la résolution en fréquence des filtres à l'échelle de Mel. La deuxième décomposition de ST permet de récupérer les informations perdues. Par conséquent, la représentation du signal de parole utilisant deux itérations de la ST étend la représentation MFCC et conserve toutes les caractéristiques du signal. Ces coefficients de ST sont calculés à l'aide du toolbox ScatNet [24]. On obtient ainsi des matrices de paramètres de dimension

$n_s \times n_f$. Pour chaque trame, le vecteur final comporte une normalisation de la transformée en logarithme des vecteurs issus de la ST (section 2). Ces vecteurs de ST jouent le rôle d'un sous-échantillonnage pour la première couche du CNN.

Layer name	ScatCNN	Output
Input	inputlayer	$n_s \times n_f \times 1$
Conv1 block	conv1D, 3×1,16 bn relu	$n_s \times n_f \times 16$
Pooling	maxpool, 2×1, stride (2,1)	$n_s/2 \times n_f/2 \times 16$
Conv2 block	conv1D, 3×1,32 bn relu	$n_s/2 \times n_f/2 \times 32$
Pooling	maxpool, 2×1, stride (2,1)	$n_s/4 \times n_f/4 \times 32$
Conv3 block	conv1D, 3×1,64 bn relu	$n_s/4 \times n_f/4 \times 64$
Pooling	maxpool, 2×1, stride (2,1)	$n_s/8 \times n_f/8 \times 64$
Embedding	fc,n_{sp}	n_{sp}
Loss	softmax	

TABLE 1 – L'architecture CNN-ST. Chaque ligne spécifie les filtres convolutifs, leur taille et les filtres.

Le CNN proposé est composé de trois couches de convolution et d'une couche entièrement connectée FC. Chaque couche de convolution est formée d'un filtre 1D de longueur 3 auquel est associée une normalisation. Chaque couche convolutionnelle est suivie d'une couche de mise en commun -max-pooling- de taille 2×1 et de pas 2×1. Ces couches de convolution disposent respectivement de 16, 32 et 64 filtres. Une couche entièrement connectée avec n_{sp} neurones, où n_{sp} est le nombre de locuteurs à identifier, est connectée à la couche softmax. Nous utilisons des unités linéaires rectifiées comme fonctions d'activation dans toutes les couches. La figure 3 et le tableau 1 montrent avec plus de détail l'architecture du réseau proposé. Le nombre de paramètres est de l'ordre 18 millions pour une entrée de dimension $433 \times 16 \times 1$.

4 Expériences

Cette section décrit les expériences et les résultats obtenus avec notre approche.

4.1 Base de données

Deux bases de données sont utilisées dans les expériences : TIMIT [32] et LibriSpeech [36]. La base TIMIT contient des enregistrements de qualité studio de 630 locuteurs (192 femmes, 438 hommes), échantillonnés à 16 kHz, et couvrant les huit principaux dialectes de l'anglais américain. Chaque locuteur lit dix phrases phonétiquement riches. Nous ne prenons en compte que 462 locuteurs de TIMIT afin de nous limiter uniquement aux discours de courte durée. Pour chaque locuteur, il y a 8 phrases avec 5 phrases "SX" et 3 phrases "SI". Les phrases "SX" ont une durée de 2s à 5s, elles sont utilisées pour entraîner le système. Les phrases "SI" ont une durée de 2s à 7s, elles sont utilisées pour le test.
La base de données LibriSpeech est constituée de livres audio lus à haute voix par 2483 locuteurs, dont 1281 hommes et 1202 femmes qui ont enregistré leur voix spontanément. Le signal de parole est généralement propre, mais le dispositif d'enregistrement et les conditions de canal varient beaucoup entre les différentes voix et les différents locuteurs. Nous avons décidé de garder 7 enregistrements de chaque locuteur pour l'entraînement, et 3 enregistrements pour le test. Pour les deux ensembles de données, la durée des phrases d'entraînement est d'environ 10 à 12 secondes pour chaque locuteur et la durée moyenne des phrases de test est de 2 à 5 secondes. Les énoncés d'une durée inférieure à 6 secondes représentent environ 87% des données. Les expériences ne sont menées que dans des conditions de discours de courte durée, et notre méthode n'est testée qu'avec des discours de courte durée (aussi bien pour le test que l'apprentissage), de $2s$ à $4s$.
Aucun pré-traitement spécifique à la parole n'a été appliqué (par exemple, suppression du silence, détection et suppression de la parole non voisée) pour les deux bases.

4.2 Cadre expérimental

La première couche de ST est composée de 8 ondelettes de type Gabor par octave. La seconde couche est formée par une seule ondelette de type Morlet par octave. La longueur de la fenêtre a été fixée à 500 ms. Par la suite, les coefficients ST sont normalisés et transformés en logarithme. Avec ce réglage, le nombre de coefficients ST correspondant à une seule trame est de 77 et 356 pour le premier et le second ordre respectivement. La descente de gradient stochastique a été utilisée pour optimiser le réseau avec un taux d'apprentissage de 0,001 et un momentum de 0,9. Le réseau est entrainé avec un nombre de cycles d'apprentissage égal à 30 et un batch de taille 64. Notre implémentation est basée sur les boîtes à outils Matlab pour le scattering (`Scatnet` [1]) et l'apprentissage profond. Le temps de calcul des coefficients ST pour chaque trame est d'environ 10,6 ms. Le temps total d'apprentissage et de test dépend du nombre de locuteurs et de discours utilisés. Les expériences dans la base de données TIMIT prennent environ 37 minutes pour l'entraînement et 25 minutes pour les tests.

4.3 Systèmes de la littérature

Dans cet article, nous proposons de comparer notre système avec les systèmes SincNet [37] et CNN-Raw [38] pour l'identification de locuteurs. SincNet est une nouvelle architecture de réseau neuronal, qui utilise directe-

1. `http://www.di.ens.fr/data/software/scatnet/`.

Layer name	Sincnet	CNN
Conv1 block	conv1D,251 $\times$ 1, 80 relu	conv1D,3 $\times$ 1, 80 relu
Pooling	maxpool,5 $\times$ 1, stride (1,2)	maxpool,3 $\times$ 1, stride (2,1)
Conv2 block	conv1D,5 $\times$ 1,60 relu	conv1D,5 $\times$ 1, 80 relu
Pooling	maxpool,3 $\times$ 1, stride(2,1)	maxpool,3 $\times$ 1, stride(2,1)
Conv3 block	conv1D,5 $\times$ 1,60 relu	conv1D,5 $\times$ 1, 60 relu
Pooling	maxpool,3 $\times$ 1, stride(2,1)	maxpool,3 $\times$ 1, stride(2,1)
Embedding Embedding Embedding Embedding	fc, 2048 bn fc, 2048 bn fc, 2048 bn fc, 2484	fc, 2048 bn fc, 2048 bn fc, 2048 bn fc, 2484
Loss	softmax	softmax

TABLE 2 – Architecture de Sincnet and CNN.

ment le signal de parole en entrée. La première couche convolutionnelle de SincNet est composée des fonctions 1D Sinc. SincNet convolue directement le signal de parole avec un ensemble de fonctions Sinc. Ces fonctions Sinc jouent le rôle des filtres passe-bande. Les filtres sont initialisés à l'aide de la banque de filtres à fréquence Mel et leurs fréquences de coupure basses et hautes sont adaptées avec une rétro-propagation standard comme toute autre couche. La première couche effectue des convolutions basées sur Sinc, en utilisant 80 filtres de longueur 251. Les deux autres couches utilisent 60 filtres de longueur 5. Ensuite, trois couches entièrement connectées, composées de 2048 neurones et normalisées, sont appliquées. Toutes les couches cachées utilisent une fonction d'activation non-linéaire ReLU. La classification binaire au niveau de la trame est effectuée en appliquant un classificateur softmax et un critère d'entropie croisée [37].

Dans le système CNN-Raw, le signal de parole est directement envoyé à la première couche du réseau. Trois couches de convolution sont utilisées pour effectuer l'extraction des caractéristiques. Chaque couche de convolution est composée de 80 filtres suivis d'un max-pooling. Ensuite, trois couches entièrement connectées, composées de 2048 neurones et normalisées, sont appliquées. Toutes les couches cachées utilisent des fonctions d'activation non-linéaire de type ReLU. La classification binaire au niveau de la trame est effectuée en appliquant un classificateur softmax et un critère d'entropie croisée [38]. Le tableau 2 résume le nombre de filtres et la taille des noyaux utilisés dans sincnet et CNN. Les deux réseaux sont entraînés avec un nombre de cycles d'apprentissage égale à 800 et un batch de taille 128.

	SincNet	CNN	CNN-ST
Paramètres $\times 10^6$	26,5	27,6	**18,1**

TABLE 3 – Nombre de paramètres d'apprentissage de l'architecture proposée et des systèmes de la littérature.

4.4 Résultats

Le tableau 3 résume le nombre de paramètres d'apprentissage de l'architecture hybride et les autres systèmes de la littérature. Nous observons une diminution de 33% du nombre de paramètres d'apprentissage par rapport aux systèmes Sincnet et CNN-raw. Ceci se traduit par la rapidité d'apprentissage de notre architecture avec un nombre de cycles d'apprentissage égal à 30. Afin de comparer les systèmes d'identification du locuteur, le taux d'identification correcte par locuteur est utilisé. Les résultats sont réunis dans le tableau 4 sur les deux ensembles de données TIMIT et Librispeech. Nous avons entrainé et testé dans un premier temps les trois systèmes avec des portions de durée 4s. Les résultats de ce tableau montrent que notre réseau hybride CNN-ST est plus performant que les systèmes SincNet et CNN-Raw sur des énoncés courts. Sur l'ensemble de données TIMIT, notre système réalise une amélioration relative d'environ 18% par rapport à CNN-Raw et 15 % par rapport à SincNet. De plus, sur l'ensemble de données LibriSpeech, notre système réalise une amélioration relative d'environ 18 % par rapport à CNN-Raw et de 23% par rapport à SincNet. Le tableau montre que le remplacement de la première couche de convolution de CNN par des filtres bien définis améliore les taux d'identification. Par la suite nous avons entrainé et testé les systèmes CNN-raw et sincnet avec la totalité du discours, nous observons toujours une meilleure performance de notre architecture malgré son utilisation dans des conditions d'apprentissage et de test limitées à 4 s.

	LibriSpeech	TIMIT
CNN-raw-4s	66,39	46.67
CNN-raw-totale	69.82	60.53
SincNet-raw-4s	70.81	49.81
SincNet-raw-totale	79.32	61.81
CNN-ST-4s	**88.04**	**64.29**

TABLE 4 – Taux d'identification correcte (%) du système proposé d'identification du locuteur et des systèmes de la littérature entrainés et testés avec des durées variables.

Nous examinons plus en détail les performances de notre système dans le tableau 5. Dans ce tableau, nous évaluons l'effet de la durée des discours des locuteurs utilisés en apprentissage sur les performances. Nous varions le nombre

Test	Durée du discours d'apprentissage		
	8s	12s	Totalité
2s	77.16	78.03	79.86
4s	86.27	87.63	88.04

TABLE 5 – Taux d'identification correcte (%) du système proposé sur l'ensemble de données LibriSpeech (2484 locuteurs) entrainé et testé avec des énoncés de durée de 4s et 2s.

de discours utilisés en apprentissage pour avoir une durée totale de 8s ou 12s par locuteur. Par exemple, un entrainement avec des discours de 2s pour une durée totale de 8s implique donc l'utilisation de 4 discours par locuteur. Ce tableau indique les taux d'identification correcte pour des discours de test de 2s et 4s. Nous remarquons que la variation du nombre de discours en phase d'apprentissage affecte peu les performances. Nous obtenons toujours des bons taux d'identification soit avec une portion des donnée d'apprentissage ou avec la totalité des discours. La durée totale des données d'apprentissage par locuteur est de l'ordre de 14s. De même nous remarquons la robustesse de l'architecture proposée à la variation de durée des données de test. Le tableau 5 montre l'effet de la durée du discours de test sur les taux d'identification, nous remarquons une amélioration de 10% avec des discours de 4s par rapport à 2s.

5 Conclusion

Dans cet article, nous avons proposé un système d'identification du locuteur qui apprend les informations discriminantes du locuteur directement à partir de discours courts en utilisant la transformée diffusion par ondelette ST et un CNN. Les caractéristiques issues de la ST fournissent une description stable des informations d'identité du locuteur. Les expériences menées sur les bases de TIMIT et Librispeech indiquent que la méthode proposée est très performante pour les tâches d'identification du locuteur avec des discours courts et des nombres d'échantillons d'apprentissages limités. Nous avons comparé l'efficacité de notre système par rapport aux méthodes SincNet et CNN-Raw dans les même conditions. Nos résultats montrent que notre méthode hybride CNN-ST apporte des améliorations significatives par rapport à ces deux méthodes. Ce travail pourrait être étendu afin de pouvoir utiliser en entrée des discours de durée variable.

Références

[1] A. Poddar, M. Sahidullah, and G. Saha, "Speaker verification with short utterances : a review of challenges, trends and opportunities," *IET Biometrics*, vol. 7, no. 3, pp. 91–101, 2018.

[2] Rohan Kumar Das and S. R. M. Prasanna, "Speaker verification for variable duration segments and the effect of session variability," *Lecture Notes in Electrical Engineering*, pp. 193–200, 2015.

[3] A. Kanagasundaram, R. Vogt, D. Dean, and S. Sridharan, "i-vector based speaker recognition on short utterances," in *Proc. of Interspeech*, 2011.

[4] D. A. Reynolds and Richard C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," in *IEEE Transactions on Speech and Audio Processing*. IEEE, 1995, vol. III, pp. 72–83.

[5] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[7] Wei Li, Tianfan Fu, and Jie Zhu, "An improved i-vector extraction for speaker verification," *EURASIP Journal on Audio, Speech and Music Processing*, pp. 1–9, 2015.

[8] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. of International Conference on Computer Vision*, 2007.

[9] X. Zhao and D. Wang, "Analyzing noise robustness of mfcc and gfcc features in speaker identification," in *Proc. of ICASSP*, 2013, pp. 7204–7208.

[10] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proc. of ISCA*, 2011, pp. 2341–2344.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of Advances in Neural Information Processing Systems*, 2012.

[12] R. Collobert and J. Weston, "A unified architecture for natural language processing : Deep neural networks with multitask learning," in *Proc. of the International Conference on Machine Learning*, 2008.

[13] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Proc. of Speaker Odyssey*, 2014.

[14] S. Yaman, J. W. Pelecanos, and R. Sarikaya, "Bottleneck features for speaker recognition," in *Proc. of Speaker Odyssey*, 2012, pp. 105–108.

[15] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small foot-print text-dependent speaker verification," in *Proc. of ICASSP*, 2014, pp. 4052–4056.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors : Robust dnn embeddings for speaker recognition," in *Proc. of ICASSP*, 2018.

[17] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," in *The hand-book of brain theory and neural networks*, 1995, vol. 3361, p. 1995.

[18] Abdel-Hamid O., Mohamed Abdel-rahman, Jiang Hui, Deng Li, Penn Gerald, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE Transactions on Audio, Signal, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[19] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. of ICASSP*, 2013.

[20] C. Zhang, K. Koishida, and J. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embedding," *IEEE Transactions on Audio, Signal, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.

[21] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb :a large-scale speaker identification dataset," in *Proc. of Interspeech*, 2017.

[22] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of CNN-based speech recognition system using raw speech as input," in *Proc. of Interspeech*, 2015.

[23] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "To-wards directly modeling raw speech signal for speaker verification using CNNs," in *Proc. of ICASSP*, 2018.

[24] Joakim Andén and Stéphane Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[25] Václav Chudáček, Joakim Andén, Stéphane Mallat, Patrice Abry, and Muriel Doret, "Scattering transform for intrapartum fetal heart rate variability fractal analysis : A case-control study," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1100–1108, 2014.

[26] E. Oyallon and S. Mallat, "Deep roto-translation scattering for object classification," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2865–2873.

[27] Stéphane Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[28] P. Fousek, P. Dognin, and V. Goel, "Evaluating deep scattering spectra with deep neural networks on large-scale spontaneous speech task," in *Proc. of ICASSP*, 2015, p. 54.

[29] V. Peddinti, T. N. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel, "Deep scattering spectrum with deep neural network," in *Proc. of ICASSP*, 2014, pp. 361–364.

[30] M. Versteegh, R. Thiollière, T. Schatz, X. Nga Cao, X. Anguera, A. Jansen, and E. Dupoux, "Deep scattering spectrum with deep neural networks," in *Proc. of Interspeech*, 2015, p. 55.

[31] N. Zeghidour, G. Synnaeve, M. Versteegh, and E. Dupoux, "A deep scattering spectrum—Deep Siamese network pipeline for unsupervised acoustic modeling," in *Proc. of ICASSP*, 2016, pp. 4965–4969.

[32] L. Lamel, R. Kassel, and S. Seneff, "Speech Database Development : Design and Analysis of the Acoustic-Phonetic Corpus," in *Proc. of DARPA Speech Recognition Work-shop*, 1986.

[33] Joan Bruna and Stéphane Mallat, "Invariant scattering convolution networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[34] Joan Bruna, Arthur Szlam, and Yann LeCun, "Learning stable group invariant representations with convolutional networks," in *arXiv preprint arXiv :1301.3537*, 2013.

[35] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko, "Scaling the Scattering Transform : Deep Hybrid Networks," in *arXiv preprint arXiv :1703.08961*, 2017.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech : An ASR corpus based on public domain audio books," in *Proc. of ICASSP*, 2015, pp. 5206–5210.

[37] M. Ravanelli and Y. Bengio, "Speaker Recognition from raw waveform with SincNet," in *Proc. of SLT*, 2018.

[38] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, "On Learning Vocal Tract System Related Speaker Discriminative Information from Raw Signal Using CNNs," in *Proc. of Interspeech*, 2018.