XXXIIe Journées d'Études sur la Parole
4-8 June 2018, Aix-en-Provence, France

Impact des techniques d'adaptation au locuteur dans l'espace des paramètres pour des modèles acoustiques purement neuronaux

Natalia Tomashenko Yannick Estève
LIUM, Le Mans Université, France
prenom.nom@univ-lemans.fr

RÉSUMÉ

Cet article explore l'utilisation de techniques d'adaptation au locuteur pour des modèles acoustiques *bidirectionnels* de type *long short term memory* (BLSTM) entrainés avec la fonction objective dite de *classification temporelle connectionniste* (CTC). Les modèles acoustiques BLSTM-CTC prennent de plus en plus d'importance dans les systèmes de reconnaissance automatique de la parole, mais peu d'études ont été menées jusqu'ici pour y appliquer des techniques d'adaptation au locuteur. Dans cet article, nous explorons l'utilisation de trois techniques différentes : l'approche par *feature space maximum likelihood linear regression* (fMLLR), celle s'appuyant sur l'utilisation de i-vectors, et une approche exploitant la technique d'adaptation *maximum a posteriori* (MAP) appliquée sur des modèles gaussiens dont sont dérivés des paramètres fournis au modèles acoustiques neuronaux. Enfin, cette étude présente une comparaison du comportement des modèles BLSTM-CTC avec celui de modèles markoviens associés à un *time-delay neural network* (TDNN).

ABSTRACT

Exploration of feature-space speaker adaptation techniques for end-to-end acoustic models.

This paper investigates speaker adaptation techniques for *bidirectional long short term memory* (BLSTM) recurrent neural network based acoustic models trained with the *connectionist temporal classification* (CTC) objective function. BLSTM-CTC AMs play an important role in end-to-end automatic speech recognition systems. However, there is a lack of research in speaker adaptation algorithms for these models. We explore three different feature-space adaptation approaches for CTC acoustic models : feature-space maximum linear regression, i-vector based adaptation, and maximum a posteriori adaptation using GMM-derived features. In addition, the adaptation behavior is compared for BLSTM-CTC models and time-delay neural network (TDNN) models trained with the cross-entropy criterion.

MOTS-CLÉS : Adaptation au locuteur, reconnaissance de la parole de bout en bout, paramètres acoustiques dérivés de GMM, réseaux de neurones profonds, modèles acoustiques.

KEYWORDS: Speaker adaptation, end-to-end speech recognition, GMM-derived features, deep neural network, acoustic model.

10.21437/JEP.2018-64

1 Introduction

Plusieurs approches neuronales de type bout en bout (*end-to-end*) ont récemment été proposées dans la littérature pour la reconnaissance automatique de la parole (Hannun *et al.*, 2014; Bahdanau *et al.*, 2016). Les modèles acoustiques (AMs) de type bout en bout tentent de produire des séquences de phonèmes ou de graphèmes à partir du signal de parole à l'aide d'architectures purement neuronales (Chorowski *et al.*, 2014; Graves & Jaitly, 2014; Miao *et al.*, 2015). Ils présentent une alternative à l'approche hybride devenue classique qui associe modèles de Markov cachés et réseaux de neurones profonds (HMM-DNNs).

L'adaptation au locuteur est un composant essentiel des modèles acoustiques hybrides HMM-DNNs à l'état de l'art, et plusieurs techniques d'adaptation ont été proposées pour les DNNs. Ces techniques comprennent la transformation linéaire, qui peut être appliquée à différents niveaux d'un système HMM-DNN (Gemello *et al.*, 2006), les techniques de régularisation, comme la régularisation L2-prior (Liao, 2013) ou la régularisation par divergence de Kullback-Leibler (Yu *et al.*, 2013), l'adaptation de l'espace du modèle (Swietojanski & Renals, 2014), l'apprentissage multitâche (Price *et al.*, 2014), l'adaptation factorisée (Li *et al.*, 2014), l'adaptation par codes de locuteurs (Xue *et al.*, 2014), l'utilisation de paramètres auxiliaires, comme les i-vecteurs (Saon *et al.*, 2013), les paramètres acoustiques dérivés de mélanges de modèles gaussiens (GMMD) (Tomashenko & Khokhlov, 2014), et bien d'autres. Pourtant, la majorité des travaux publiés au sujet des modèles acoustiques de bout en bout n'utilise pas de techniques d'adaptation au locuteur.

L'objectif de cet article est d'étudier l'impact de cette adaptation lorsqu'elle est utilisée pour des modèles acoustiques neuronaux de bout en bout. Dans notre étude, nous prenons l'exemple des modèles acoustiques *bidirectionnels* de type *long short term memory* (BLSTM) entrainés avec la fonction objective dite de *classification temporelle connectionniste* (CTC). Dans la suite de cet article nous nommerons ces modèles les modèles CTC. Pour évaluer cet impact, nous avons implémenté trois différentes techniques d'adaptation au locuteur pour ce type de modèles acoustiques, et avons mis en place une analyse expérimentale de ces méthodes. De plus, nous souhaitons comparer l'impact de ces techniques d'adaptation sur les modèles CTC à leur impact sur des modèles de Markov cachés associés à un *time-delay neural network* (TDNN) appris à l'aide du critère d'entropie croisée (CE).

La suite de l'article est organisée comme suit. Un rapide survol des modèles acoustiques neuronaux de bout en bout est présenté en section 2, et les résultats expérimentaux sont donnés en section 3. Enfin, une conclusion est fournie en section 4.

2 Reconnaissance de la parole neuronale de bout en bout

Une des premières avancées qui a permis de se rapprocher de systèmes de reconnaissance de la parole de type bout en bout a été proposée dans (Graves *et al.*, 2013) où, pour la tâche de reconnaissance de phonèmes, un réseau de neurones récurrent profond de type BLSTM est appris afin de transformer directement les séquences d'observations acoustiques en phonèmes. Cette proposition s'appuie sur la fonction de coût objective CTC (Graves *et al.*, 2006). C'est ce type de modèle CTC qui est utilisé dans notre étude. D'autres approches de type bout en bout ont également été présentées dans la littérature, comme les systèmes de type encodeur/décodeur avec mécanisme d'attention (Chorowski *et al.*, 2014; Bahdanau *et al.*, 2016), ou les réseaux de neurones convolutifs (Collobert *et al.*, 2016).

2.1 LSTMs bidirectionnels profonds

Les réseaux de neurones récurrents (RNNs) sont une extension des réseaux de neurones profonds (DNN) sur lesquels sont ajoutées des connections entre différents types d'unités neuronales, en particulier des connections vers des couches cachées d'états antérieurs. L'utilisation de la récurrence à travers la dimension temporelle permet aux RNNs de modéliser la dynamique du comportement d'un phénomène dans le temps. Afin de capturer l'information sur l'ensemble d'une séquence d'entrée, une architecture de réseau de neurones récurrent bidirectionnel (BRNN) a été proposée dans (Schuster & Paliwal, 1997). Dans les BRNNs, les données sont traitées dans deux directions (avant et arrière) à l'aide de deux couches cachées (une pour le traitement vers l'avant, l'autre vers l'arrière) qui alimentent la même couche de sortie. Les systèmes de reconnaissance de la parole à l'état de l'art utilisent des architectures neuronales profondes avec plusieurs couches cachées. Les sorties des couches cachées *forward* (vers l'avant) et *backward* (vers l'arrière) à l'instant t sont concaténées : cette concaténation devient l'entrée des prochaines couches récurrentes. L'apprentissage des modèles RNN s'effectue généralement en appliquant l'algorithme d'apprentissage de rétropropagation à travers le temps (BPTT). Cependant, faire apprendre à des RNNs les dépendances temporelles longue distance peut devenir difficile en raison des problèmes de disparition et d'explosion du gradient (Bengio *et al.*, 1994). Pour éviter ce problème, les unités neuronales de type *long short-term memory* (LSTM) ont été introduites dans (Hochreiter & Schmidhuber, 1997). Dans le cadre de la reconnaissance automatique de la parole de bout en bout, les unités LSTM sont utilisées comme éléments de base des BRNNs (Miao *et al.*, 2015, 2016; Graves *et al.*, 2013).

2.2 Classification temporelle connectionniste

Avec l'approche CTC, l'alignement entre les éléments d'entrée et les étiquettes de sortie est inconnu. L'approche CTC peut être implémentée avec une couche de sortie de type *softmax* qui utilise une unité supplémentaire pour l'étiquette vide $\emptyset$. Le symbole $\emptyset$ correspond à l'émission d'aucune sortie et est utilisé pour estimer la probabilité de ne pas proposer d'étiquette de sortie à un instant donné. Le réseau de neurones est alors appris de manière à maximiser sur les données d'apprentissage la log-probabilité de toutes les séquences de sortie valides. L'ensemble des séquences valides d'étiquettes pour une séquence d'entrée est défini par l'ensemble de toutes les séquences possibles d'étiquettes telles que ces séquences soient construites dans le bon ordre, tout en acceptant les répétitions et l'étiquette $\emptyset$ entre deux étiquettes. Ces cibles pour l'apprentissage CTC peuvent être calculées en utilisant des transducteur a états finis (FSTs) et l'algorithme *forward-backward* peut être utilisé pour calculer la fonction de coût CTC. Aucune probabilité de transition d'états ou d'états initiaux n'est nécessaire pour l'approache CTC, au contraire de l'approche hybride DNN-HMM.

3 Résultats expérimentaux

Cette étude concerne principalement les techniques d'adaptation dans l'espace des paramètres pour les modèles neuronaux de bout en bout. Trois techniques d'adaptation des AMs ont été explorées dans nos expériences :

1. l'approche par fMLLR (Gales, 1998),
2. l'utilisation de i-vecteurs (Senior & Lopez-Moreno, 2014),
3. l'utilisation de l'adaptation MAP (Gauvain & Lee, 1994) appliquées à des paramètres GMMD (Tomashenko & Khokhlov, 2014, 2015; Tomashenko *et al.*, 2016b,c,a).

3.1 Données expérimentales

Les expériences ont été menées sur le corpus TED-LIUM (Rousseau *et al.*, 2014). Nous avons utilisé la dernière (seconde) version de ce corpus. Ce jeu de données publiquement disponible contient 1495 présentations des conférences TED, correspondant à 207 heures de parole pour 1242 locuteurs, enregistrées en 16kHz. Pour les expériences avec l'approche SAT (*speaker adaptative training*) et pour l'adaptation, nous avons retiré du corpus original les données correspondant à des locuteurs y apparaissant moins de 5 minutes. Le reste du corpus a été divisé en 4 parties : corpus d'apprentissage, corpus de développement, et deux corpus de test. Les caractéristiques générales des ensembles de données obtenus sont présentées dans la table 1. Pour l'évaluation, un modèle de langage 4-gram

Caractéristiques	Train	Dev.	Test_1	Test_2
Durée totale, en heures	171,66	3,49	3,49	4,90
Durée moyenne par locuteur, en minutes	10,0	15,0	15,0	21,0
Nombre de locuteurs	1 029	14	14	14
Nombre de mots	-	36 672	35 555	51 452

TABLE 1 – Statistiques du corpus de données.

avec un vocabulaire de 152k mots a été utilisé. Ce modèle de langage est proche du modèle "*small*" actuellement fourni dans la recette Kaldi *tedlium s5_r2*. Plus de détails sur ces données expérimentales sont fournis dans (Tomashenko *et al.*, 2016b).

3.2 Systèmes de référence (*baselines*)

Pour les expériences décrites dans cet article, nous avons utilisé la boîte à outils logicielle Kaldi (Povey *et al.*, 2011) et le système Eesen (Miao *et al.*, 2015), qui se distinguent principalement par leur modélisation acoustique (Eesen est dérivé de Kaldi).

Trois modèles acoustiques indépendants du locuteur ont été estimés avec le système Eesen, qui ne diffèrent que par les paramètres acoustiques utilisés pour représenter le signal de parole. Ces trois types de paramètres sont :

1. *fbanks* $\oplus\Delta \oplus \Delta\Delta$ *(dimension* $= 120$*)* : les paramètres de type bancs de filtre à 40 dimensions, concaténés avec leurs dérivées premières et secondes ;
2. les *paramètres MFCC à haute résolution (dimension* $= 40$*)* : ce sont des paramètres MFCC calculés sans réduction de dimension, en conservant les 40 cepstres ;
3. les paramètres de type *bottleneck (BN) (dimension* $= 40$*)*.

Le premier type de paramètres correspond à celui proposé dans la version originale de Eesen pour la recette liée au corpus TED-LIUM. Pour les modèles acoustiques des deux autres types de paramètres, 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 comme proposé dans (Peddinti *et al.*, 2015).

Le premier modèle acoustique de référence a été appris de la manière décrite dans (Miao *et al.*, 2015), en utilisant le critère CTC et la même architecture BLSTM profonde. Le réseau de neurones BLSTM contient cinq couches bidirectionnelles d'unités BLSTM contenant 320 unités pour chaque

sous-couche *forward* (avant) ou *backward* (arrière). Les paramètres acoustiques, fournis en entrée du réseau de neurones, ont été normalisés à l'aide de la soustraction de la moyenne par locuteur, et à l'aide de la normalisation de la variance. La couche de sortie est une couche de type *softmax* à 41 dimensions qui correspondent à 39 phonèmes indépendants du contexte, un modèle de bruit et au symbole $\emptyset$.

Comme évoqué plus haut, le troisième modèle acoustique indépendant du locuteur a été appris sur des paramètres BN. Pour calculer ces paramètres, un modèle de type DNN (*feedforward Deep Neural Network*) a été construit avec l'architecture suivante : une couche d'entrée de 440 dimensions, quatre couches cachées à 1500 dimensions, sauf la troisième qui ne contient que 40 neurones et dont seront extraits les paramètres BN, et une couche de sortie de 4025 dimensions. Les paramètres acoustiques utilisés pour l'apprentissage de cet extracteur de *bottlenecks* sont le résultat de la concaténation (dimension par dimension : *splicing*) de 11 trames consécutives de paramètres MFCC de 40 dimensions chacune, pour un total de 440 paramètres.

3.3 Modèles acoustiques adaptés au locuteur

Trois techniques d'adaptation des modèles acoustiques ont été expérimentées de manière empirique dans cette section : fMLLR, adaptation par i-vecteur, et adaptation MAP appliquée sur des mélanges de modèles gaussiens dérivés. Les mêmes stratégies d'augmentation des données d'apprentissage telles qu'évoquées plus haut ont été appliquées pour tous les modèles acoustiques adaptés. Tous les modèles SAT ont été appris avec la même architecture neuronale (à l'exception de la couche d'entrée) et le même critère d'apprentissage, comme décrit dans la section 3.2 pour les modèles indépendants du locuteur (non adaptés). Les six modèles acoustiques SAT on été estimés avec les paramètres suivants :

4. *MFCC* $\oplus$ *i-vectors (dimension* $= 140$*)*;
5. *BN* $\oplus$ *i-vectors (dimension* $= 140$*)*;
6. *BN avec fMLLR (dimension* $= 40$*)*;
7. *MFCC* $\oplus$ *GMMD (dimension* $= 167$*)*;
8. *BN* $\oplus$ *GMMD (dimension* $= 167$*)*;
9. *BN avec fMLLR* $\oplus$ *GMMD (dimension* $= 167$*)*.

Pour les modèles acoustiques utilisant les paramètres #4 et #5, les i-vecteurs de 100 dimensions ont été calculés *online* comme présenté dans (Peddinti *et al.*, 2015), et les statistiques utilisées pour le calcul des i-vecteurs ont été mises à jour toutes les deux phrases durant l'apprentissage. Pour les modèles acoustiques de #7 à #9, nous avons appliqué les paramètres BN pour l'apprentissage du modèle auxiliaire de type GMM (modèle de mélange gaussien) utilisé pour l'extraction des paramètres GMMD. Les paramètres GMMD adaptés au locuteur ont été obtenus de la manière décrite dans (Tomashenko *et al.*, 2016b).

3.4 Résultats expérimentaux pour les modèles acoustiques de type CTC

Pour l'ensemble des résultats proposés dans cette section, les techniques d'adaptation ont été appliquées de manière non supervisée lors du décodage des données de test en utilisant les transcriptions automatiques générées par le meilleur modèle acoustique générique (non adapté). Les taux d'erreurs

sur les mots obtenus par les différents systèmes sont présentés dans le tableau 2. Les trois premières lignes du tableau (#1–#3) correspondent aux modèles acoustiques génériques de référence (*baselines*), qui ont été construits selon les approches décrites en section 3.2. La première ligne représente le système Eesen tel qu'il est distribué (Miao *et al.*, 2015). Les six lignes suivantes (#4–#9) présentent les résultats des modèles adaptés. La numérotation utilisée dans la table 2 coïncide avec la numérotation des sections 3.2 et 3.3. Les deux dernières lignes du tableau (#10 et #11) ont été obtenues en utilisant les modèles acoustiques des lignes #8 et #9, mais l'adaptation des paramètres GMMD (notés alors GMMD* dans les tableaux 2, 3) de #10 et #11 a été effectuée à partir des transcriptions obtenu à l'aide du modèle adapté #6, alors que pour toutes les autres expériences l'adaptation a été réalisée à partir des transcriptions obtenues grâce au modèle générique #2. Parmi les systèmes CTC #1–#9, c'est le système #9 qui obtient les meilleurs résultats. Ce système correspond à l'utilisation de paramètres GMMD adaptés par MAP, et concaténés avec des paramètres BN adaptés par fMLLR. Une légère amélioration (#11) peut être obtenue en réadaptant les paramètres du modèle à partir des transcriptions automatiques que ce modèle obtient dans une première passe. De toutes les techniques d'adaptation appliquées séparément (#4–#8), c'est l'adaptation des paramètres GMMD par MAP qui obtient les meilleures performances, que ce soit avec les paramètres BN ou les paramètres MFCC.

#	Paramètres	**CTC** : WER,%			**TDNN** : WER,%		
		Dev.	Test_1	Test_2	Dev.	Test_1	Test_2
1	fbanks $\oplus\Delta \oplus \Delta\Delta$	14,57	11,71	15,29	-	-	-
2	MFCC	13,21	11,16	14,15	13,69	11,34	14,38
3	BN	13,63	11,84	15,06	12,32	10,48	14,00
4	MFCC $\oplus$ i-vecteurs	12,92	10,45	14,09	11,63	9,62	13,28
5	BN $\oplus$ i-vecteurs	13,47	11,37	14,31	11,62	9,75	13,30
6	BN-fMLLR	12,45	10,96	13,79	**10,70**	**9,28**	**12,84**
7	MFCC $\oplus$ GMMD	11,95	10,20	14,04	11,30	9,75	13,74
8	BN $\oplus$ GMMD	11,66	10,14	13,88	11,07	9,75	13,55
9	BN-fMLLR $\oplus$ GMMD	**11,63**	**9,91**	**13,58**	10,92	9,54	13,27
10	BN $\oplus$ GMMD*	11,67	10,11	13,70	10,29	9,20	13,04
11	BN-fMLLR $\oplus$ GMMD*	**11,41**	9,93	**13,47**	**10,15**	**9,06**	**12,84**

TABLE 2 – Taux d'erreurs sur les mots (WER) en fonction des paramètres acoustiques, de la technique d'adaptation (aucune de #1 à #3) et du type de modèles acoustiques (CTC ou TDNN). GMMD* correspond aux paramètres GMMD adaptés à partir des transcriptions produites par un modèle SAT (par défaut ces transcriptions proviennent de l'utilisation d'un modèle générique).

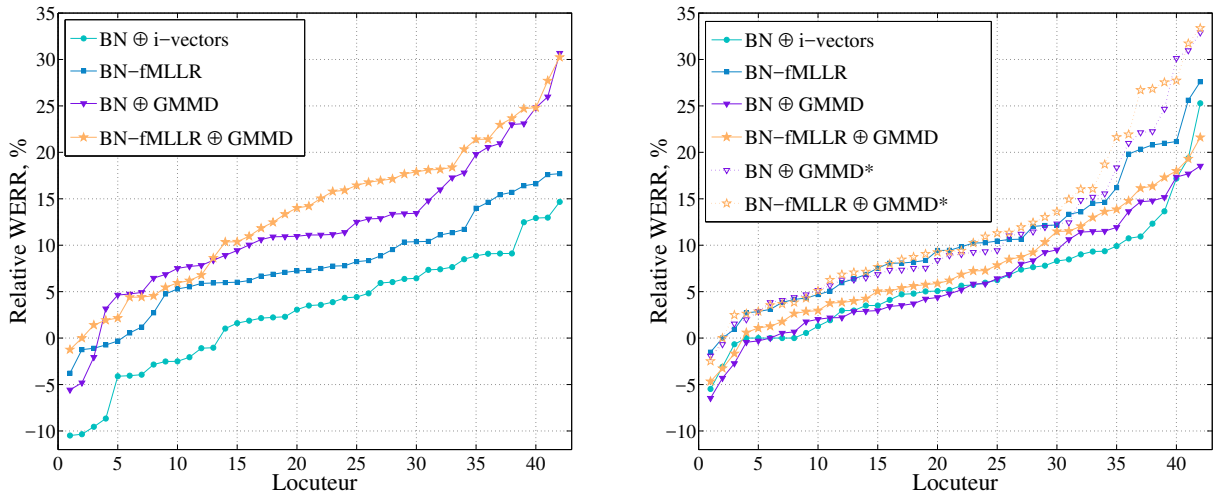
3.5 Comparaison de l'impact des techniques adaptation en fonction de la nature des modèles acoustiques : CTC ou TDNN

Dans cette section, nous souhaitons comparer le comportement des techniques d'adaptation lorsqu'elles sont appliquées à des modèles CTC à leur comportement lorsqu'elles sont appliquées à des modèles hybrides combinant modèles de Markov cachés et réseau de neurones. Pour cela, nous avons choisi de refaire les mêmes expériences que pour les modèles CTC en utilisant un modèle de type TDNN, car ces modèles sont actuellement souvent utilisés dans les systèmes à l'état de l'art (Peddinti *et al.*, 2015). Ces modèles acoustiques ont été estimés avec le critère d'entropie croisée (CE). Pour réaliser cette comparaison, nous avons construit le même ensemble de modèles acoustiques génériques

#	Paramètres	**CTC** : rel. WERR,%			**TDNN** : rel. WERR,%		
		Dev.	Test_1	Test_2	Dev.	Test_1	Test_2
4	MFCC $\oplus$ i-vecteurs	2,2	6,4	0,4	5,6	8,2	5,1
5	BN $\oplus$ i-vecteurs	-2,0	-1,9	-1,1	5,7	7,0	5,0
6	BN-fMLLR	5,8	1,8	2,5	13,2	11,5	8,3
7	MFCC $\oplus$ GMMD	9,5	8,6	0,8	8,3	7,0	1,9
8	BN $\oplus$ GMMD	11,7	9,1	1,9	10,2	7,0	3,2
9	BN-fMLLR $\oplus$ GMMD	12,0	**11,2**	4,0	11,4	9,0	5,2
10	BN $\oplus$ GMMD*	11,7	9,4	3,2	16,5	12,2	6,9
11	BN-fMLLR $\oplus$ GMMD*	**13,6**	11,0	**4,8**	**17,6**	**13,6**	**8,3**

TABLE 3 – Réduction relative du taux d'erreurs sur les mots (WERR) pour les modèles acoustiques adaptés de type CTC et TDNN, par rapport au meilleur module générique (non adapté) de chaque type de modèle (#2 pour CTC et #3 pour TDNN). Ces valeurs sont calculées à partir des résultats du tableau 2.

et adaptés que nous avions construit pour les modèles CTC (voir sections 3.2 et 3.3), à l'exception du modèle #1. Tous les modèles TDNN ont été appris de la même manière et utilisent la même architecture neuronale. Ils ne diffèrent que par la nature des paramètres acoustiques qu'ils doivent traiter. L'architecture des modèles TDNN est semblable à celle décrite dans (Peddinti *et al.*, 2015). Le contexte temporel est de la forme $[t-16, t+12]$, alors que les index de concaténation (splicing) que nous avons utilisés sont les suivants : $[-2, 2]$, $\{-1, 2\}$, $\{-3, 3\}$, $\{-7, 2\}$, $\{0\}$, $\{0\}$. Ces modèles utilisent des couches cachés de 850 dimensions avec fonctions d'activation ReLU (Rectified Linear Units), et une couche de sortie d'environ 4000 dimensions.



(a) CTC (b) TDNN

FIGURE 1 – Réduction relative du taux d'erreurs sur les mots (WERR) pour les locuteurs des corpus de test et de développement en fonction de la technique d'adaptation utilisée, et comparativement au modèle acoustique utilisant des paramètres BN (#3). Pour chaque modèle acoustique, les résultats sont classés par ordre croissant du WERR.

Comme les résultats des modèles CTC, les résultats expérimentaux liés aux modèles TDNN sont

indiqués dans le tableau 2 et la figure 1. Pour les modèles TDNN, l'adaptation est effectuée à partir des transcriptions automatiques générées avec le modèle utilisant des paramètres BN. Dans la figure 1b, pour les modèles TDNN nous avons ajouté l'utilisation de modèles SAT (GMMD*) pour générer les transcriptions pour l'adaptation, car elle permet une amélioration plus conséquente des performances que ce que nous avions observé avec les modèles CTC. Le tableau 3 montre les réductions relatives du taux d'erreurs en fonction des techniques d'adaptation pour les modèles CTC et TDNN, par rapport au meilleur modèle acoustique générique correspondant (#2 pour CTC et #3 pour TDNN). Comme nous pouvons le constater, le choix des paramètres optimaux dépend de la nature du modèle acoustiques. Pour les modèles TDNN, nous observons dans nos expériences que les paramètres BN donnent de meilleurs résultats que les MFCC, alors que pour les modèles CTC la situation est inversée. De plus, nous remarquons que le classement des systèmes en fonction de leur taux d'erreurs diffère selon la nature (CTC ou TDNN) des modèles acoustiques.

4 Conclusions

Cet article porte sur l'étude du bénéfice potentiel apporté par les techniques d'adaptation au locuteur aux systèmes de reconnaissance de la parole neuronaux de bout en bout. Il montre que l'adaptation au locuteur reste un mécanisme essentiel pour l'amélioration des performances dans ce nouveau paradigme pour la reconnaissance de la parole. Les résultats expérimentaux sur le corpus TED-LIUM montrent que, dans un mode non supervisé, les techniques d'adaptation et d'augmentation des données d'apprentissage peuvent apporter une réduction relative du taux d'erreurs sur les mots comprise entre 10 et 20%, par exemple en se comparant à un modèle CTC générique utilisant des paramètres de type banc de filtres. En moyenne pour les modèles CTC, les meilleurs résultats sont obtenus en utilisant des paramètres dérivés de GMMs adaptés par MAP, en les combinant à des *bottlenecks* adaptés par fMLLR. Nous avons montré que les performances obtenues par les différentes techniques d'adaptation dépendent de la nature de l'architecture neuronale d'un modèle acoustique. Enfin, cet article présente des résultats expérimentaux qui permettent de comparer dans des conditions réalistes les performances des approches BLSTM-CTC à celles des approches à l'état de l'art comme les HMM-TDNN. Le taux d'erreurs sur les mots du meilleur modèle générique de type TDNN est relativement plus bas de 1 à 7% par rapport à celui du meilleur modèle CTC. Avec les modèles adaptés, cet écart augmente et le meilleur modèle TDNN commet entre 5 à 13% d'erreurs de moins que le meilleur modèle CTC.

Références

BAHDANAU D., CHOROWSKI J., SERDYUK D., BRAKEL P. & BENGIO Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *ICASSP*, p. 4945–4949 : IEEE.

BENGIO Y., SIMARD P. & FRASCONI P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, **5**(2), 157–166.

CHOROWSKI J., BAHDANAU D., CHO K. & BENGIO Y. (2014). End-to-end continuous speech recognition using attention-based recurrent NN : First results. *arXiv preprint arXiv :1412.1602*.

COLLOBERT R., PUHRSCH C. & SYNNAEVE G. (2016). Wav2letter : an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv :1609.03193*.

GALES M. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech and language*, **12**(2), 75–98.

GAUVAIN J.-L. & LEE C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Trans. Speech and Audio Proc.*, **2**, 291–298.

GEMELLO R., MANA F., SCANZIO S., LAFACE P. & DE MORI R. (2006). Adaptation of hybrid ANN/HMM models using linear hidden transformations and conservative training. In *ICASSP*, p. 1189–1192.

GRAVES A. *et al.* (2006). Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, p. 369–376 : ACM.

GRAVES A. & JAITLY N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, volume 14, p. 1764–1772.

GRAVES A., MOHAMED A.-R. & HINTON G. (2013). Speech recognition with deep recurrent neural networks. In *ICASSP*, p. 6645–6649 : IEEE.

HANNUN A., CASE C., CASPER J., CATANZARO B., DIAMOS G., ELSEN E., PRENGER R. *et al.* (2014). Deep speech : Scaling up end-to-end speech recognition. *arXiv preprint arXiv :1412.5567*.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long short-term memory. *Neural computation*, **9**(8), 1735–1780.

LI J., HUANG J.-T. & GONG Y. (2014). Factorized adaptation for deep neural network. In *ICASSP*, p. 5537–5541 : IEEE.

LIAO H. (2013). Speaker adaptation of context dependent deep neural networks. In *ICASSP*, p. 7947–7951.

MIAO Y. *et al.* (2016). An empirical exploration of CTC acoustic models. In *ICASSP*, p. 2623–2627 : IEEE.

MIAO Y., GOWAYYED M. & METZE F. (2015). Eesen : End-to-end speech recognition using deep rnn models and wfst-based decoding. In *ASRU*, p. 167–174 : IEEE.

PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH*, p. 3214–3218.

POVEY D. *et al.* (2011). The Kaldi speech recognition toolkit. In *ASRU*.

PRICE R., ISO K. & SHINODA K. (2014). Speaker adaptation of deep neural networks using a hierarchy of output layers. In *SLT*, p. 153–158 : IEEE.

ROUSSEAU A., DELÉGLISE P. & ESTÈVE Y. (2014). Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks. In *LREC*, p. 3935–3939.

SAON G., SOLTAU H., NAHAMOO D. & PICHENY M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *ASRU*, p. 55–59.

SCHUSTER M. & PALIWAL K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, **45**(11), 2673–2681.

SENIOR A. & LOPEZ-MORENO I. (2014). Improving DNN speaker independence with i-vector inputs. In *ICASSP*, p. 225–229.

SWIETOJANSKI P. & RENALS S. (2014). Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In *SLT*, p. 171–176 : IEEE.

TOMASHENKO N. *et al.* (2016a). Exploration de paramètres acoustiques dérivés de GMMs pour l'adaptation non supervisée de modèles acoustiques à base de réseaux de neurones profonds. In *JEP*, p. 337–345.

TOMASHENKO N. & KHOKHLOV Y. (2014). Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. In *INTERSPEECH*, p. 2997–3001.

TOMASHENKO N. & KHOKHLOV Y. (2015). GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models. In *INTERSPEECH*, p. 2882–2886.

TOMASHENKO N., KHOKHLOV Y. & ESTEVE Y. (2016b). On the use of Gaussian mixture model framework to improve speaker adaptation of deep neural network acoustic models. In *INTERSPEECH*, p. 3788–3792.

TOMASHENKO N., KHOKHLOV Y., LARCHER A. & ESTEVE Y. (2016c). Exploring GMM-derived features for unsupervised adaptation of deep neural network acoustic models. In *SPECOM*, p. 304–311.

XUE S. *et al.* (2014). Fast adaptation of deep neural network based on discriminant codes for speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Trans. on*, **22**(12), 1713–1725.

YU D., YAO K., SU H., LI G. & SEIDE F. (2013). KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition. In *ICASSP*, p. 7893–7897.