Exploration de paramètres acoustiques dérivés de GMMs pour l'adaptation non supervisée de modèles acoustiques à base de réseaux de neurones profonds

Natalia Tomashenko[1, 2, 3] Yuri Khokhlov[3] Anthony Larcher[2] Yannick Estève[2]
(1) ITMO University, Saint-Pétersbourg, Russie
(2) LIUM, Le Mans, France
(3) STC-innovations Ltd, Saint-Pétersbourg, Russie
prenom.nom@univ-lemans.fr
khokhlov@speechpro.com

RÉSUMÉ

L'étude présentée dans cet article améliore une méthode récemment proposée pour l'adaptation de modèles acoustiques markoviens couplés à un réseau de neurones profond (DNN-HMM). Cette méthode d'adaptation utilise des paramètres acoustiques dérivés de mixtures de modèles Gaussiens (*GMM-derived features*, *GMMD*). L'amélioration provient de l'emploi de scores et de mesures de confiance calculés à partir de graphes construits dans le cadre d'un algorithme d'adaptation conventionnel dit de *maximum a posteriori* (MAP). Une version modifiée de l'adaptation MAP est appliquée sur le modèle GMM auxiliaire utilisé dans une procédure d'apprentissage adaptatif au locuteur (*speaker adaptative training*, SAT) lors de l'apprentissage du DNN. Des expériences menées sur le corpus Wall Street Journal (WSJ0) montrent que la technique d'adaptation non supervisée proposée dans cet article permet une réduction relative de $8,4\%$ du taux d'erreurs sur les mots (WER), par rapport aux résultats obtenus avec des modèles DNN-HMM indépendants du locuteur utilisant des paramètres acoustiques plus conventionnels.

ABSTRACT

Exploring GMM-derived features for unsupervised adaptation of deep neural network acoustic models

In this paper we investigate GMM-derived features recently introduced for adaptation of context-dependent deep neural network HMM (CD-DNN-HMM) acoustic models. We present an initial attempt of improving of the previously proposed adaptation algorithm by applying lattice scores and by using confidence measures in the traditional maximum a posteriori (MAP) adaptation algorithm. Modified MAP adaptation is performed for the auxiliary GMM model used in a speaker adaptive training (SAT) procedure for a DNN. Experimental results on the Wall Street Journal (WSJ0) corpus show that the proposed adaptation technique can provide, on average, an $8,4\%$ relative word error rate (WER) reduction under an unsupervised adaptation setup, compared to speaker independent DNN-HMM systems built on conventional features.

MOTS-CLÉS : adaptation au locuteur, réseaux de neurones profonds, MAP, CD-DNN-HMM, paramètres acoustiques dérivés de GMM (GMMD), apprentissage adaptatif au locuteur (SAT).

KEYWORDS : speaker adaptation, deep neural networks (DNN), MAP, CD-DNN-HMM, GMM-derived parameters (GMMD), speaker adaptive training (SAT).

1 Introduction

Aujourd'hui, les réseaux de neurones profonds (DNNs) ont détrôné les modèles GMM-HMMs dans la plupart des systèmes état-de-l'art de reconnaissance de la parole (RAP), depuis qu'il a été montré que les modèles DNN-HMM obtiennent de meilleurs résultats dans plusieurs tâches de RAP (Hinton *et al.*, 2012). De nombeux algorithmes d'adaptation ont été développés pour les modèles GMM-HMM (Gales, 1998 ; Gauvain & Lee, 1994) mais ne peuvent pas être facilement appliqués aux DNNs en raison des différentes natures de ces systèmes. Les GMMs sont des modèles génératifs appris par maximisation de la vraisemblance sur les données d'apprentissage alors que les DNNs sont des modèles discriminant, dont les paramètres sont estimés pour minimiser les erreurs de classification. Puisque l'estimation des paramètres des DNNs est basée sur un critère discriminant, elle est plus sensible aux erreurs d'étiquettes et moins pertinente pour une adaptation non supervisée.

Plusieurs méthodes d'adaptation ont récemment été proposées pour les DNNs, et quelques unes (Rath *et al.*, 2013 ; Seide *et al.*, 2011 ; Lei *et al.*, 2013 ; Tomashenko & Khokhlov, 2014, 2015 ; Liu & Sim, 2014 ; Kanagawa *et al.*, 2015) tirent avantage de l'adaptabilité des GMMs. Cependant il n'existe pas de méthode universelle pour transférer de manière efficiente les algorithmes d'adaptation du paradigme gaussien vers celui des DNNs. L'objectif de cette étude est de faire un pas dans cette direction en utilisant des paramètres acoustiques dérivés de GMM pour estimer des modèles DNN.

La plupart des méthodes existantes pour adapter les modèles DNN peuvent être classées en quatre catégories : (1) les transformations linéaires, (2) les techniques de régularisation, (3) les paramètres auxiliaires, (4) les combinaisons de GMM et DNN. **La transformation linéaire** est une des approches les plus populaires pour l'adaptation des réseaux de neurones (ANN). Elle peut être appliquée à différents niveaux d'un système ANN-HMM : sur les paramètres d'entrée, avec une transformation linéaire (*linear input network transformation*, LIN) (Gemello *et al.*, 2006 ; Neto *et al.*, 1995 ; Li & Sim, 2010 ; Trmal *et al.*, 2010) ou avec une régression linéaire discriminante sur l'espace des paramètres (*feature-space discriminative linear regression, fDLR)* (Seide *et al.*, 2011 ; Yao *et al.*, 2012) ; sur les activations des couches cachées (*linear hidden network transformation*, LHN) (Gemello *et al.*, 2006 ; Neto *et al.*, 1995) ; sur la couche softmax, comme avec LON (Li & Sim, 2010) ou avec une régression linéaire discriminante sur les paramètres de sortie (*output-feature discriminative linear regression*) (Yao *et al.*, 2012). L'adaptation de modèles acoustiques hybrides par partage de probabilités *a posteriori* (Stadermann & Rigoll, 2005) peut également être considérée comme une transformation linéaire de ces probabilités. Enfin, les auteurs de (Dupont & Cheboub, 2000) décrivent une méthode qui repose sur une transformation linéaire dans l'espace des paramètres et sur une analyse en composantes principales. La seconde catégorie de méthodes d'adaptation consiste à réestimer complètement le réseau ou seulement une partie à l'aide de **techniques de régularisation** spécifiques pour améliorer la généralisation, comme la régularisation *L2-prior* (Liao, 2013), la régularisation par divergence de Kullback-Leibler (Yu *et al.*, 2013), ou l'apprentissage conservatif (Albesano *et al.*, 2006). Dans (Stadermann & Rigoll, 2005), seul un sous-ensemble des neurones cachés avec une variance maximale (calculée sur les données d'adaptation) est réestimé. Le nombre de paramètres spécifiques au locuteur est réduit dans (Xue *et al.*, 2014) à travers une factorisation basée sur une décomposition en valeurs singulières, et un apprentissage adaptatif régularisé d'un sous-ensemble de paramètres de DNN est exploré dans (Ochiai *et al.*, 2014).

L'utilisation de paramètres auxiliaires est une autre approche pour laquelle les vecteurs de paramètres acoustiques sont augmentés de paramètres additionnels spécifiques au locuteur ou au canal. Ces paramètres sont calculés pour chaque locuteur ou pour chaque phrase, à la fois pendant l'apprentissage

et pendant le décodage. Les i-vecteurs sont un exemple de paramètres auxiliaires efficaces (Senior & Lopez-Moreno, 2014 ; Saon *et al.*, 2013) ; il a été montré qu'ils étaient complémentaires avec une adaptation fMLLR. L'adaptation par codes de locuteurs (Abdel-Hamid & Jiang, 2013) et l'adaptation factorisée (Li *et al.*, 2014) sont des méthodes alternatives qui prennent en compte les facteurs sous-jacents qui contribuent à dégrader le signal de parole.

Le moyen le plus courant pour **combiner les modèles GMM et DNN** à des fins d'adaptation consiste à utiliser comme entrées, lors de l'apprentissage du DNN, des paramètres GMM ayant été adaptés, par exemple par fMLLR (Rath *et al.*, 2013 ; Seide *et al.*, 2011 ; Kanagawa *et al.*, 2015). Dans (Lei *et al.*, 2013), les scores de vraisemblance des modèles DNN et GMM, adaptés au niveau de l'espace des paramètres en utilisant la même transformation fMLLR, sont combinés au niveau des états pendant le décodage. Les auteurs de (Liu & Sim, 2014) proposent de combiner les modèles GMM et DNN en utilisant une approche par régression des poids variant dans le temps (*temporally varying weight regression, TVWR).* Dans la méthode d'apprentissage de la couche *bottleneck* dépendant du locuteur décrite dans (Doddipatla *et al.*, 2014), les paramètres du *bottleneck* normalisé par locuteur sont calculés et utilisés pour l'apprentissage de modèle GMM-HMM. Dans les systèmes de type *tandem*, les paramètres dérivés des réseaux de neurones sont également utilisés pour l'apprentissage des GMM (Ellis & Reyes-Gomez, 2001).

Une autre approche d'adaptation basée sur les modèles s'appuie sur l'analyse des contributions des neurones cachés spécifiques au locuteur (*learning speaker-specific hidden unit contributions*, LHUC). Dans (Siniscalchi *et al.*, 2013), la forme de la fonction d'activation est modifiée pour mieux correspondre aux caractéristiques liées au locuteur.

Dans le passé, il a été montré dans (Pinto & Hermansky, 2008) que les log-vraisemblances de GMM peuvent être utilisées avec succès comme paramètres pour estimer un système de reconnaissance de phonèmes construit sur la base d'un MLP (*multi-layer perceptron*). Dans notre étude, nous explorons une nouvelle approche de type *speaker adaptative training* (SAT) appliquée sur les DNN et basée sur l'utilisation de paramètres dérivés de GMM (GMMD) en entrée du réseau de neurones (Tomashenko & Khokhlov, 2014, 2015). Notre approche s'appuie également sur l'utilisation de techniques d'adaptation propres aux GMMs, appliquées aux GMMD.

Dans cet article, nous présentons une première tentative d'amélioration de cette approche en utilisant des graphes de reconnaissance durant l'adaptation MAP. La suite de l'article est organisée comme suit. Dans la Section 2, l'approche SAT pour les DNN-HMM basés sur les paramètres dérivés de GMM est introduite. La Section 3 décrit l'algorithme d'adaptation MAP utilisant des scores de graphes. Les résultats expérimentaux sont donnés en Section 4. Enfin, une conclusion est fournie en Section 5.

2 Apprentissage adaptatif au locuteur pour les modèles DNN-HMM basés sur des paramètres dérivés de GMMs

La construction de paramètres dérivés de GMMs pour l'adaptation de DNNs a été proposée dans (Tomashenko & Khokhlov, 2014, 2015), où il a été montré que ces paramètres rendent possible l'utilisation de techniques d'adaptation de HMM-GMMs dans le paradigme DNN, par exemple au travers d'une adaptation MAP ou fMLLR . Nos paramètres sont obtenus comme suit (voir Figure 1) : tout d'abord, 13 coefficients cepstraux de fréquence Mel (MFCC) et leurs coefficients Δ et $\Delta\Delta$ (au total 39 coefficients) sont extraits, et une normalisation par la moyenne cepstrale (CMN) est appliquée par

locuteur. Ensuite, un modèle GMM-HMM auxiliaire monophone indépendant du locuteur est utilisé pour transformer les vecteurs de paramètres cepstraux en vecteurs de vraisemblances. Au cours de cette étape, une adaptation au locuteur du modèle GMM-HMM est réalisée pour chaque locuteur du corpus d'apprentissage et le nouveau modèle SA GMM-HMM (SA : adapté au locuteur) créé est utilisé pour obtenir les paramètres dérivés de GMMs adaptés au locuteur. Dans le modèle GMM-HMM

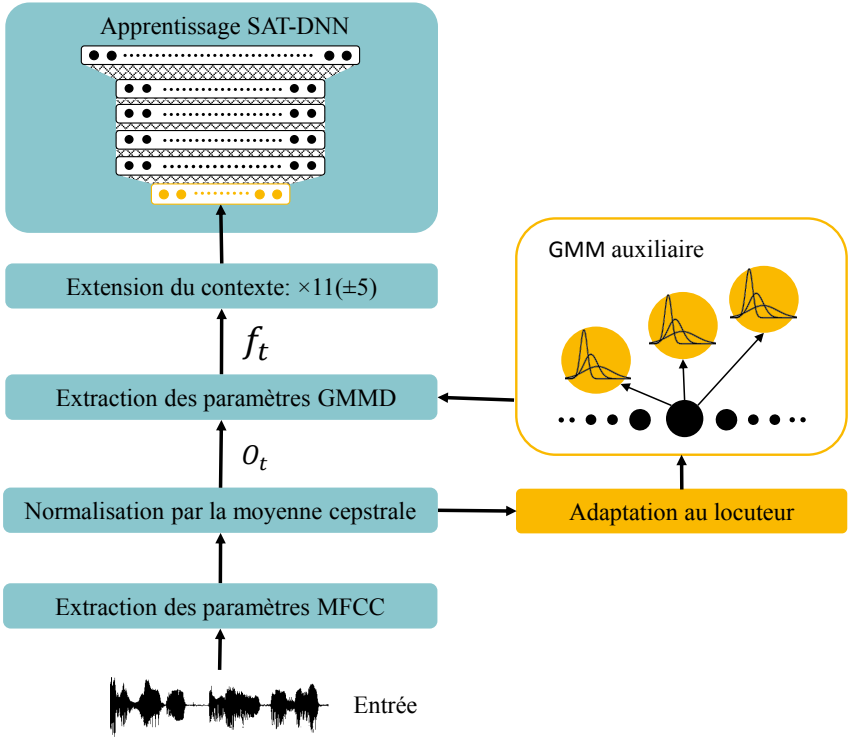

FIGURE 1 – Utilisation de paramètres dérivés de GMMs adaptés pour une apprentissage SAT de DNN-HMM.

auxiliaire, chaque phonème est modélisé indépendamment du contexte avec un modèle de Markov à trois états. Pour un vecteur de paramètres MFCC donné, un nouveau vecteur de paramètres dérivés de GMMs est obtenu en calculant les log-vraisemblances de tous les états des modèles monophones GMM-HMM. En supposant que o_t est un paramètre acoustique à l'instant t, alors le vecteur f_t de paramètres dérivés de GMMs est calculé comme suit :

$$f_t = [p_t^1, \ldots, p_t^n] \tag{1}$$

où n est le nombre d'états du modèle monophone auxiliaire GMM-HMM, et :

$$p_t^i = \log\left(P(o_t \mid s_t = i)\right) \tag{2}$$

est la log-vraisemblance estimée par le GMM-HMM. Ici, s_t est l'indice de l'état du HMM à t. Dans notre cas, n est égal à 132 ($= 39 \times 3 + 3 \times 5$) : 39 HMM à trois états (un HMM par phonème), un

modèle de silence modélisé par un HMM à cinq états, et deux modèles HMM de bruits (un parole et un non-parole) à 5 états. Nous obtenons ainsi, pour chaque trame, un vecteur à 132 dimensions de paramètres dérivés de GMMs.

Enfin, ces paramètres sont concaténés sur une fenêtre temporelle de 11 trames (± 5 autour de la trame visée). Ces 1452 (11×132) paramètres par trame, appelés GMMD, sont utilisés par la suite comme entrée pour l'apprentissage des DNN .

3 Utilisation de scores de graphes de décodage pour l'adaptation MAP

L'utilisation d'informations et de mesures de confiance provenant de graphes est une méthode connue pour améliorer les performances d'une adaptation non supervisée (Uebel & Woodland, 2001 ; Gollan & Bacchiani, 2008). Dans cette étude, nous utilisons l'algorithme d'adaptation MAP pour adapter des modèles GMM-HMM indépendants du locuteur (Gauvain & Lee, 1994). L'adaptation au locuteur d'un modèle DNN-HMM construit à partir de paramètres GMMD est réalisé au moyen d'une adaptation MAP des modèles monophones GMM-HMM auxiliaires, qui ont été utilisés pour calculer les paramètres GMMD.

Nous avons modifié l'algorithme classique d'adaptation MAP en utilisant des graphes au lieu d'un alignement sur la meilleure hypothèse, comme expliqué ci-dessous. Notons m l'indice d'une Gaussienne dans un modèle acoustique indépendant du locuteur, et μ_m la moyenne de cette Gaussienne. Alors l'estimation MAP du vecteur des moyennes est :

$$\widehat{\mu}_m = \frac{\tau \mu_m + \sum_t \gamma_m(t) p_s(t) o_t}{\tau + \sum_t \gamma_m(t) p_s(t)} \tag{3}$$

où τ est le paramètre qui contrôle l'équilibre entre l'estimation de la moyenne par maximum de vraisemblance et sa valeur *a priori*, $\gamma_m(t)$ est la probabilité *a posteriori* du composant gaussien m à l'instant t, et $p_s(t)$ est la mesure de confiance pour l'état s à l'instant t obtenue en calculant les probabilités *a posteriori* des arcs dans le graphe des états de la première passe de décodage.

L'algorithme *forward-backward* est utilisé pour calculer ces probabilités *a posteriori* à partir du graphe. Notons que lorsque $p_s(t) = 1$ pour tous les états, la formule (3) represente l'adaptation MAP classique.

En plus de cette pondération au niveau de la trame, nous appliquons une stratégie de sélection basée sur la mesure de confiance en n'utilisant dans la formule (3) que les observations dont les scores de confiance dépassent un seuil fixé *a priori*.

Pour l'adaptation des modèles acoustiques DNN, l'adaptation MAP est d'abord appliquée sur les modèles monophones GMM-HMM indépendant du locuteur pour créer un modèle adapté au locuteur SA GMM-HMM, comme nous venons de le décrire. Ensuite, au moment de reconnaissance de la parole, les GMMD sont calculés à partir de ce modèle SA GMM-HMM. L'approche proposée peut être considérée comme une technique de transformation dans l'espace des paramètres, puisque les DNN-HMMs sont appris sur des paramètres dérivés de GMMs.

4 Résultats expérimentaux

Dans ce travail préliminaire, nous présentons les résultats obtenus en suivant le protocole standard WSJ0 **si_et_20** (Paul & Baker, 1992) qui comporte 333 phrases lues (5645 mots) par 8 locuteurs. Nous utilisons un modèle de langage trigramme contenant 20000 mots (*open NVP LM*). Le taux de mots hors vocabulaire est de $1,5\%$. Ce modèle de langage est réduit selon le procédé décrit dans la recette KALDI pour WSJ (Povey *et al.*, 2011) avec un seuil à 10^{-7}.

Les modèles acoustiques sont estimés sur 7138 phrases prononcées par 83 locuteurs extraits des données d'apprentissage standard SI-84, soit environ 13 heures de parole et 2 heures de silence enregistrées en 16kHz. Le jeu de 39 phonèmes est complété par un modèle de silence et deux modèles de bruit. Les modèles acoustiques sont appris avec la plateforme KALDI (Povey *et al.*, 2011) en suivant la recette WSJ fournie, excepté pour l'extraction des paramètres GMMD et l'adaptation du modèle. Notre système de référence utilise 13 paramètres MFCC avec leurs dérivées premières et secondes. Ces 39 paramètres, utilisés avec leur contexte de 11 trames (5 avant et après) sont comparés aux paramètres GMMD proposés.

Trois réseaux de neurones profonds (DNN) sont estimés : un modèle de référence, indépendant du locuteur (SI) utilisant 11×39MFCC ; deux modèles utilisant les paramètres GMMD : l'un indépendant du locuteur (SI) et l'autre adapté au locuteur (SAT). Ces trois DNNs partagent la même topologie (exceptée la dimension des entrées) et sont entrainés avec les mêmes données. Un système GMM auxiliaire est également appris sur les mêmes données. L'apprentissage du modèle DNN SAT est décrit dans la Section 2 et la valeur de τ est fixée empiriquement à 5. L'apprentissage du modèle DNN SI avec les paramètres GMMD suit le processus décrit par la Figure 1 sans appliquer l'étape d'adaptation au locuteur. Les trois modèles disposent de 6 couches cachées à 2048 neurones et une couche de sortie de dimension 2355 correspondant aux états dépendant du contexte obtenus par une classification hiérarchique pour le système CD-GMM-HMM. Les DNN sont initialisés en empilant des machines de Boltzman restreintes. L'apprentissage final optimise une fonction d'entropie croisée et se termine par 5 itérations d'apprentissage séquentiel discriminatif avec un critère sMBR (*state Minimum Bayes Risk*).

Les expériences d'adaptation sont réalisées de façon non-supervisée sur les données d'évaluation en utilisant les transcriptions ou les graphes obtenus après la première passe. Deux expériences sont réalisées en adaptant le modèle GMM auxiliaire selon un critère : (1) *maximum a posteriori* standard, (2) *maximum a posteriori* utilisant les scores de graphes et la mesure de confiance décrite en Section 3 pour laquelle le seuil de confiance est fixé à $0,6$. Les performances des différentes adaptations sont reportées dans le Tableau 1 en terme de taux d'erreur mot pour les systèmes avec et sans adaptation.

Type of Features	Adaptation	WER, %	Δ WER, %
11×39MFCC	SI	7,51	référence
GMMD	SI	7,83	-
GMMD	SAT (MAP alignment)	7,09	5,6
GMMD	SAT (MAP lattice-based)	6,93	8,4

TABLE 1 – Taux d'erreur mots (%) évalué sur la tâche WSJ0 **si_et_20** et amélioration relative Δ WER par rapport à la référence.

5 Conclusions

Dans cet article, nous avons étudié des paramètres dérivés de GMM introduits récemment pour l'adaptation de modèles acoustiques DNN-HMM. Notre avons présenté une amélioration de l'approche précédemment proposée en appliquant le concept d'adaptation au locuteur au modèle DNN appris sur des paramètres dérivés de GMM et utilisant une mesure de confiance. Nous avons adapté un système GMM auxiliaire avec un critère de *maximum a posteriori* pour adapter le modèle DNN. Les résultats expérimentaux préliminaires obtenus sur le corpus WSJ0 démontrent que pour une adaptation non supervisée, la méthode proposée diminue relativement le taux d'erreur mots de $8,4\%$ par rapport à un système DNN-HMM indépendant du locuteur appris sur des paramètres MFCC standards. Il est important de noter que dans l'approche proposée, l'adaptation MAP du modèle GMM auxiliaire peut être remplacée par d'autres méthodes. Cette méthode fournit donc un cadre général pour transférer les algorithmes d'adaptation des modèles GMM-HMM pour les systèmes DNN.

Remerciements

Ce travail a ete partiellement finance par la commission europeenne a travers le projet EUMSSI, sous le numero de contrat 611 057, dans le cadre de lappel FP7-ICT-2013-10.

Références

ABDEL-HAMID O. & JIANG H. (2013). Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7942–7946 : IEEE.

ALBESANO D., GEMELLO R., LAFACE P., MANA F. & SCANZIO S. (2006). Adaptation of artificial neural networks avoiding catastrophic forgetting. In *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, p. 1554–1561 : IEEE.

DODDIPATLA R., HASAN M. & HAIN T. (2014). Speaker dependent bottleneck layer training for speaker adaptation in automatic speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*, p. 2199–2203.

DUPONT S. & CHEBOUB L. (2000). Fast speaker adaptation of artificial neural networks for automatic speech recognition. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, p. 1795–1798 : IEEE.

ELLIS D. P. & REYES-GOMEZ M. (2001). Investigations into tandem acoustic modeling for the aurora task. In *Eurospeech 2001 : Scandinavia : 7th European Conference on Speech Communication and Technology : September 3-7, 2001, Aalborg Congress and Culture Centre, Aalborg-Denmark : proceedings*, p. 189–192 : ISCA-Secretariat.

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

GAUVAIN J.-L. & LEE C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and audio processing, IEEE transactions on*, **2**(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 *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, p. I–I : IEEE.

GOLLAN C. & BACCHIANI M. (2008). Confidence scores for acoustic model adaptation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, p. 4289–4292 : IEEE.

HINTON G., DENG L., YU D., DAHL G. E., MOHAMED A.-R., JAITLY N., SENIOR A., VANHOUCKE V., NGUYEN P., SAINATH T. N. *et al.* (2012). Deep neural networks for acoustic modeling in speech recognition : The shared views of four research groups. *Signal Processing Magazine, IEEE*, **29**(6), 82–97.

KANAGAWA H., TACHIOKA Y., WATANABE S. & ISHII J. (2015). Feature-space structural maplr with regression tree-based multiple transformation matrices for DNN.

LEE L. & ROSE R. C. (1996). Speaker normalization using efficient frequency warping procedures. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 1, p. 353–356 : IEEE.

LEI X., LIN H. & HEIGOLD G. (2013). Deep neural networks with auxiliary gaussian mixture models for real-time speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7634–7638 : IEEE.

LI B. & SIM K. C. (2010). Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. p. 526–529.

LI J., HUANG J.-T. & GONG Y. (2014). Factorized adaptation for deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 5537–5541 : IEEE.

LIAO H. (2013). Speaker adaptation of context dependent deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7947–7951 : IEEE.

LIU S. & SIM K. C. (2014). On combining DNN and GMM with unsupervised speaker adaptation for robust automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 195–199 : IEEE.

NETO J., ALMEIDA L., HOCHBERG M., MARTINS C., NUNES L., RENALS S. & ROBINSON T. (1995). Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system.

OCHIAI T., MATSUDA S., LU X., HORI C. & KATAGIRI S. (2014). Speaker adaptive training using deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 6349–6353 : IEEE.

PAUL D. B. & BAKER J. M. (1992). The design for the wall street journal-based csr corpus. In *Proceedings of the workshop on Speech and Natural Language*, p. 357–362 : Association for Computational Linguistics.

PINTO J. P. & HERMANSKY H. (2008). *Combining evidence from a generative and a discriminative model in phoneme recognition*. Rapport interne, IDIAP.

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.

RATH S. P., POVEY D., VESELỲ K. & CERNOCKỲ J. (2013). Improved feature processing for deep neural networks. In *INTERSPEECH*, p. 109–113.

SAON G., SOLTAU H., NAHAMOO D. & PICHENY M. (2013). Speaker adaptation of neural network acoustic models using i-vectors. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, p. 55–59 : IEEE.

SEIDE F., LI G., CHEN X. & YU D. (2011). Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, p. 24–29 : IEEE.

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

SINISCALCHI S. M., LI J. & LEE C.-H. (2013). Hermitian polynomial for speaker adaptation of connectionist speech recognition systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, **21**(10), 2152–2161.

STADERMANN J. & RIGOLL G. (2005). Two-stage speaker adaptation of hybrid tied-posterior acoustic models. In *ICASSP (1)*, p. 977–980.

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

TOMASHENKO N. & KHOKHLOV Y. (2014). Speaker adaptation of context dependent deep neural networks based on MAP-adaptation and GMM-derived feature processing. In *Fifteenth Annual Conference of the International Speech Communication Association*, p. 2997–3001.

TOMASHENKO N. & KHOKHLOV Y. (2015). GMM-derived features for effective unsupervised adaptation of deep neural network acoustic models. In *Sixteenth Annual Conference of the International Speech Communication Association*, p. 2882–2886.

TRMAL J., ZELINKA J. & MÜLLER L. (2010). Adaptation of a feedforward artificial neural network using a linear transform. In *Text, Speech and Dialogue*, p. 423–430 : Springer.

UEBEL L. & WOODLAND P. C. (2001). Improvements in linear transform based speaker adaptation. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 1, p. 49–52 : IEEE.

XUE J., LI J., YU D., SELTZER M. & GONG Y. (2014). Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, p. 6359–6363 : IEEE.

YAO K., YU D., SEIDE F., SU H., DENG L. & GONG Y. (2012). Adaptation of context-dependent deep neural networks for automatic speech recognition. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, p. 366–369 : IEEE.

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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, p. 7893–7897 : IEEE.