Nouvelle méthode d'entraînement des systèmes hybrides HMM/ANN à base d'une segmentation floue

Application pour la reconnaissance automatique de la parole

Lilia Lazli et Mohamed Tayeb Laskri

Laboratoire de Recherche en Informatique (LRI)
Groupe de Recherche en Intelligence Artificielle (GRIA)
Département d'Informatique
Faculté des Sciences de l'Ingénieur
Université Badji Mokhtar d'Annaba
B.P.12 Sidi Amar 23200 Annaba – Algérie

L_Lazli@Yahoo.fr mtlaskri@wissal.dz

RÉSUMÉ. De nombreuses expériences ont déjà montré qu'une forte amélioration du taux de reconnaissance des systèmes HMM traditionnels est observée lorsque plus de données d'entraînement sont utilisées. En revanche, l'augmentation du nombre de données d'entraînement pour les modèles hybrides HMM/ANN s'accompagne d'une forte augmentation du temps nécessaire à l'entraînement des modèles mais pas ou peu des performances du système. Pour pallier cette limitation, nous rapportons dans ce papier les résultats obtenus avec une nouvelle méthode d'entraînement basée sur la fusion de données. Cette méthode a été appliquée dans un système de reconnaissance de la parole arabe, basé sur une segmentation floue.

MOTS-CLEFS : Reconnaissance de la parole arabe, segmentation floue, modèles de Markov cachés, réseaux de neurones artificiels, méthode de fusion de données.

1. Introduction

Plusieurs résultats récents en reconnaissance automatique de la parole (obtenus sur différentes bases de données allant des petits lexiques aux très grands lexiques) ont montré que les systèmes hybrides HMM/ANN combinant la technologie des modèles de Markov cachés (Hidden Markov Models – HMM) et des réseaux de neurones artificiels (Artificial Neural Networks – ANN), conduisent généralement à des performances de reconnaissance équivalentes ou meilleures que celles des systèmes HMM utilisés dans les mêmes conditions, avec cependant plusieurs avantages supplémentaires au niveau des besoins en CPU et mémoire [1][4][8]. En effet, des modèles hybrides HMM\ANN ont été conçus ces dernières années pour la parole : pour l'Anglais et pour le Français afin d'additionner les qualités de chacun des modèles fusionnés mais sans réellement homogénéiser l'architecture.

Néanmoins l'un des principaux défauts liés à ces modèles hybrides réside dans le fait que le nombre de paramètres est en quelque sorte borné. En effet, aucune amélioration n'est généralement observée (comme habituellement pour les HMM continus) lorsque le nombre des données d'entraînement et / ou de paramètres est fortement augmenté. Le tableau 1 reporte les résultats obtenus sur un corpus personnel avec deux réseaux de neurones de type MLP (Multi-Layer Perceptron) entraînés sur 1000 phrases pour le premier et environ 4000 pour le second.

Table 1. *Taux d'erreur au niveau du mot pour les trois dictionnaires de 60, 150, et 700 mots et deux modèles hybrides entraînés avec 1000 et 4000 phrases (paramètres log RASTA-PLP).*

Taille	*MLP (1000 phrases)*	*MLP (4000 phrases)*
60 mots	1.3%	1.5%
150 mots	5.0%	4.8%
700 mots	23.0%	24.2%

Les résultats reportés sur le tableau1 montrent que l'augmentation du nombre des données d'entraînement n'est pas réellement utiles pour améliorer le système hybride de base. Ceci est probablement du au faible nombre de paramètres associés au système indépendant du contexte. Ces paramètres sont bien estimés à partir des 1000 phrases et donc l'augmentation du nombre des données d'entraînement n'a aucun effet bénéfique sur le système. Nous proposons dans ce papier une nouvelle méthode visant à explorer ce problème. Cette méthode est basée sur des expériences qui ont déjà montré qu'il est possible d'améliorer sensiblement les performances des systèmes en combinant plusieurs modèles.

2. Description de la procédure de fusion

Cette procédure vise à éclater les données d'entraînement en plusieurs parties pour entraîner plusieurs réseaux et les recombiner lors de la phase de reconnaissance. Cet éclatement est réalisé par une simple classification des trames acoustiques (celles qui ont été incorrectement classées sont réutilisées pour entraîner un autre réseau). Cette procédure a été testée sur la base de données personnelle pour des vocabulaires de 60, 150 et 700 mots. N'ayant pas observé d'amélioration significative du taux de reconnaissance en utilisant un réseau MLP entraîné sur la totalité des données d'entraînement (4000 phrases) par rapport à celui entraîné sur un plus petit ensemble d'entraînement (voir tableau 1), nous avons alors cherché à tirer mieux parti de ces données supplémentaires dont nous disposions, pour améliorer sensiblement les taux de reconnaissance.

En premier lieu, nous entraînons un réseau MLP classique estimant les probabilités a posteriori de mots sur une petite partie des données d'entraînement (ce réseau sera par la suite dénommé MLP1). Ce réseau est alors utilisé pour filtrer le reste des données d'entraînement pour un second réseau. Les données conservées pour entraîner le second réseau sont celles pour lesquelles le premier réseau MLP1 s'est trompé dans la classification. Ainsi pour toutes les données d'entraînement, nous comparons pour chaque trame acoustique, les sorties du réseau MLP1 (correspondant aux probabilités a posteriori) et nous sélectionnons celles correspondant à la probabilité la plus élevée. Si la sortie sélectionnée est la bonne (celle correspondant à l'alignement Viterbi forcé), la donnée est écartée du nouvel ensemble d'entraînement, sinon elle est gardée. Dans un esprit de simplification et pour éviter la suppression de trames acoustiques correspondant à un même mot sur une grosse partie de la base d'entraînement, nous avons décidé, pour chaque mot de la base d'entraînement, de calculer un pourcentage d'erreur de classification des trames acoustiques (nombre de trames mal classées / nombre de trames). Si ce taux d'erreur est supérieur à un seuil fixé, le mot est gardé pour l'entraînement du deuxième réseau ; sinon, il n'est plus pris en compte. Il va de soi que plus le seuil fixé est élevé, moins il y aura de mots gardés et donc moins la subdivision des données d'entraînement sera bonne. Le seuil a été fixé de manière à obtenir un nombre suffisant de trames acoustiques (environ le nombre de trames utilisées pour entraîner MLP1) et pour s'assurer de la validité de l'entraînement des différents réseaux. Nous avons ainsi filtré une première fois les données d'entraînement pour créer un nouvel ensemble d'entraînement qui sera utilisé pour entraîner un second réseau (noté MLP2). Enfin, le reste des données d'entraînement est passé au travers des deux réseaux MLP1 et MLP2. Si ces deux réseaux sont en désaccord sur la classification des exemples présentés à l'entrée, cet exemple est alors ajouté aux données d'entraînement du troisième réseau. En revanche, si les deux réseaux sont d'accord, l'exemple est écarté. Là encore, un taux d'erreur par mot est calculé et

comparé à un seuil pour décider ou non de l'insertion dans les données d'entraînement du troisième réseau (noté MLP3). Le processus de division des données d'entraînement a été stoppé après le troisième réseau, mais il faut noter qu'il est possible de continuer le processus décrit dans ce paragraphe jusqu'à ce que l'ensemble des données d'entraînement ait été utilisé. L'entraînement des réseaux est réalisé de manière classique par propagation arrière du gradient de l'erreur quadratique. Ces réseaux sont ensuite combinés par différentes méthodes pour estimer les probabilités utilisées par les HMMs.

3. Les différentes méthodes de combinaison

En admettant que les trois réseaux aient été entraînés par la méthode décrite dans le paragraphe précédent, il faut pouvoir utiliser ces réseaux de manière efficace pour la reconnaissance. Chacun des trois réseaux MLP1, MLP2, MLP3 est composé de 10 sorties (correspondant aux 10 états stationnaires des HMM), 288 nœuds cachés et 2880 nœuds en entrée correspondant à 9 trames acoustiques de 26 paramètres, quantifiés par l'algorithme FCM (Fuzzy C-Means)[1] (voir [5], [6] pour plus de détails). Le système utilisé lors de la reconnaissance est décrit sur la figure 1.

Ainsi, après extraction des paramètres acoustiques classiques (log RASTA-PLP pour les expériences décrites ci-après, voir [3], [7] pour plus de détails) et quantification à l'aide de l'algorithme FCM, un passage par chacun des trois réseaux MLP1, MLP2 et MLP3 est effectué, nous disposons donc pour chaque trame acoustique de 3 séries de probabilités qu'il nous faut combiner. Il existe plusieurs techniques pour combiner les probabilités ; nous allons en décrire quelques-unes qui ont été testées sur la base de données personnelle.

[1] Les vecteurs acoustiques ont été quantifiés en 4 dictionnaires selon le principe de l'algorithme FCM comme suit :

- 128 prototypes pour les coefficients log RASTA-PLP
- 128 prototypes pour les Δ log RASTA-PLP
- 32 prototypes pour la dérivée première de l'énergie ΔE
- 32 prototypes pour la dérivée seconde de l'énergie ΔΔE

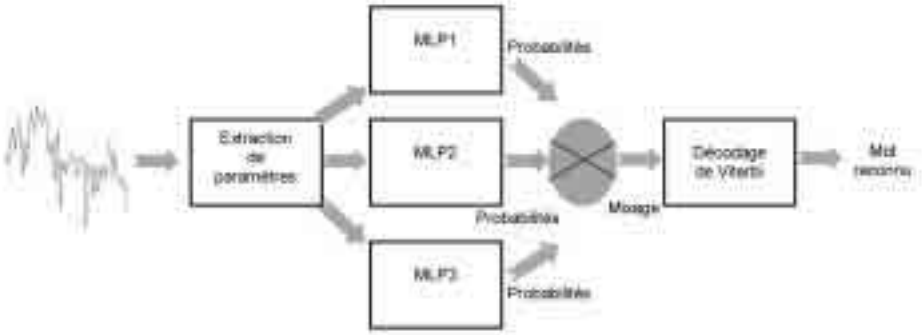

Figure1. *Le processus de reconnaissance pour la méthode de fusion.*

3.1. Combinaison linéaire

Il s'agit de la combinaison la plus simple. Chacune des composantes des 3 vecteurs de probabilités pour chaque trame acoustique est « moyennée » selon la formule classique :

$$OUT[i] = \frac{1}{N}\sum_{j=1}^{N} Out_j[i] \qquad (1)$$

où :

- N est le nombre d'experts (nombre de réseaux) utilisé (3 dans notre cas).
- $Out_j[i]$ est la composante i du vecteur observé à la sortie de l'expert j.

3.2. Combinaison linéaire dans le domaine logarithmique

C'est le même type de combinaison que celle décrite précédemment si ce n'est que nous avons utilisé $Out_j[i] = log[P_{MLP}[q_i|X]]$. Ce type de combinaison a déjà été utilisé avec succès dans des recombinaisons de systèmes à bandes multiples [2], [8] ou pour le calcul du score du « garbage »[1] en reconnaissance de mots clés par modèles hybrides [1]. Le vecteur de sortie fourni aux modèles HMM est donc la moyenne des log probabilités de chacune des sorties des réseaux.

3.3. Combinaison basée sur le critère entropique

Dans le type de combinaison précédente, le critère de sélection des vecteurs de sortie était basé sur l'accord ou le désaccord entre les deux premiers réseaux MLP1 et MLP2.

[1] Le « garbage » est un modèle qui prend en compte l'ensemble des mots prononcés par un locuteur qui n'appartiennent pas au lexique utilisé

Ici, nous nous sommes basés sur un critère du type entropique. Pour chaque trame acoustique et pour chaque réseau, nous calculons l'entropie du réseau selon la formule suivante :

$$Entropie = -\sum_{k=1}^{Noutputs} p(q_k \setminus X) * \log(p(q_k \setminus X)) \quad (2)$$

L'entropie est une mesure de la validité de l'information ou de l'incertitude d'une donnée. Le vecteur de sortie fourni au décodage est celui correspondant au réseau pour lequel l'entropie est la plus petite. En effet, dans le cas extrême où le réseau est absolument sûr (1 pour une sortie, 0 pour les autres), l'entropie est alors nulle. Dans le cas où le réseau n'est pas capable de se décider (même probabilité *1 \ N outputs* pour chacune des sorties), l'entropie vaut alors – log(*1\N outputs*) > 0.

3.4. Combinaison par l'intermédiaire d'un MLP

Ce type de combinaison fournit généralement des résultats assez bons. Il suffit d'entraîner un réseau MLP classique avec en entrée, un vecteur composé des probabilités de sortie de chacun des trois MLP et un contexte acoustique quelconque. Ainsi dans les expériences menées sur la base de données personnelle, nous avons entraîné un réseau MLP avec 3*10 composantes en entrée, correspondant aux probabilités des 3 réseaux et un contexte de 3 trames acoustiques, soit 90 nœuds d'entrée, 288 nœuds cachés, 10 nœuds de sorties.

4. Expériences et résultats

La procédure décrite dans ce papier a été testée sur une base de données personnelle. Les paramètres utilisés sont les log RASTA-PLP calculés toutes les 10 ms sur des fenêtres d'analyse de 30 ms. L'ordre de l'analyse LPC est fixé à 10. Pour la classification des trames acoustiques, nous avons utilisé l'algorithme FCM. Sur 3900 sons composant les données d'entraînement (environ 437 000 de trames acoustiques) les trois réseaux ont été entraînés sur les différentes parties décrites dans le tableau 2.

Table 2. *Nombres de trames acoustiques utilisées pour chacun des réseaux*

MLP	*Trames (train)*	*Trames (cross)*[1]
1	150 000	30 000
2	120 000	15 000
3	167 000	22 000
Réseau de base	437 000	67 000

[1] La phase de « cross validation » est utilisée pour adapter le taux d'apprentissage du MLP.

A titre de comparaison, nous mettrons dans chacun des tableaux de résultats ceux correspondant au modèle hybride HMM/ANN de base dans les mêmes conditions. Les premiers tests effectués sur cette méthode ont été réalisés au niveau de la trame acoustique (tableau 3) en premier, puis au niveau du mot (tableau 4).

Table 3. *Taux de reconnaissance au niveau de la trame acoustique pour les différentes méthodes de combinaisons : paramètres log RASTA-PLP et distributions discrètes floues.*

	Linéaire	*Log linéaire*	*Entropie*	*MLP*	*MLP de base*
Train	75.8 %	76.0 %	75.1 %	**77.2 %**	75.3 %
Cross	**74.7 %**	74.6 %	70.9 %	72.1 %	67.4 %

Tab 4. *Taux d'erreur au niveau du mot pour les trois dictionnaires de 60, 150, et 700 mots et les différentes méthodes de combinaisons : paramètres log RASTA-PLP et distributions discrètes floues.*

Taille	*Linéaire*	*Log linéaire*	*Entropie*	*MLP*	*MLP de base*
60 mots	**0.8 %**	1.1 %	1.5 %	0.9 %	2.1 %
150 mots	4.8 %	5.4 %	5.2 %	**4.3 %**	5.4 %
700 mots	16.7 %	16.8 %	18.5 %	**15.2 %**	18.7 %

5. Conclusion et perspective

Nous avons défini une méthode permettant de diviser en plusieurs parties l'ensemble d'entraînement et d'entraîner plusieurs MLP sur chacune de ces parties. Nous espérons ainsi tirer profit de l'entraînement des réseaux sur des données filtrées par la procédure de fusion mettant en exergue des propriétés différentes du signal. Différents types de combinaisons des systèmes ont été testés :

- La combinaison linéaire.
- La combinaison linéaire dans le domaine logarithmique.
- La combinaison par le critère entropique.
- La combinaison par un MLP.

Une réduction significative du taux d'erreur a pu être observée en utilisant la méthode de fusion décrite dans ce papier (44% pour 60 mots, 17% pour 150 mots et 14% pour 700 mots). Cette procédure nous a permis de tirer au mieux parti des nombreuses données d'entraînement dont nous disposions. Cette amélioration, obtenue dans le cadre d'une reconnaissance de mots isolés arabe, devrait aussi être constatée

pour un système de reconnaissance de la parole continue. Dans cette optique, la même procédure décrite dans ce papier pourra être appliquée et le même système utilisé.

Il semble que la méthode de combinaison des sorties des MLP la plus efficace (du moins pour l'expérience décrite ici), consiste à combiner les sorties des trois réseaux de neurones par le biais d'un MLP classique. Cette combinaison linéaire est très facile à mettre en œuvre.

6. Références

[1] J.M. BOITE et *al*, "Task independent and dependent training : Performance and comparison of HMM and hybrid HMM/MLP and approaches". *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol.1, pp. 617-620, April 1994.

[2] S. DUPONT et *al*, " Multi-stream speech recognition ". *Tech Rep IDIAP-RR 96-07*. IDIAP, Martigny, 1996.

[3] H. HERMANSKY, et *al*, "RASTA Processing of speech". *IEEE Trans. On Speech and Audio Processing*, vol.2, no.4, pp. 578-589, 1994.

[4] J.L. GAUVAIN et *al*, " Speaker-independent continuous speech dictation". *Proc. Speech Communication*, November 1994.

[5] L. LAZLI et *al*, "Modèle hybride HMM-MLP basé flou : appliqué à la reconnaissance de la parole arabe". *SETIT2003/IEEE : conférence internationale sur les Sciences Electroniques, Technologie de l'Information et des Télécommunications*, pp.104, 17-21 Mars, Sousse, Tunisie, 2003.

[6] L. LAZLI et *al*, "Connectionist probability estimators in HMM arabic speech recognition using fuzzy logic". *MLDM'03: the 3rd international conference on Machine Learning and Data Mining in pattern recognition, LNAI 2734, Springer-verlag*, pp. 379-388, juillet 5-7, Leipzig, Allemagne, 2003.

[7] L. LAZLI, "Discriminant learning for hybrid HMM-ANN system using a fuzzy clustering for arabic speech recognition". *JIEEE'03: the 5th Jordanian International Electrical and Electronics Engineering conference*, pp?, octobre 14-16, Amman, Jordan, 2003.

[8] A. MORRIS et *al*, "MAP combination of multi-stream HMM or HMM/ANN experts". *In Eurospeech, Special Event Noise Robust Recognition*, Aalborg, Denmark, 2001.