Combinaison de différents jeux de paramètres acoustiques pour la reconnaissance de la parole.

Loic Barrault, Driss Matrouf, Georges Linarès, Renato De Mori

LIA - Université d'Avignon, BP 1228
84911 Avignon Cedex 9 - France
{loic.barrault, driss.matrouf, georges.linares, renato.demori}@univ-avignon.fr

ABSTRACT

With the purpose of improving Automatic Speech Recognition (ASR) systems performance, many different approaches on combining them have been studied. In this paper, a combination of state *a posteriori* probabilities given by different feature sets is proposed. In order to perform a coherent combination of state posterior probabilities, the acoustic models trained on different feature sets must have the same topology (*i.e.* same set of states). For this purpose, a fast and efficient *twin* model training protocol is proposed. Then, two different strategies for combining probabilities are presented : the linear and the log linear interpolation. By using log linear interpolation, a relative Word Error Rate (WER) reduction of about 15% on MEDIA and 14% on ESTER corpora have been respectively observed.
Keywords: speech recognition, posterior probabilities combination

1. Introduction

Les systèmes de reconnaissance automatique de la parole (RAP) commettent des erreurs dues à l'imperfection des modèles utilisés, aux limitations des paramètres acoustiques extraits à partir du signal de parole et aux approximations faites par les décodeurs. Dans le but d'accroître les performances des systèmes de reconnaissance, il a été proposé de les combiner. La combinaison peut se faire aux différents niveaux du système, à savoir, les paramètres acoustiques, les probabilités générées par les modèles acoustiques, le décodage et les hypothèses de reconnaissance.

L'utilisation de réseaux de neurones, d'arbres de décision et d'autres techniques d'apprentissage automatique ont été utilisées pour combiner les résultats de plusieurs systèmes de RAP afin de réduire le taux d'erreur mot (WER) [10]. Dans [12], la combinaison log-linéaire des probabilités des mots issues de différents modèles acoustiques fournit une amélioration significative des performances. Dans [9], des modèles acoustiques différents sont obtenus en utilisant des méthodes de *tying* de gaussiennes aléatoires. Chaque modèle utilise le même jeu de paramètres acoustiques. Les hypothèses de reconnaissance obtenues sont ensuite combinées avec ROVER. Une manière efficace de combiner les résultats de différents systèmes consiste à effectuer une combinaison par réseaux de confusion (CNC, [8]). Cette technique permet d'obtenir une meilleure approximation des probabilités *a posteriori* des mots.

L'utilisation de plusieurs jeux de paramètres différents repose sur l'hypothèse que certaines caractéristiques du signal de parole sont capturées par certains jeux de paramètres et ignorées par d'autres. Cela motive donc l'idée de vouloir combiner ces flux d'observations acoustiques dans le but de capturer l'information complémentaire présente dans chacun d'eux. Des paramètres spécifiques (comme par exemple le voisement, [11]) ont été intégrés dans un flux de traits acoustiques afin d'apporter de l'information supplémentaire dans le vecteur de paramètres acoustiques. Une généralisation de cette approche consiste à concaténer différents jeux de paramètres acoustiques en un seul flux d'observations. Afin de réduire la complexité de modélisation inhérente à ce type de combinaison, des algorithmes ont été développés afin de sélectionner des sous-ensembles d'observations acoustiques parmi un long flux. Cette sélection se base sur un critère qui optimise la classification automatique des données de parole en phonèmes ou traits phonétiques [6]. Une solution consiste à sélectionner un ensemble de mesures acoustiques qui garantissent une grande valeur de l'information mutuelle entre ces mesures et des paramètres phonétiques caractéristiques.

Lorsque plusieurs jeux de paramètres sont utilisés en vue de les combiner, le problème de la comparaison des probabilités calculées avec les différents modèles se pose. Dans [6], une pondération dépendante de l'état des log-vraisemblances relative à différents éléments du vecteur de paramètres est proposé. Une autre approche consiste à combiner les probabilités au niveau de la trame [5]. Il est montré que ce type de combinaison produit une amélioration plus grande des résultats que la combinaison par CNC.

Dans cet article, nous proposons d'effectuer une combinaison basée sur la trame avant le décodage. Des modèles acoustiques ayant la même topologie (*i.e.* même ensemble d'états) mais utilisant des paramètres acoustiques différents sont considérés. Chaque modèle utilise un jeu de paramètres différent Y_n^i pour la $n^{\text{ième}}$ trame de parole afin de produire la vraisemblance d'un état $L(Y_n^i|q)$ pour chaque état q. Ces vraisemblances sont normalisées et combinées trame à trame pour produire un rapport de vraisemblances normalisées. Des réductions significatives du WER ont été obtenues sur les corpus de parole téléphonique français à grand et très grand vocabulaires, MEDIA et ESTER.

La section 2 décrit la procédure d'entraînement spécifique utilisée pour générer les modèles acoustiques utilisés pour les expériences de reconnaissance de la parole. L'architecture et les combinaisons linéaire et log-linéaire des rapports de vraisemblances sont présentées dans la section 3. La section 4 contient les résultats expérimentaux obtenus.

2. Apprentissage de modèles *jumeaux*

La parole est une source d'information produisant un signal dans lequel des symboles sont encodés. Comme le signal affiche une grande variabilité pour une même phrase, les séquences d'échantillons sont transformées en vecteurs de paramètres acoustiques, plus stables et plus pertinents pour la reconnaissance de la parole. Chaque trame Y_n est transformée en un vecteur de paramètres acoustiques représenté dans un espace acoustique. Considérons $\Im^i, i = \{1, \ldots, I\}$, un ensemble d'espaces acoustiques correspondant à différents jeux de paramètres $\{Y^i\}$, et $Y_n^i, i = \{1, \ldots, I\}$ les instances de la trame Y_n dans ces espaces acoustiques. Les vecteurs de paramètres acoustiques sont utilisés pour calculer la vraisemblance qu'un symbole q appartenant à un vocabulaire Q soit présent dans la trame. Considérons maintenant des modèles acoustiques dépendant du contexte composés de modèles de Markov cachés (HMM) dans lesquels une mixture de gaussiennes modélise la densité de probabilité pour chaque état q. La génération d'hypothèses de mots est effectuée avec une stratégie de décodage qui estime les probabilités postérieures $P(q|Y_n)$ des états du modèle avec des vraisemblances normalisées.

Afin de combiner au niveau de la trame les probabilités calculées avec plusieurs jeux de paramètres, il est nécessaire d'entraîner des modèles acoustiques différents mais ayant la même topologie. Nous proposons donc une technique permettant d'entraîner des modèles acoustiques *jumeaux*, possédant le même ensemble d'états correspondant respectivement au même contexte phonétique (par exemple, les modèles auront tous un état équivalent pour modéliser la partie centrale du phonème [i])

Considérons un modèle source M^0 utilisant un jeu de paramètres acoustiques Y^0. Le but est de créer des nouveaux modèles M^i ayant le même ensemble d'états que M^0, mais chacun utilisant un jeu de paramètres Y^i. Pour cela, l'alignement forcé du corpus d'entraînement est effectué avec le modèle source M^0. Chaque mixture de gaussienne (GMM) associé à chaque état de M^i est entraîné en utilisant l'algorithme Expectation-Maximization (EM) en suivant les étapes suivantes :

- L'étape *Expectation* est effectuée en utilisant le jeu de paramètres Y^0 sur le GMM correspondant de M^0
- L'étape *Maximization* est effectuée en utilisant le jeu de paramètres Y^i
- Les paramètres du modèle M^i sont ré-estimés en utilisant plusieurs itérations d'adaptation par maximum *a-posteriori* (MAP). La segmentation du corpus d'entraînement est mise à jour en utilisant le M^i obtenu à chaque itération.

Le modèle source doit avoir de bonnes performances puisqu'il est utilisé pour fixer la variable cachée de chaque trame Y_n^i. Il est important de remarquer que cette procédure assure que chaque état correspondant des modèles correspondent à la même unité acoustico-phonétique. Au final, le modèles source et les modèles jumeaux font une partition équivalente de leurs espaces acoustiques (même nombre de gaussiennes), mais les distributions des symboles au sein de ces zones sont différentes puisqu'elles ont été ré-estimées en utilisant d'autres jeux de paramètres acoustiques.

3. Architecture pour la combinaison de probabilités *a posteriori* au niveau de la trame

Les modèles sont utilisés dans l'architecture représentée dans la figure 1.

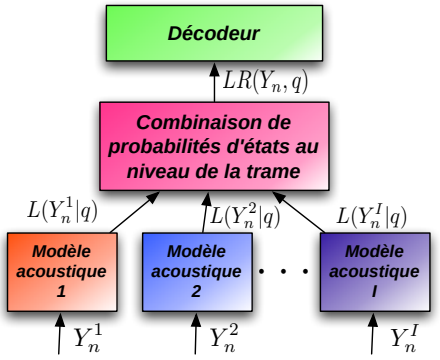

Fig. 1: *Architecture du système pour la combinaison de probabilités au niveau de la trame.*

Les vraisemblances $L(Y_n^i|q)$ sont calculées de manière synchrone pour chaque jeu de paramètres acoustiques. Ensuite, pour chaque trame, un rapport de vraisemblances $LR(n, q)$ est calculé. Plusieurs manières pour combiner les probabilités postérieures peuvent être utilisées.

La combinaison linéaire des probabilités postérieures présuppose que chaque jeu de paramètres fournit une quantité d'information proportionnelle à leur probabilité postérieure correspondante. La relation suivante peut alors être utilisée pour la combinaison trame à trame des probabilités *a posteriori* des états :

$$\hat{P}(q|Y_n) = \sum_{i=1}^{I} \alpha_i P(q|Y_n^i) \text{ avec } \sum_{i=1}^{I} \alpha_i = 1 \quad (1)$$

où α_i est un facteur de pondération utile pour introduire la confiance que l'on confère à un jeu de paramètres. Les modèles acoustiques ont des GMM associés aux états permettant d'obtenir des vraisemblances $L(Y_n^i|q)$. En admettant que la probabilité *a priori* $P(q)$ d'un état q est la même pour tous les états, les probabilités $P(q|Y_n)$ et $P(Y_n|q)$ sont proportionnelles à la combinaison linéaire de vraisemblances suivante :

$$LCLR(n, q) = \sum_{i=1}^{I} \left[\alpha_i \frac{L(Y_n^i|q)}{\sum_{g \in Q} L(Y_n^i|g)} \right] \quad (2)$$

Une autre manière de combiner les probabilités postérieures trame à trame consiste à supposer que le système est dans l'état q seulement si tous les jeux de paramètres sont d'accord avec cela.

$$\hat{P}(q|Y_n) = P(q|Y_n^1, \ldots, Y_n^I) = \prod_{i=1}^{I} P(q|Y_n^i) \qquad (3)$$

En faisant les mêmes hypothèses que pour l'équation 2, la combinaison log-linéaire de rapports de vraisemblances suivante est utilisée :

$$LLCLR(n,q) = \sum_{i=1}^{I} \alpha_i \log \left[\frac{L(Y_n^i|q)}{\sum_{g \in Q} L(Y_n^i|g)} \right] \qquad (4)$$

Si on ne fait aucune distinction entre les jeux de paramètres, alors on peut utiliser $\alpha_i = \frac{1}{I}$. Cette hypothèse a produit de bons résultats dans les expériences décrites dans la section suivante.

4. Expériences

Trois jeux de paramètres correspondant à trois manières sensiblement différentes de transformer le signal de parole ont été considérés. Les coefficients obtenus par Prédiction Linéaire Perceptuelle (PLP) [3] ont été utilisés pour construire le premier modèle. Le second jeu de paramètres est obtenu en ajoutant le filtrage RASTA [4] aux paramètres PLP. Ce jeu sera dénoté RPLP. Le troisième jeu de paramètres est calculé avec l'Analyse à Résolution Multiple (MRA) suivie d'une analyse en composantes principales [2]. Tous les vecteurs de paramètres contiennent les dérivées premières et secondes.

Le système fondé sur les HMMs utilisé pour les expériences décrites dans cette section est SPEERAL, décrit dans [7]. Il possède un vocabulaire de 65K mots, 10040 modèles de phonèmes dépendant du contexte, 3600 états émetteurs pouvant être partagés parmi les modèles ayant le même phonème central et 232716 gaussiennes. Les modèles acoustiques ont été entraînés séparément en utilisant l'approche d'apprentissage de modèles jumeaux avec les 82 heures de parole téléphonique du corpus d'entraînement d'ESTER [1]. Le corpus d'apprentissage composé de 82639 mots d'un autre corpus français, MEDIA a été utilisé pour adapter les modèles avec l'adaptation MAP.

Un ensemble de résultats en terme de WER sont présentés dans le tableau 1. Ils ont été obtenus avec le corpus de test de MEDIA. MEDIA est un corpus de dialogues enregistré en utilisant le protocole du Magicien d'Oz : 250 locuteurs ont effectué des réservations d'hôtels en suivant 5 scénarios. Ce corpus de parole téléphonique est composé de 3771 phrases et 26092 mots. Les résultats obtenus avec la combinaison trame à trame des probabilités postérieures calculées avec des modèles utilisant les paramètres MRA, RPLP et PLP sont présentés.

Les combinaisons linéaire et log-linéaire des probabilités postérieures fournissent une réduction conséquente du WER. Les meilleurs résultats ont été obtenus en effectuant une combinaison log-linéaire des probabilités issues des modèles utilisant les trois jeux de paramètres disponibles. On observe une réduction du WER d'environ 14% relativement au meilleur système utilisant un seul jeu de paramètres.

Tab. 1: *Résultats de la combinaison trame à trame sur le corpus de test de MEDIA. (G.R. : Gain relatif, I.C. : intervalle de confiance)*

Paramètres	WER (%)	G.R. (%)	I.C. (%)
MRA	33.2	-	0.57
RPLP	32.2	-	0.57
PLP	32.1	-	0.57
En utilisant la combinaison linéaire			
MRA+RPLP	29.5	8.4	0.55
MRA+PLP	28.2	12.1	0.55
RPLP+PLP	28.0	13.0	0.54
MRA+RPLP+PLP	28.1	12.7	0.55
En utilisant la combinaison log linéaire			
MRA+RPLP	29.2	9.3	0.55
MRA+PLP	28.2	12.1	0.55
RPLP+PLP	28.2	12.1	0.55
MRA+RPLP+PLP	**27.6**	**14.0**	0.54
Rover	29.3	8.7	0.55
Oracle	25.4	20.8	0.52

L'Oracle consiste à sélectionner le vecteur de probabilités proposé par le modèle fournissant la plus grande probabilités pour l'état qui a effectivement émis la trame. Cet état est déterminé par alignement forcée de la référence sur le signal de parole.

Des expériences de reconnaissance ont été effectuées sur la partie téléphonique du corpus de test d'ESTER (512 phrases, 4813 mots). Les résultats obtenus avec les mêmes types de combinaison sont présentés dans le tableau 2.

Tab. 2: *Résultats de la combinaison trame à trame sur ESTER. (G.R. : Gain relatif, I.C. : intervalle de confiance)*

Paramètres	WER (%)	G.R. (%)	I.C. (%)
MRA	41.1	-	1.39
RPLP	37.9	-	1.37
PLP	46.6	-	1.40
En utilisant la combinaison linéaire			
MRA+RPLP	35.2	7.1	1.35
MRA+PLP	33.0	12.9	1.33
RPLP+PLP	33.7	11.1	1.34
MRA+RPLP+PLP	35.1	7.4	1.35
En utilisant la combinaison log linéaire			
MRA+RPLP	35.5	6.3	1.35
MRA+PLP	34.8	8.2	1.35
RPLP+PLP	35.9	5.3	1.36
MRA+RPLP+PLP	**32.2**	**15.0**	1.32

Les meilleurs résultats ont également été obtenus avec la combinaison log linéaire des trois jeux de paramètres. Une réduction du taux d'erreur d'environ 15% relativement au meilleur système utilisant un seul jeu de paramètres a été observé. Les bénéfices de l'approche proposée pour la combinaison trame à trame de probabilités *a posteriori* est évidente même dans le cadre de très grand vocabulaires.

Analyse des résultats. Les résultats expérimentaux montrent que la combinaison trame à trame des probabilités des états calculées avec des modèles utilisant des paramètres acoustiques différents mène à une réduction conséquente du WER. On observe la même tendance pour les deux corpus exploités.
Lorsque deux modèles sont utilisés, la combinaison linéaire des probabilités produit des résultats légèrement meilleurs que la combinaison log-linéaire, même si l'avantage est peu significatif. Cependant, lorsque l'on combine les trois systèmes à notre disposition, la combinaison log-linéaire surpasse significativement la combinaison linéaire. On peut d'ailleurs remarquer que la combinaison linéaire des probabilités issues des trois modèles dégrade le WER par rapport à la combinaison de deux modèles.

Ces résultats ont été générés en utilisant un poids égal pour chaque jeu de paramètres acoustiques ($\alpha_i = \frac{1}{I}$ dans les équations 2 et 3). Nous avons tenté d'introduire une mesure de confiance afin de pondérer les vecteurs de probabilités produits par chaque modèle. Plusieurs expériences ont été menées avec différents poids de combinaison, mais aucune n'a produit de WER plus faible que la combinaison utilisant des poids égaux. En particulier, l'utilisation de poids inversement proportionnels à l'entropie du vecteur de probabilités augmente légèrement le taux d'erreur. L'analyse de l'entropie des vecteurs de probabilités montre qu'elles sont très semblables entre les différents jeux de paramètres, mais très hétérogènes comparées à celles des vecteurs de probabilités des combinaisons considérées. Un fait remarquable est que les combinaisons linéaire et log-linéaire provoquent une réduction de l'entropie moyenne des vecteurs de probabilités.

5. Discussion et conclusions

La combinaison des probabilités *a posteriori* des états offre une solution à plusieurs problèmes pouvant être rencontrés avec d'autres approches. Les méthodes de combinaison après le décodage, telles que les réseaux de confusion CNC ou ROVER, sont limitées par le fait qu'elles opèrent sur des sorties asynchrones qui ne sont plus reliées au signal et qui ont été obtenues en utilisant une sélection préliminaire des hypothèses les plus probables se basant sur de l'information et un savoir partiel. Par ailleurs, ce genre de combinaison ne reconsidère pas les hypothèses de mots trouvées par les différents systèmes et ne produit pas de nouvelles hypothèses. Elle espère qu'un ou plusieurs systèmes propose l'hypothèse correcte et tente de l'extraire en utilisant des mesures de confiance. Dans notre protocole, le processus de décodage est directement influencé par une combinaison effectuée en amont de sorte que les hypothèses qui auraient pu être élaguées à cause d'une faible probabilité peuvent être réévaluées. Cette combinaison à bas niveau ne compte pas sur les hypothèses faites par les systèmes et tente de capturer l'information complémentaire au niveau de la trame avant que les hypothèses de mots ne soient produites.

Lorsque trois jeux de paramètres sont combinés, la combinaison log-linéaire surpasse la combinaison linéaire. Dans le but d'expliquer ce fait, une analyse détaillée des distributions des probabilités des états doit être effectuée. L'ajout d'information complémentaire comme par exemple d'autres modèles appris avec de nouveaux jeux de paramètres est simplifié par l'apprentissage de modèles jumeaux, et doit être considéré pour des études futures. Une perspective à ce travail consiste également à évaluer d'autres types de combinaisons, éventuellement hiérarchiques, permettant de tirer profit de cette architecture.

Références

[1] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, JF. Bonastre, and G. Gravier. The ester phase ii evaluation campaign for the rich transcription of french broadcast news. In *EUROSPEECH*, Lisbon, Portugal, September 2005.

[2] R. Gemello, F. Mana, D. Albesano, and R. De Mori. Multiple resolution analysis for robust automatic speech recognition. *Computer Speech and Language*, 20(1) :2–21, 2006.

[3] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *The Journal of the Acoustical Society of America*, 87 :1738–1752, 1990.

[4] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4) :578–589, October 1994.

[5] B. Hoffmeister, T. Klein, R. Schluter, and H. Ney. Frame based system combination and a comparison with weighted rover and cnc. In *ICSLP*, pages 537–540, 2006.

[6] M. Kamal Omar and M. Hasegawa-Jonhson. Maximum mutual information based acoustic-features representation of phonological features for speech recognition. In *ICASSP*, volume 1, pages 81–84, Orlando, FL, 2002.

[7] G. Linarès, P. Nocera, D. Massonié, and D. Matrouf. The lia speech recognition system : From 10xrt to 1xrt. *Lecture Notes in Computer Science*, 4629/2007 :302–308, 2007.

[8] L. Mangu, E. Brill, and A. Stolcke. Finding consensus among words : Lattice-based word error minimization. In *EUROSPEECH*, volume 1, pages 495–498, 1999.

[9] O. Siohan, B. Ramabhadran, and B. Kingsbury. Constructing ensembles of asr systems using randomized decision trees. In *ICASSP*, volume 1, pages 197–200, Philadelphia, PA, March 2005.

[10] T. Utsuro, Y. Kodama, T. Watanabe, H. Nishizaki, and S. Nakagawa. Confidence of agreement among multiple lvcsr models and model combination by svm. In *ICASSP*, volume 1, pages 16–19, Hong Kong, China, 2003.

[11] A. Zolnay, R. Schluter, and H. Ney. Robust speech recognition using a voiced-unvoiced feature. In *ICSLP*, volume 2, pages 1065–1068, Denver, CO, 2002.

[12] A. Zolnay, R. Schluter, and H. Ney. Acoustic feature combination for robust speech recognition. In *ICASSP*, volume 1, pages 457–460, Philadelphia, PA, March 2005.