Comparaison de différentes méthodes de classification pour la détection de mots clés en parole continue

Yassine BenAyed* – Dominique Fohr* – Jean Paul Haton* – Gérard Chollet**

* Laboratoire Lorrain de Recherche en Informatique et ses Applications
LORIA/INRIA Lorraine BP239
54506, Vandœuvre-les-Nancy, France
(ybenayed, fohr, jph)@loria.fr

** Ecole Nationale Supérieure des Télécommunications
ENST, CNRS-LTCI 46 rue Barrault
75634 Paris, France
chollet@tsi.enst.fr

RÉSUMÉ. Cet article s'inscrit dans le cadre de la détection de mots clés dans un flux de parole. Nous présentons le problème de détection comme un problème de classification où chaque mot clé peut appartenir à deux classes différentes, à savoir "correct" et "incorrect". Cette classification est réalisée tout d'abord, par l'utilisation des Réseaux de Neurones Artificiels (RNA) en particulier le Perceptron Multi-Couches (PMC). Ensuite, nous proposons l'utilisation des SVM comme technique de classification innovante et efficace et qui a fait ses preuves dans plusieurs domaines de recherche. Chaque mot clé reconnu est représenté par un vecteur caractéristique qui constitue l'entrée du classifieur. Pour déterminer ce vecteur, nous proposons trois représentations vectorielles basées sur l'emploi des probabilités d'observations acoustiques locales et de la durée de chaque état.

ABSTRACT. This paper deals about the detection of keywords in a speech flow. We present the problem of detection as a problem of classification where each keyword can belong to two different classes, i.e., "correct" and "incorrect". This classification is carried out by the use of Artificial Neural Networks, in particular Multi-Layer Perceptron. Then we propose the use of the SVM, an innovating technique for classification which proved reliable in several research fields. Each recognized keyword is represented by a characteristic vector which constitutes the entry of the classifior. To determine this vector, we propose three vectorial representations based on the probability of local acoustic observation and the duration of each state.

MOTS-CLÉS : Détection de mots clés, Perceptron multi-couches, support vector machines

KEYWORDS : Keyword detection, multi-layer Perceptron, support vector machines

1. Introduction

Les utilisateurs d'un système de reconnaissance automatique de la parole sont souvent peu conscients des contraintes du système et s'expriment généralement dans des environnements bruités, comme c'est le cas, par exemple des services vocaux interactifs par téléphone. Les systèmes de reconnaissance doivent être capables de rejeter les entrées incorrectes, les mots hors-vocabulaire et les diverses perturbations accidentelles (hésitations, bruit, etc.).

Plusieurs recherches ont été menées sur la détection de mots clés dans un flux de parole. Les modèles "poubelles" sont généralement utilisés dans les systèmes de reconnaissance afin d'absorber les mots hors-vocabulaire [1] [2]. Une mesure de confiance peut également être utilisée, puisqu'elle représente la fiabilité des hypothèses de reconnaissance [3] [4]. Au cours d'une procédure de post-traitement, les hypothèses les moins fiables seront rejetées. Cette dernière étape consiste en un processus de vérification des mots reconnus, il s'agit donc d'une classification de ces mots en deux classes : correct et incorrect. C'est pour cette raison que nous avons pensé à utiliser une méthode de classification plus efficace que le simple seuil de la mesure de confiance. Cette approche présente aussi l'avantage de pouvoir utiliser un ensemble de caractéristiques relatives à chaque mot reconnu et de les présenter au classifieur afin de décider de l'acceptation ou du rejet du mot considéré. Tout d'abord, nous commençons par l'utilisation des Réseaux de Neurones Artificiels (RNA) en particulier le Perceptron Multi-Couches (PMC). Ensuite nous présentons les SVM (*Support Vevtor Machines*, en français Machine à vecteur support) comme une technique de classification innovante et efficace et qui a fait ses preuves dans plusieurs domaines de recherche.

Les bases des SVM ont été développées par Vapnik [5]. Les SVM ont gagné de la popularité grâce à plusieurs caractéristiques attractives et prometteuses [6]. Leur formulation repose sur le principe de minimisation du risque structurel qui a été montré meilleur que la minimisation du risque empirique traditionnellement employée dans les RNA conventionnels. La minimisation du risque structurel minimise la limite supérieure du risque prévu, ce qui s'oppose à la minimisation du risque empirique qui minimise l'erreur sur la base d'apprentissage. C'est cette différence qui donne aux SVM une plus grande capacité de généralisation, but de l'apprentissage statistique. Les SVM ont été développés afin de résoudre le problème de classification, mais récemment ils ont été étendus aux problèmes de régression [7].

Cet article est organisé comme suit : nous présentons dans la deuxième section notre système de reconnaissance. Dans la troisième section nous décrivons la base de données utilisée. Nous détaillons dans le quatrième section les différentes représentations vectorielles proposées pour réaliser la phase de classification suivies des résultats expérimentaux et enfin, nous discutons ces résultats dans la cinquième partie.

2. Description du système de reconnaissance

Dans cet article, nous utilisons le système de reconnaissance automatique de la parole ESPERE développé au LORIA [8]. Pour faire l'apprentissage de notre système, nous utilisons une modélisation contextuelle des phonèmes. Ces modèles de phonèmes dépendent du contexte et permettent ainsi de tenir compte des phénomènes de coarticulation. Il s'agit des triphones, un triphone tient compte de deux phonèmes, le phonème précédent et le phonème suivant afin de fixer le contexte. Chaque triphone est modélisé par un HMM à 3 états (gauche-droite) à densités continues, avec 4 gaussiennes par état.

L'extraction des paramètres a été faite en se basant sur les coefficients Mel-Cepstre (MFCC). Nous avons utilisé une fenêtre d'analyse de 256 échantillons de signal et 24 filtres triangulaires. À chaque trame, correspond un vecteur de 35 coefficients : 11 MFCC (le premier coefficient C_0 est enlevé), 12 dérivées premières et 12 dérivées secondes. Pendant la phase d'apprentissage, les phonèmes sont appris dépendamment du contexte et, en reconnaissance, chaque mot clé est obtenu par la concaténation des modèles de triphones qui les composent.

3. Base de données

Pour l'apprentissage du système de reconnaissance, nous avons utilisé un corpus extrait de la base de données française SPEECHDAT 1000. Le corpus utilisé est composé de 8836 phrases phonétiquement riches, prononcées par 1000 locuteurs et enregistrées à travers le réseau téléphonique. Cette base est utilisée pour l'apprentissage de 9562 modèles de triphones. La base de test utilisée contient 2245 phrases lues représentant une durée de 3 heures et 23 minutes. Ces phrases appartiennent à la base de données française SPEECHDAT 5000. Ces phrases, phonétiquement riches sont prononcées par 900 locuteurs, différents de ceux qui ont participé à l'enregistrement de la base d'apprentissage. Afin de fixer les paramètres de nos différentes méthodes, nous avons utilisé une base de développement composée de 250 phrases extraites de la base de données française SPEECHDAT 1000. Ces phrases sont différentes de celles utilisées dans la phase d'apprentissage.

Dans le corpus de test utilisé, nous disposons de 15820 mots dont 3220 sont des occurrences de mots clés et 12600 sont des mots hors-vocabulaire. Notre tâche consiste à détecter les occurrences des 20 mots clés choisis dans cette base de test.

4. Représentation vectorielle des mots

D ans un processus de classification, nous avons besoin de deux éléments importants. Le premier est le choix du classifieur à utiliser et le deuxième est le choix de la construction des vecteurs représentatifs des données. Ces vecteurs serviront comme entrée au classifieur. Nous avons choisi d'utiliser les classifieurs PMC et SVM. Il nous reste alors à choisir le deuxième élément qui consiste à représenter chaque mot clé reconnu par un vecteur caractéristique.

Dans une représentation vectorielle, nous pouvons introduire plusieurs informations qui caractérisent au mieux un mot clé. Nous disposons essentiellement de deux types d'informations qui sont les probabilités acoustiques locales de chaque phonème et la durée de ses états. Ces deux informations sont obtenues après une phase d'alignement de la sortie du système de reconnaissance sur les modèles HMM des phonèmes et elles représentent respectivement une caractéristique acoustique et une caractéristique temporelle. Nous allons tester ces deux types d'informations afin de bien représenter les mots clés. Dans une étape ultérieure, nous combinons aussi ces deux informations dans l'optique d'une amélioration au niveau de la représentation des mots et au niveau des performances du système de détection.

4.1. Utilisation des probabilités d'observations acoustiques locales

Un mot est composé d'une séquence de phonèmes et à chaque phonème correspond une probabilité d'observation acoustique locale. Cette probabilité constitue une information acoustique importante permettant de bien caractériser les mots à classer. En adoptant une telle représentation, chaque mot clé sera caractérisé par un vecteur contenant les probabilités acoustiques locales des phonèmes qui le constituent. L'utilisation d'un tel vecteur aidera sans doute le classifieur à bien discriminer ses classes puisqu'il ne s'agit pas d'une seule mesure, mais au contraire, d'un vecteur de probabilités. Pour ces raisons, nous avons adopté cette représentation dans laquelle un vecteur caractéristique d'un mot clé donné peut être représenté comme suit :

$$\vec{V}_{Mot} = [P(O_1|Ph_1), P(O_I|Ph_i), ..., P(O_T|Ph_N)]$$

Cette représentation vectorielle a été testée avec les deux classifeurs PMC et SVM. Après une série d'expériences en utilisant la base d'apprentissage, le nombre de neurones dans la couche cachée de PMC a étét fixé à 4. Le nombre de neurones dans la couche d'entrée est fixé à 10 représentant le nombre maximal de phonèmes dans un mot clé. La sortie du réseau contient un seul neurone. Concernant le classifieur SVM, nous avons utilisé deux fonctions noyaux : linéaire et RBF (fonctions radiales de base).

Pour évaluer les méthodes proposées, nous avons utilisé la courbe ROC (Receiver Operating Characteristic) appelée aussi *courbe caractéristique d'opération du récepteur* qui consiste à représenter le taux de détection (TD) en fonction du nombre de Fausses

Acceptations par Mot Clé et par Heure (FA/MC/H). Cette disposition de courbe *ROC* prend mieux en compte la notion du temps puisqu'au niveau de l'axe des abscisses le nombre de fausses acceptations est normalisé par le nombre d'heures. Ce type de courbe est considéré comme étant la plus efficace pour la comparaison de différents systèmes de détection. La figure 1 présente les courbes ROC correspondantes à l'utilisation du PMC et des SVM avec une représentation vectorielle des mots à base des probabilités d'observations acoustiques locales.

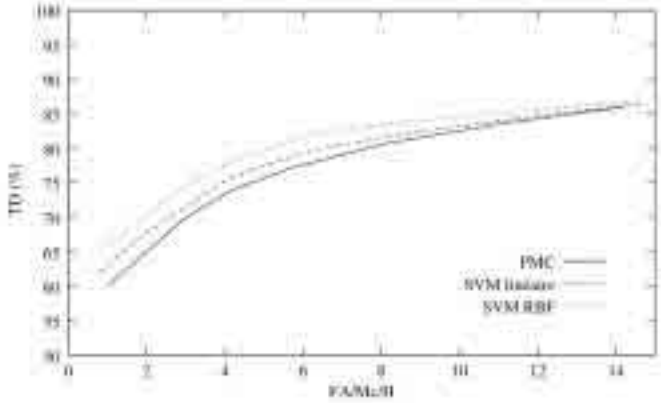

Figure 1. *Courbes ROC obtenues par une représentation vectorielle à base de probabilités acoustiques locales des phonèmes.*

Pour obtenir une valeur unique décrivant cette courbe, nous utilisons la moyenne des probabilités de détection pour un taux de fausses acceptations par mot clé et par heure variant entre 0 et 10. Cette valeur est appelée *Valeur de mérite* ou FOM pour "Figure Of Merit".

En adoptant cette représentation des mots, le PMC nous a donné une valeur de FOM égale à 74%. Les SVM linéaires ont amélioré les résultats en augmentant la valeur de FOM à 76%. Les meilleures performances ont été enregistrées en utilisant les SVM à noyau RBF avec un FOM de 78.4%.

4.2. Utilisation du nombre de trames par état

Nous proposons ici une autre représentation vectorielle, différente de celle des probabilités où nous utilisons le nombre de trames par état dans chaque phonème. Un mot est alors représenté par un vecteur à 30 éléments correspondant aux nombres de trames de chaque état dans chaque phonème du mot en question.

Nous avons expérimentallement fixé pour le PMC le nombre de neurones dans la couche cachée à 10, le nombre de neurones dans la couche d'entrée à la valeur de 30 et un

seul neurone pour la sortie. Pour le classifieur SVM, nous avons utilisé les deux mêmes fonctions noyaux, linéaire et RBF.

Les résultats obtenus par les deux méthodes de classification PMC et SVM en utilisant la représentation vectorielle à base de nombre de trames par état dans chaque phonème constituant le mot clé sont présentés sur la figure 2. Les courbes correspondent aux courbes ROC obtenues par nos deux types de classifieurs.

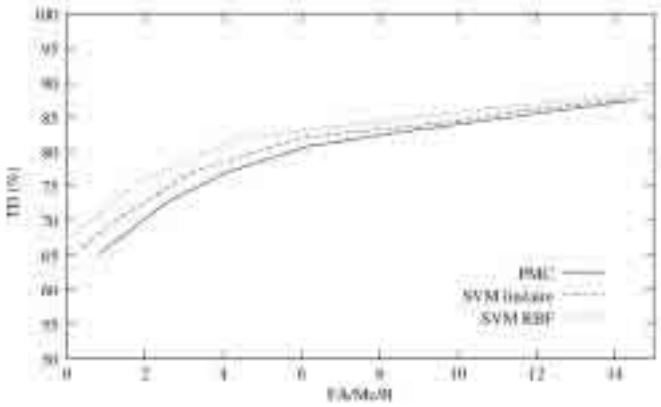

Figure 2. *Courbes ROC obtenues par une représentation vectorielle à base du nombre de trames par état dans chaque phonème.*

Les valeurs de FOM obtenues en utilisant cette représentation vectorielle est de 77.4% pour le PMC, de 79.2% pour les SVM linéaires et de 81.2% pour les SVM RBF.

4.3. Représentation vectorielle mixte

Afin de mieux représenter les mots en sortie du système de reconnaissance, nous avons voulu exploiter le maximum d'information pouvant les caractériser, en l'occurrence les deux types de données déjà utilisés et qui représentent deux informations complémentaires au niveau de chaque mot : les probabilités d'observations acoustiques locales caractérisant le mot d'un point de vue acoustique et le nombre de trames dans chaque état, qui constitue une information d'ordre temporel. Ces deux types de représentations ont déjà donné de très bons résultats, leur combinaison ne peut qu'améliorer les performances. Un mot sera donc représenté par un vecteur de 40 éléments composé des probabilités acoustiques locales des phonèmes et du nombre de trames de chaque état dans chaque phonème.

Nous avons évalué les capacités de cette représentation mixte avec les deux types de classifieurs PMC et SVM. Le PMC utilisé est un réseau à 40 neurones d'entrée, 12 neurones dans la couche cachée et un neurone de sortie.

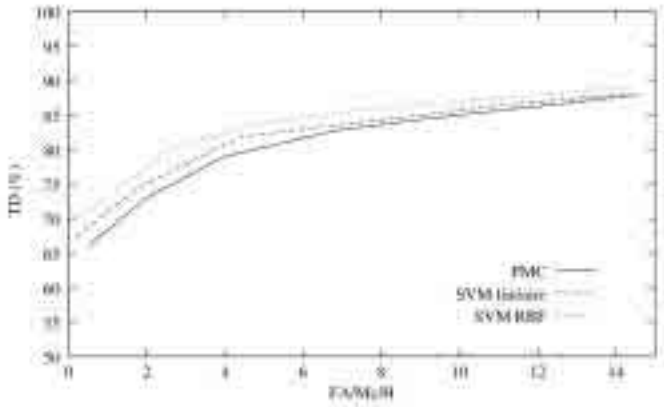

Figure 3. *Courbes ROC obtenues par une représentation vectorielle mixte.*

Les résultats obtenus par cette représentation mixte avec les deux types de classifieurs sont présentés sur la figure 3 avec les courbes ROC correspondantes. Les meilleurs performances ont été réalisées par les SVM à noyau RBF où on atteint les 82.9% comme valeur de FOM. Les SVM à noyau linéaire ont donné aussi de bons résultats en atteignant les 81.5% comme valeur de FOM. Le PMC aussi a réalisé les meilleurs de ses performances en utilisant cette représentation mixte. En effet, nous obtenons à ce niveau un FOM de l'ordre de 79.4%.

5. Conclusion

Nous avons présenté dans cet article les performances obtenues par les deux classifieurs PMC et SVM pour la classification des mots clés reconnus afin de détecter les mots clés réellement prononcés. Nous avons proposé trois représentations vectorielles de l'entrée du classifieur à base de probabilité d'observation acoustique locale et du nombre de trames par état dans chaque phonème dans un mot.

Pour bien représenter le mot en entrée du classifieur, nous nous sommes orientés d'abord vers l'utilisation des probabilités d'observations acoustiques locales des phonèmes. Afin d'introduire l'aspect temporel, nous avons ensuite introduit une représentation vectorielle à base du nombre de trames par état dans chaque phonème d'un mot clé. Enfin, nous avons proposé une représentation vectorielle plus complète qui utilise les probabilités acoustiques et le nombre de trames par état. Les résultats obtenus prouvent, en premier lieu, que les représentations à base de vecteur mixte sont les meilleures et, en second lieu, que les SVM à base de fonction noyau RBF donnent toujours les meilleurs résultats et enfin que les SVM linéaires sont plus performants que le PMC. Ceci peut être expliqué d'une part par le fait que les SVM réalisent une projection des entrées dans un

espace de caractéristiques de plus grande dimension ce qui facilite la discrimination des données. D'autre part, le PMC utilise le principe de minimisation du risque empirique alors que les SVM se basent sur le principe de minimisation du risque structurel qui est plus efficace que le premier.

L'utilisation d'un classifieur a prouvé son efficacité en améliorant les performances de notre système de détection. En effet, nous avons atteint la valeur de 82.9% au niveau FOM en utilisant les SVM à noyau RBF. Les meilleurs résultats obtenus pour les SVM linéaires sont de l'ordre de 81.5% comme valeur de FOM. Concernant le PMC, les meilleures performances réalisées sont de l'ordre de 79.4% comme FOM.

Parmi les solutions envisageables pour améliorer le taux de reconnaissance, nous proposons de perfectionner les représentations des mots en ajoutant un maximum d'informations d'ordre acoustique et éventuellement linguistique. Il serait aussi rentable de rechercher d'autres fonctions noyaux pour les SVM, autres que la fonction linéaire et la fonction RBF et qui soient adaptées à la nature des données.

6. Bibliographie

[1] BENAYED Y., FOHR D., HATON J.P., CHOLLET G., « A New Keyword Spotting Approach Based on Reward Function », *Seventh International Symposium on Signal Processing and Its Applications, pp.405-408, Paris, 2003.*

[2] CAMINERO J., TORRE C., VILLARRUBIA L., MARTÍN C., HERNÁNDEZ L., « On-line garbage modeling with discriminant analysis for utterance verification », *International Conference on Spoken Language Processing, pp. 2111-2114, Philadelphia/PA, 1996.*

[3] BENAYED Y., FOHR D., HATON J.P., CHOLLET G., « Confidence Measures for Keyword Spotting Using Support Vector Machines », *IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 588-591, Hong Kong, 2003.*

[4] BENAYED Y., « Détection de mots clés dans un flux de parole », *Ecole Nationale Supérieure des Télécommunications, 2003.*

[5] VAPNIK V., « The nature of statistical learning theory », *Springer Verlag, New York, 1995.*

[6] CHAPELLE O., VAPNIK V., BOUSQUET O., MUKHERJEE S., « Choosing Multiple Parameters for Support Vector Machines », Machine Learning, *vol. 46, n° 3, pp. 131-159, 2002.*

[7] VAPNIK V., « Statistical learning theory », *John Wiley and Sons, 1998.*

[8] FOHR D., MELLA O., ANTOINE C., « The automatic speech recognition engine ESPERE experiments on telephone speech », *International Conference on Spoken Language Processing, pp. 246-249, Beijing, 2000.*