ADAPTATION SPECTRALE PAR QUANTIFICATION VECTORIELLE : EXEMPLE DE LA RAP À FRÉQUENCES D'ÉCHANTILLONNAGE MULTIPLES

Richard LAMY, Laurent BESACIER.

Equipe *GEOD* (Groupe d'Etude sur l'Oral et le Dialogue) - Laboratoire *CLIPS*
Université Joseph Fourier - BP 53 - 38041 GRENOBLE Cedex 9
Tél : 04 76 63 55 81 – Fax : 04 76 63 55 52
Mél : Richard.Lamy@imag.fr ; Laurent.Besacier@imag.fr

ABSTRACT

This paper presents a non linear approach for spectral adaptation based on Vector Quantization. The idea is to transform feature vectors extracted from signals of one quality to feature vectors of another quality. Our method is applied to the particular case of speech recognition at multiple sampling rates. Such a method, which can be applied to other adaptation problems, allows very acceptable correspondence between two considered feature spaces. Thus, a generic ASR system trained on 16kHz signals is able to recognize lower sampling rate signals without any adaptation of its acoustic models.

1. INTRODUCTION

La reconnaissance automatique de la parole s'applique à ce jour sur de nombreux signaux de qualité différente (fréquence d'échantillonnage, quantification, codage, conditions d'enregistrement). Pour faire face à cette variabilité, on peut limiter l'apprentissage à une qualité de signal donnée et obtenir les modèles acoustiques adaptés à celle-ci. Le défaut de cette approche vient du fait que, parfois, les modèles seront appris sur des signaux de mauvaise qualité. Par ailleurs, on ne peut pas prévoir à l'avance toutes les propriétés des signaux qui seront rencontrées par le système au cours de la phase de reconnaissance.

Une autre approche est d'utiliser un modèle acoustique appris sur des données enregistrées dans de très bonnes conditions et de transformer les signaux que l'on veut reconnaître et/ou le modèle acoustique de référence pour réduire le « mismatch » entre apprentissage et test. Cette transformation peut avoir lieu au niveau *signal*, au niveau des *paramètres* ou par *transformation du modèle* lui-même. Dans cette dernière catégorie se trouvent notamment les méthodes d'adaptation du type MLLR qui effectuent un traitement sur les gaussiennes du modèle acoustique.

Partons d'un « cas d'école » où l'on dispose d'un modèle acoustique (issu de l'apprentissage) appris sur des signaux propres échantillonnés à 16kHz, et de signaux de test à bande limitée (du type téléphonique) de fréquence d'échantillonnage 8kHz. Une première solution consiste à dégrader les signaux utilisés pour l'apprentissage, c'est-à-dire à les sous-échantillonner à une fréquence d'échantillonnage de 8kHz, et à réapprendre un modèle acoustique sur ces nouvelles données. On peut également envisager une approche du type *signal* où les signaux de test seront sur-échantillonnés à 16kHz avant d'être comparés au modèle acoustique de référence. Cependant, ce type de traitement donne des résultats peu satisfaisants. En effet d'après le théorème de *Shannon*, les fréquences au delà de 4kHz (du signal échantillonné à 8kHz) ne sont pas pour autant retrouvées lors d'un sur-échantillonnage à 16kHz. On peut encore envisager l'approche *paramètres* qui consiste à modifier les vecteurs acoustiques extraits des signaux de test pour réduire le « mismatch » avec le modèle acoustique de référence. C'est cette dernière approche que nous nous proposons de décrire dans cet article.

Notre approche se range donc dans cette catégorie appelée « adaptation spectrale » [Ham99]. L'adaptation spectrale, contrairement à l'adaptation de modèle, ne vise pas à améliorer la précision des modèles, mais à rendre les données de test comparables aux modèles d'origine. Le principal avantage de l'approche d'adaptation spectrale est que, généralement, le système de reconnaissance initial n'est pas modifié. Les applications se trouvent par exemple au niveau de l'adaptation au locuteur, à des signaux bruités de tout genre, à des qualités de signal différentes.

Les techniques les plus courantes pour ce genre d'approche reposent sur des transformations linéaires des paramètres, souvent améliorées [Cou01]. Nous présentons ici une autre approche, exploitant le principe général de la quantification vectorielle (VQ : Vector Quantization) [Ger92].

La première partie de cet article sera dédiée au principe général ainsi qu'aux méthodes envisagées. Dans une seconde partie, nous présenterons les expérimentations qui ont été menées et les résultats de celles-ci. Enfin les perspectives et les études envisagées à court terme concluront cet article.

2. TRANSFORMATION DE PARAMETRES PAR ADAPTATION SPECTRALE

2.1 Méthode générale

Notre méthode nécessite que l'on dispose d'un même corpus décliné en plusieurs qualités de signal. Dans notre cas, nous disposons d'un corpus de N signaux (initialement 16kHz-16Bits) ainsi que de sa version sous-échantillonnée à 8kHz. A ce niveau, nous ne

parlons pas d'unités phonétiques, mais uniquement de deux ensembles de vecteurs acoustiques d'un espace de dimension p. Le but, alors mathématique, consiste à trouver une correspondance analytique **f** entre ces deux ensembles (*figure 1*).

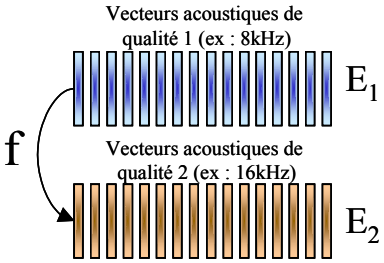

Figure 1 : Principe général

2.2 Transformation par Quantification Vectorielle

Pour trouver la fonction de correspondance f, nous utilisons le principe de la quantification vectorielle.
Un quantificateur vectoriel Q de façon générale est composé d'un encodeur E et d'un décodeur D en cascade, comme montré dans la *figure 2* .

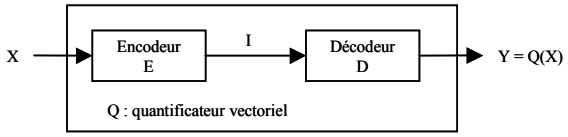

Figure 2: Structure générale d'un quantificateur vectoriel [Ger92].

L'encodeur trouve l'indice du mot du dictionnaire, correspondant à l'entrée X. Le décodeur utilise l'indice I fourni par l'encodeur pour reconstituer Y = Q(X). Dans notre cas, le problème est plus compliqué. Nous souhaitons trouver une correspondance entre deux espaces spectraux. Dans ce cas, le schéma général du quantificateur est modifié (*figure 3*).

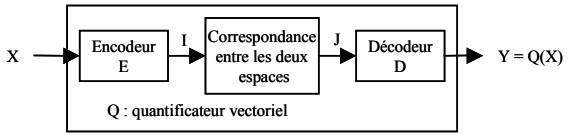

Figure 3: Structure de nos quantificateurs.

Ce qui pose donc trois problèmes majeurs : Comment classifier efficacement l'espace E1 ? Comment trouver la correspondance entre les classes des deux espaces ? Comment reconstruire un vecteur en fonction de la classe de E2 ?

2.3 Classification par kmeans et reconstruction par moyenne

Le but ici est de caractériser les vecteurs acoustiques source en un certain nombre de classes, puis d'associer un vecteur *représentant* dans le format cible à chaque classe. Nous ne produisons pas de signal, nous quantifions le signal source puis l'associons au plus proche du format cible. Cette approche a été abordée dans le cadre d'une adaptation spectrale générale proposée par [Yao96].

Tout les coefficients ne sont pas exprimés dans des unités homogènes et ne représentent pas les mêmes caractéristiques [Kin01]. La distance euclidienne n'a donc plus vraiment de sens. Nous choisissons alors la distance utilisée couramment en ACP, ce qui revient en fait à centrer et réduire le tableau de données. Pour la classification de ces vecteurs centrés réduits, nous utilisons l'algorithme « binary split » des k-means [Rab93] pour regrouper les vecteurs en classes et pour associer un vecteur moyen cible à chacune des classes (*figure 4*). Des tests comparatifs entre plusieurs algorithmes de classification [Kin00] confirment ce choix.

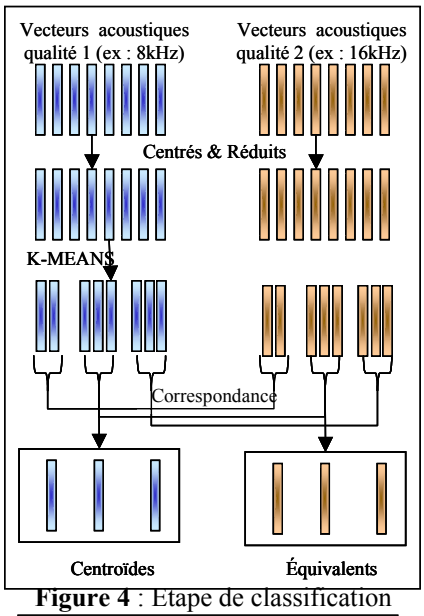

Figure 4 : Etape de classification

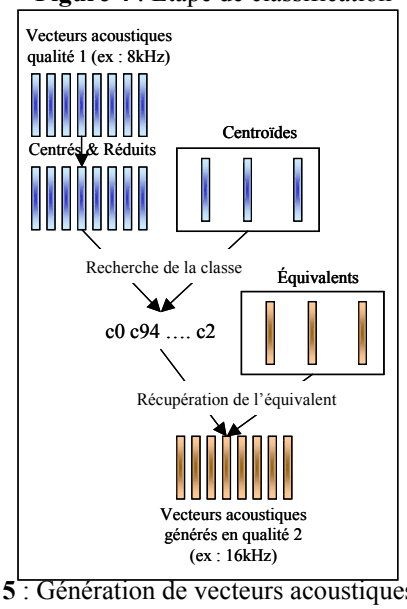

Figure 5 : Génération de vecteurs acoustiques pendant la phase de reconnaissance

A ce niveau, nous disposons des centroïdes et des équivalents pour chaque classe. La *figure 5* montre la transformation opérée lorsqu'un signal test est rencontré. Les vecteurs acoustiques du signal à reconnaître sont d'abord centrés et réduits, puis comparés aux centroïdes pour connaître leur classe d'appartenance. Cette phase réalisée, ces vecteurs sont

remplacés par les vecteurs équivalents (dans l'autre espace) de leur classe d'appartenance. La phase de reconnaissance s'effectue sur ces nouveaux vecteurs.

3. EXPERIMENTATIONS

3.1 Système de reconnaissance de la parole

Nous décrivons ici les systèmes de reconnaissance utilisés, les corpus et les vecteurs acoustiques considérés pour nos expérimentations.

Reconnaissance de parole continue du Français
Notre système de reconnaissance de parole continue du français RAPHAEL utilise la boîte à outils Janus-III [Wos93] du CMU. Les modèles acoustiques dépendants du contexte (750 modèles contextuels, 16 gaussiennes chacun) sont appris sur un corpus contenant 12 heures de parole continue de 72 locuteurs extrait de la base Bref80 [Lam91]. Le vocabulaire contient près de 5500 variantes phonétiques de 2900 mots distincts ; Il est spécifique à la réservation et l'information touristique. Le modèle de langage trigramme a été calculé à partir de documents d'Internet.

Reconnaissance de digits Anglais
Afin de tester également notre approche sur une base de signaux connue de tous, un système de reconnaissance de mots isolés anglais, dont les modèles acoustiques sont appris sur la partie « *train* » de TIDIGITS [Leo84] (6700 signaux), est également utilisé.

3.2 corpus de test et de développement

Pour notre étude, nous disposons de plusieurs corpus de parole « stéréos », français et anglais, sur lesquels nous pourrons mener nos recherches. Ici stéréo est à prendre dans le sens où les signaux d'un même corpus sont déclinés en différentes qualités et alignés au vecteur près (cf. *figure 1*).
Ces deux corpus ont été fractionnés en deux parties distinctes. Une partie « développement » pour le calcul des centroïdes et équivalents, et une partie « test » pour effectuer les tests de reconnaissance.
CSTAR120 : corpus français de 120 phrases dans le domaine du tourisme [Bla00]. La partie développement comprend 99 signaux, la partie test comprend les 21 signaux restants.
TIDIGITS : corpus de digits connectés en anglais. Notre partie développement comprend 2501 signaux de la partie « *test* » de TIDIGITS, et notre partie test comprend 3450 autres signaux de la partie « *test* » de TIDIGITS.

3.3 Vecteurs caractéristiques

Pour les vecteurs acoustiques, nous utilisons 13 coefficients cepstraux de type MFCC, leurs dérivées premières et secondes, l'énergie, sa dérivée première et seconde, et le taux de passage par zéro (zerocrossing).
Ces vecteurs de dimension 43 sont ensuite réduits à une dimension de 24 par application d'une LDA (Linear Discriminant Analysis). Cependant, nous n'appliquons pas notre technique d'adaptation spectrale sur les vecteurs de dimension 43, mais uniquement sur les coefficients statiques (MFCC, zerocrossing et énergie). Les dérivées premières et secondes sont ensuite recalculées à partir du vecteur statique généré, puis la LDA est appliquée.

3.4 Résultats

Adaptation complète pour le passage 8khz→16khz en reconnaissance de parole continue du français
Nous effectuons un changement d'espace spectral, c'est-à-dire que nous classifions l'espace spectral 8kHz et associons à chaque classe obtenue un vecteur équivalent dans l'espace spectral 16kHz. Les performances de reconnaissance en fonction du nombre de classes utilisées pour caractériser l'espace spectral sont présentées sur le graphe de la *figure 6*.

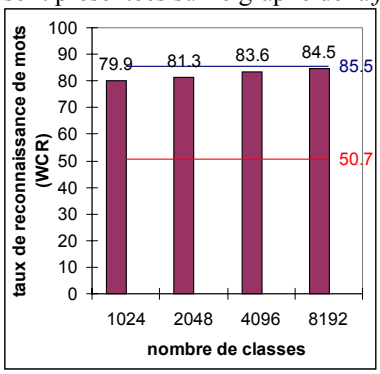

Figure 6 : Résultats de notre méthode d'adaptation sur la reconnaissance de parole du français.

Les deux références sur le graphique sont respectivement : le taux obtenu pour des signaux 8kHz sur-échantillonnés à 16kHz, et reconnus par des modèles 16kHz. (référence 'basse' 50.7%), et le taux obtenu pour des signaux 8kHz reconnus par des modèles 8kHz (référence 'haute' 85.5%).
Notre méthode effectue une bonne classification de l'espace, dans le sens où les signaux échantillonnés à 8khz sont bien reconnus par un système de reconnaissance n'ayant jamais rencontré de signaux de cette qualité lors de la phase d'apprentissage.

Adaptation complète pour le passage 8khz→ 16khz en reconnaissance de digits anglais
Cette fois-ci, dans le souci de ne pas avoir d'influence non mesurable du modèle de langage, nous testons un système de reconnaissance de chiffres sur le corpus anglais TIDIGITS. L'adaptation concerne encore le passage de la qualité 8kHz à la qualité 16kHz. Les résultats sont présentés sur la *figure 7*, avec les références 'basse' et 'haute' suivantes : le taux obtenu pour des signaux 8kHz sur-échantillonnés à 16kHz, et reconnus par des modèles 16kHz. (81.9%), et le taux obtenu pour des signaux 8kHz reconnus par des modèles 8kHz (96.7%). Cette seconde série de tests nous permet de confirmer les résultats précédemment obtenus. En effet, d'un point de vue acoustique, notre

méthode de transformation spectrale permet de passer d'un espace à l'autre en codant les vecteurs de paramètres sur 10 à 13 bits (1024 à 8192 classes).

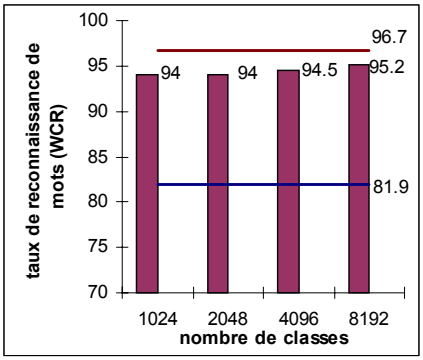

Figure 7 : Résultats de notre méthode d'adaptation sur la reconnaissance de digits anglais.

Notons que les tests d'un quantificateur développé sur le corpus français CSTAR120, et testé sur le corpus anglais TIDIGITS, ont donné des résultats médiocres. Il semblerait que notre méthode, caractérisant un espace acoustique capture également des informations comme les conditions d'enregistrements (microphone, environnement sonore, etc.).

CONCLUSION & PERSPECTIVES

Nous avons présenté ici une approche non linéaire d'adaptation spectrale, dans le cas particulier du passage de la qualité 8kHz à la qualité 16kHz. Nous montrons que cette technique permet, dans la plupart des cas, une bonne transformation de l'espace des paramètres acoustiques. Cependant, cette méthode, basée sur les principes de la quantification vectorielle, peut être généralisée à d'autres espaces. Le test de cette méthode dans le cas de signaux bruités est en cours et nous avons d'ores et déjà quelques résultats prometteurs.

Notons que le fait de disposer d'un corpus « stéréo » (double qualité) est assez contraignant. Nous travaillons actuellement sur une méthode similaire qui ne nécessite pas de corpus stéréo, en classifiant les deux espaces et en faisant correspondre les classes.

Pour confirmer de manière définitive les choix quant à notre technique de classification, une étude sur la qualité de la classification selon la taille du dictionnaire est en cours, fondée sur les idées de [Kan00]. Pour cela, nous exploitons différentes mesures de distorsion telles que l'erreur quadratique moyenne (MSE : Mean Squared Error) [Kinn00]. La recherche d'un vecteur dans des dictionnaires de taille importante doit également être optimisée, en donnant au dictionnaire une structure [Ger92].

Enfin, nous ajoutons une question aux trois problèmes présentés auparavant : que doit-on classifier ? En effet, la suite de vecteurs générée ne contient sans doute plus la corrélation entre un vecteur i et les vecteurs i-1 et i+1. D'où l'idée, présentée dans [And01] de création de « super-vecteurs » regroupant les vecteurs voisins.

BIBLIOGRAPHIE

[And01] Andrassy B., Vlaj D., Beaugeant C. (2001), " recognition performance of the siemens front-end with and without frame dropping on the aurora 2 database ", Eurospeech 2001, Aalborg, Denmark.

[Bla00] Blanchon, H. & Boitet, C. (2000). " Speech Translation for French within the C-STAR II Consortium and Future Perspectives " . Proc. ICSLP 2000. Beijing, China. 16-20 October, 2000. vol. 4/4 : pp. 412-417.

[Cou01] Couveur L., Dupont S., Ris C., Boite J-M., Couvreur C., (2001), " Fast Adaptation for Robust Speech Recognition in Reverberant Environments ", Proc. ITRW 2001 Sophia-Antipolis, pp 85-88.

[Ger92] Gersho A., Gray R.M. (1992) , " Vector quantization and signal compression ", Kluwer Academic Publishers, BOSTON.

[Ham99] Hamaker J, (1999) , " MLLR : a speaker adaptation technique for LVCSR ", cours au sein de ISIP - Institute for Signal and Information Processing - , Department of Electrical and Computer Engineering, novembre 1999.

[Kan00] Kannan R., Vempala S., Vetta A. (2000), " On clusterings – good, bad and spectral ", Proceedings of the 41st Symposium on the Foundations of Computer Science, pp367-77, 2000.

[Kin00] Kinnunen T., Kilpeläinen T., Fränti P. (2000), " Comparison of clustering algorithms in speaker identification ", SPC 2000, pp. 222-227

[Kin01] Kinnunen T., Kärkkäinen I., Fränti P. (2001), " Is speech data clustered ? – statistical analysis of cepstral features ", Eurospeech 2001,vol. 4, pp. 2627-2630.

[Lam91] Lamel, L.F., Gauvain, J.L., Eskénazi, M. " BREF, a Large Vocabulary Spoken Corpus for French ", Eurospeech, Gênes, Italy, Vol 2, pp. 505-508, 24-26 September 1991.

[Leo84] R. G. Leonard. "A database for speaker-independent digit recognition.",Proceedings of ICASSP 1984, vol. 3, 1984.

[Rab93] Rabiner, L. & Juang, B-H. (1993) " Fundamentals of speech recognition ", pp 126-127.

[Wos93] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki-Waibel, N., Waibel, A., and Ward, W. (1993) " Recent Advances in JANUS : A Speech Translation System ". *Eurospeech*, 1993, volume 2, pp 1295-1298.

[Yao96] Yao L., Yu D., Huang T. (1996), " a Unified Spectral Transformation Adaptation Approach for Robust Speech Recognition ",ICSLP,Vol. 2