Compensation en milieu variant abruptement

Vincent Barreaud, Irina Illina, Dominique Fohr et Vincent Colotte

LORIA/INRIA
615 rue du jardin botanique 54602 Villers-lès-Nancy FRANCE
Tél. : ++33 (0)3 83 59 20 74 - Fax : ++33 (0)3 83 27 83 19
Mél : [barreaud,illina,fohr,colotte]@loria.fr

ABSTRACT

An improvement of a previously proposed frame-synchronous noise compensation algorithm based on *Stochastic Matching* is presented. This version takes into account the abrupt changes in the acoustical environment with no *a priori* hypothesis. In this method, the parameters of a compensation function are estimated during the recognition process and reinitialized at each detected change. It has been experimentally shown that a change in acoustic environment induced a shift in the distribution of a variable derived from the compensation process. The changes in the environment are detected using three different detection algorithms (Shewart algorithm, Bayesian Information Criterion and Spectral Variation Function) that search for the disruption in this distribution. The use of such a detection process improves the word recognition scores up to 30% on sentences corrupted by abruptly occurring noises.

1. INTRODUCTION

Les systèmes de reconnaissance automatique de la parole (SRAP) voient leurs performances diminuer de manière significative lorsque les environnements dans lesquels ils ont été entraînés et ceux dans lesquels ils sont utilisés diffèrent. La différence entre les deux milieux est d'abord dû à des sources de bruits extérieurs qui s'ajoute au signal de parole ainsi qu'à des variations dans le canal de transmission. Ainsi, la combinaison de ces différences transforme une séquence de parole X qui serait parfaitement reconnue par le système en une séquence de parole corrompue Y d'une façon difficilement modélisable. Enfin, les sources de distorsion peuvent varier au cours de la production d'une phrase, de manière imprévisible et leur nature exacte est rarement connue.

Plusieurs types de techniques ont été proposées pour rendre la RAP plus robuste au bruit. La méthode exposée dans cet article entre dans le cadre de la compensation : les vecteurs de données corrompus sont modifiés par une transformation dont les paramètres sont estimés d'après les caractéristiques du bruit. Cet ensemble de méthodes regroupe des techniques comme la soustraction du cepstre moyen (Mean Cepstre Removal, MCR) ou l'association stochastique (Stochastic Matching, SM) [5].

Les algorithmes séquentiels synchrones sont très prometteurs lorsqu'il s'agit de traiter des sources de bruit non-stationnaires, même si ce type d'algorithme souffre souvent de problèmes de convergence liés au manque de données. Notre travail se fonde sur l'algorithme développé dans [2]. Celui-ci permet de réactualiser les paramètres d'une fonction de compensation affine à chaque trame. Ainsi, cet algorithme temps réel autorise une compensation en parallèle avec le processus de reconnaissance. De fait, il permet d'effectuer une compensation de milieu variant lentement. De plus, il ne nécessite aucune information *a priori* sur la nature du bruit.

Au cours d'une utilisation réelle, un SRAP peut être confronté à des apparitions soudaines d'un bruit dans l'environnement acoustique. Dans ce cas, aucune information n'est disponible sur le moment d'apparition, le niveau ou la nature de ce bruit. Par conséquent, un algorithme de compensation devrait réagir à cet événement et réactualiser sa politique de compensation instantanément. Dans cet optique, deux problèmes sont à résoudre. Le premier concerne la détection du changement d'environnement et le deuxième la stratégie à adopter pour s'adapter à la nouvelle source de bruit.

La détection de changement dans un signal est un sujet fréquemment abordé dans bien des domaines. Mais dans la plupart des cas, lorsqu'une détection temps réel est envisagée, un modèle de la perturbation qui apparaît est nécessaire ce qui implique une hypothèse *a priori* dont cherche à s'affranchir notre solution. Dans [1] a été présentée une amélioration de notre algorithme de base prenant en charge les variations brusques de l'environnement. A chaque trame, en parallèle avec l'algorithme de Viterbi, la distance entre une observation et l'état le plus probable est calculée. La distribution de cette distance (appelé δ par la suite) est Gaussienne dans un environnement stable. Une brusque variation de ce dernier peut donc être repérée en surveillant toute rupture dans le comportement de cette variable. Le changement d'environnement repéré, le processus de compensation est réinitialisé afin de ne plus prendre en compte les caractéristiques passées.

Nous proposons ici un approfondissement de cette méthode. Précédemment, nous avons utilisé l'algorithme de Shewart pour surveiller les changements rapides de δ. Ici, nous introduisons deux autres possibilités. La première est le critère BIC (Bayesian Information Criterion). La deuxième est l'approche SVF (Spectral Variation Function) qui n'exploite plus la forme gaussienne de la distribution de δ mais compare à chaque instant les contextes droit et gauche de δ.

Cet article s'articule de la manière suivante : Dans un premier temps nous rappellons brièvement le cadre théorique décrivant notre processus de compensation et nous mettrons en évidence le caractère gaussien de la distribution

de δ. Par la suite, nous présentons les trois algorithmes de détection envisagés. Enfin nous concluons après avoir présenté les tâches difficiles mises en place afin de tester ces algorithmes et les résultats associés.

2. CADRE THÉORIQUE

Le détail des calculs et approximations utilisées se trouvent dans [2]. Voici le résultat principal. Considérons un système de reconnaissance utilisant le paradigme des modèles de Markov cachés (HMM). Ces modèles comportent N états et utilisent des matrices de covariance diagonales. Chaque état n est caractérisé par un mélange de K fonctions de probabilité gaussiennes de moyenne $\mu_{(n,k)}$ et variance $\sigma_{(n,k)}$ et de poids $w_{(n,k)}$.

Considérons une fonction de compensation simple $f_B(y_{t+1}) = y_{t+1} + b_t$. Il a été montré (voir [2]) qu'une séquence de paramètres de biais $B_t = \{b_0, \ldots, b_t\}$ peut être estimée grâce à la séquence optimale d'états donnée par l'algorithme de Viterbi :

$$b_{t+1} = b_t \quad \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{t+1|t+1,B_t}(n,k)\frac{y_{t+1}+b_t \quad \mu_{(n,k)}}{\sigma^2_{(n,k)}}}{\sum_{\tau=1}^{t+1}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\gamma_{\tau|t+1,B_\tau \ \ 1}(n,k)}{\sigma^2_{(n,k)}}} \tag{1}$$

où

$$\gamma_{\tau|t+1,B_\tau \ \ 1}(n,k) = p(s_\tau = n, g_\tau = k|Y_{t+1}, B_\tau \ \ 1)$$

C'est à dire que $\gamma_{\tau|t+1,\Theta_\tau \ \ 1}(n,k)$ est la probabilité que le τ-iéme état de la séquence optimale d'états globale s_τ soit celui d'indice n et que sa principale composante gaussienne g_τ soit celle d'indice k, sachant la séquence partielle d'observations et $Y_{t+1} = \{y_1, \ldots, y_{t+1}\}$ la séquence des estimations précédentes $\Theta_\tau \ \ 1$. Les probabilités $\gamma_{\tau|t+1,B_\tau \ \ 1}(n,k)$ sont indisponibles lors de l'alignement de Viterbi puisqu'il est nécessaire de connaître la séquence complète des états pour la calculer. Dans notre algorithme, nous avons fait l'hypothèse que la "forward probability"

$$\alpha_{\tau|B_\tau \ \ 1}(n,k) = p(Y_\tau, s_\tau = n, g_\tau = k|B_\tau \ \ 1)$$

pouvait être utilisée en place de γ dans l'équation 1 et donner la nouvelle expression de θ :

$$b_{t+1} = b_t \quad \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} \alpha_{t+1|B_t}(n,k)\frac{y_{t+1}+b_t \quad \mu_{(n,k)}}{\sigma^2_{(n,k)}}}{\sum_{\tau=1}^{t+1}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\alpha_{\tau|B_\tau \ \ 1}(n,k)}{\sigma^2_{(n,k)}}} \tag{2}$$

Cette expression se simplifie si l'on suppose que la somme sur l'ensemble des états possibles et de leur composante gaussienne au temps τ peut être approximée par la seule contribution du couple (n,k) qui maximise $\alpha_{\tau|B_\tau \ \ 1}(n,k)$. Dans ce cas, l'équation devient :

$$b_{t+1} = b_t \quad \frac{\delta_{t+1}}{\sum_{\tau=1}^{t+1}\frac{1}{\sigma^2_{(n,k)_\tau}}} \tag{3}$$

où

$$\delta_{t+1} = \frac{y_{t+1} + b_t \quad \mu_{(n,k)_{t+1}}}{\sigma^2_{(n,k)_{t+1}}} \tag{4}$$

La figure 1 représente la distribution de la variable δ, pour la seconde dimension cepstrale, sur l'ensemble des phrases de test de VODIS,. La distribution concernant les phrases bruitées (enregistrées par un micro distant dans l'habitacle d'une voiture) apparaît en pointillés tandis que la distribution se rapportant à la parole propre (même condition mais avec un micro proche de la bouche) est représentée en trait plein. Cette figure met en évidence que la distribution de δ obtenue dans l'environnement bruité est décalée par rapport à celle obtenue dans l'environnement calme. Ce décalage n'est pas d'amplitude comparable sur toutes les dimensions. Suite à cette observation, nous avons conclus qu'un changement soudain de l'environnement en cours de reconnaissance entraînerai un changement abrupte dans la distribution de δ. Par conséquent, un algorithme surveillant toute rupture dans les valeurs de δ pourrait déclencher une ré-initialisation du dénominateur de l'équation 3 qui représente l'historique caduque de l'environnement, s'adaptant ainsi aux nouvelles conditions.

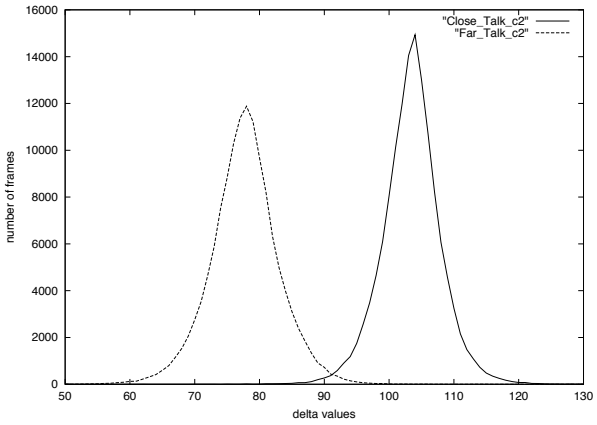


FIG. 1: Distribution de δ pour la seconde dimension cepstrale pour la parole propre (*close talk*) et bruite (*far talk*).

3. DÉTECTION DE CHANGEMENT

Cette section est dédiée au repérage de fracture dans le comportement de δ. Considérons que l'environnement acoustique change brusquement à l'instant t de env_0 à env_1. La distribution de δ avant t (resp. après) peut être modélisée par une Gaussienne $mode_0$ (resp. $mode_1$) de moyenne et variance (m_0, v_0) (resp. (m_1, v_1)). Notre but est de détecter, de façon réactive, le basculement de env_0 vers env_1.

3.1. L'algorithme de Shewart

Comme décris dans [3] l'algorithme de surveillance de *Shewart* permet de repérer le moment où une variable cesse de suivre une distribution gaussienne. Cette méthode suppose la connaissance des paramètres de cette distribution (m_0, v_0). Dans notre approche, ces paramètres sont estimés pendant les premières trames de chaque phrase et affinés tant qu'aucun changement d'environnement n'est repéré. Le changement de comportement est détecté lorsque la différence entre la moyenne de δ calculée sur une fenêtre de taille M et m_0 devient trop importante par rapport à la variance v_0. Le test repose classiquement sur la comparaison, à chaque trame t, de la vraisemblance de deux hypothèses :

H_0 l'environnement est env_0 jusqu'en t.

H_1 l'environnement env_1 a succédé à env_0 dans les M dernières trames précédant t.

Il existe donc un temps de latence M entre l'apparition du changement et son éventuelle détection. Le seuil de déclenchement est réglé empiriquement et dépend du niveau de bruit. Dans l'implémentation choisie, chaque dimension est décorrélée : un basculement peu s'opérer sur une dimension sans qu'il soit repéré dans d'autres.

3.2. Le critère d'information bayesien (BIC)

C'est un critère de vraisemblance pénalisé par la complexité du modèle, c'est à dire son nombre de paramètres ([3]). Le critère BIC permet de sélectionner un modèle parmi plusieurs pour modéliser la séquence d'observation. A chaque instant t, on évalue le rapport de vraisemblance entre deux hypothèses pouvant modéliser la séquence des δ jusqu'à l'instant t. Ces hypothèses sont :

H_0 La séquence est générée par un seul processus Gaussien multi-dimensionnel.

H_1 La séquence est générée par deux processus Gaussiens multi-dimensionnels successifs.

Les paramètres des modèles utilisés dans les hypothèses sont réestimés à chaque trame. Cette technique ne fait pas intervenir, en théorie, de seuil puisque le rapport décrit est comparé à l'unité. Dans l'implémentation proposée, les dimensions sont corrélées : les modèles sont multi-dimensionnels et le repérage d'un basculement s'opère sur toutes les dimensions au même instant.

3.3. La fonction de variation spectrale (SVF)

La fonction de variations spectrale (SVF) a été proposée dans [4] pour exprimer le taux de variation acoustique du signal à un instant donné. Cette méthode n'utilise pas la nature gaussienne de la distribution de δ. A chaque instant t, on considère un contexte droit et un contexte gauche L valeurs de δ. On effectue ensuite une comparaison entre les vecteurs moyens de chaque contexte à l'aide d'un produit scalaire normalisé. Il s'agit ensuite de mesurer la dissemblance entre chacun des vecteurs d'un contexte avec ceux du contexte opposé. Ceci est fait par l'intermédiaire du cosinus de ces deux vecteurs.

$$cos(\delta_i, \delta_j) = \frac{< \delta_i, \delta_j >}{|\delta_i||\delta_j|}$$

Avec $t \quad L \leq i \leq t$ et $t \leq j \leq t + L$. Ainsi, une valeur proche de 1 devrait témoigné que les deux vecteurs sont très semblables. Inversement, une valeur proche de $\quad 1$ indiquerai que les deux vecteurs ont des caractéristiques éloignées. On somme ensuite chaque cosinus calculé pour toutes les associations possibles de deux vecteurs (l'un du contexte droit et l'autre du contexte gauche) pour obtenir la valeur de la fonction :

$$SVF(t) = \frac{1}{2}(1 \quad \frac{1}{L^2} \sum_{i=t}^{t} \sum_{j=t}^{t+L} cos(\delta_i, \delta_j))$$

Ainsi, la fonction SVF donne des valeurs proches de 0 à l'intérieur d'un segment de signal stable et proche de 1 pour une partie transitoire. Les pics de la fonction donneront une estimation de l'instant de basculement. Cette approche est intéressante puisqu'elle s'affranchit de l'estimation des paramètres d'un modèle d'environnement pour la variable δ. Dans l'implémentation proposée, les dimensions sont liées dans le processus de détection de basculement.

4. CADRE EXPÉRIMENTAL

Afin de tester notre approche dans des environnements adverses, nous avons bruité artificiellement deux bases de test. La base VODIS (Voice-Operated Driver Information Systems) regroupe 200 locuteurs francophones. Les enregistrements ont été effectués en voiture par un micro proche de la bouche (*close talk*) et sont donc considérés comme peu bruités. Le signal a été échantilloné à la fréquence de 11025Hz et encodé sur des séquences de cepstres à 36 dimensions (12 coefficients statiques comprenant l'énergie, 12 Δ et 12 $\Delta\Delta$). Les phrases enregistrées par *close talk* de 175 locuteurs ont été utilisées pour construire des modèles de phonèmes à trois états, caractérisés par 8 Gaussiennes de matrice de covariance diagonales. La base de test est constituée de nombres enregistrés par microphone *close talk* et prononcés par les 25 autre locuteurs et corrompus comme décrit ci-après.

Les expériences ont aussi été conduites sur la partie finnoise de la base Aurora3 en utilisant la même paramétrisation. Nous avons utilisé les modèles de mots entraînés pour la tâche HM (Highly Mismatch), c'est à dire des modèles entraînés dans un milieu peu bruité. Ces modèles sont constitués de 16 états, caractérisés par 3 Gaussiennes de matrice de covariance diagonale. La base de test est constituée de chiffres prononcés en milieu peu bruité par les autres locuteurs.

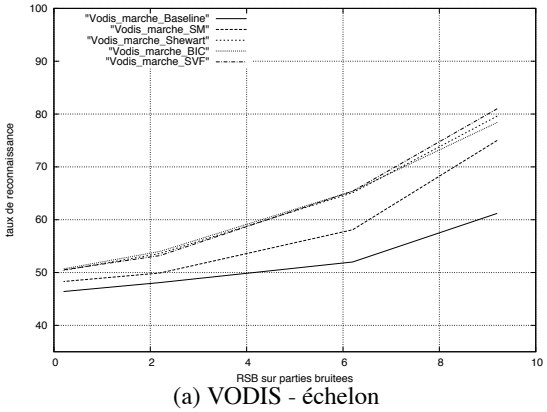
Les bases de test ont été bruitées artificiellement par addition d'un bruit d'avion (bucanneer2.wav de NOISEX) à différents rapport signal à bruit et de deux façons. Premièrement (épreuve : *échelon*) le bruit est ajouté à partir du milieu de chaque phrase de test. Deuxièmement (épreuve : *aléatoire*) le bruit est ajouté pendant une durée de 300ms, aléatoirement à chaque phrase de test, deux apparitions du bruit ne pouvant être séparées de moins de 300ms.

5. RÉSULTATS

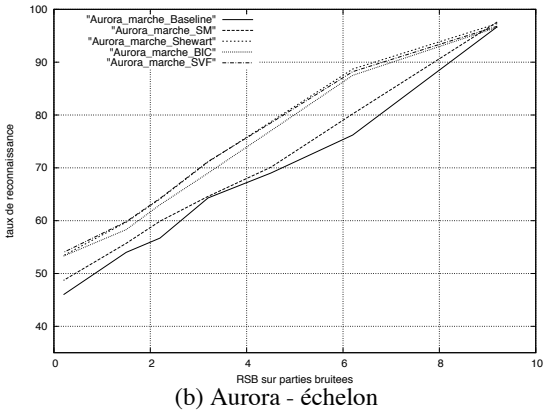
Les figures 2 et 3 représentent l'évolution des taux de reconnaissance en mot en fonction du rapport signal à bruit (RSB) dans les parties bruitées artificiellement. Le taux de reconnaissance de référence (Baseline) est obtenu par le système ESPERE sans aucun processus de compensation (ligne pleine). Les taux obtenus par l'algorithme de compensation proposé dans la section 2 (SM) sont rapportés. Enfin, les résultats obtenus par l'algorithme de la section 2 intégrant les trois méthodes de détection (Shewart, BIC et SVF de la section 3) sont présentés.

La figure 2 donne les résultats pour l'épreuve *échelon* sur les bases Aurora et VODIS. On remarque ainsi que les trois méthodes de détection donnent des résultats équivalents, quelque soit le niveau de bruit. La détection de changement d'environnement introduit une amélioration relative (par rapport à Baseline) qui décroît en fonction du niveau de bruit de 32.4% à 9.3% pour VODIS et de 16.4% à 8% pour Aurora.

La figure 3 donne les résultats pour l'épreuve *aléatoire* sur les bases Aurora et VODIS. On remarque que, pour ce genre d'environnement très défavorable, l'apport de la

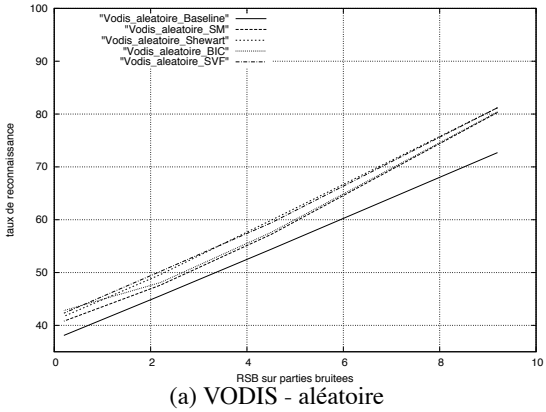


(a) VODIS - échelon

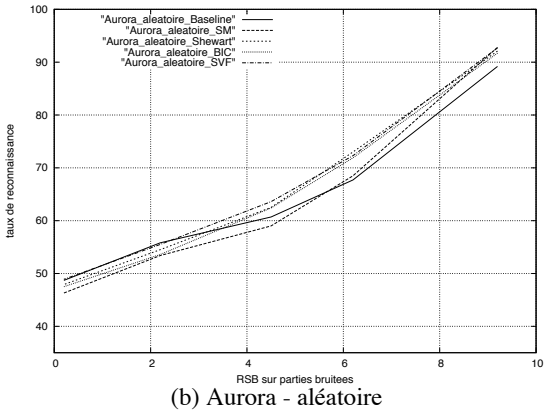


(b) Aurora - échelon

FIG. 2: Taux de reconnaissance pour l'épreuve *échelon*



(a) VODIS - aléatoire



(b) Aurora - aléatoire

FIG. 3: Taux de reconnaissance pour l'épreuve *aléatoire*

détection de changement est réduit. Dans le cas de VODIS, l'amélioration relative reste constante à 12% quel que soit le niveau de bruit. Pour Aurora, l'amélioration relative n'est que de 7.8% et décroit. De plus, la méthode utilisant BIC, ne donne pas de résultats supérieurs à la méthode n'utilisant pas de système de détection. En effet, les environnements se succèdent rapidement et l'estimation des paramètres des gausssiennes nécessaires au calcul du BIC n'est pas parfaite. Par contre, la méthode de Shewart qui utilise elle aussi une telle estimation ne semble pas affectée, ce qui prouverai l'intérêt d'une détection décorélée sur chaque dimension comme le fait la méthode de Shewart. Enfin, la méthode SVF ne demandant l'estimation de ces paramètres obtient, en toute circonstance, les meilleurs résultats.

6. CONCLUSION

Cet article présente une amélioration d'un algorithme de compensation en ligne synchrone à la trame utilisant le cadre de travail de l'association stochastique (*stochastic matching*). Nous avons intégré un système de détection de changement d'environnement ne se basant que sur l'étude d'une variable dérivée du processus de compensation. Par conséquent cet algorithme ne nécessite aucune information *a priori* sur l'environnement d'utilisation. Ainsi la reconnaissance est améliorée dans un milieu pouvant varier brusquement et apériodiquement. Nous avons évalué notre algorithme sur des tâches de reconnaissance de nombres pour des phrases issues des bases VODIS et Aurora artificiellement bruitées. La suite de notre travail portera sur la précision de la détection de changement ainsi que sur la corrélation des changements sur chaque dimension.

REMERCIEMENTS

Cet article a été publié grâce au soutient du projet OZONE (IST-2000-30026).

RÉFÉRENCES

[1] V. Barreaud, I. Illina, and D. Fohr. On-line compensation for non-stationary noise. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 2003.

[2] V. Barreaud, I. Illina, and D. Fohr. On-Line Frame-Synchronous Compensation of Non-Stationary noise. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2003.

[3] M. Basseville and I.V. Nikiforov. *Detection of Abrupt Changes : Theory and Application*. Prentice-Hall, 1993.

[4] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo. Improved connected digit recognition using spectral variation functions. In *Proceedings of the International Conference on Spoken Language Processing*, 1992.

[5] A. Sankar and C.H. Lee. A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition. *IEEE Transaction on Speech and Audio Processing*, 1996.