Une nouvelle architecture de compensation du bruit pour la reconnaissance robuste de la parole

Khalid Daoudi et Murat Deviren

INRIA-LORIA, équipe Parole
B.P. 101 - 54602 Villers les Nancy, France
Tél. : +33 (0)3 83 59 20 22 - Fax : +33 (0)3 83 59 19 27
Mél : daoudi,deviren@loria.fr - http ://www.loria.fr/equipes/parole

ABSTRACT

We present a novel noise compensation architecture which makes no assumptions on how the noise sources alter the speech data and which do not rely on clean speech models. Rather, this new architecture makes the (realistic) assumption that speech databases recorded under different background noise conditions are available. Its main principle is to process individually each database and to construct a parametric representation which describes the variation of acoustic models w.r.t. noise models. This representation is then used during recognition to estimate the acoustic models in the new environment. We evaluate the performance of this new compensation scheme on a connected digits recognition task and show that it can perform significantly better than multi-conditions training, which is the most widely used technique in these kind of scenarios.

1. INTRODUCTION

La robustesse au bruit est un problème très difficile auquel sont confrontés les systèmes de reconnaissance de la parole dans les applications concrètes. Plusieurs techniques [7] ont été proposées pour améliorer les performances de la reconnaissance en présence de mismatch entre les conditions d'apprentissage et celles de l'application. Ces techniques peuvent être classifiées en deux catégories : celles fondées sur le pré-traitement du signal de la parole (RASTA [4] ou amélioration de l'intelligibilité [1] par exemple) et les techniques de compensation. Dans ces dernières, des modèles acoustiques initiaux (généralement les modèles de parole propre) sont transformés pour représenter le nouvel environnement. Les techniques de compensation peuvent elles mêmes être classifiées en deux catégories : les schémas *adaptatifs* et *prédictifs*. La compensation adaptative, telle que MAP[2] et MLLR [5], consiste à utiliser la parole bruitée observée (au test) pour transformer les modèles acoustiques initiaux. Un inconvénient de cette approche est qu'elle nécessite au moins quelques dizaines de secondes de parole pour fournir de bonnes estimations des nouveaux modèles. Un autre inconvénient plus important est la nécessité d'une transcription lorsque le mode d'adaptation est non supervisé. Ceci est généralement fait à l'aide des modèles initiaux, ce qui limite les performances si la précision de la transcription est faible. La compensation prédictive (PMC [6] par exemple) ne repose pas sur les données de parole bruitée mais utilise plutôt les observations du bruit pour estimer les modèles dans le nouvel environnement. Les nouveaux modèles de la parole sont alors une combinaison entre les modèles initiaux et un modèle (paramétrique) du bruit estimé à partir d'observations de ce bruit. La combinaison est obtenue par une fonction qui modélise la contamination de la parole par les sources de bruit (bruit additif ou/et convolutif...). Un inconvénient est que cette modélisation peut être irréaliste et conduire ainsi à des estimations erronées des distributions de la parole bruitée. Un autre inconvénient de la compensation prédictive est sa forte dépendance de la paramétrisation (MFCC en général) et du type de modèles probabilistes (HMMs en général) utilisés.

Il est important de souligner la complémentarité entre la compensation adaptative et prédictive surtout lorsque la parole est contaminée par un bruit de fond. En effet, dans ce cas les observations du bruit peuvent être utilisées dans un schéma prédictif pour fournir un "meilleur" modèle initial (que celui de la parole propre) pour la compensation adaptative.

Outre les inconvénients que nous avons mentionnés, une importante limitation de la compensation prédictive est la nécessité de disposer d'un modèle initial spécifique (généralement celui de la parole propre), supposant ainsi que les données d'apprentissage ont été collectées dans un environnement spécifique. Cependant, de nos jours il est plus courant de disposer de données enregistrées dans plusieurs environnements acoustiques. Par exemple, pour les systèmes de RAP en voiture, les données sont collectées en utilisant différentes voitures, avec fenêtres ouvertes et fermées, avec radio allumée et éteinte...etc. Dans ces scénarios, la compensation prédictive n'est pas le meilleur moyen pour traiter le problème de mismatch. La technique la plus utilisée dans ces cas (car elle donne en général les meilleurs résultats) est celle de la reconnaissance *multi-conditions*. Cette dernière consiste à utiliser lors du test les modèles appris sur toutes les données disponibles (dans toutes les conditions), ces modèles représentent alors un environnement acoustique "moyen" qu'on espère proche de celui du test.

Dans cet article nous considérons que les données de parole ont été enregistrées dans différentes conditions de bruit. Nous développons ensuite une nouvelle architecture de compensation prédictive, que nous appelons compensation *supervisée* et qui est bien adaptée à ce type de scénarios. Le principe de cette nouvelle architecture est le suivant. En premier lieu, pour chaque condition de bruit, les modèles acoustiques sont appris en utilisant les données correspondantes à cette condition, nous les appelons modèles *matchés*. Ensuite, nous utilisons les observations du bruit correspondant (extraites de la base d'apprentissage) pour construire un modèle probabiliste qui tente

de représenter la distribution du bruit, nous l'appelons modèle du *bruit*. Après avoir traité toutes les conditions, nous procédons à l'apprentissage supervisé d'un modèle paramétrique (prédéfini) qui tente de décrire la variation des modèles matchés par rapport à ceux du bruit. Lors du test, les observations du bruit de l'environnement de l'application sont utilisées pour estimer le modèle du bruit correspondant, ce dernier est ensuite donné comme argument au modèle paramétrique pour fournir des modèles estimés que nous espérons proches des modèles matchés (à l'environnement de l'application).

Ce papier est organisé de la façon suivante. Dans la prochaine section, nous commençons par développer l'architecture générale de la compensation supervisée. Nous définissons ensuite la procédure particulière que nous utilisons dans cet article. Dans la section 3, nous posons le cadre expérimental pour nos évaluations. La section 4 est consacrée aux résultats et leur analyse.

2. ARCHITECTURE DE LA COMPENSATION SUPERVISÉE

Nous avons décidé de considérer la compensation supervisée comme faisant partie de la classe des techniques de compensation prédictive en raison de deux propriétés fondamentales qu'elles partagent. Tout d'abord, les deux se basent sur l'hypothèse (très réaliste) que les meilleures performances de reconnaissance sont atteintes lorsqu'il n'y a pas de mismatch entre les conditions d'apprentissage et celles du test. Ensuite, les deux requièrent des modèles du bruit pour faire la compensation, *aucune* donnée de parole dans le nouvel environnement n'est nécessaire. Nous développons maintenant l'architecture générale de la compensation supervisée.

Nous supposons la donnée d'un ensemble $D_1, ..., D_K$ de bases d'apprentissage enregistrées respectivement sous les conditions de bruit $n_1, ..., n_K$. Soit W l'ensemble des unités acoustiques (mots, phonèmes...) que nous cherchons à modéliser. Pour chaque $w \in W$ et$k \in \{1, ..., K\}$, nous notons $\lambda_w(k)$ le modèle matché pour w appris en utilisant la base D_k, c.à.d. sous la condition du bruit n_k. Si N représente la variable bruit, l'idée principale derrière notre architecture est de supposer qu'il existe une fonction paramétrique f_w telle que :

$||\lambda_w(N) - f_w(N)|| \cong 0,$ pour une certaine norme $||.||$.

Les paramètres de f_w sont estimés par un apprentissage supervisé en utilisant l'ensemble des entrées-sorties $\{n_k, \lambda_w(k) : k = 1, ..., K\}$. Lors de la reconnaissance, un nouveau bruit n^* est observé et est donné comme entrée à chaque f_w pour calculer le modèle, que nous appelons *f-estimé*, $\tilde{\lambda}_w(n^*) = f_w(n^*)$. Notre but étant que $\tilde{\lambda}_w(n^*)$ soit "assez proche" du modèle $\lambda_w(n^*)$ matché au bruit n^*. Cette nouvelle architecture possède plusieurs avantages par rapport aux techniques de compensation usuelles :

- Elle ne fait aucune hypothèse sur la façon dont le bruit contamine la parole.
- Elle est complètement indépendante de la paramétrisation et du type de modèles probabilistes utilisés pour traiter et modéliser la parole.
- Elle ne requière pas la disponibilité d'un modèle initial de la parole. Seuls les paramètres de f_w sont requis pour estimer les modèles de la parole dans le nouvel environnement.
- Elle est algorithmiquement très rapide.
- Elle peut être facilement incrémentée dans le sens où si une nouvelle base D_{K+1} est disponible, seul $\{n_k, \lambda_w(k) : k = 1, ..., K\}$ doit être stocké en mémoire pour tenir en compte cette nouvelle information et mettre à jour l'estimation de f_w. Ceci est un avantage majeur par rapport à l'apprentissage multi-conditions. En effet, pour tenir en compte D_{K+1}, ce dernier nécessiterait de stocker tout $\{D_1, ..., D_K\}$ et réapprendre les modèles en utilisant $\{D_1, ..., D_K, D_{K+1}\}$. Ceci est évidemment très lourd en terme de complexité et de mémoire.

Le choix de la forme des f_w est défini par l'utilisateur selon ses connaissances a priori, les contraintes algorithmiques et les spécificités de l'application. Dans le reste de ce papier nous nous proposons de considérer que f_w est linéaire. Nous verrons que même avec un choix aussi simple, les résultats sont intéressants et prometteurs. Nous utilisons des HMM et des mono-Gaussiennes pour modéliser respectivement la parole et le bruit. Pour compléter la définition de notre architecture, nous devons spécifier des représentations pour $\lambda_w(k)$ et n_k. Pour ce dernier, nous choisissons simplement le vecteur moyenne ν_{n_k} de la Gaussienne qui modélise la distribution des MFCC statiques de n_k. Pour $\lambda_w(k)$, nous choisissons les moyennes du mélange de Gaussiennes comme étant les paramètres à transformer (comme c'est le cas généralement dans les techniques de compensation). Formellement, pour chaque $w \in W$, si $\mu_{wls}(n_k)$ est le vecteur moyenne de la composante l du mélange à l'état s, nous définissons f_w comme étant l'ensemble $\{A_{wls}, B_{wls}\}$ de matrices de régression et de biais du modèle linéaire :

$$\mu_{wls}(n_k) = A_{wls}\nu_{n_k} + B_{wls}. \quad (1)$$

L'apprentissage de $\{A_{wls}, B_{wls}\}$ est fait de façon supervisée par le critère des moindres carrés. Lors de la reconnaissance, un nouveau ν_{n^*} est estimé à partir d'observations de n^*, le modèle (1) est ensuite appliqué pour calculer la f-estimation $\tilde{\mu}_{wls}(n^*)$. Pour les matrices de transitions, les poids des composantes du mélange et les covariances, nous gardons ceux du modèle de la parole propre. Nous soulignons ici que ceci n'implique en aucune façon que le modèle de la parole propre est un prérequis dans notre architecture. Ceci est juste un choix dans le cadre de ce travail préliminaire pour montrer le potentiel de notre approche.

3. CADRE EXPÉRIMENTAL

Nos expérimentations sont effectuées sur la partie "man" du corpus de chiffres connectés TIDIGITS. La base d'apprentissage (resp. test) est constituée de 4235 (resp. 428) séquences prononcées par 55 (resp. 56) locuteurs. Chaque séquence contient entre 1 et 7 chiffres. Pour créer la base de parole bruitée (à l'apprentissage et au test), nous utilisons 15 enregistrements différents de bruit de la base NOISEX (http ://spib.rice.edu/spib/select_noise.html). Les séquences bruitées sont obtenues en additionnant des segments de bruit à la parole propre. Pour chaque séquence, un segment est choisi aléatoirement à partir de l'enregistrement complet du bruit. Le niveau du bruit est ajusté par un facteur positif α. Précisément, si $s(t)$ est le signal de parole propre et $n(t)$ est le bruit sélectionné, le signal de parole bruitée est :

$$y(t) = s(t) + \alpha n(t). \quad (2)$$

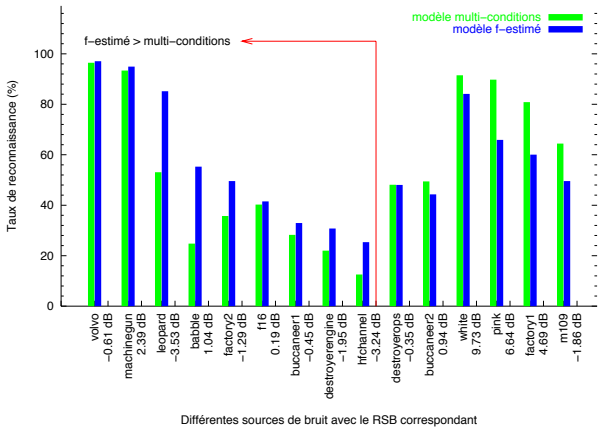

FIG. 1: *Comparaison entre les performances du modèle multi-conditions et celles du modèle f-estimé pour $\alpha = 0.3$.*

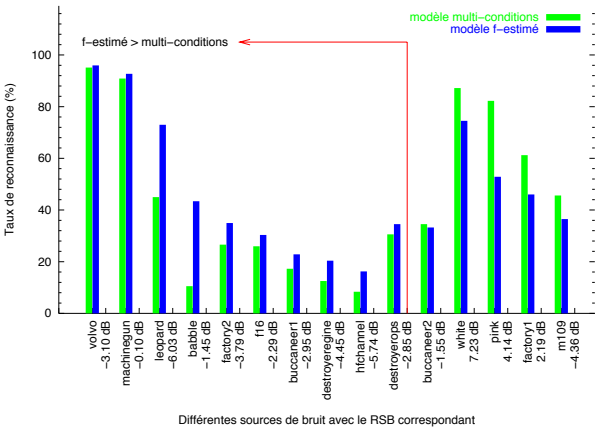

FIG. 2: *Comparaison entre les performances du modèle multi-conditions et celles du modèle f-estimé pour $\alpha = 0.4$.*

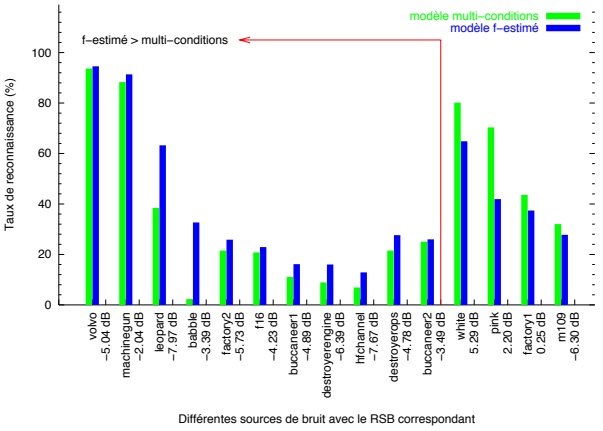

FIG. 3: *Comparaison entre les performances du modèle multi-conditions et celles du modèle f-estimé pour $\alpha = 0.5$.*

Le choix de α détermine le rapport signal sur bruit (RSB) moyen de la base bruitée. Le RSB est défini comme étant 10 fois le rapport logarithmique entre la puissance du signal et celle du bruit. Quand cette définition est appliquée directement pour calculer le RSB d'une séquence de parole, les segments de silence faussent le calcul. En effet, comme le bruit est très dominant dans ces segments, le RSB calculé ainsi sera faible pour les séquences contenant beaucoup de silence. Pour remédier à ce problème, nous calculons le RSB sur les seuls segments de parole, c.à.d., en excluant les régions de silence. Ceci nécessite une segmentation du corpus en régions de parole et de silence. Cette segmentation est obtenue par alignement forcé en utilisant les HMMs appris sur le corpus de la parole propre.

Le front-end et le back-end sont basée sur les spécifications du consortium AURORA3 [3]. Les vecteurs acoustiques sont obtenus utilisant le front-end standard ETSI SQL W1007 [3] qui est une analyse cepstrale où 13 MFCC sont extraits à partir de 23 bancs de filtre Mel. Cette analyse est effectuée en appliquant une fenêtre de Hamming sur des trames de 25ms et un shift de 10ms. Un filtre de préaccentuation avec $a = 0.97$ est appliqué. Les vecteurs sont ainsi constitués de 12 MFCC (c_0 est exclu) plus la trame d'énergie logarithmique, les Δ et $\Delta\Delta$ sont aussi inclus. Les onze chiffres (de 0 à 9 plus le "oh") sont modélisés comme mots par des HMMs à 16 états émetteurs par mot avec une topologie gauche-droite sans saut. La densité de chaque état est modélisée par une mélange de 3 Gaussiennes à matrices de covariance diagonales. Deux modèles de silence sont utilisés : un HMM à 3 états pour les pauses avant et après chaque séquence, et à un seul état pour les pauses entre mots. Nous utilisons la même paramétrisation pour les signaux de bruit, et nous modélisons chaque source $\alpha n(t)$ par une mono-Gaussienne apprise sur l'enregistrement complet du bruit $n(t)$. Dans toutes nos expérimentations, le software HTK est utilisé pour la paramétrisation, l'apprentissage des modèles et la reconnaissance.

4. RÉSULTATS ET DISCUSSION

Dans cette section nous évaluons les performances de notre architecture de compensation prédictive. Dans toutes les expérimentations nous avons choisi $\alpha = 0.05$ pour construire les bases d'apprentissage $D_1, ..., D_{K=15}$ (avec $\alpha = 1$ les bruits dominent la parole ce qui conduit à des performances très faibles des modèles matchés). Nous commençons l'évaluation par le cas où le mismatch est en terme du RSB, c.à.d., lorsque le type du bruit de testa été rencontré à l'apprentissage mais le RSB est différent (une situation souvent rencontrée dans les applications réelles). Le test est effectué en utilisant des facteurs de bruit plus élevés ($\alpha = 0.3, 0.4$ et 0.5) ce qui conduit à des niveaux de RSB plus faibles comparé aux conditions d'apprentissage. Les figures 1,2 et 3 montrent les performances de la reconnaissance multi-conditions et celles de la compensation supervisée. Sur l'axe des abscisses, les sources de bruits sont groupées en deux ensembles : le premier (resp. second) est composé de bruits pour lesquels le modèle f-estimé (resp. multi-conditions) donne de meilleurs taux de reconnaissance que le modèle multi-conditions (resp. f-estimé). La première observation frappante est que notre modèle f-estimé donne de meilleurs résultats (parfois de façon considérable) que le multi-conditions dans la majorité des cas. Ceci montre que, dans cette application, notre architecture est "globalement" plus robuste au bruit que la reconnaissance multi-conditions. En outre, plus le RSB décroît plus le nombre de cas où le modèle f-estimé est meilleur augmente. En effet, parmi 15 bruit il y a 9, 10 puis 11 cas pour lesquels le f-estimé est meilleur que le multi-conditions (voir figures 1, 2 et 3, respectivement).

Ceci suggère que la compensation supervisée est plus robuste à la décroissance du RSB que la reconnaissance multi-conditions. On peut aussi noter que plus le RSB décroît plus cette dernière "perd" un bruit en faveur de notre approche, ce qui suggère une certaine stabilité de notre procédure.

Dans la dernière expérimentation, nous évaluons les propriétés de généralisation de notre architecture lorsque le mismatch est en termes de RSB **et** de type de bruit. Pour ce faire, nous sélectionnons un bruit cible sur lequel les tests seront effectués. Le modèle matché à ce bruit est alors exclu de l'ensemble des entrées-sorties pour l'apprentissage de la régression (1). Le corpus correspondant à ce bruit (c.à.d. les données contaminées par ce bruit) est aussi exclu lors l'apprentissage du modèle multi-conditions. La table 1 montre les taux de reconnaissance obtenus à différents niveaux de RSB pour le bruit cible *machinegun*. Les taux des modèles matchés (à la condition du bruit *machinegun*) correspondants sont donnés comme références. Ces résultats montrent que, même si aucune information sur le bruit cible n'a été utilisée lors del'apprentissage, les performances de la compensation supervisée sont nettement supérieure à celles de la reconnaissance multi-conditions. Ceci suggère que la compensation supervisée peut être plus robuste (que la reconnaissance multi-conditions) non seulement en terme de variation du RSB mais aussi en terme de nouveaux types de bruit non rencontrées lors de l'apprentissage.

TAB. 1: *Comparaison entre les taux de reconnaissance (%) du modèle matché, multi-conditions et f-estimé pour le bruit cible machinegun.*

RSB	matché	multi-conditions	f-estimé
$\alpha = 0.05$	98.33	95.54	**97.56**
$\alpha = 0.3$	98.05	82.38	**88.86**
$\alpha = 0.4$	97.84	78.41	**84.75**
$\alpha = 0.5$	97.49	75.49	**82.03**

5. CONCLUSION

Nous avons développé une nouvelle architecture de compensation prédictive qui présente plusieurs avantages par rapport aux techniques classiques. Nous avons montré que malgré des choix très simples, cette architecture peut conduire à des performances supérieures à celles de la reconnaissance multi-conditions. Évidemment, en vue des expérimentations préliminaires que nous avons effectuées, nous ne pouvons pas faire de conclusions définitives. Nous pouvons cependant affirmer que cette nouvelle technique paraît prometteuse et mérite d'être plus approfondie. En effet, plusieurs pistes restent encore à explorer pour exploiter le potentiel de cette technique : choix de la forme de f_w et du critère d'estimation de ses paramètres, choix de la représentation pour $\lambda_w(k)$ et n_k, choix des caractéristiques à transformer, choix du modèle de bruit...etc. Ceci fera l'objet de nos travaux futurs.

RÉFÉRENCES

[1] Y. Ephraim. Gain-adapted hidden markov models for recognition of clean and noisy speech. *IEEE Trans. Signal Processing*, 40 :1303–1316, 1992.

[2] J.L. Gauvain et C.H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech and Audio Processing*, 2 :291–298, 1996.

[3] H.G. Hirsch et D. Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ISCA ITRW ASR2000*, Paris, 2000.

[4] H. Hermansky et N. Morgan. Rasta processing of speech. *IEEE Trans. Speech and Audio Processing*, 2 :578–589, 1994.

[5] C.J. Leggetter et P.C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hmms. *Computer Speech and Language*, 9 :171–186, 1995.

[6] M.J.F. Gales. Predictive model-based compensation schemes for robust speech recognition. *Speech Communication*, 3 :49–74, 1998.

[7] Y. Gong. Speech recognition in noisy environments : A survey. *Speech Communication*, 16 :261–291, 1995.