Etudes comparatives des robustesses au bruit de l'approche 'Full Combination' et de son approximation

Astrid Hagen[§] *et Hervé Glotin*[§‡]

§ IDIAP, Martigny, Suisse
‡ ICP, Grenoble, France
hagen,glotin@idiap.ch - glotin@icp.fr

Abstract

Sub-band based ASR aims to use reliable information from uncorrupted sub-bands only. For this one can select the best combination over all 2^d combinations of d sub-bands, without any need of independence assumption between sub-bands. We show that this approach can actually be set up by a combination function called the "Full Combination (FC)", constituting the fullband *posteriors* decomposed into a sum of weighted *posteriors* from 2^d experts. This approach is itself often not realizable as it encorporates training of too many experts. An approximation can estimate the *posteriors* for each combination based on the d sub-bands only, with a weak sub-band independence assumption. Tests under equal weights comparing FC and its approximation on band-limited noise show that for these noises the approximation performs as well as and sometimes better than FC, and both better than the baseline fullband system.

1. Introduction

La RAP (Reconnaissance Automatique de la Parole) multi-bandes exploite la redondance spectrale dans le but d'augmenter sa robustesse à l'inadéquation des données tout en faisant un minimum d'hypothèses sur le bruit interférant [?, ?, ?]. Nous verrons que les experts des bandes non bruitées fournissent suffisamment d'information pour permettre un décodage robuste. Dans ce papier nous développons le modèle *Full Combination* (FC) qui s'inscrit dans le paradigme multi-bande de la RAP et dont les relations avec la perception humaine de la parole sont reprises dans [?]. Des expériences précédentes en RAP ont montré que le traitement indépendant des sous-bandes peut faire chuter les performances en parole claire. Une alternative consiste à travailler sur les estimations phonétiques des 2^d combinaisons (ou flux) des d sous-bandes, (y compris le flux vide correspondant aux probabilités *a priori*). Dans une première approche ces estimations ont été calculées pour chaque flux puis sélectionnées une à une [HTP96]. Dans notre approche FC, leurs estimations sont pondérées et sommées. Comme l'entraînement d'un expert par combinaison, soient 2^d experts pour d sous-bandes, est rapidement irréalisable, il est préférable de travailler avec des approximations de ces combinaisons. Nous montrons alors qu'il est possible d'obtenir des résultats similaires ou meilleurs au modèle FC avec son approximé «AFC», que nous comparerons sur des bruits de bande(s) idéaux, stationnaires centrés sur une des d sous-bandes, ou changeant de fréquence centrale toutes les 125ms. Après la description de la théorie et de l'implémentation des approches FC et AFC, les paramètres des systèmes hybrides HMM/ANN (ANN pour réseaux neuronaux ou Artifical Neural Network) ainsi que les données de test sont présentés. Puis suivent les expériences et leurs discussions.

2. Les modèles FC et AFC

2.1. L'approche «Full Combination»

Les systèmes multi-bandes pour la RAP décomposent le domaine spectral en plusieurs sous-bandes, qui sont traitées indépendemment, et dont les paramètres caractéristiques x sont passés aux reconnaisseurs correspondants. Les probabilités *a posteriori* $P(q_k|x)$ des sous-bandes sont recombinées dans le processus de reconnaissance. Dans notre approche les 2^d flux des combinaisons des d sous-bandes sont intégrés suivant ces évènements j_{propre} collectivement exhaustifs et mutuellement exclusifs : «la j^{ieme} combinaison de sous-bande est le flux qui produit la meilleure reconnaissance parmi tous les flux possibles». Considérant que les données bruitées hors du j^{ieme} flux propre sont négligeables dans l'estimation des probabilités *a posteriori* [?, ?], nous posons $P(q_k|x, j_{propre}) \simeq P(q_k|x_j)$, où x_j est le vecteur acoustique du j^{ieme} flux. Nous avons alors:

$$P(q_k|x) \simeq \sum_{j=1}^{2^d} P(j_{propre}|x) P(q_k|x_j) \qquad (1)$$

Les probabilités dénotant les données claires $P(j_{propre}|x)$ dans (1) peuvent être estimées de différentes façons comme cela est démontré dans [BG99, HMB99, ?, ?]. Pour les expériences présentées dans cet article, les $P(j_{propre}|x)$ sont équiprobables, ce qui est déjà fort intéressant en terme de robustesse comme nous allons le montrer.

Dans l'approche FC les termes $P(q_k|x_j)$ sont donnés en sortie du réseau de neuronne qui est entraîné et testé sur les paramètres acoustiques x_j. Dans l'approche AFC les termes $P(q_k|x_j)$ sont estimés à partir des sorties des ANN relatifs uniquement aux sous-bandes contenues dans la j^{ieme} combinaison.

2.2. L'Approximation du FC : l'AFC

Pour éviter l'entraînement de 2^d ANNs on peut estimer les probabilités $P(q_k|x_j)$ des combinaisons en utilisant uniquement les probabilités $P(q_k|x_i)$ issues des observations $x_{i,i\in\{1..d\}}$ des d sous-bandes qui composent cette combinaison (on notera J cet ensemble de sous-bandes, de cardinal $|J|$). Ce modèle ne requiert qu'une hypothèse d'indépendance des observations des sous-bandes conditionnellement à chaque classe phonétique, hypothèse plus faible que l'indépendance absolue [?]. Nous avons alors $P(x_j|q_k) \simeq \prod_{i\in J} P(x_i|q_k)$, donc

$$P(q_k|x_j)\frac{p(x_j)}{p(q_k)} \simeq \prod_{i\in J} P(q_k|x_i)\frac{p(x_i)}{p(q_k)} \qquad (2)$$

$$P(q_k|x_j) \simeq \frac{\prod_{i\in J} P(q_k|x_i)}{p^{|J|-1}(q_k)} \cdot \Theta \qquad (3)$$

avec $\Theta = \frac{\prod_{i\in J} p(x_i)}{p(x_j)}$, qui disparaît par normalisation sur toutes les classes phonétiques pour obtenir des estimations telles que : $\sum_k P(q_k|x_j) = 1$.

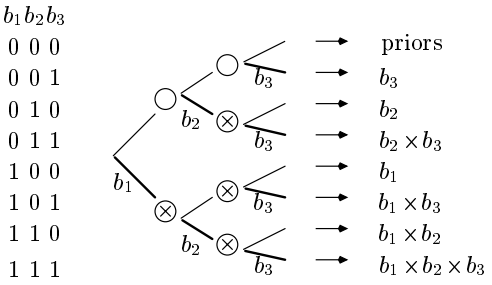

Figure 1: La valeur obtenue dans chaque feuille de l'arbre correspond à la multiplication des valeurs des branches en gras du chemin parcouru. On observe que le nombre de multiplications dans l'arbre (4) est inférieur à celui indiqué dans la colonne de droite (5), où le calcul est effectué indépendamment pour chaque combinaison. Cette différence augmente proportionnellement au nombre de bandes considérées.

Le calcul de l'approximation des probabilités $P(q_k|x_j)$ pour les 2^d combinaisons j peut se faire efficacement avec une procédure récursive qui réutilise les multiplications des composantes partagées par plusieurs combinaisons, au lieu d'effectuer un calcul indépendant pour chaque J. Ceci réduit considérablement le nombre de calcul lorsque le nombre de bandes considérées est grand, car par trame et par phonème 2^d–d–1 multiplications sont nécessaires au lieu de d×2^{d-1}–2^d+1, ce qui apporte une réduction d'un facteur $\approx d/2$ quand $d \to \infty$. Cette procédure peut être illustrée par une structure en arbre où les valeurs à multiplier se trouvent dans les branches et où chaque noeud équivaut à un appel de la fonction récursive qui accumule les multiplications précédentes et effectue le branchement respectif, comme indiqué dans la figure 1 pour le cas $d = 3$ et où $b_i = P(q_k|x_i)\ \forall i \in J$ pour simplifier la notation. Même si dans cet exemple le gain n'est pas significatif (4 multiplications au lieu de 5), dans le cas de 8 bandes il y aura 247 multiplications au lieu de 769, 65519 contre 458753 avec 16 bandes.

3. Expériences

3.1. Les reconnaisseurs spectre entier

Les paramètres d'entrée des ANNs sont du type PLP, avec un prétraitement J-RASTA [HM94] pour la moitié de l'expérience. Les entrées de l'ANN comportent 9 vecteurs acoustiques consécutifs, fournissant une information contextuelle importante au système. Les sorties correspondent aux 27 phonèmes significatifs de la base de données. Les vecteurs acoustiques pour le spectre entier comprennent 12 coefficients PLP (ou J-RASTA-PLP) et l'énergie (ainsi que les dérivées premières et deuxièmes de ces paramètres). Pour les ANNs du spectre entier nous avons choisi 1750 unités cachées. En clair le taux d'erreur au niveau du mot est de 8.0 % pour le système spectre entier sur les J-RASTA-PLPs et de 7.1 % pour le système spectre entier sur les PLPs.

3.2. Les 4 reconnaisseurs sous-bandes

Nous avons travaillé avec $d = 4$ sous-bandes. Pour ne pas rajouter des termes de dépendance inter-bandes dans nos modèles FC et AFC, il est précieux de veiller à choisir des bandes spectrales du signal qui se recouvrent au minimum mais dont l'union représente le spectre entier. Dans ce but, nous avons redéfini les sous-bandes qui ne tenaient pas compte de ces contraintes d'indépendance dans des études précédentes [BG99, HMB99]. De plus, nos expériences ont confirmé que la première bande critique n'est pas pertinente en parole téléphonique, nous l'avons donc supprimé. Ainsi nous avons choisi un ensemble homogène de quatre sous-bandes décrit dans la table 1 et qui permet toujours de modéliser un formant par sous-bande. L'ordre des analyses LPC et le nombre de coefficients extraits ont été optimisés sur plusieurs expériences. Dans le cas des fusions de sous-bandes i en une combinaison J, les ordres LPC ainsi que le nombre de coefficients extraits sont la somme de ceux des sous-bandes contenues dans J. Ainsi le nombre de paramètres dans le modèle FC et le modèle AFC sont identiques.

sous-bandes	en Hz	LPC	# coeff.
1	115-629 Hz	3	5
2	565 1370 Hz	3	5
3	1262 2292 Hz	2	3
4	2122 3769 Hz	2	3
134	115-629 Hz, 1262-3769 Hz	7	11

Table 1: Définition des 4 sous-bandes (coupure à 3dB) et des paramètres extraits. Exemple de combinaison : 134, montrant le calcul du nombre de paramètres des flux : somme de ceux des sous bandes, garantissant un nombre de paramètres constant entre FC et AFC. Le faible recouvrement fréquentiel entre sous-bandes est du aux filtres PLP des bandes critiques.

Les systèmes HMM/ANN hybrides pour les sous-bandes correspondent aux systèmes HMM/ANN hybrides spectre entier de base, la seule différence étant le nombre d'entrées, le nombre d'unités cachées restant proportionnel au nombre d'entrée. Les ANNs

des sous-bandes et des combinaisons de sous-bandes ont entre 666 et 1750 unités cachés.

3.3. Base de Données et Bruitage

Base de Données Numbers'95 Nous travaillons sur la base NUMBERS'95 qui contient des chiffres prononcés en continu en anglais, provenant de lignes téléphoniques. Pour l'entraînement de nos ANN nous avons utilisé 3590 phrases, et pour les expériences 200 phrases de l'ensemble de test.

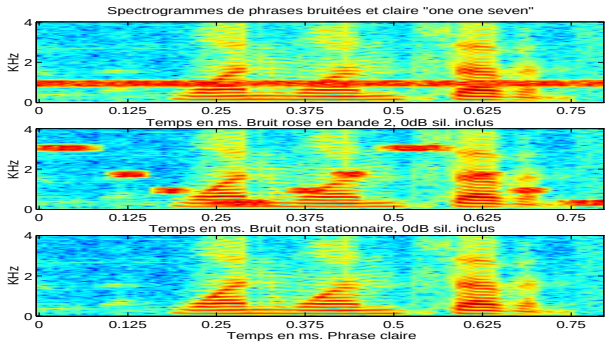

Figure 2: Illustration de bruitage par "bruit coloré" dans la bande 2 (haut), du bruit non stationnaire (milieu), de la parole non-bruitée pour la même phrase «one one seven» (bas).

Définition des bruits idéaux Afin de tester nos modèles dans des conditions idéales nous avons généré une série de quatre "bruits colorés" centrés sur les sous-bandes i ($i = 1..4$) du modèle et n'affectant qu'une sous-bande à la fois (générés par filtres trapézoidaux, 300 Hz de large, fréquence centrale égale à celle des sous-bandes du modèle hors recouvrement). Enfin pour tester l'effet de la robustesse des modèles FC et AFC suivant la distribution du bruit, nous construisons un bruit non stationnaire à partir des mêmes bruits colorés en conservant une répartition homogène du bruit sur les sous-bandes. Des pavés de 125 ms sont régulièrement tirés des sous-bandes 1, 2, 3, 4, 4, 3, 2 et 1 (voir figure 2), comme dans [BGTB98]. Pour clarifier les expériences les niveaux des bruits ajoutés ont été calculés silences inclus et ajoutés phrase par phrase.

3.4. Résultats et discussions

Nous présentons à la figure 3 les résultats des tests sur des bruits colorés dans les différentes sous-bandes 1 à 4, en utilisant les approches FC et AFC et des paramètres PLP. Pour comparaison, cette figure montre également les courbes correspondantes au système spectre entier testé dans les mêmes conditions. Nous constatons que non seulement l'approche FC mais aussi son approximation présentent de très bonnes propriétés de robustesse aux bruits colorés. En effet la reconnaissance peut s'effectuer de façon très fiable sur les composantes non bruitées et la contribution des flux bruités intégrés dans le FC ou AFC est peu perturbante car leur distribution de probabilités *a posteriori* a une forte entropie (distribution plus uniforme). Nous reviendrons sur cette analyse dans le cas de l'AFC.

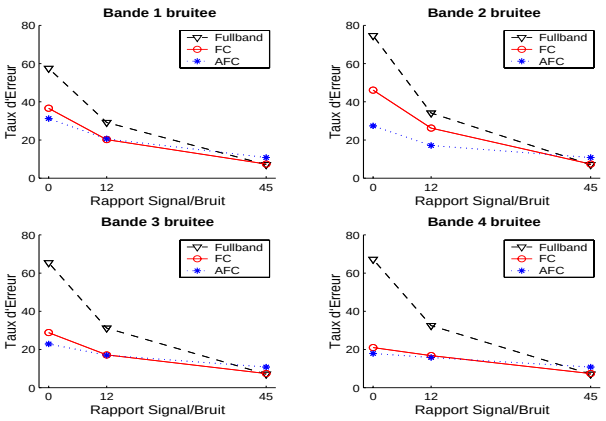

Figure 3: Taux d'erreur des systèmes *spectre entier*, *FC* et *AFC* en utilisant des **PLP**. Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée dans des sous-bandes 1 à 4.

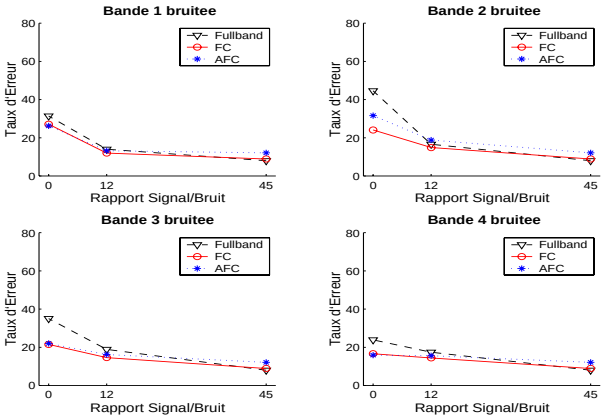

Figure 4: Taux d'erreur des systèmes *spectre entier*, *FC* et *AFC* en utilisant des **J-RASTA**. Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée stationnaire dans des sous-bandes 1 à 4.

La figure 4 présente les résultats avec les paramètres caractérisques J-RASTA-PLP. On voit que le filtre J-RASTA est capable de supprimer une partie des interférences dues au bruit coloré ce qui améliore les résultats sur tous les bruits colorés comparés aux résultats avec PLP seul. Mais là encore le FC en tout RSB est plus performant que le système spectre entier (ou égale en claire). Il en est de même pour l'AFC sauf en parole claire où les performances sont moindres car l'expert spectre entier est estimé, la perte des termes de dépendance inter-bande se faisant alors plus ressentir. De même les résultats FC ou AFC obtenus avec les PLPs et J-RASTA-PLPs dans le cas du bruit additif non-stationnaire (fig. 5) indiquent une robustesse plus élevée que celle du spectre entier, même si on note une hausse génèrale des taux d'erreur par rapport aux deux expériences précèdentes (il y a en effet plus de trames bruitées par effet de bord). Dans ces conditions le processus J-RASTA est mis en défaut par rapport au PLP simple, ce qui montre l'inadéquation de J-RASTA à un bruit si non-stationnaire.

Une propriété intéressante de nos modèles est que l'AFC montre une robustesse égale ou supérieure au

FC, sauf dans le cas où la bande 2 est bruitée. Nous avons mesuré que la bande 2 est la plus performante en RAP propre parmi les 4 sous-bandes, l'AFC est donc particulièrement pénalisée dans ce cas. Dans tous les autres cas, les bonnes performances de l'AFC sont dues à une meilleure exploitation de la redondance du signal. En effet en condition bruitée les entropies des vecteurs de probabilités *a posteriori* augmentent autrement dit ces vecteurs ont une distribution plus applatie : les probabilités *a posteriori* tendent vers l'équiprobabilité. Donc dans le modèle AFC qui procède par produits et normalisations, les sous-bandes bruitées affectent peu la distribution des vecteurs porteurs d'information correcte qui eux sont trés discriminants. Le modèle AFC joue donc le rôle d'un filtre : l'information provenant d'une sous-bande de données claires est mieux conservée en sortie du modèle AFC, alors qu'elle est noyée avec les données bruitées dans le cas du modèle FC. En FC dès qu'un flux est partiellement bruité, les probabilités *a posteriori* issues directement de l'expert correspondant sont globalement détériorées et irrécupérables comme le montre [?], ce qui défavorise le FC.

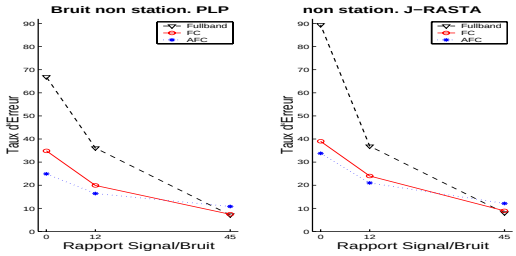

Figure 5: Taux d'erreur des systèmes *spectre entier*, *FC* et *AFC* en utilisant des **PLP** (à gauche) et des **J-RASTA** (à droite). Parole propre (RSB = 45 dB) et bruitée à 0 et 12 dB RSB. Bruit à bande limitée non-stationnaire variant entre les sous-bandes 1 à 4.

4. Conclusions et Perspectives

Nous avons montré qu'en utilisant l'approche sous-bandes FC ou son approximation AFC, même dans le cadre le plus simple de la pondération équiprobable, la robustesse du RAP est plus élevée que celle d'un modèle spectre entier J-RASTA pour les bruits à bande limitée stationnaires et non-stationnaires. Avec des modèles à résolution spectrale supérieure des résultats identiques sont attendus en bruits naturels, ce qui est réalisable efficacement avec la procédure récursive présentée dans ce papier. Une amélioration en parole propre du modèle AFC est accessible en utilisant le vrai estimateur spectre entier, les performances en claires pour AFC et FC sont alors comparables suivant nos récentes expériences. Cette étude comparative entre FC et AFC a mis en évidence que le modèle AFC a des performances sensiblement égales en conditions bruitées, et parfois même supérieures au FC, ce qui a été discuté. Des études publiées ou en cours montrent un accroissement de robustesse des modèles lorsque les poids sont variables selon les performances relatives des flux en condition claire comme les poids « Expectation Maximization » et « Least Mean Square Error » [?]. Mais les gains en robustesse sont plus grands avec l'usage de poids adaptatifs au bruit. Ces derniers peuvent être basés sur un classique RSB [HMB99] ou suivant une approche «CASA» à partir d'indices d'harmonicité ou de localisation spatiale, poids développés et testés en AFC dans [BG99, GBT99, ?].

Remerciements:

Ce travail a été soutenu par les projets Européens TMR SPHEAR et LTR RESPITE, et l'office Fédéral de l'Education et de la Science (OFES).

Bibliographie

[BG99] F. Berthommier and H. Glotin. A new snr-feature mapping for robust multistream speech recognition. In *Proc. Int. Congress on Phonetic Sciences (ICPhS)*, 1999.

[BGTB98] F. Berthommier, H. Glotin, E. Tessier, and H. Bourlard. Interfacing of casa and partial recognition based on amultistream technique. In *Int. Conf. on Spoken Language Processing*, volume 4, pages 1415–1419, 1998.

[GBT99] H. Glotin, F. Berthommier, and E. Tessier. A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition. In *Proc. Euro. Conf. on Speech Com. and Tech. (EUROSPEECH)*, volume 5, pages 2351–2354, sept. 1999.

[HM94] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[HMB99] A. Hagen, A. Morris, and H. Bourlard. Different weighting schemes in the full combination subbands approach in noise robust asr. *Proc. ESCA Workshop on Robust Methods for Speech Reco. in Adverse Conditions*, 1999.

[HTP96] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. *Int. Conf. on Spoken Language Processing*, pages 462–465, 1996.