Reconnaissance robuste de la parole par segmentation signal/bruit en sous-bandes

Hervé Glotin[†,‡], *Emmanuel Tessier*[‡], *Hervé Bourlard*[†], *Frédéric Berthommier*[‡]

[‡]Institut de la Communication Parlée, 46 avenue Félix Viallet, 38000 Grenoble, France
[†]IDIAP, rue du Simplon 4, 1920 Martigny, Suisse
e-mail : (glotin, bourlard)@idiap.ch, (tessier, bertho)@icp.inpg.fr

RÉSUMÉ

Nous proposons un modèle fondé sur le couplage d'un système de reconnaissance de la parole et de l'analyse de scène auditive (modèle CASA). Pour la reconnaissance de la parole, nous faisons appel à un système hybride HMM/ANN (Modèles de Markov cachés / réseaux de neurones artificiels), modifié de façon à classifier des signaux dont la représentation temps-fréquence est partielle. Le but est d'obtenir une robustesse de l'identification d'un signal de parole malgré l'addition d'un bruit de bande intense. Le modèle CASA identifie à chaque instant la sous-bande dans laquelle le bruit est présent de façon à l'exclure. Nous examinons différents modes de sélection fondés sur un indice de confiance calculé AVANT ou APRES l'entrée dans l'étape de reconnaissance. L'indice AVANT basé sur l'entropie de Shannon, est évalué au niveau primitif. Il tient compte de la structure harmonique du signal. L'indice APRES, également entropique, est calculé sur les sorties des MLP. L'indice AVANT est le plus efficace au vu des performances de reconnaissance de parole continue établies sur NUMBERS93.

MOTS CLEFS

Reconnaissance robuste de la parole continue, analyse de scène auditive, système multi-bandes, système hybride HMM/ANN, entropie.

1. INTRODUCTION

Tous les algorithmes de reconnaissance de la parole (ASR en anglais, pour *Automatic Speech Recognition*) ont une grande sensibilité au bruit. Or, l'oreille humaine est très robuste à ces perturbations. Les signaux de parole occupent une bande de fréquence utile qui s'étale de 100 Hz à 4 kHz et présentent une certaine redondance dans le domaine fréquentiel. Allen [All94] suggère que le processus de reconnaissance humain est basé sur l'exploitation de cette redondance : entre 3 et 4 sous-bandes peuvent être traitées indépendamment avec une reconnaissance de leur contenu propre. *A priori*, ceci permet de mieux résister à la présence d'un bruit masquant complètement le contenu de l'une de ces bandes (*i.e.*, c'est une amputation de la représentation spectrale du signal) à condition de pouvoir sélectionner les sous-bandes dans lesquelles le signal est dominant. Les modèles classiques ne comportent pas une telle étape, dite de segmentation primitive, au cours de laquelle on sépare les signaux à reconnaître des bruits parasites. Pour cela, nous utilisons un système de sélection, dit AVANT, fondé sur une étape de traitement intermédiaire tenant compte du degré d'harmonicité du signal. Nous couplons ce processus avec un modèle de reconnaissance approprié, de type HMM/ANN [MB95]. Pour comparaison, nous étudions aussi un mode de sélection dit APRES, fondé sur la sélection de la sous-bande bruitée à partir des distributions de sortie des ANN.

2. LE MODÈLE DE RECONNAISSANCE DE LA PAROLE

2.1. *Le système hybride HMM/ANN*

Nous utilisons un système hybride HMM/ANN. Pour un vecteur acoustique donné, l'estimateur de sa probabilité d'appartenance à l'une des classes de sortie (*i.e.*, les états du modèle HMM) est un réseau de neurones de type perceptron multicouche (MLP). Les MLP sont entraînés et testés à partir de spectres Log-Rasta-LPC (LPC pour "Linear Prediction Coding") calculés indépendamment sur chaque sous-bande ou groupe de sous-bandes [HTP96]. Le découpage fréquentiel en 4 sous-bandes est : [0, 901] Hz, [797, 1661] Hz, [1493, 2547] Hz et [2298, 4000] Hz. Pour la reconnaissance, les fenêtres d'analyse du signal sont de 25 ms et se recouvrent sur 12,5 ms. Les MLP génèrent pour chaque fenêtre les probabilités des états HMM (états "phonétiques", au nombre de 58). Celles ci sont données au décodeur acoustico-phonétique de type Viterbi qui réalise leur mise en séquence. Les MLP possèdent tous une seule couche cachée de 400 unités. Pour rendre compte du contexte, 9 fenêtres consécutives sont présentées simultanément à l'entrée du MLP. Nous notons MLP(x) les 4 MLP "partiels" entraînés sur une seule sous-bande. Le vecteur caractérisant chaque fenêtre comporte les dérivées première et seconde de l'énergie du signal (30 entrées), plus le nombre de coefficients cepstraux choisis, qui varie pour chaque sous-bande (respectivement 8, 5, 3 et 3). Le nombre d'entrées est respectivement pour $x = \{1, 2, 3, 4\}$: 162,135,117,117. Nous notons MLP(xyz) les 4 MLP partiels, chacun entraîné et testé sur les combinaisons (xyz) de trois sous-bandes. Ils sont classés suivant le numéro de la bande manquante et ils ont respectivement 369, 396, 414, 414 entrées. Le reconnaisseur "classique" en pleine bande est noté MLP(1234). Il est entraîné et testé avec des entrées

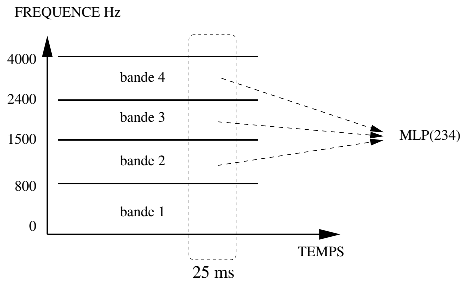

Figure 1: Découpage temps fréquence, et analyse d'une fenêtre de temps par le MLP(234).

Log-Rasta-PLP à 12 coefficients. Avec les 14 coefficients d'énergie, ce MLP(1234) possède 234 entrées. La Figure 1 montre l'analyse d'une fenêtre par le MLP(234).

2.2. Estimation des fonctions de vraisemblances

Soit X une séquence de N vecteurs acoustiques $X = \{x_n\}_{n\in[1,N]}$. Chaque modèle HMM de mot, M_i, est construit à partir d'un ensemble de C classes $\Omega = \{\omega_c\}_{c\in[1,C]}$, une classe par phonème. Chaque topologie HMM M_i est définie comme un graphe orienté contenant K états $q^k{,}_{k\in[1,K]}$ chacun associé à une classe ω_{q^k} de Ω. Les MLP sont de bons estimateurs de probabilités *a posteriori* d'appartenance à une classe [MB95]. Chaque MLP possède $C = 58$ unités de sortie (une par classe ω_c) et il est entraîné, puis testé sur les vecteurs acoustiques de 25 ms pour générer les probabilités *a posteriori* $P(\omega_c|x_n,\Theta)_{\forall c\in[1,C]}$, où Θ représente l'ensemble des paramètres du MLP. Or, les probabilités *a priori* $P(\omega_c)$ sont connues et, d'après le théorème de Bayes nous avons :

$$\frac{P(\omega_c|x_n,\Theta)}{P(\omega_c)} = \frac{\mathcal{P}(x_n|\omega_c,\Theta)}{\mathcal{P}(x_n)}$$

avec $\mathcal{P}$ des densités de probabilité. Les $\mathcal{P}(x_n)$ étant constants et indépendants de la classe, le terme de gauche est une grandeur proportionnelle à la vraisemblance $P(x_n|\omega_c,\Theta)$ [Bou96]. Cette valeur sera utilisée comme probabilité d'appartenance à une classe dans l'algorithme de décodage.

2.3. Environnement

Nous utilisons l'environnement STRUT intégrant les étapes de traitement du signal, de décodage acoustico-phonétique et de statistique après correction orthographique. La phase d'apprentissage et les tests sont réalisés en parole continue sur NUMBERS93 [OGI Numbers'93 DB]. Celle-ci est constituée de 2167 phrases téléphonées de nombres produits par 1132 locuteurs. Nous avons utilisé 1534 phrases pour l'entraînement et 384 phrases pour le test. Nous ne réalisons pas d'entraînement à partir de signaux bruités. Le bruit additif utilisé pour les tests est synthétisé à partir d'un bruit blanc gaussien de bande [0,430] Hz, donc seule la première bande est significativement bruitée. Le rapport signal sur bruit est de 0 dB RMS en moyenne phrase par phrase (silences inclus et en peline bande).

2.4. Les MLP Associatifs et Combinatoires

Nous proposons deux modèles :

1. l'un "associatif" est fondé sur la sélection et l'association des MLP(x) par produit,
2. l'autre "combinatoire" est basé sur la sélection des MLP(xyz).

Dans le cas d'un bruit de bande étroit stationnaire dans chaque fenêtre à court terme (*i.e.*, ne contaminant qu'une seule sous-bande à la fois), un sélecteur guidant le choix du meilleur MLP permet de conserver un taux de reconnaissance optimum. Tout d'abord, nous évaluons systématiquement les performances de tous les MLP(xyz) et MLP(x) avec et sans bruit dans la sous-bande 1 (voir Table 1). Le modèle associatif résulte du produit des réponses des MLP(x), fenêtre par fenêtre. Il est intéressant de noter que l'association des MLP(xyz) conduit à 11,5 % d'erreur en signal propre, ce qui est équivalent au MLP(1234). Cela suggère qu'un modèle, également de type "associatif", mais avec des groupes de sous-bandes (i.e., à partir des MLP(xyz), plus robustes) est à envisager.

type du MLP	% err. signal propre	% err. signal bruité	MLP associés	% err. signal propre
(1)	38.9	84.8		
(2)	39.3	40.5		
(3)	55.3	56.8	Π_4 MLP(xyz)	11.5
(4)	65.2	65.7		
(1234)	11.3	55.6	(1)*(2)*(3)*(4)	27.6
(234)	19.0	19.2	(2)*(3)*(4)	38.6
(134)	14.9	55.9	(1)*(3)*(4)	32.7
(124)	12.2	48.6	(1)*(2)*(4)	27.8
(123)	12.3	50.5	(1)*(2)*(3)	24.9

Table 1: Taux d'erreur en pourcentage de mots continus reconnus: signal propre, signal bruité en sous-bande 1, pour les MLP(x), MLP(xyz), MLP(1234) et les produits de trois ou quatre MLP(x) (par ex. (1)*(2)*(3)). Le produit des 4 MLP(xyz) est noté Π_4 MLP(xyz). Les taux d'erreur des modèles de droite en signal bruité n'ont pas été calculés.

Nous voyons immédiatement table 1 que le modèle "associatif" conduit à de mauvais résultats par rapport au modèle "combinatoire". Cela traduit la perte des informations de covariance entre sous-bandes. Nous développons un sélecteur adapté à ces deux types de modèles.

3. LE PRINCIPE DE SÉLECTION

Il est fondé sur l'utilisation d'une mesure d'entropie, soit appliquée sur une représentation intermédiaire du signal AVANT reconnaissance, soit évaluée sur les fonctions de vraisemblance associées à chaque MLP (sélecteur dit APRES). Il permet dans les deux cas la sélection du ou des meilleurs MLP pour chaque fenêtre de temps de 25 ms. La figure 2 en donne le schéma général.

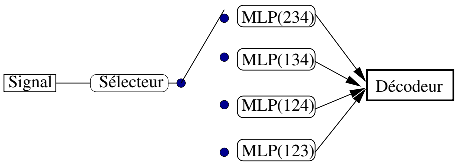

Figure 2: Principe de sélection des MLP combinatoires dans le processus de reconnaissance. Ici sélection du MLP(234) pour une fenêtre donnée.

3.1. Le sélecteur AVANT

L'entropie d'un système est une mesure quantitative de son degré de désordre. Appliquée sur une représentation du signal, elle est potentiellement capable de nous indiquer l'existence de structures, par opposition au bruit, dont l'entropie est maximale. En effet l'auto-corrélation d'un bruit est une distribution quasi-uniforme, donc son entropie est élevée, alors que l'entropie de l'auto-corrélation d'un signal non bruité périodique ou harmonique sera plus faible.

Nous voulons obtenir des sélecteurs aptes à travailler en bruit non-stationnaire, de plus l'analyse de périodicité d'un signal de parole réclame des fenêtres temporelles à moyen terme. Nous calculons donc pour chaque fenêtre de 50 ms, glissantes de 12,5 ms, l'entropie de l'auto-corrélation en sous-bandes ou en groupes de sous-bandes.

Après démodulation du signal (rectification + passe-bande, selon [BL96]), puis normalisation, nous calculons l'entropie $H(AC)$ de l'auto-corrélogramme :

$$H(AC) = -\sum_{t=1}^{N}\left(\frac{e(t)}{\sum_{t=1}^{N} e(t)}\right)log_2\left(\frac{e(t)}{\sum_{t=1}^{N} e(t)}\right) \quad (1)$$

avec : $e(\tau) = ac(\tau) - \min(ac(\tau)) + 1$ et τ l'axe des délais, N étant le nombre d'échantillons de l'auto-corrélation pris en compte et ac(τ) l'auto-corrélogramme.

Deux entrées sont utilisées pour ce calcul de l'entropie d'autocorrélation : les groupes de sous-bandes (xyz), ou les sous-bandes (x) prises indépendamment. Dans le premier cas, l'identification de la Bande Bruitée (IBB) est effectuée fenêtre par fenêtre en sélectionnant le groupe de sous-bandes dont l'entropie est minimale : sélecteur H(AC(xyz)). Dans le second cas, la sous-bande identifiée comme étant bruitée est celle qui possède l'entropie maximale. Ce second mode de sélection est noté H(AC(x)).

3.2. Le sélecteur APRES

Pour définir ce sélecteur, nous appliquons une mesure d'entropie sur la distribution des sorties des MLP associée à chaque fenêtre. En effet en s'appuyant sur les représentations utilisées au cours de l'étape de reconnaissance, cette mesure signale si l'information d'entrée s'apparie correctement avec l'information mémorisée, qui décrit l'ensemble des structures devant être reconnues. Le bruit, l'absence de structure, ou bien des distorsions importantes seront aussi diagnostiqués à ce niveau. Ainsi, nous pourrons optimiser la reconnaissance en choisissant le MLP le plus discriminant.

Notre fonction de transfert des MLP étant la sigmoïde, nous pouvons faire les calculs de l'entropie sur les sorties normalisées ou non. Au vu des résultats la normalisation est nécessaire car la bonne Identification de la Bande Bruitée (IBB, voir définition plus bas) est augmentée de 14 points : 44.6 % contre 30.7 %. De plus afin de rester homogène avec le sélecteur AVANT nous avons élargi la fenêtre d'intégration à 50 ms en moyennant pour une frame donnée son entropie avec celle de sa voisine précédente et suivante. Compte tenu de leur recouvrement, l'intégration est faite sur 50 ms. Le gain de bonne IBB ainsi obtenu est assez faible : 1.4 points (46.0 % contre 44.6 %). Nous avons mesuré que pour du bruit stationnaire plus la fenêtre est large plus le taux de bonne IBB augmente, mais notre but est de rester dans la perspective d'une application en bruit non stationnaire, ce qui requiert des fenêtres d'intégration temporelle de l'ordre de 50 ms.

L'entropie d'une fenêtre F est donc calculée sur une moyenne de 3 entropies calculées chacune après normalisation des sorties associées des MLP. Finalement le mode de sélection APRES est défini par le choix pour la fenêtre F du MLP(xyz) dont l'entropie est la plus faible. Ou inversement par le choix du MLP(x) qui conduit à l'entropie maximale.

4. RÉSULTATS

4.1. Les taux d'IBB

Nous évaluons l'IBB avec du bruit de bande stationnaire additif en bande 1. Nous montrons (Table 2) les taux d'identification corrects de la bande 1, et ce pour les sélecteurs AVANT et APRES.

bande d'entrée	AVANT H(AC)	APRES H(MLP)
(1)	60.1	34.4
(234)	81.5	46.0

Table 2: Scores d'identification de la bande bruitée (toujours la bande 1) sur l'ensemble des fenêtres de la base de test. Conditions : signal bruité (0 dB), modes de sélection AVANT et APRES, à partir des bandes d'entrées (1) ou (234) pour AC(1) ou AC(234), et MLP(1) ou MLP(234).

Nous avons illustré avec la figure 3 la sortie de ce processus de sélection. La bande sélectionnée dans chaque fenêtre de temps est indiquée, pour les différents modes de sélection décrits dans l'article.

4.2. Les scores de reconnaissance

Les scores sont exprimés en terme de taux d'erreur. Le taux d'erreur est la proportion de mots incorrects en faisant la somme (délétions + insertions + substitutions). En pleine bande et sans bruit, il est de 11,3 %. Les résultats

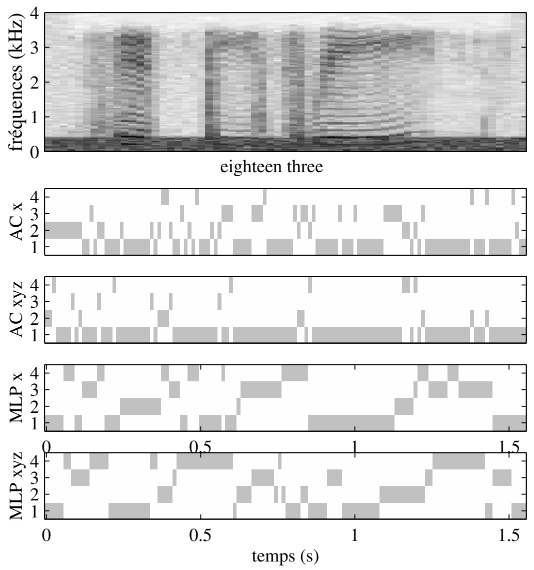

Figure 3: Analyse d'un échantillon bruité de la base NUMBERS93 ("eighteen three", de durée 2.5s). De haut en bas : spectre du signal bruité à 0 dB par du bruit Gaussien de bande [0, 430] Hz. Pour chaque fenêtre de 25 ms, la sélection est indiquée en noir. De haut en bas, elle est établie: par entropie d'auto-corrélation mono-bande (x) ou sous-bandes groupées (xyz), et par entropie des sorties normalisées des MLP(x) ou les MLP(xyz).

de tous les MLP, avec et sans bruit, sont dans la Table 1. Nous vérifions que le MLP(234) présente les meilleurs scores, avec environ 19 % d'erreur, que le signal soit bruité ou non. Ce score est une borne inférieure qui correspond à une identification toujours correcte de la bande bruitée (taux d'IBB=100%).

Table 3, nous indiquons les taux d'erreur de reconnaissance après sélection automatique (i.e., sensible à l'IBB). Les scores sont établis uniquement pour les MLP combinatoires MLP(xyz), pour les deux modes de sélection AVANT et APRES. La sélection elle-même est réalisée par groupe (xyz), ou bien par sous-bandes indépendantes (x). Le mode AVANT avec H(AC(xyz)) est le plus performant, avec 22.0 % d'erreur. L'écart observé, de 2.8 points par rapport à la borne idéale, est imputable aux erreurs d'IBB (voir Table 2), peu fréquentes avec ce mode de sélection (81.5 % de bonne IBB). Celui-ci parvient à différencier correctement le bruit et le signal de parole. Nous n'avons pas calculé les scores de reconnaissance du modèle associatif après sélection automatique, car il est d'emblée moins performant (la borne de sélection idéale en signal propre est de 38.6 %, voir Table 1).

5. CONCLUSION ET PERSPECTIVES

Nous montrons un premier exemple performant de couplage entre un niveau d'analyse primitif et un système de reconnaissance de la parole HMM/ANN, testé sur une base de données de référence. Les résultats obtenus avec le sélecteur AVANT sont très prometteurs (22.0 % contre 55.6 %) et ils confirment la possibilité de séparer un

	AVANT : H(AC)	APRES : H(MLP)
sur (x)	27.5	42.5
sur (xyz)	22.0	37.6

Table 3: Taux d'erreur en pourcentage de mots continus reconnus sur le signal bruité avec sélection automatique des MLP(xyz) "combinatoires", suivant les 2 modes AVANT et APRES, avec les deux modes de calcul sur les mono-bandes (x) ou groupes de bandes (xyz). La référence du reconnaisseur classique MLP(1234) est 55.6 % d'erreur.

signal de parole et une source interférente au cours d'une étape primitive, et selon un mode AVANT ascendant ("bottum-up"). L'application de notre modèle à du bruit non stationnaire est assez immédiate grâce à la nature dynamique de nos sélecteurs. L'usage d'autres indices primitifs (ITD : Différence de Temps Interaurale, ou AM : Modulation d'Amplitude) de façon à mieux couvrir les zones non voisées, et à améliorer le modèle de sélection AVANT sont en cours.

De même, pour améliorer l'adaptativité du modèle de reconnaissance tout en gardant la possibilité de pré-calculer les représentations, nous pourrons tester des combinaisons de sous-bandes plus étroites (mais plus nombreuses) ou qui se superposent fréquentiellement. Dans cette perspective, ce sont les MLP de type associatifs MLP(x) qui seraient le mieux placés. Nous testerons également la possibilité de pondérer les reconnaisseurs sous-bandes plutôt que de les sélectionner.

Remerciements

Ce travail a été soutenu par le projet COST249, par Eurodoc (co-tutelle de thèse IDIAP-ICP) et il entre dans le cadre du contrat TMR SPHEAR.

BIBLIOGRAPHIE

[All94] B. Allen. How do human process and recognize speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, 1994.

[BL96] F. Berthommier et C. Lorenzi. Precise and perceptually relevant processing of amplitude-modulation in the auditory system: Physiological and functional models. *Neurobiology*, pages 139–153, 1996.

[Bou96] H. Bourlard. Reconnaissance automatique de la parole : modélisation ou description ? Dans *Journées Etude Parole '96*, pages 263–272, 1996.

[GTBB98] H. Glotin, E. Tessier, H. Bourlard, et F. Berthommier. Reconnaissance multi-bandes de la parole bruitée par couplage entre les niveaux primitifs et d'identification. Dans *Journées Etude Parole '98, à paraître*, 1998.

[HTP96] H. Hermansky, S. Tibrewala, et M. Pavel. Towards ASR using partially corruped speech. Dans *Intl. Conf. On Spoken Language Processing*, pages 458–461, Oct 1996.

[MB95] N. Morgan et H. Bourlard. Continuous speech recognition. *IEEE signal processing*, pages 25–42, May 1995.