Reconnaissance multi-bandes de la parole bruitée par couplage entre niveaux primitif et d'identification

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

[†] ICP 46 avenue Félix Viallet, 38000 Grenoble, France
[‡] IDIAP Simplon 4, Case Postale 592, 1920 Martigny, Suisse
e-mail: (glotin, tessier, bertho)@icp.inpg.fr, (glotin, bourlard)@idiap.ch

ABSTRACT

We propose to interface primitive and speech recognition stages. The goal is to improve identification of speech corrupted by a high level band-limited noise. We use primitive information linked to the pitch, which primarily allows us to detect the noisy subband. A multi-stream speech recognition system is applied to the selected sub-bands. We have tested different protocols to select sub-bands where the speech signal is dominant. One is using a confidence level calculated AFTER entry to a recognition stage. This is based on the Shannon entropy. We extend the use of entropy estimation to detect the corrupted sub-band from primitive information, BEFORE recognition. This is compared with the more classical R_1/R_0 periodicity estimation. Recognition performances, established on NUMBERS93, are better (error rate at 27.5 %) for BEFORE which is a bottom-up interfacing method, when the lower frequency band is noisy (0 dB SNR).

1. INTRODUCTION

Les modèles de reconnaissance de la parole ont à présent des performances excellentes lorsque le signal n'est pas bruité. Leur fiabilité découle d'une représentation appropriée des signaux, ainsi que de leur redondance. Néanmoins, celle-ci n'est pas exploitée correctement car les performances chutent beaucoup lorsqu'un signal interférant est ajouté. Afin de lever certains types d'interférences, dites focalisées (bruit de bande étroite ou sinusoïde), l'usage des SBSR ("Reconnaissance de la Parole en Sous-Bandes") a été proposé par Bourlard et coll. [B+96]. Leur principe est de construire une représentation (statistique ou MLP) des signaux de parole dans laquelle un niveau d'indépendance entre des zones fréquentielles est acquis au moment de la phase d'apprentissage. Pour la parole, chacune de ces zones spectrales, dites sous-bandes, code un formant en moyenne. Le but est de limiter la représentation des associations entre régions spectrales éloignées (i.e., de rendre indépendantes les analyses de ces régions, en éliminant l'utilisation des termes de covariance), tout en préservant une propriété de reconnaissance partielle d'une sous-bande ou d'un ensemble de sous-bandes.

Un modèle, dit "de reconnaissance partielle", a été proposé par Green et coll. [G+95], dont la propriété essentielle est de reconnaître des régions présélectionnées dans le plan temps-fréquence, ceci à partir de l'apprentissage en pleine bande. Mais ce modèle n'est applicable (voir [C+96]) qu'à condition de disposer d'un tel sélecteur, puis de restituer en temps réel la représentation utilisée pour la reconnaissance (i.e., en pratique, avec un modèle statistique, d'inverser des matrices de variance covariance partielles). Cette seconde contrainte, de nature algorithmique, n'est pas levée actuellement, mais nous souhaitons développer l'usage du principe de sélection, car il apporte une solution simple au problème du couplage entre le niveau dit primitif et l'étape de reconnaissance, ceci en vue d'acquérir une robustesse vis à vis de signaux interférents de nature variée.

Or, nous disposons de modèles inspirés d'études sur la perception humaine, de type CASA ("*Computational Auditory Scene Analysis*"), capables d'extraire des informations à partir du signal en tenant compte de ses propriétés structurelles selon différentes résolutions temporelles. Les indices primitifs sont : l'harmonicité / périodicité, la modulation d'amplitude (AM), les moments de début et de fin, la différence de temps interaurale (ITD) pour la localisation spatiale. L'extraction de ces indices primitifs repose sur l'utilisation du spectrogramme et de représentations intermédiaires en fonction du facteur de résolution temporelle ou spectrale.

Nous montrerons comment ces algorithmes de type CASA peuvent répondre au principe de sélection (étape de sélection dite AVANT) en prenant comme exemple l'attribut de périodicité ("*pitch*"), extrait par auto-corrélation. Avec ces méthodes, nous isolons les zones du plan temps-fréquence dans lesquelles un signal de parole est dominant sur d'autres signaux interférents (i.e., le rapport signal sur bruit est localement élevé dans le plan temps-fréquence), ou bien, inversement, nous assurons l'identification de la bande la plus bruitée. Puis, nous proposons d'associer une méthode de type CASA et un SBSR qui reconnaît partiellement la parole lorsqu'une seule sous-bande parmi 4 est exclue (i.e., 3 sous-bandes sur 4 sont sélectionnées). Nous nous inspirons d'une étude précédente [TB97] pour présenter une comparaison entre plusieurs modes de sélection, à partir d'indices établis AVANT ou APRES l'entrée dans l'étape de reconnaissance. Pour APRES, nous mettons en œuvre un nouvel indice global proposé par Bourlard et coll., fondé sur la mesure d'entropie. Cet indice permet de choisir la meilleure observation partielle réalisée sur le signal, ceci de façon autonome.

2. ENVIRONNEMENT

Nous utilisons l'environnement STRUT intégrant les étapes de traitement du signal, de décodage acoustico-phonétique et de statistique après correction orthographique. Celui-ci permet aussi de construire des SBSR. 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 du système hybride ANN/HMM [MB95] implémenté dans STRUT génèrent les probabilités des états HMMs (au nombre de 58), ceci pour chaque fenêtre, et l'enchaînement est réalisé par le Viterbi. Les MLP sont entraînés à partir de spectres LPC ("*Linear Predictive Coding*") [H+96] calculés sur chaque sous-bande indépendamment.

Nous notons MLP(xyz) les 4 combinaisons, chacune entraînée sur trois bandes x, y, z, avec pour comparaison un MLP pleine bande (entraîné sur le Log-Rasta PLP). L'une des 4 combinaisons sera sélectionnée pour chaque fenêtre de temps. La phase d'apprentissage et les tests sont réalisés sur NUMBERS93, sans bruitage. 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 est synthétisé à partir d'un bruit blanc Gaussien de bande [0, 430] Hz, donc seule la bande 1 est significativement bruitée sur la base de test. Le rapport signal sur bruit est de 0 dB RMS en moyenne phrase par phrase (silences inclus, et en pleine bande).

3. METHODES

3.1 Principe de sélection

Il est essentiellement fondé sur l'utilisation d'un index permettant de différencier le signal et le bruit. Cet index permet de sélectionner les zones du plan temps-fréquence dans lesquelles le rapport signal sur bruit (RSB) est trop défavorable pour conduire à une reconnaissance. Appliqué sur une représentation intermédiaire du signal AVANT reconnaissance, celui-ci est soit une mesure d'entropie, soit un index de périodicité R_1/R_0. Une mesure d'entropie est appliquée sur les distributions des sorties de chaque MLP (sélecteur dit APRES). L'index entropique est testé dans les deux niveaux afin d'homogénéiser les méthodes de sélection AVANT et APRES.

De façon générale, 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. L'utilisation de l'entropie sur le spectrogramme a été proposée par [A+97]. Appliquée sur les représentations utilisées durant l'étape de reconnaissance, elle 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 grâce à cet index global. Ainsi, nous pourrons optimiser la reconnaissance en choisissant le MLP(xyz), effectuant une reconnaissance partielle avec le RSB le plus favorable.

3.2. Sélecteur AVANT reconnaissance à partir de l'auto-corrélogramme

Identification de la bande bruitée (IBB) : l'auto-corrélation est évaluée dans chaque sous-bande sur des fenêtres glissantes de 50 ms, toutes les 12.5 ms (l'analyse de périodicité d'un signal de parole réclame des fenêtres temporelles à moyen terme), après démodulation (rectification + passe-bande, selon [BL96]). Nous appliquons une mesure d'entropie pour identifier la sous-bande bruitée. En effet l'auto-corrélation d'un bruit est une distribution quasi-uniforme, donc d'entropie élevée, alors que l'auto-corrélation d'un signal périodique ou harmonique sera plus faible. Nous renormalisons l'autocorrélogramme ac(τ), afin de calculer l'entropie sur une distribution:

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

avec $e(\tau) = ac(\tau) - \min(ac(\tau)) + 1$ et τ l'axe des délais, T étant le nombre d'échantillons de l'auto-corrélation pris en compte. L'indice R_1/R_0 est le rapport entre (1) R_1=amplitude du premier pic de l'autocorrélogramme dans l'intervalle correspondant aux fréquences [90, 250] Hz et (2) R_0=amplitude en τ=0.

L'IBB est effectuée par comparaison entre sous-bandes, en choisissant celle dont l'entropie d'auto-corrélation est la plus forte ou bien celle dont le rapport R_1/R_0 est le plus faible. Le processus de sélection AVANT consiste à choisir le MLP(xyz) qui ne comprend pas cette sous-bande. A priori, ce modèle sera plus satisfaisant dans les zones voisées, où la différence d'entropie périodicité-bruit est plus forte, alors que dans les zones non voisées, la distinction sera plus difficile (par ex. avec les fricatives).

3.3 Sélecteur APRES l'entrée dans l'étape de reconnaissance

Nous appliquons une mesure d'entropie après normalisation de la distribution des sorties *i* d'amplitude s_i associées à chaque MLP (dont la fonction de transfert est une sigmoïde). Ces sorties sont assimilables à des probabilités [MB95]. L'entropie permet d'associer une confiance à ce MLP en se basant sur la distribution de toutes ses sorties. Plus la distribution des probabilités est uniforme, plus l'entropie sera grande, moins la confiance pour ce MLP sera grande. Or, l'entropie augmente lorsque du bruit est présent dans l'entrée, car les traits phonétiques sont masqués, et de plus, le bruit lui-même correspond à une réponse du réseau dont l'entropie est forte (ici, le réseau n'est pas adapté au bruit). Pour une frame Fc, le mode de sélection APRES est défini par le choix du MLP dont l'entropie moyenne sur Fc et ses 2 voisines est la plus faible (fenêtre de 50 ms comme AVANT), car les autres combinaisons sont supposées recevoir le bruit dans leurs entrées. L'entropie est évaluée avec :

$$H(MLP) = -\sum_{i=1}^{N} s_i \log_2(s_i)$$

avec N = nombre d'états de sorties (58).

3.4 Performances d'IBB

Nous établissons les performances d'IBB des différents modes de sélection AVANT et APRES, ainsi que leurs recouvrements (Figure 1). Nous remarquons que AVANT et APRES sont complémentaires (les succès de l'un peuvent correspondent aux erreurs de l'autre). A priori, le sélecteur AVANT, en particulier R_1/R_0, fonctionnerait plutôt lorsque le signal est voisé, alors que le sélecteur APRES serait plus performant dans les segments consonantiques. D'autre part l'ensemble des fenêtres correctement détectées par R_1/R_0 le sont aussi par H(AC) à 5 % près. Les distributions des sélections sont indiquées Table 1 où les erreurs se repartissent d'une façon homogène dans les autres sous-bandes. La Figure 2 illustre la répartition temporelle des IBB. On observe que les deux sélecteurs AVANT ont un comportement similaire pour cet exemple.

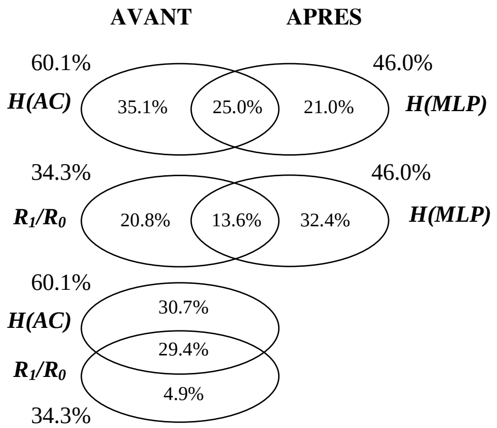

Figure 1 : Taux de succès IBB pour les méthodes AVANT R_1/R_0, H(AC) et APRES H(MLP)**.**

Table 1: Répartition de l'IBB, détaillant le choix de chaque sous-bande sur l'ensemble des fenêtres de temps de la base bruitée (silences compris).

Sous-bande sélectionnée	AVANT *R1/R0*	AVANT *H(AC)*	APRES *H(MLP)*
1	**34.3 %**	**60.1 %**	**46.0 %**
2	20.1 %	15.0 %	15.2 %
3	21.0 %	10.6 %	17.6 %
4	24.6 %	14.4 %	21.2 %

4. PERFORMANCE DE RECONNAISSANCE

Le taux d'erreur est la proportion de mots incorrects après décodage acoustico-phonétique et correction orthographique en faisant la somme (délétions + insertions + substitutions). En pleine bande et sans bruit, le taux d'erreur est de 11.3 %. Les résultats de tous les MLP, avec et sans bruit, sont dans la Table 2. Un MLP(xyz) réalise une reconnaissance partielle excluant l'une des sous-bandes parmi 4. Nous vérifions que MLP(234) présente les meilleurs scores, avec 19 % d'erreur, que le signal soit bruité ou non.

Ce taux d'erreur de 19 % est une borne inférieure qui correspond à une identification parfaite de la bande bruitée (toujours la bande 1 dans notre cas). Les écarts observés par rapport à ce score idéal sont imputables aux erreurs d'IBB. Tout d'abord, remarquons qu'une erreur d'IBB se traduit par une double pénalité : la réduction des données et l'introduction de bruit. Dans ce cas, la reconnaissance partielle s'effectue sur 2 bandes seulement (à 0 dB RMS, l'information transmise par la bande 1 est presque complètement masquée) sans que la bande bruitée ne soit exclue. En comparant les Tables 1 et 2, nous pouvons constater que les taux d'erreur obtenus avec les modes AVANT et APRES sont corrélés négativement au taux d'IBB. Néanmoins, les scores obtenus avec R_1/R_0 et H(MLP) sont similaires, ce qui indique une meilleure efficacité de R_1/R_0 peut être liée à une meilleure IBB dans les segments parole+bruit ou bien lorsque les signaux sont voisés.

Table 2: Taux d'erreur de reconnaissance en mots continus suivant le MLP utilisé, sur les phrases test de NUMBERS93. Sélections ad hoc (*), puis sélections AVANT et APRES. Un bruit gaussien de bande [0, 430] Hz à 0 dB est appliqué en sous-bande 1.

<table>
<tr><th>MLP sélectionné</th><th>Signal Propre (*)</th><th>Signal bruité (*)</th><th>Signal bruité AVANT R₁/R₀</th><th>Signal bruité AVANT H(AC)</th><th>Signal bruité APRES H(MLP)</th></tr>
<tr><td>MLP(123)</td><td>12.3 %</td><td>50.5 %</td><td rowspan="5">37.9 %</td><td rowspan="5">27.5 %</td><td rowspan="5">37.6 %</td></tr>
<tr><td>MLP(124)</td><td>12.2 %</td><td>48.6 %</td></tr>
<tr><td>MLP(234)</td><td>19.0 %</td><td>19.2 %</td></tr>
<tr><td>MLP(134)</td><td>14.9 %</td><td>55.9 %</td></tr>
<tr><td>MLP(1234)</td><td>11.3 %</td><td>55.6 %</td></tr>
</table>

5. CONCLUSION ET PERSPECTIVES

Nous exhibons un premier exemple performant de couplage entre un niveau d'analyse primitif et un système de reconnaissance de la parole, testé sur une base de données de référence. Les résultats obtenus avec le sélecteur AVANT sont très prometteurs et ils confirment la possibilité de séparer un signal de parole et une source interférente au cours d'une étape primitive, et selon un mode AVANT ascendant ("bottum-up"). Les autres solutions semblent moins performantes dans le contexte que nous avons choisi mais elles pourraient apporter des solutions lorsque le signal interférant est de nature différente. Pour progresser dans l'élaboration d'un modèle de reconnaissance robuste à des bruits de nature variée, plusieurs méthodes sont envisagées :

• En premier lieu, nous vérifierons ces conclusions en bruitant les bandes 2, 3 et 4. Nos premiers résultats indiquent une différence de comportement des sélecteurs AVANT, avec de meilleures performances de R_1/R_0 par rapport à H(AC), tout en gardant des taux d'IBB satisfaisants.

• Test du même modèle avec des bruits non stationnaires (à bande étroite mais dont la fréquence centrale change). Dans ce but, nous avons limité l'usage d'informations *a priori* sur le bruit (la continuité temporelle n'est pas une contrainte forte, puisque l'IBB est réalisée indépendamment dans chaque fenêtre temporelle).

• Usage d'autres indices primitifs (ITD, AM) de façon à mieux couvrir les zones non voisées, et à améliorer le modèle de sélection AVANT. Remarquons que ces indices sont extraits à partir d'autres représentations intermédiaires exprimées selon des échelles de temps différentes (plus petite pour l'ITD, plus grande pour l'AM).

• Amélioration de l'adaptativité du modèle de reconnaissance de façon à préserver sa capacité de reconnaissance partielle tout en gardant la possibilité de pré-calculer les représentations. Pour cela, nous pourrons tester des combinaisons de sous-bandes plus étroites (mais plus nombreuses) ou qui se superposent fréquentiellement, ainsi que les modèles associatifs à points de recombinaisons syllabiques [MB95].

• Des modes de fusions autre que la sélection / recombinaison de sous-bandes sont envisageables par pondération des probabilités sous-bandes à partir des indices primitifs.

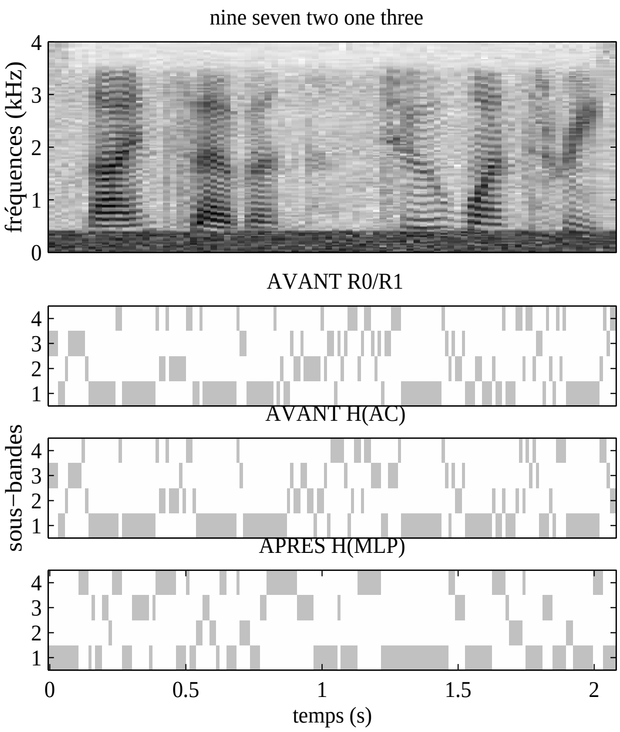

Figure 2 : Analyse d'une séquence bruitée de la base NUMBERS93. De haut en bas : spectrogramme du signal bruité à 0 dB. Dans les 3 figures du bas, pour chaque fenêtre de 25 ms, la bande sélectionnée est indiquée en grisé pour AVANT R_1/R_0, H(AC) et APRES H(MLP).

REMERCIEMENTS

Ce travail a été soutenu par le projet COST249, par Eurodoc, et il entre dans le cadre de SPHEAR, contrat TMR coordonné par Phil Green. Nos remerciements à Stéphane Dupont pour son aide et ses conseils dans l'élaboration du système multi-bandes que nous utilisons pour cet article.

BIBLIOGRAPHIE

[A+97] I. Abdallah, S. Montresor & M. Baudry, "Un algorithme récursif pour la segmentation des signaux de la parole basé sur un critère entropique local", *Actes du 4ième congrès français d'Acoustique,* Marseille, 1997, pp. 85-88

[BL96] F. Berthommier & C. Lorenzi, "Precise and perceptually relevant processing of amplitude-modulation in the auditory system: physiological and functional models", In *Neurobiology*, V, Torre & F. Conti Eds, New-York, Plenum Press, 1996, pp. 139-153

[B+96] H. Bourlard, S. Dupont, H. Hermansky & N. Morgan, "Towards subband-based speech recognition", in *Proc. of European Signal Processing Conference,* Sept. 1996, pp. 1579-1582

[C+96] M. P. Cooke, A. Morris & P. Green , "Recognising occluded speech", in *Proc. of Workshop on the auditory bases of speech processing,* Keele, 1996, pp. 297-300

[G+95] P. D. Green, M. P. Cooke & M. D. Crawford, "Auditory scene analysis and HMM recognition of speech in noise", in *Proc. ICASSP,* 1995, pp. 401-404

[H+96] H. Hermansky, M. Pavel & S. Tibrewala, "Towards ASR using partially corrupted speech", in *Proc. of Intl. Conf. On Spoken Language Processing,* Oct. 1996, pp. 458-461.

[MB95] N. Morgan & H. Bourlard, "Continuous Speech Recognition - An Introduction to the Hybrid HMM/Connectionist Approach", *IEEE Signal Processing Magazine,* Invited Paper, vol 12, n°3, May 1995

[TB97] E. Tessier & F. Berthommier, "A model of the cumulated effect of pitch and interaural delay differences for double vowel segregation", in *Proc. of ICSP'97*, Séoul, 1997, pp. 753-758