COMPARAISON DE TECHNIQUES DE PARAMÉTRISATION SPECTRALE POUR LA RECONNAISSANCE VOCALE EN MILIEU BRUITÉ.

*G. BAUDOIN, *P. JARDIN, **G. CHOLLET, ***J. GROSS

*ESIEE, BP99, Noisy Le Grand, 93162 CEDEX, France - **IDIAP, CP 609, 1920 Martigny, Suisse
***UNIVERSITÉ DE LJUBLJUANA, Slovénie

RÉSUMÉ

Plusieurs paramétrisations du signal de parole sont testées et comparées pour la tâche de reconnaissance de mots isolés monolocuteur en présence de bruit. Les paramètres de type coefficients cepstraux pondérés (coefficients de retard de groupe) s'avèrent les plus performants lors d'une dégradation du rapport signal sur bruit. La prise en compte d'une échelle "perceptive" pour les coefficients MFCC pondérés explique la supériorité de ceux-ci par rapport aux autres coefficients de même type. Les coefficients LSP, satisfaisants dans un environnement idéal donnent des résultats très dégradés dans le bruit.

ABSTRACT

Several speech spectral representations are tested and compared for the isolated word speaker dependent recognition task. Weighted cepstral coefficients (group delay coefficients) were found to be the most efficient in presence of noise. Performances are even better when a MEL frequency scale is used for the cepstral weighted coefficients. The LSP coefficients quite good in a quiet environment, become inefficient with noise.

Les techniques de reconnaissance vocale utilisées aujourd'hui ont de très bonnes performances lorsque le signal vocal est enregistré dans de bonnes conditions : peu de bruit, bon microphone. Mais les résultats sont encore insuffisants dans de nombreuses situations réelles où le bruit n'est pas négligeable (automobile par exemple).

Le travail présenté a consisté à comparer trois techniques de paramètrisation spectrale du point de vue de leurs performances pour la reconnaissance vocale et plus particulièrement à évaluer leur robustesse en présence de bruit. Cette étude s'inscrit dans le cadre d'un projet GRECO communication parlée (GDR - PRC - CHM), portant sur la comparaison de techniques d'analyse et de paramètrisation pour la reconnaissance. Le système s'appuie sur le programme existant SAMREC1 et les bases de données EUROM0 et NOISEROM (RSG_10) (puis BDBRUIT quand elle sera disponible). Les techniques sont jugées à travers leur score de reconnaissance monolocuteur en accès lexical à l'aide de méthodes de type DTW (Dynamic Time Warping).

Paramètres spectraux comparés

Les paramètres de référence sont les MFCC (Mel Frequency Cepstrum Coefficients) utilisés originellement dans SAMREC1. Nous les confrontons aux LSP (Line Spectrum Pairs) et aux coefficients de retards de groupe.

Les MFCC sont des coefficients cepstraux obtenus à partir des sorties d'un banc de filtres répartis sur une échelle MEL (un banc de 24 filtres a été utilisé ici).

Les LSP ou paires de raies spectrales permettent de décrire l'enveloppe spectrale à partir d'une analyse LPC. Ces paramètres sont très efficaces pour le codage à bas débit. Ils ont d'autre part été mis en oeuvre dans un système de reconnaissance [Paliwal,88], [Furui,90] où, en absence de bruit, ils ont donné des résultats un peu supérieurs à ceux obtenus avec une mesure par distance cepstrale "liftrée".
En ce qui concerne la perception de la parole, la position fréquentielle des formants est plus importante que leur largeur de bande. Aussi les LSP sont-ils de bons candidats pour la reconnaissance puisqu'ils sont formés de fréquences qui représentent assez bien les positions des formants. Mais leur robustesse au bruit apparait plus incertaine, à moins d'utiliser des techniques de

soustraction spectrale non linéaire avant l'analyse LPC [Mokbel,92] en vue d'atténuer l'influence du bruit.

Les coefficients de retard de groupe sont obtenus par pondération de coefficients cepstraux. Etant donnée une fonction de transfert rationnelle H(z), les coefficients cepstraux Cn sont les coefficients de Fourier de $\log(|H(\omega)|)$. On appelle coefficients de retard de groupe les coefficients Gn de la transformée de Fourier de la fonction de retard de groupe $t(\omega) = -\delta\phi(\omega)/d\omega$ (où $\phi(\omega) = \arg(H(\omega))$. Ils sont reliés aux coefficients cepstraux par: $Gn = nCn$. Cette pondération revient donc à négliger les coefficients Cn d'indices faibles c'est-à-dire qu'une importance moins grande est accordée à la pente globale du spectre. D'autre part, suite aux travaux de [Murthy et Yegnanarayana,91] on sait qu'il est possible d'exploiter les fonctions de retards de groupe pour l'analyse spectrale des signaux de parole et que ces fonctions sont robustes en présence de bruit.

Distances utilisées

Plusieurs distances locales sont utilisées pour la reconnaissance vocale.
La distance de Mahalanobis est définie par:

$$d_M(x_t, x_r) = (x_t - x_r)^T C (x_t - x_r)$$

C est la matrice de covariance ,$x_t$le vecteur de test et x_r le vecteur de référence.
La matrice de covariance des coefficients cepstraux étant presque diagonale ,d_M est équivalente dans ce cas à une distance euclidienne pondérée par l'inverse de la variance des coefficients:

$$d_{eiv} = \sum_1^L \frac{(x_{t,i} - x_{r,i})^2}{\sigma_i^2}$$

Pour les coefficients LPC qui sont assez fortement corrélés la distance d'Itakura est la plus efficace.

D'autres distances ont été proposées pour les coefficients cepstraux afin d'obtenir de meilleurs résultats dans le bruit:

- Distance de projection cepstrale (Mansour,Juang,89):
- Plusieurs pondérations (liftrages) des coefficients cepstraux ont été proposés ,en particulier les distances "à quéfrence pondérée" ou Root Power Sum (les c_n sont multipliés par leur indice n).

Ce liftrage remplace les coefficients cepstraux par les coefficients de retard de groupe.

Nous avons utilisé les distances suivantes:
- Pour les coefficients cepstraux, SAMREC1 utilise la distance euclidienne.
- Pour les LSP ,conformément aux travaux de Paliwal [Paliwal,88] nous avons considéré une distance euclidienne pondérée par la densité spectrale de puissance $P(\omega)$ élevée à la puissance c=0.15 ($P(\omega)$ est obtenue à partir du filtre A(z) de prédiction linéaire par :$P(\omega)=1/|A(\exp(j\omega))|$).
- Pour les coefficients de retard de groupe ,suivant les travaux de [Itakura,87], nous avons choisi une distance euclidienne pondérée par $w_n = n^s \exp(-n^2/2\tau^2)$.

Les coefficients de retard de groupe ont été calculés par deux méthodes différentes:
- les LPCC_RG obtenus en pondérant les coefficients LPCC (Linear Prediction Cepstrum Coefficient) par w_n.
- les MFCC_RG qui sont les coefficients MFCC (Mel Frequency Cepstrum Coefficient) pondérés de même par w_n.

(remarque:les LPCC pourraient être transposés sur une échelle MEL par transformation bilinéaire.)

Nous avons recherché les valeurs des paramètres τ et s qui assurent les meilleurs taux de reconnaissance. Lorsqu'on augmente τ et s les pics spectraux deviennent plus aigüs et la pente globale du spectre s'atténue. Les valeurs optimales obtenues sont:

τ=5 et s=2 pour les LPCC_RG.
τ=10 et s=1 pour les MFCC_RG.

Le tableau 1 donne le taux de reconnaissance moyen obtenu pour les LPCC_RG avec un rapport signal sur bruit de 10dB pour différentes valeurs de s et τ=5.Le tableau 2 donne les résultats pour les LPCC_RG dans les mêmes conditions de RSB pour différentes valeurs de τ et s=1.

s	score moyen (%)
0	31
1	56
2	75
3	51

tableau 1 :avec τ=5

τ	score moyen (%)
3	30
4	41
6	44
7	37
12	12

tableau 2 : avec s=1

Données utilisées pour l'expérimentation

Nous avons utilisé une base de signaux de parole et une base de signaux de bruit.

- La base pour la parole est extraite de EUROM0. Les expériences effectuées portent sur les chiffres en Français prononcés par quatre locuteurs différents. Il existe trois séquences de cent chiffres pour chaque locuteur. Vingt chiffres d'une séquence sont utilisés comme référence et les deux autres séquences complètes fournissent les mots de test.

- La base de bruits est la base NOISEROM RSG_10.Les tests ont été effectués avec du bruit blanc (signal.003) et du bruit de voiture (signal.023, volvo roulant à 130 km/h). Le spectre de ces bruits est donné figure 1.

Le bruit est ajouté aux signaux de parole suivant le rapport signal sur bruit spécifié. Dans nos expériences c'est un bruit multiplicatif dépendant de l'amplitude du signal. Nous sommes ainsi en mesure de comparer nos résultats avec ceux obtenus par Itakura [Itakura,87].

La fréquence d'échantillonnage de EUROM0 est de 16 kHz,celle de RSG_10 a été convertie de 19.98 kHz à 16 kHz.

La longueur des trames est de 20 ms, leur rythme est de 10 ms.

Résultats

Pour chaque méthode étudiée nous avons conservé le taux de reconnaissance et la matrice de confusion pour chaque séquence et chaque locuteur. Les tableaux suivants donnent les résultats (moyenne des taux de reconnaissance sur les séquences et sur les locuteurs) pour les différentes méthodes:

- MFCC10 : 10 MFCC avec une distance euclidienne.
- LSP20 : 20 LSP avec une distance euclidienne .
- LSP20w : 20 LSP avec une distance euclidienne pondérée.
- LPCC : 30 LPCC avec une distance euclidienne.
- LPCC_RG : 30 LPCC (à partir de 10 LPC) pondérés par $w_n(\tau=5,s=2)$ avec une distance euclidienne.
- MFCC_RG : 10 MFCC pondérés par $w_n(\tau=10,s=1)$ avec une distance euclidienne.

rsb	30	10	02
mfcc10	99.1	60.9	23.9
lsp20	94	34	13.2
lsp20w	79.9	29.4	17.2
lpcc	88	30	19
lpcc_rg	96.5	74.6	46.9
mfcc_rg	99.5	94.9	67.9

Tableau 3 : score moyen de reconnaissance : bruit blanc

rsb	20	0	-8
mfcc10	99.4	99.2	93.7
lsp20	98.9	98.1	87.5
lsp20w	96.7	93.1	79.2
lpcc	98	97.2	84.5
lpcc_rg	98.9	98.1	93.9
mfcc_rg	99.6	99.2	93.6

Tableau 4 : score moyen de reconnaissance : bruit de voiture

Conclusions

A rapport signal sur bruit identique, le bruit de la voiture roulant à 130 km/h sur route asphaltée est moins génant qu'un bruit blanc, la mesure de la puissance du bruit pondérée par une courbe psycho acoustique le laissait prévoir (A sur la figure 1)

Les paramètres MFCC_RG donnent les meilleurs résultats de reconnaissance en présence de bruit. Leur performance est encore satisfaisante pour un rapport signal sur bruit de 2 dB dans le cas d'un bruit blanc.

La différence entre MFCC_RG et LPCC_RG provient de l'emploi d'une échelle MEL pour les MFCC_RG, cette échelle pouvant aussi être utilisée pour les LPCC_RG au moyen d'une transformation bilinéaire.

Les paramètres LSP utilisés avec une distance euclidienne sont très sensibles au bruit. La pondération par la densité spectrale n'améliore pas les résultats

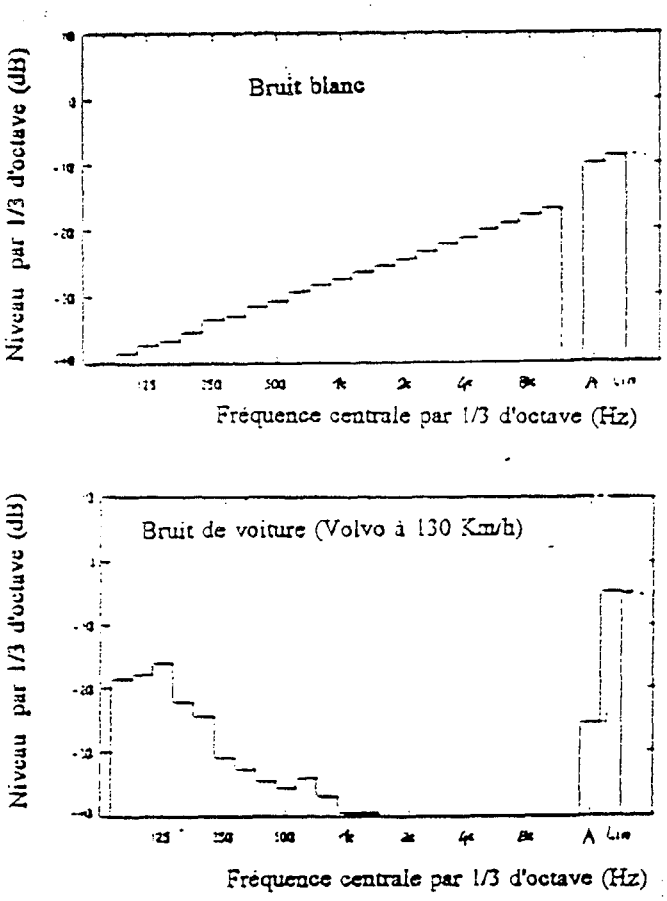

Figure 1

Références

[Itakura 87] Itakura F., Umezaki T., "Distance measure for speech recognition based on the smoothed group delay spectrum",ICASSP Proc.,87,pp. 1257-1260

[Mokbel 92] Mokbel C.,"Reconnaissance des mots en ambiance bruitée",Thése ENST 1992.

[Murthy 91] Murthy H., Yegnanarayana, "Speech processing using group delay functions", Signal Processing vol 22,1991, pp. 257-267

[Paliwal 88] Paliwal, "Perception based distance LSP measure for speech recognition", JASA, sup 1., vol 84, S15.

[Furui 90] Furui, S. Sagayama, S. Gurgen, "Line Spectrum Pairs based distance measures for specch recognition", International conference on spoken language processing, Kobe, Japan.