ANALYSE DES PERFORMANCES DU SYSTEME DE RECONNAISSANCE SYRIL 2*

B. Flocon, P. Lockwood, L. Sauter

LABORATOIRES de MARCOUSSIS
CR-C.G.E. - Route de Nozay - 91460 Marcoussis

ABSTRACT

SYRIL 2 is a speaker independent isolated word recognizer based on cepstral analysis, and dynamic time warping algorithm. Clustering techniques applied on a large data base are used during the building of the reference set. In its original version, SYRIL 2 is working with 5 representatives for each word of the vocabulary. 10 parameters are calculated every 16 milliseconds and stored on a 16-bit word. Different tests have been performed in order to see the influence of the reduction of the number of parameters or the number of bit for storing the data. Comparison between different ways for the calculation of the cepstra coefficients have been made. Finally a simple building of the reference set by choosing 5 speakers among 40 has been made : results have been compared with those obtained with the use of a clustering procedure on these 40 representatives.

INTRODUCTION

Les systèmes de reconnaissance de mots isolés multilocuteur donnent à l'heure actuelle des performances suffisamment bonnes pour que la réalisation de maquettes ou prototypes puisse être envisagée. Un certain nombre de contraintes liées à la conception des systèmes de reconnaissance rendent ces maquettes chères. Disposant d'un système de reconnaissance de mots multilocuteur simulé sur ordinateur (SYRIL 2), nous avons pensé qu'il serait intéressant de chercher à réduire le "coût" de certains éléments de notre système. Dans un premier temps, les éléments que nous avons cherché à simplifier concernent, d'une part, la partie "analyse du signal", et d'autre part, le principe utilisé pour la création de l'ensemble de références.

Les résultats présentés montrent que certaines simplifications n'entrainent pas de dégradation et ne remettent donc pas en cause les performances du système. D'autres, par contre, peuvent être jugées inacceptables car elles entrainent trop de différences avec le système original SYRIL 2.

*Etude partiellement financée par un contrat ESPRIT.

LE SYSTEME SYRIL 2

Ce système a été développé à partir de SYRIL [1]. Il s'agit d'un système de reconnaissance de mots isolés multilocuteur. Le vocabulaire disponible sur SYRIL 2 comporte 130 mots sans hiérarchisation du vocabulaire ; ces 130 mots correspondent aux commandes d'une machine de traitement de textes ; ils n'ont pas été particulièrement choisis et comportent certains groupes de mots voisins ("placer", "classer" ; "mot", "non" ...). Afin de resituer SYRIL, nous donnons brièvement une description des modules qui le composent :

1. Le module d'extraction d'extraction des paramètres de SYRIL et SYRIL 2 est effectué sur un array processor qui réalise en temps réel le calcul de 10 paramètres toutes les 16 millisecondes, après filtrage passe-bas à 4000 Hz, lissage par une fenêtre de Hamming de largeur 32 millisecondes et codage sur 12 bits. Les 10 paramètres utilisés dans SYRIL 2 sont 9 coefficients du cepstre (MFCC) calculés à l'aide de l'échelle Mel [2] et un paramètre représentant la puissance du signal (nous utilisons les MFCC à partir de MFCC(1) ; il n'y a donc pas directement de terme représentant l'énergie dans les 9 MFCC que nous utilisons).

2. L'ensemble de référence de SYRIL 2 comporte 5 références pour chaque mot du vocabulaire. Ces références ont été obtenues après traitement de 40 élocutions de chaque mot du vocabulaire provenant de 40 locuteurs différents (20 hommes et 20 femmes). Un algorithme de classification [1] utilisant un seuil adaptatif est utilisé pour déterminer 10 classes à l'aide des 40 éléments ; seules les 5 classes les plus représentatives sont conservées. Un algorithme de calcul de moyenne au sein d'une classe est appliqué afin d'obtenir un représentant unique pour chaque classe.

3. L'algorithme de reconnaissance de SYRIL 2 est basé sur la programmation dynamique [3] ; la règle du plus proche voisin est appliquée afin de décider quel est le mot reconnu. Dans la version de SYRIL 2 utilisée dans les expériences qui vont être décrites, aucun seuil de rejet ou de proximité n'est utilisé.

Les données utilisées dans SYRIL 2 sont stockées sur 16 bits.

MODIFICATIONS PAR RAPPORT AU SYSTEME INITIAL

Nous avons réalisé un certain nombre de modifications afin de tester les possibilités de simplifications qui pourraient être apportées au système SYRIL, sans entraîner une dégradation trop importante des performances. Ces modifications ont été effectuées à trois niveaux :

1. Modification sur le calcul des paramètres d'analyse

. calcul des paramètres cepstraux à l'aide de la prédiction linéaire (utilisation des LPCC : Linear Prediction Cepstrum Coefficients).

. utilisation d'un banc de filtres simulés pour calculer les coefficients cepstraux, au lieu d'une transformée de Fourier suivie d'un filtrage. Des bancs de 19 et 16 filtres ont été simulés ; les sorties de ces filtres ont été moyennées deux à deux, de manière à effectuer un lissage comparable à celui obtenu lors de l'utilisation de la transformée de Fourier.

2. Réduction de la quantité d'information représentant l'ensemble des données : deux orientations ont été testées :

. calcul de 5 paramètres pour représenter le signal toutes les 16 millisecondes (4 MFCC et un paramètre de puissance)

. compression des données sur 8 bits au lieu de 16 (après recentrage de la puissance) ; les essais de compression ont été effectués sur 5 et 10 paramètres.

3. Utilisation de 5 locuteurs pris au hasard pour constituer l'ensemble de références : au lieu d'enregistrer 40 personnes et d'effectuer les traitements de classification et de moyennage décrits plus haut, nous avons pris aléatoirement 5 locuteurs et nous avons utilisé leurs élocutions pour constituer l'ensemble de références, sans effectuer de traitement particulier sur les données correspondantes.

VOCABULAIRES DE TEST

Les 3 séries de tests décrits ci-dessus ont été réalisées sur différents types de vocabulaire. Dans tous les cas, l'ensemble de références est constitué de 5 références pour les 130 mots du vocabulaire de traitement de texte.

1. Premier vocabulaire de test : il s'agit d'utiliser 57 élocutions des 130 mots en tant que base de test. Cela représente une base de données de 130 x 57 = 7410 mots. Les performances obtenues correspondent à des performances d'un système fonctionnant pour une centaine de mots ; la difficulté de ce vocabulaire peut être considérée comme moyenne.

2. Second vocabulaire de test : le test de reconnaissance est effectué en comparant les 10 chiffres (0, 1, ... 9) aux 130 mots précédemment cités. Là encore 57 locuteurs ont été testés. La base de test comporte donc 10 x 57 = 570 mots. Le fait de comparer ces 10 mots aux 130 ne présume pas des performances réelles d'un système qui reconnaîtrait uniquement les 10 chiffres ; mais ce test permet de se placer dans des conditions où le nombre d'erreurs est suffisamment important pour que l'on puisse mesurer l'impact d'une modification au niveau de la détérioration des performances du système.

3. Troisième vocabulaire de test : un sous-vocabulaire de 18 mots dont la liste est donnée dans la figure 1 a été extrait des 130 mots. Ce sous-vocabulaire comporte un certain nombre de mots présentant plus de confusions que la plupart des autres mots. La figure 1 présente les mots avec lesquels sont confondus le plus fréquemment ces 18 mots. Ce 3è test consiste également à comparer ces 18 mots aux 130. La base de test comporte 18 x 57 locuteurs, soit 1026 mots. La liste de ces mots est donnée dans la figure 1.

La composition des vocabulaires de tests est rappelée dans la figure 5.

MODE	
MATHEMATIQUE	AUTOMATIQUE
MULTIPLIER	
MOT	NON
MODIFIER	
NUMERO	ZERO
NOUVEAU	
NON	MOT
OPERER	
OPTION	
OUI	
POSITION	
POURCENTAGE	
PAGE	
PRECEDENT	
PARTAGER	
PETIT	
PLACER	CLASSER

Figure 1 : Liste des mots du 3ème vocabulaire de test avec, en face, les confusions les plus fréquentes.

RESULTATS

Nous présenterons les résultats en différents tableaux. Dans chaque tableau, nous présenterons en première ligne le système de références (SYRIL 2) pour les 3 vocabulaires de tests définis plus haut. Les résultats sont donnés en nombres d'erreurs avec, entre parenthèses, l'augmentation du taux d'erreurs en pourcentage par rapport au système de référence (1ère ligne).

Une analyse de chacun des tests est faite après la présentation de chaque table de résultats.

- **Influence du type d'analyse**

	Vocab. 1	Vocab. 2	Vocab. 3
MFCC 256 pts	175	37	18
LPCC 256 pts	344 (+97%)	51 (+38%)	19
19 filt.	213 (+22%)	31 (-16%)	27 (+50%)
16 filt.	230 (+31%)	31 (-16%)	31 (+72%)

Figure 2 : résultats suivant le type d'analyse

Les résultats ci-dessus montrent que l'utilisation d'un banc de 19 filtres ou de 16 filtres pour calculer les MFCC au lieu d'une transformée de Fourier dégrade légèrement les performances. Par contre, le calcul des coefficients cepstraux par prédiction linéaire fait pratiquement doubler le nombre d'erreurs. Ceci semble confirmer la supériorité des MFCC [4] par rapport à d'autres types de paramètres.

- **Influence du nombre de bits et du nombre de paramètres utilisés pour la représentation des données**

énergie +	Vocab. 1	Vocab. 2	Vocab. 3
9 MFCC (16 bits)	175	37	18
9 MFCC (8 bits)	183 (+5%)	35 (-5%)	19 (+5%)
4 MFCC (16 bits)	235 (+26%)	49 (+32%)	27 (+50%)
4 MFCC (8 bits)	574 (+228%)	131 (+254%)	55 (+205 %)

Figure 3 : résultats suivant le nombre de bits et/ou paramètres

On voit clairement apparaître, sur les résultats ci-dessus, le fait que l'utilisation de 8 bits suffit largement pour la représentation des 10 paramètres utilisés dans le système de reconnaissance. Le fait de n'utiliser que 5 paramètres, au lieu de 10, dégrade légèrement les performances : par contre, le cumul des deux opérations (réduction du nombre de paramètres et du nombre de bits) apporte une dégradation très importante.

- **Influence de la méthode utilisée pour choisir les représentants**

Pour réaliser ces tests, nous avons effectué successivement 4 tirages aléatoires de 5 locuteurs parmi les 40 dont nous disposions. Les tests ont été effectués pour les 3 vocabulaires définis précédemment.

	Vocab. 1	Vocab. 2	Vocab. 3
5 classes	175	37	18
tirage 1	307 (+75%)	71 (+92%)	34 (+89%)
tirage 2	351 (+100%)	67 (+81%)	34 (+89%)
tirage 3	341 (+95%)	44 (+19%)	34 (+89%)
tirage 4	354 (+100%)	68 (+84%)	27 (+50%)

Figure 4 : résultats suivant la création des références

Le fait d'utiliser des références calculées par moyennage dans 5 classes obtenues après classification automatique appliquée à un ensemble de 40 éléments permet d'obtenir 2 fois moins d'erreurs que lorsque l'on prend aléatoirement 5 locuteurs pour créer l'ensemble de références. Rappelons que 175 erreurs pour le vocabulaire 1 correspondent à un taux d'erreurs de 2,3 %.

vocabulaire numéro	nombre mots	nombre locuteurs	nombre mots test
1	130	57	7410
2	10	57	570
3	18	57	1026

Figure 5 : vocabulaires de test

CONCLUSION

Comme cela a été dit plus haut, il ne s'agit pas ici de donner des performances du système SYRIL 2 dans son mode de fonctionnement normal. Le but est de comparer l'influence de certaines simplifications ou réductions de la complexité de SYRIL 2 sur les performances du système. Il en ressort un certain nombre d'informations intéressantes à différents niveaux : coût de la création de l'ensemble de références (multiplication du nombre d'erreurs par 2 si l'on n'utilise pas la classification), coût du stockage des données de l'ensemble de références (peu de changements en diminuant par 2 la quantité à stocker, mais forte dégradation quand on divise par 4), coût du calcul des paramètres (diminution des performances si on préfère un banc de 19 filtres à une transformée de Fourier), coût de la prédiction linéaire (diminution des performances par rapport aux résultats obtenus par la FFT et le filtrage selon l'échelle Mel).

Tous ces tests montrent bien que les systèmes SYRIL et SYRIL 2 doivent une part non négligeable de leurs performances à certains éléments qui les rendent "chers". Bien évidemment, dans le cas d'une application pratique, on cherchera à réduire les coûts ; mais un compromis coût/performance devra être choisi ; nous pensons que les résultats ci-dessus pourront jouer un rôle dans ce choix.

REFERENCES

[1] B. FLOCON, N. BRIANT
"SYRIL : Système temps réel de reconnaissance de mots isolés indépendant du locuteur", 4ème congrès AFCET RFIA, Paris 1984.

[2] S.B. DAVIS, P. MERMELSTEIN, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences", : IEEE-ASSP 28 n° 4, Août 1980.

[3] H. SAKOE, S. CHIBA, "Dynamic programming optimization for spoken word recognition", IEEE - ASSP 26 n° 1, Février 1978.

[4] C. GAGNOULET, M. COUVRAT, "SERAPHINE : a connected word speech recognition system", ICASSP - vol 2, p. 887, Paris, 1982.