Journées d'électroniques, EPFL, oct. 1985
"Traitement en temps réel",

RECONNAISSANCE DU LOCUTEUR ET DE MOTS ISOLES PAR DES SYSTEMES MINIATURISES: UNE COMPARAISON

H. Hügli et A. Mokeddem

INSTITUT DE MICROTECHNIQUE DE L'UNIVERSITE, CH-2000 Neuchâtel, Suisse

RESUME

La reconnaissance du locuteur et celle de mots isolés s'effectue pratiquement selon des méthodes et au moyen de systèmes très semblables.
Ce papier effectue une comparaison des méthodes et des performances de ces deux types de reconnaissance dans le cas de systèmes miniaturisés, ceci dans une optique d'intégration des deux fonctions dans un même système.
Sur la base d'expériences de reconnaissance avec un système à analyse spectrale, on mesure l'effet sur les performances quand on fait varier les facteurs de design principaux que sont la capacité mémoire et les besoins en calcul de comparaison de mots.
La comparaison fournit des règles de design. En particulier, elle fait apparaître la vérification et l'identification du locuteur comme une opération sensiblement plus critique que celle de reconnaître des mots isolés.

ABSTRACT

As speaker recognition and isolated word recognition are practically implemented on very similar systems, we compare in this paper the performances of both types of recognition when changing the methods and parameters of processing. This comparison is made in the context of small e.g. single chip recognition systems. We address the question whether and how both types of recognition can be brought on a same recognition system.
Results are given based on recognition tests performed using a recognizer with spectral analyser. Main design factors affecting memory and computing time are analysed.
This comparison gives design guidelines. Speaker recognition shows more constraining for the recognition system than isolated word recognition.

INTRODUCTION

Beaucoup d'études sont disponibles en reconnaissance de mots isolés et du locuteur, la plupart se basant sur la prédiction linéaire ou l'analyse spectrale de la parole et effectuant la comparaison de mots au moyen de la programmation dynamique (DTW).
Récemment, des systèmes de reconnaissance de mots isolés intégrés ont été proposés et réalisés. Il existe des contraintes importantes dans de tels systèmes, surtout si l'on vise un système monopuce: contraintes de capacité mémoire et de calcul [4].
Il existe deux motivations pour envisager ensemble la reconnaissance du locuteur et celle de la parole. Celle de réaliser ces deux fonctions avec un même circuit soit alternativement dans le temps, soit en même temps [2].
Dans la conception de systèmes de reconnaissance, les reconnaissances de la

parole et du locuteur ont la plupart du temps été considérées de manière séparée. Ce papier cherche à les comparer dans des mêmes conditions, ceci dans l'optique de trouver des règles de design de système tenant compte des exigences des deux domaines.

SYSTEME DE RECONNAISSANCE

La figure 1 illustre le système de reconnaissance utilisé pour les tests. Ce système à analyseur spectral est décrit dans [5] et est en outre très voisin du système décrit dans [1]. Rappelons quelques caractéristiques.

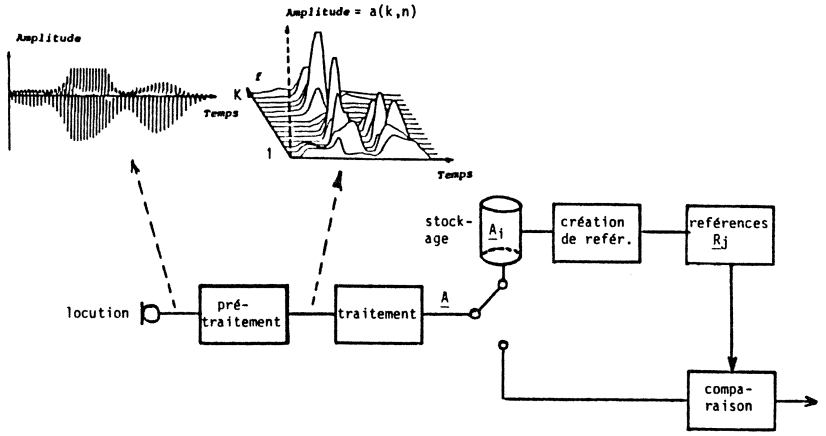

Fig. 1 Système de reconnaissance de mots et du locuteur

Analyse

Le prétraitement consiste en un étage de préaccentuation, suivi d'un analyseur spectral à K=14 canaux répartis de manière classique selon les bandes critiques de l'oreille, puis d'un circuit de détection de début et de fin de mot.

Le traitement comporte trois phases. Premièrement, la normalisation temporelle est une normalisation du spectrogramme [a(k,n)], n= 1..N à une longueur temporelle fixe N. Deux méthodes de normalisation temporelle sont en général utilisées: la normalisation temporelle linéaire et la normalisation temporelle curviligne [3]. La différence est que la première s'effectue dans l'espace temps alors que la deuxième se déroule dans l'hyperespace des composantes k= 1..K de a(k,n).
Deuxième phase, la normalisation d'amplitude s'effectue séparément pour chaque temps n. En définissant le signal moyen au temps n:

$$\bar{a}_n = \frac{1}{K} \sum_{k=1}^{K} a(k,n) \qquad (1)$$

on peut écrire la normalisation d'amplitude par l'équation:

$$a'(k,n) = a(k,n)/\bar{a}_n \qquad (2)$$

Troisième phase, la quantification par laquelle on quantifie le domaine borné [0..1] en 2^b niveaux discrets. Par:

$$a''(k,n) = s*a(k,n)/\bar{a}_n \qquad (3)$$

où le scalaire s est un bouton de réglage, on ajuste le domaine de a' sur le domaine [0..1] significatif pour la quantification. Nous verrons dans les résultats l'influence de s.

Comparaison

La comparaison entre les matrices test $\underline{A} = [a''(k,n)]$ et de référence $\underline{R} = [r(k,n)]$ de dimension K*N utilise la distance de Chebyshev entre deux colonnes de $\underline{A}$ et $\underline{R}$:

$$d(m,n) = \sum_{k=1}^{K} \left| a''(k,m) - r(k,n) \right| \qquad (4)$$

et la distance entre les deux matrices recalées temporellement:

$$D(\underline{A},\underline{R}) = \min_{(\text{chemin})} \sum_{n=1}^{N} d(m,n) \qquad (5)$$

Le recalage entre $\underline{A}$ et $\underline{R}$ peut se faire par décalage constant entre m et n, avec $|m-n| =< R$, et l'on parle alors de comparaison linéaire. R est l'excursion maximale du décalage [5].
Il peut aussi se faire de manière dynamique en considérant tous les décalages variables entre m et n situés dans un domaine d'excursion [-R..R]. On parle alors de comparaison dynamique (DTW) [5].

Création de références

Nous considérons une seule référence par locution de manière à limiter les besoin en capacité mémoire. Une référence $\underline{R}$ est créée à partir de trois locutions $\underline{A}_1$, $\underline{A}_2$, $\underline{A}_3$ par moyennage:

$$\underline{R} = \sum_{i=1}^{3} \underline{A}i \qquad (6)$$

Une autre technique consiste à aligner temporellement sur une même locution, toutes les locutions $\underline{A}i$ avant d'effectuer le moyennage. Elle utilise la comparaison dynamique pour trouver l'alignement optimal et est appelée création de références avec alignement dynamique.

FACTEURS DE DESIGN

Pour un système de reconnaissance miniaturisé, les facteurs de design les plus importants sont la capacité mémoire et les besoins en calcul, notamment le besoin en calcul pour effectuer la comparaison.
La capacité mémoire requise est directement proportionnelle au nombre de références $\underline{R}$, à la taille K*N de chaque référence et à b, le nombre de bits utilisé pour coder chaque r(k,n). Des nombres qu'il faudra maintenir petit.

Les besoins en calcul concernent surtout la comparaison dynamique (DTW)

puisque c'est le calcul le plus lourd de toute la reconnaissance. Une comparaison linéaire est moins lourde.
Notons que le calcul de la comparaison par recalage, aussi bien linéaire que dynamique, peut être réduit substantiellement si b=1, c'est-à-dire si $\underline{A}$ et $\underline{R}$ sont binaires, car le calcul de la distance d(m,n) devient alors une opération simple [6].

EXPERIENCES ET RESULTATS

Nos expériences auront donc essentiellement pour but de mesurer les performances de reconnaissance de mots et du locuteur en fonction de ces facteurs importants et surtout de déceler, au-delà des performances absolues, les différences entre ces deux types de reconnaissance.
Nous chercherons bien sûr aussi à mettre en évidence tous les autres facteurs importants.

Tests

Les tests sont des test de reconnaissance de mots ou du locuteur effectués selon les caractéristiques des tables 1 et 2. Notons que la reconnaissance du locuteur comprend deux formes distinctes, la vérification et l'identification du locuteur.

RECONNAISSANCE DE MOTS		
Locutions:	zéro	en avant
	un	en arrière
	deux	terminer
	trois	
	quatre	
	cinq	
	six	
	sept	
	huit	
	neuf	
Locuteurs:	5 hommes et 5 femmes	
Nbre de tests:	390	

Table 1: tests de reconnaissance de mots

RECONNAISSANCE DU LOCUTEUR	
Locutions:	/1 vous arrivez plus tôt demain
	/2 les petits oiseaux
	/3 tigre
	/4 rhinocéros
	/5 éléphant
	/6 kangourou
	/7 pélican
Locuteurs:	15 hommes et 4 femmes
Nbre. de tests:	en moyenne 350

Table 2: tests de reconnaissance du locuteur

Réduction des données

Les figures 2, 3 et 4 illustrent les performances de reconnaissance de mots et de locuteur quand on varie b, le nombre de bits utilisé pour coder $\underline{A}$ et $\underline{R}$.

Considérons d'abord la figure 2, la reconnaissance de mots avec la normalisation temporelle linéaire. Le taux d'erreur est pratiquement constant quand b dépasse 4 bits, c'est-à-dire que 4 bits suffisent. Observons maintenant la forte variation du taux d'erreur aux environs de 1 et de 2 bits, variation liée à s, le bouton d'ajustement du domaine d'amplitude. On voit l'importance de cet ajustement de domaine qui permet d'obtenir des performances de reconnaissance de mots aussi bonnes avec 1 bit que pour plusieurs bits.

Les figures 2 et 3 permettent de comparer les performances de reconnaissance selon les méthodes de normalisation temporelle linéaire et curviligne. Elles montrent l'avantage de la normalisation curviligne pour la reconnaissance de mots. Notons qu'il est largement fait usage de cet avantage dans la pratique[3].

En substance, la reconnaissance de mots utilise avantageusement une normalisation temporelle curviligne et peut profiter de l'ajustement de domaine pour quantifier les locutions à 1 bit.

En comparaison, les tests en reconnaissance du locuteur montrent que l'on ne peut profiter ni de l'un ni de l'autre effet. En effet, nous n'avons mesuré aucune modification mesurable des performances quand on passe de la normalisation temporelle linéaire à la curviligne.

La figure 4 montre en outre les variations de performances en reconnaissance du locuteur quand on fait varier le nombre de bit b. Ici, même dans le cas d'un ajustement de s optimal, il subsiste une forte dépendance entre taux d'erreur et b, l'erreur augmentant sensiblement pour b devenant petit. Ces résultats indiquent un choix nécessaire de $b \geq 4$.

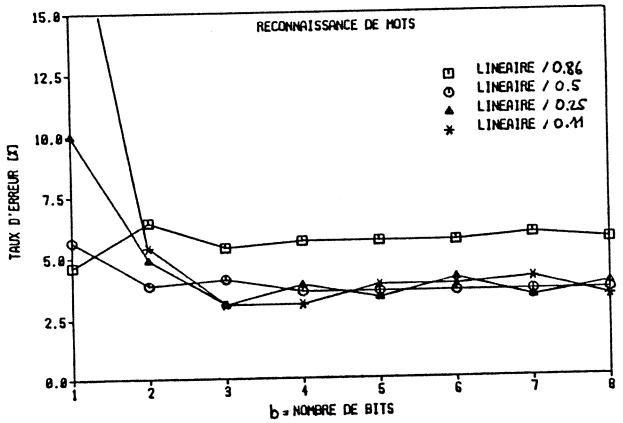

Fig. 2 Reconnaissance de mots avec normalisation temporelle linéaire. Effets de la quantification (1..8 bits) et de l'ajustement de l'amplitude (s = 0,86 .. 0,11); avec K*N = 14*20

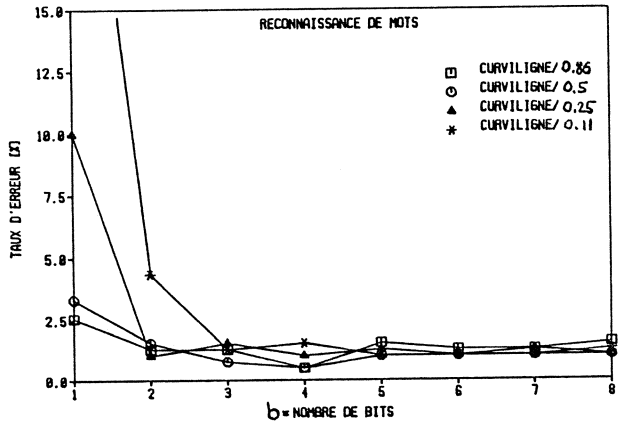

Fig. 3 Reconnaissance de mots avec normalisation temporelle curviligne. Effets de la quatification (1..8 bits) et de l'ajustement de l'amplitude (s = 0,86 .. 0,11); avec K*N = 14*20

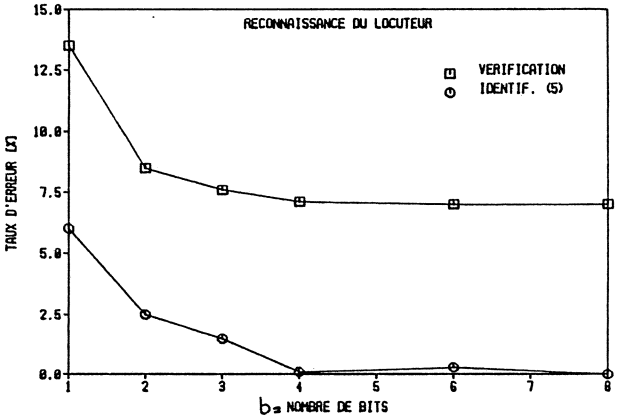

Fig. 4 Reconnaissance du locuteur. Effets de la quantification (1..8 bits) en vérification et en identification (5 locuteurs); avec K*N = 14*20

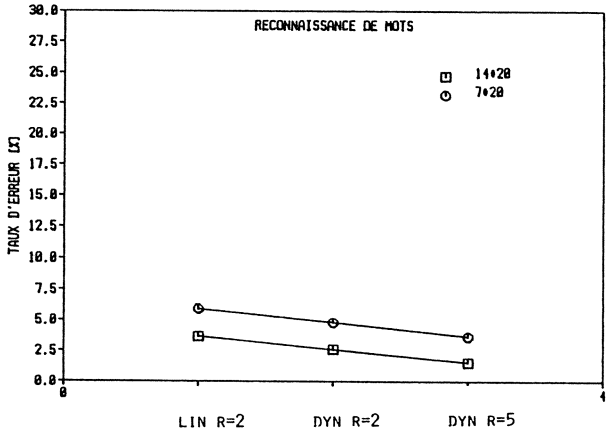

Fig. 5 Reconnaissance de mots avec des tableaux K*N de 14*20 et 7*20:
- comparaison linéaire, R=2
- comparaison dynamique, R=2
- comparaison dynamique, R=5

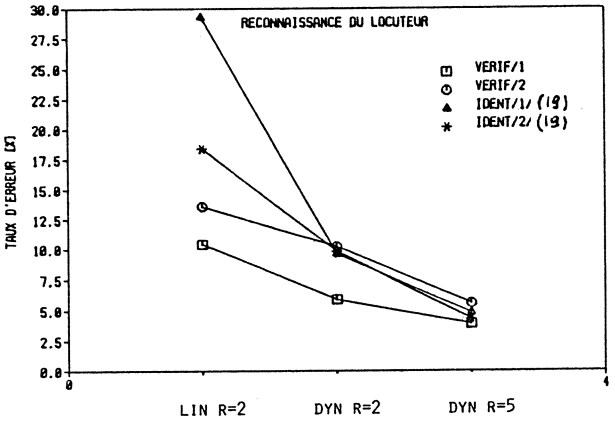

Fig. 6 Reconnaissance du locuteur avec les locutions /1 et /2, vérification et identification (19 locuteurs):
- comparaison linéaire, R=2
- comparaison dynamique, R=2
- comparaison dynamique, R=5

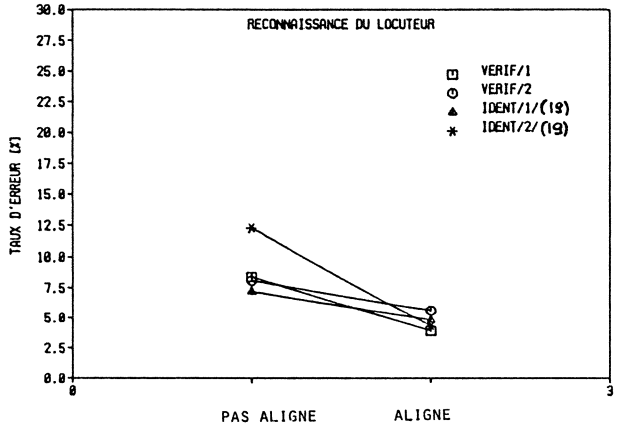

Fig. 7 Reconnaissance du locuteur avec les locutions /1 et /2, vérification et identification (19 locuteurs). Création des références par moyennage de trois locutions:
- sans alignement
- avec alignement dynamique

Méthodes de comparaison

Les figures 5 et 6 comparent les performances de reconnaissance de mots et du locuteur pour trois méthodes de comparaison: la comparaison linéaire avec une excursion maximale de R=2 et la comparaison dynamique avec R=2 puis R=5. Beaucoup plus qu'avec la reconnaissance de mots, on assiste avec la reconnaissance du locuteur à une dégradation importance des performances quand on renonce à la DTW. Notons que cette dégradation est encore plus marquée en identification qu'en vérification.
Nous pouvons conclure que la DTW est souhaitable pour reconnaître les mots et qu'elle est nécessaire pour reconnaître les locuteurs.

Création des références

La figure 7 met en relief les différences de performances quand les références sont créées par moyennage des trois Ai ou par moyennage des trois Ai alignés dynamiquement au préalable.
Ceci est donc une autre caractéristique de la reconnaissance du locuteur, cette différence n'ayant pas d'effet mesurable en reconnaissance de mots.

En substance on constate en reconnaissance du locuteur une grande sensibilité à des variations des signaux, alors que cette sensibilité est bien moins grande en reconnaissance de mots.
Au-dela de l'expérience, on peut donner une explication de plausibilité: les différences inter-mots donnent lieu à des fortes variations d'amplitude alors que les différences inter-locuteurs s'appliquent à des mêmes locutions et provoquent donc des petites différences d'amplitude. Variations importantes et quantification grossière dans un cas, variations subtiles et quantification fine dans l'autre cas.

CONCLUSIONS

Nous avons mesuré les performances en reconnaissance de mots et en reconnaissance du locuteur d'un système à analyse spectrale de la parole dans des tests de reconnaissance typiques pour des systèmes miniaturisables. Partout, la reconnaissance du locuteur est plus exigeante que la reconnaissance de mots isolés. En particulier, la reconnaissance du locuteur est plus exigeante pour le système à cause de sa grande sensibilité à: 1) une quantification à peu de bits, 2) à des erreurs de décalage se manifestant en reconnaissance ou lors de la création de références.
Lors de réalisations pratiques, il faut s'attendre, pour la reconnaissance du locuteur, à une rapide dégradation des performances quand on renonce à la programmation dynamique DTW ou qu'on utilise moins de 4 bits pour coder les références.

REMERCIEMENTS

Ce travail a été réalisé dans le cadre du projet CERS 1158 de la Commission pour l'Encouragement des Recherches Scientifiques soutenu par les maisons suivantes: ASULAB S.A., AUTOPHON A.G., CEH S.A., CIR S.A., HASLER A.G. et METTLER A.G.

REFERENCES:

[1] Billi R. & al., "Performance Analysis of Speaker-Trained Isolated Word Recognition System", Int. Zurich Seminar 1982

[2] Rosenberg A.E. & Shipley K.L., "Speaker Identification and Verification Combined with Speaker Independant Word Recognition", Proc. ICASSP 81

[3] Gauvain J.L, Mariani J. & Lienard J.S., "On the Use of Time Compression for Word-Based Recognition", Proc. ICASSP 83

[4] Bui N. & al., "An Integrated Voice Recognition System", IEEE Trans. on ASSP, Vol. ASSP-31, Feb. 1983

[5] Mokeddem A., Hugli H. & Pellandini, "Evaluation of Criterion Based Clustering Procedures for Generating Multiple Reference Templates in SISR", Seventh Int. Conf. on Pattern Recognition, Montreal, Aug. 1984

[6] Dijkstra E. & Piguet C., "An orthogonal array of systolic processors for Dynamic Time Warping", Proc. Journées d'électronique 85