Adaptation au locuteur de systèmes de reconnaissance. Régression linéaire multiple et perceptrons multicouches

Speaker adaptation for speech recognition systems. Multiple linear regression and multilayer perceptrons

J. P. TUBACH
TÉLÉCOM Paris, Département Signal
(CNRS, URA 820)
46, rue Barrault, 75013 PARIS

Jean-Pierre TUBACH est né le 15 septembre 1942, à Paris. Ingénieur Civil des Mines de Paris (1964), il obtient le doctorat ès Sciences à Grenoble en 1970, avec la première thèse française en reconnaissance de la Parole.

Entré à TÉLÉCOM Paris 1983, il a dirigé de 1986 à 1989 son équipe « Reconnaissance des Formes et Traitement de la Parole ». Jean-Pierre Tubach est actuellement Adjoint Directeur Scientifique de TÉLÉCOM Paris, responsable du programme doctoral.

G. CHOLLET
TÉLÉCOM Paris, Département Signal
(CNRS, URA 820)
46, rue Barrault, 75013 PARIS

Gérard CHOLLET est né à Paris le 13 juillet 1947. Après des études à l'Université Paris VI (Maîtrise de Physique et DEA d'Informatique), puis à Santa Barbara (Ph. D. de l'Université de Californie), il a été Professeur à l'Université de Floride jusqu'en 1978. Il est chercheur CNRS depuis cette date. De 1981 à 1983, il a dirigé l'équipe « Parole » de CIT-Alcatel, puis a rejoint l'URA-820 à TÉLÉCOM Paris dès sa création. Il dirige actuellement l'équipe « Parole » de cette URA.

K. CHOUKRI*
TÉLÉCOM Paris, Département Signal
(CNRS, URA 820)
46, rue Barrault, 75013 PARIS
* Maintenant à Cap Sogeti Innovation.

K. CHOUKRI est né à Karia Arekmane, Nador (Maroc) en 1960. Il est titulaire d'un diplôme d'ingénieur de l'École Nationale de l'Aviation civile (ENAC), d'un master de l'École Nationale Supérieure des Télécommunications (ENST) et a soutenu une thèse de doctorat à l'ENST effectuée en collaboration avec les laboratoires de Marcoussis (CGE). Il a rejoint CAP GEMINI INNOVATION où il est responsable de projets de recherche en communication orale homme-machine (notamment dans le cadre du programme européen ESPRIT).

C. MONTACIE
TÉLÉCOM Paris, Département Signal
(CNRS, URA 820)
46, rue Barrault, 75013 PARIS

C. MONTACIE est né à Nice le 6 août 1961. Il a reçu le diplôme d'Ingénieur SUPELEC en 1986. Depuis 1986, il travaille sur sa thèse de doctorat au département Signal de l'École Nationale Supérieure des Télécommunications. Ses domaines de recherche concernent la reconnaissance de la parole et plus particulièrement la décomposition temporelle de l'évolution spectrale et les transformations non linéaires par des techniques connexionnistes.

Adaptation au locuteur de systèmes de reconnaissance

C. MOKBEL
TÉLÉCOM Paris, Département Signal
(CNRS, URA 820)
46, rue Barrault, 75013 PARIS

Chafic MOKBEL est né à Beyrouth en 1966. Ingénieur, il est diplômé de l'Université Libanaise, Faculté de Génie (Roumieh), département d'électronique et d'électricité, en 1988. Titulaire d'un DEA systèmes électroniques de l'INP de Grenoble, ENSERG, en 1989, il est actuellement, inscrit en thèse à TÉLÉCOM Paris, département Signal.

H. VALBRET
TÉLÉCOM Paris, Département Signal
(CNRS, URA 820)
46, rue Barrault, 75013 PARIS

Hélène VALBRET est née le 13 juillet 1966 à Paris. Elle a obtenu le diplôme d'Ingénieur de TÉLÉCOM Paris en 1989, et commence sa seconde année de thèse au département Signal de cette École.

RÉSUMÉ

La variabilité interlocuteur est une source majeure d'erreurs en reconnaissance automatique de la parole (RAP). Cet article décrit une série d'expériences, menées par l'Équipe « Reconnaissance des Formes et Traitement de la Parole » de TÉLÉCOM Paris, dans le but de contrôler certains aspects de cette variabilité, et permettre ainsi une adaptation au locuteur des systèmes actuels de reconnaissance de parole.
Les premières expériences utilisent une technique linéaire empruntée à l'analyse des données, la régression linéaire multiple.
Les secondes font appel aux perceptrons multicouches, et fournissent des résultats légèrement meilleurs, grâce à la prise en compte de phénomènes non linéaires.
L'amélioration des taux de reconnaissance obtenue est, en moyenne, de 16 % pour les secondes, contre 15 % pour les premières.
Ces techniques peuvent également être utilisées pour l'adaptation des reconnaisseurs à de nouveaux environnements acoustiques ou conditions de prise de son.

MOTS CLÉS

Reconnaissance de mots ; adaptation au locuteur ; transformations spectrales ; régression linéaire multiple ; perceptrons multicouches, évaluation.

SUMMARY

Interspeaker variability is a major source of errors in automatic speech recognition. This paper describes a series of experiments, conducted at TELECOM Paris by the « Pattern Recognition and Speech Processing » Group, for controlling some aspects of this variability, thus allowing for the adaptation of speech recognition systems to new users.
The first experiments are based on a linear data analysis technique : multiple linear regression (MLR).
The second set uses multilayer perceptrons, and yields slightly better results, because non linear phenomena are taken into account.
The average improvement of recognition scores is 16 % with the second approach, versus 15 % with the first one.
Those techniques can also be used for the adaptation of recognizers to new acoustical environments and recording conditions.

KEY WORDS

Word recognition ; speaker adaptation ; spectral transforms ; multiple linear regression ; multilayer perceptron ; assessment.

1. Introduction

La variabilité interlocuteur se manifeste de façon fort évidente sur les représentations temps × fréquence des sons de la parole. Cette variabilité est principalement due à des différences fonctionnelles et anatomiques entre locuteurs. Elle comporte des aspects temporels et des aspects fréquentiels. Nous ne traitons ici que des aspects fréquentiels, car on peut estimer que les premiers sont assez bien pris en compte par les algorithmes utilisés le plus souvent dans les reconnaisseurs actuels.

Nous rappelons d'abord brièvement le principe de fonctionnement d'un système très conventionnel de reconnaissance de mots isolés ou enchaînés, et plusieurs approches possibles pour parvenir à un fonctionnement correct en mode multilocuteur.

Il apparaît alors intéressant de définir des transformations entre espaces spectraux de locuteurs, pouvant être utilisées dans un but de normalisation ou/et d'adaptation. Des techniques linéaires, et d'autres, non linéaires, sont disponibles pour réaliser cette opération.

La régression linéaire multiple est une approche linéaire. L'utilisation de ce type de méthodes dans ce domaine est

apparue, à TÉLÉCOM Paris, dès 1977 (Grenier [1]). Elle transforme tous les spectres de la même façon, utilisant un seul opérateur de transformation. Or, même s'il s'agit de reconnaissance « globale » de mots, les mots du vocabulaire sont composés de sons du langage de caractéristiques différentes (voyelles, consonnes fricatives, liquides, occlusives, etc.) et une transformation spécifique non seulement du locuteur, mais aussi de la classe de son serait a priori plus judicieuse.

Des transformations différentes pour différentes classes de segments (voisés/non voisés par exemple) teneraient en fait de réaliser une approximation « linéaire par morceaux » de la transformation optimale. Cette approche nécessite malheureusement une détermination explicite et fiable des types de segments, ce qui rend sa mise en œuvre problématique.

C'est pourquoi nous avons expérimenté l'utilisation de perceptrons multicouches pour réaliser simplement une transformation spectrale non linéaire apprise.

Pour procéder à une évaluation comparative de ces deux méthodes, nous avons utilisé un sous-ensemble de la base de données de mots isolés de Texas Instruments (TI), et le système de référence que nous avons rendu disponible sur la station de travail d'évaluation développée dans le cadre du projet Esprit SAM.

2. Reconnaissance monolocuteur, multilocuteur et adaptation au locuteur : état de l'art

2.1. RECONNAISSANCE MONOLOCUTEUR ET MULTILOCUTEUR

Le système le plus classique de reconnaissance de mots isolés ou connectés, dont de nombreuses versions ont été développées et commercialisées depuis le début des années 70, peut être brièvement décrit, à titre de rappel, de la façon suivante.

Lors d'une phase d'apprentissage, un locuteur prononce une ou plusieurs fois les mots constituant le vocabulaire à reconnaître. Les résultats d'une analyse de ces mots dans le domaine spectral constituent un dictionnaire de références. En phase de reconnaissance, un mot inconnu est comparé à ces références, et identifié à celle dont il est le plus proche, au sens d'une « distance » (dissimilarité) définie entre de telles formes.

Une distance locale entre échantillons spectraux étant définie, l'algorithme de programmation dynamique (Dynamic Time Warping, DTW) (Vintsuk [2]) permet le calcul d'une mesure de dissimilarité globale en cherchant le chemin optimal mettant en correspondance les vecteurs de paramètres représentant les deux formes acoustiques « mot test » et « mot référence ». (Pour une bibliographie plus complète sur ce sujet, nous renvoyons au chapitre XVIII (techniques de reconnaissance globale) de (Calliope [3]).)

Un schéma de fonctionnement est donné par :

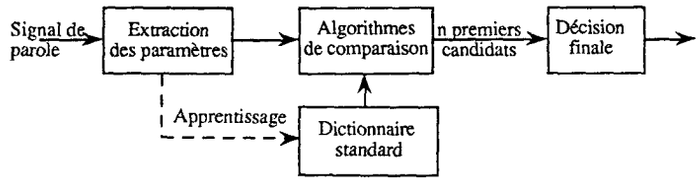

Figure 1. — Principe d'un système de reconnaissance globale à base de « DTW ».

Un fonctionnement satisfaisant est obtenu si le mot inconnu provient du locuteur qui a servi à l'apprentissage. Une dégradation apparaît dès qu'il y a un nouveau locuteur.

Pour parvenir à des performances plus satisfaisantes en multilocuteur, plusieurs approches ont été considérées. La plus classique se fonde sur des références multiples : un dictionnaire de références est construit à partir de nombreuses élocutions de chaque mot, provenant de locuteurs différents, supposés constituer un échantillon représentatif de l'ensemble des locuteurs. Des techniques de « clustering » et d'analyse discriminante (Rabiner *et al.* [4]) sont utilisées pour obtenir ce dictionnaire.

La modélisation stochastique, introduite par J. Baker et F. Jelinek au milieu des années 70 constitue une alternative remarquable, et est le plus fréquemment utilisée sous la forme des Modèles de Markov Cachés (HMMs) (Levinson [5]). Programmation dynamique et modélisation Markovienne réalisent en fait des opérations plus proches que ne le laisserait a priori penser la dissimilarité des formalismes employés (Juang [6]).

L'enregistrement et le codage des élocutions, la détermination des références multilocuteur ou l'apprentissage des modèles de Markov sont des processus très coûteux en temps de calcul.

La quantification vectorielle constitue une autre approche statistique. Elle consiste à construire un livre de codes quantifiant le domaine de représentation. Ceci s'effectue en appliquant des algorithmes de quantification sur l'ensemble des données d'apprentissage. Statistiquement ces données doivent être variées et nombreuses pour représenter l'ensemble de classes possibles (Burton *et al.* [7]).

Une approche radicalement différente consiste à rechercher des traits invariants, au niveau acoustique, avec référence aux mécanismes de production et de perception de la parole, et à les utiliser pour la représentation du vocabulaire et sa reconnaissance (Stevens *et al.* [8]). Bien que très séduisante, cette méthode n'a pas encore donné de résultats opérationnels dans le domaine qui nous intéresse, et nécessitera sans doute encore un important effort de recherche.

Une méthode différente pour tenter de maîtriser les aspects spectraux de la variabilité interlocuteur est celle qui fait l'objet de cet article : l'adaptation au locuteur. Elle s'inspire du comportement humain, dans la mesure où il n'est pas déraisonnable de considérer qu'un auditeur, percevant la parole d'un locuteur inconnu, utilise le début de leur dialogue pour s'adapter à cette voix nouvelle.

Les techniques d'adaptation sont liées en quelque sorte aux systèmes de reconnaissance correspondants. La modélisation Markovienne se prête bien à l'apprentissage multilocuteur, mais par contre les systèmes correspondants sont difficilement adaptables à des nouveaux locuteurs. L'adaptation utilisant la quantification vectorielle se fait en mettant en correspondance le nouveau livre de codes et celui servant comme référence (Shikano [9]). Cette adaptation demande beaucoup de données pour construire un livre de codes assez robuste dans l'espace de représentation du nouveau locuteur. Dans la suite nous décrivons différentes techniques d'adaptation au locuteur pour les systèmes de reconnaissance à base de DTW.

2.2. PRINCIPES DE L'ADAPTATION AU LOCUTEUR

Avant une définition plus formelle au paragraphe suivant, on peut donner de façon intuitive les principes de l'adaptation au locuteur des systèmes de reconnaissance.

Sur les formes temps × fréquence que sont les références (définies au paragraphe précédent) et les mots à reconnaître, on peut envisager plusieurs opérations.

Dans une approche dite de *normalisation*, les paramètres d'entrée subissent, lors de la reconnaissance, une transformation, définie lors d'un bref apprentissage, de façon à ce que les mots prononcés deviennent aussi proches que possible de ceux du dictionnaire d'un locuteur standard.

Le schéma correspondant est donné ci-dessous :

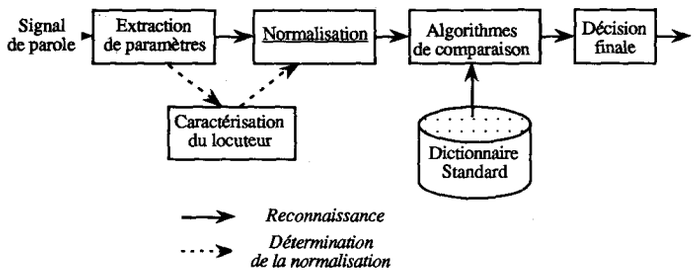

Figure 2. — Reconnaissance multilocuteur par normalisation du locuteur.

La seconde approche (*adaptation*) consiste à utiliser une procédure qui modifie les références du locuteur standard pour qu'elles soient adaptées au nouveau locuteur. Le schéma est le suivant :

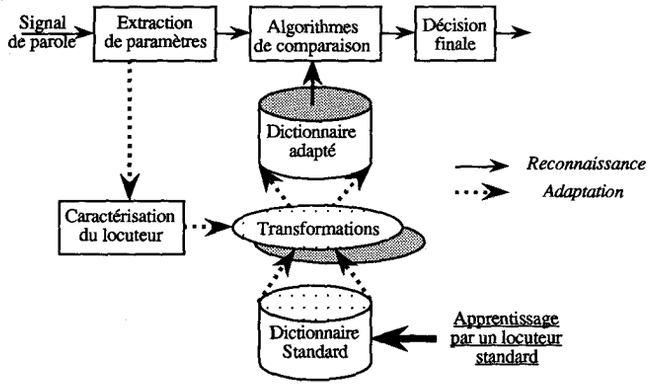

Figure 3. — Principe général de l'adaptation.

Normalisation et adaptation proprement dite sont susceptibles de fournir les mêmes résultats. Mais l'adaptation apparaît comme préférable, puisqu'elle n'impose pas de calculs supplémentaires en phase de reconnaissance, en utilisation statique tout au moins.

2.3. STRATÉGIE DYNAMIQUE

Un fonctionnement dynamique signifie que l'adaptation peut être effectuée interactivement pendant une session de reconnaissance. L'utilisateur fournit alors une réaction lorsqu'une erreur de reconnaissance survient. L'adaptation est alors effectuée sous la supervision de l'utilisateur et sur sa décision. Elle consiste à mettre à jour les opérateurs de transformation utilisés par la procédure de transformation d'espace spectral.

Un mode plus avancé consiste à ajouter à ce qui précède une adaptation systématique, dans les cas non signalés en erreur. Elle est apparue récemment dans le système Dragon Dictate 30000 (Baker [10]).

3. Transformations spectrales

3.1. FORMULATION GÉNÉRALE

Sous la forme la plus générale, les transformations pouvant être appliquées aux productions d'un nouveau locuteur et d'un locuteur de référence sont décrites par la figure suivante :

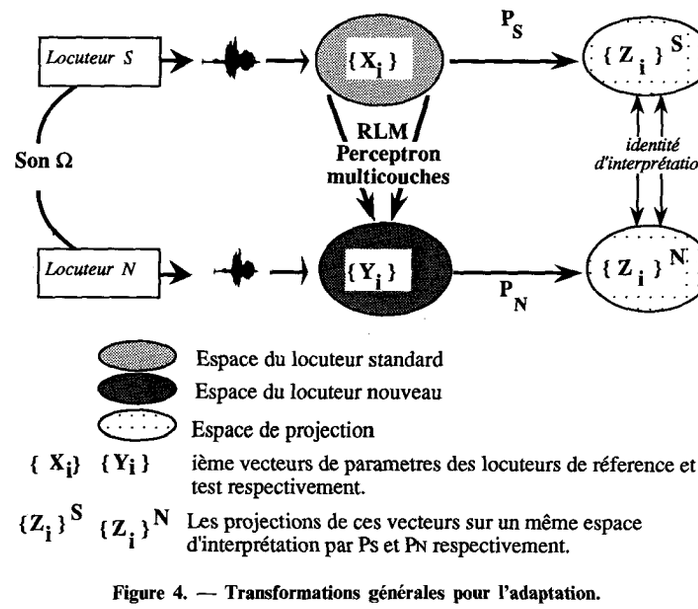

Figure 4. — Transformations générales pour l'adaptation.

Les transformations P_S et P_N doivent être telles que l'erreur quadratique moyenne $E(e_i^2)$ soit minimale pour un ensemble représentatif de sons. Formellement :

$$P_N(Y_i) = P_S(X_i) + e_i .$$

Il faut remarquer que la transformation P_S peut être utilisée pour adapter tous les spectres de référence, à la fin de la phase d'adaptation. La transformation P_N doit être appliquée, en tant que procédure de normalisation, pendant la phase de reconnaissance.

Considérons successivement le cas linéaire et le cas non linéaire.

3.2. CAS LINÉAIRE

Si l'on revient à la figure précédente, il paraît naturel de rechercher un opérateur qui permette une transformation directe des spectres du locuteur standard en ceux du locuteur courant. Alors, P_N se réduit à l'identité, et P_S, linéaire, est la régression linéaire multiple (RLM) qui fait coïncider les spectres standard et les nouveaux spectres. On dit que l'espace du nouveau locuteur est l'espace de projection.

Lorsqu'un nouveau locuteur veut utiliser le système, il prononce un nombre limité de mots prédéfinis. Nous supposerons que ces mots sont représentatifs de l'espace fréquentiel décrit acoustiquement par ce locuteur. Par DTW ces mots sont alignés avec leurs équivalents du vocabulaire standard. Nous obtenons ainsi un ensemble de vecteurs « spectraux » en correspondance. Soient Y_i et X_i ces vecteurs des locuteurs nouveau et standard respectivement. L'adaptation revient à superposer ces vecteurs. La RLM consiste à chercher une matrice constante P qui minimise $E[\| Y_i - P\{X_i\} \|^2]$, où $P\{X_i\}$ est le vecteur X_i adapté et E l'espérance mathématique.

Une partie importante de cette erreur est due au fait que les distributions de ces vecteurs ne sont pas centrées et leurs variances ne sont pas normalisées (*fig.* 5). Ainsi une procédure de centrage et de normalisation est effectuée pour obtenir les vecteurs $\{x_i\}$ et $\{y_i\}$.

La détermination de P se fait en multipliant l'ensemble des vecteurs $\{y_i\}$ par la pseudo-inverse de $\{x_i\}$ calculée par une décomposition en valeurs singulières (solution se basant sur une minimisation de l'erreur quadratique moyenne).

$$P = \{y_i\}\,\{x_i\}^{\dagger}.$$

Ainsi la transformation spectrale sera déterminée et tout le dictionnaire de référence (standard) sera centré, normalisé et transformé par la matrice P. Pendant la reconnaissance, une phase de centrage et de normalisation est nécessaire.

Cette technique sera utilisée comme référence dans cette étude.

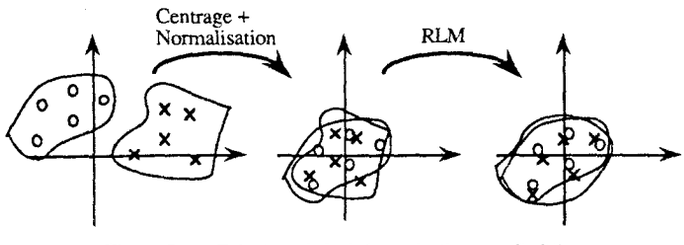

Figure 5. — Principe général de projection par RLM.

Pour tenter d'améliorer les résultats ainsi obtenus, nous avons alors recherché une approche qui permette de prendre en compte des phénomènes non linéaires.

Plusieurs méthodes permettent de déterminer une fonction non linéaire F : filtrage polynômial (Duvaut [11]), décomposition en série de Volterra (Kohn [12]) etc. L'utilisation de perceptrons multicouches constitue une alternative simple, et une initialisation linéaire est possible. Ils permettent l'extraction implicite d'une grande quantité d'informations sur la variabilité entre les deux locuteurs en se basant sur un nombre limité de données.

3.3. TRANSFORMATIONS SPECTRALES NON LINÉAIRES UTILISANT UN PERCEPTRON MULTICOUCHES

Les perceptrons multicouches (MLP) constituent un outil de calcul puissant, dont l'utilisation en traitement de la parole est récent (voir bibliographie dans Lippman [13]). Nous en supposons connus les principes.

Ils permettent l'apprentissage d'une vaste classe de fonctions analytiques. Un perceptron possédant au moins une couche cachée peut approximer n'importe quelle fonction continue. Ils peuvent être vus comme une extension de la régression linéaire multiple. En fait, un perceptron sans couche cachée et la régression linéaire multiple sont équivalents pour des données distribuées selon une loi normale.

L'ajout de niveau(x) caché(s) ne peut qu'améliorer les performances. En effet, dans le cas le plus défavorable où une fonction linéaire donne une meilleure représentation de la transformation cherchée qu'une sigmoïde, l'algorithme d'apprentissage basé sur la minimisation de l'erreur quadratique va obliger les nœuds du réseau à fonctionner dans la partie linéaire de la sigmoïde.

Dans le cadre de cette étude, nous avons testé un perceptron multicouches avec un seul niveau caché. Nous avions à évaluer 128 poids de connexions. Le corpus d'apprentissage étant constitué par un ensemble de 200 couples spectres d'entrée/spectres de sortie, il n'était pas réaliste de chercher à estimer plus de connexions.

Pour les raisons discutées dans le cas de la RLM, les paramètres d'entrée et de sortie ont été centrés et normalisés. Pour chiffrer l'apport des MLP par rapport à la RLM, nous avons calculé les déviations quadratiques entre les spectres transformés et les spectres visés. Ces déviations ont été moyennées et normalisées sous forme d'un pourcentage de déviation. C'est ce qu'on appelle erreur de reconstruction :

$$\text{Erreur de reconstruction} = \frac{E[|Y - \text{transf}(X)|^2]}{E[Y^2]}$$

où transf(X) représente la transformation des vecteurs de référence, et Y les vecteurs visés.

L'évaluation montre que l'erreur de reconstruction passe de 45 % (RLM) à 30 % avec les MLP, l'erreur initiale (c'est-à-dire sans adaptation) étant de 100 %.

4. Évaluation

4.1. DONNÉES ET RÉSULTATS

Nous avons choisi pour l'évaluation comparative de nous référer à l'algorithme de reconnaissance et à la base de données les plus répandus, de façon à faciliter d'autres comparaisons.

La base de données de mots isolés de TI fournit suffisamment d'élocutions pour nos expériences. Elle contient 20 mots anglais (y compris les chiffres), enregistrés 26 fois par chaque locuteur ; nous disposons de 14 locuteurs différents. La première élocution des 20 mots est prise comme référence monolocuteur, les tests portent sur les 25 répétitions.

Le système de reconnaissance de référence (Chollet *et al.* [14]), que nous avons rendu disponible sur la station de travail d'évaluation développée dans le cadre du projet Esprit SAM, est utilisé en mode mots isolés. L'analyse du signal de parole fournit 8 coefficients cepstraux en échelle Mel (MFCCs) toutes les 20 millisecondes. L'algorithme de comparaison aux références utilise la programmation dynamique (DTW).

L'adaptation est toujours faite sur cinq mots (one, five, two, six, no) de la première élocution de chaque locuteur. Ces mots d'adaptation sont alignés temporellement avec leurs homologues de référence par programmation dynamique. De cette façon, en moyenne 200 spectres sont appairés, et utilisés pour l'apprentissage du perceptron multicouches, ou la régression linéaire multiple.

Le perceptron multicouches comporte 8 nœuds dans son niveau caché. 100 cycles d'apprentissage sont nécessaires.

Les scores de reconnaissance sont présentés dans les tableaux suivants :

— taux de reconnaissance sans adaptation (pour référence),
— taux de reconnaissance après adaptation par régression linéaire multiple,
— taux de reconnaissance après adaptation par perceptron multicouches.

Dans tous les cas sont fournies des moyennes par rapport au critère masculin (M)/féminin (F), croisé ou non.

4.2. DISCUSSION

Il faut tout d'abord indiquer les intervalles de confiance (à 95 %) de ces observations. Pour des moyennes globales (91 000 reconnaissances de mots), il est de ± 0,25 %. Pour des taux moyens HH, HF, etc., il est de ± 0,5 %. Pour un couple de locuteurs donné, il est de ± 3,5 %.

On constate une bonne amélioration d'ensemble (en moyenne globale + 15 % avec la RLM et + 16 % avec le perceptron multicouches).

REF \ TEST		F							H						
		alk	cjp	dfg	gnl	jws	sas	sjn	grd	kab	msw	reh	rld	tbs	wmf
F	alk	99.2	73.5	79.6	87.1	79.4	76.3	82.7	38.7	59.8	67.5	63.1	58.1	50.6	53.1
	cjp	84.8	98.1	77.9	87.1	89.2	85.4	85.6	39.6	45.8	57.3	56.5	43.1	55.8	43.7
	dfg	86.5	85.4	96.5	81.2	79.6	86.9	78.5	47.3	74.8	76.5	75.0	58.3	48.5	65.0
	gnl	89.2	78.1	91.5	96.9	88.7	77.3	88.1	47.5	53.1	66.3	57.7	45.8	43.3	61.3
	jws	92.3	93.5	71.7	95.2	99.2	91.9	90.8	46.0	48.1	65.0	77.1	56.2	51.3	56.5
	sas	87.5	88.3	91.0	83.8	93.1	98.1	86.9	49.4	59.2	87.7	88.1	76.0	61.5	72.7
	sjn	91.5	83.1	72.9	92.1	85.8	81.2	92.1	37.9	45.8	62.5	53.5	41.7	49.2	45.8
H	grd	41.2	22.5	37.3	30.2	37.7	40.6	34.4	95.0	77.9	75.2	80.6	72.5	83.8	80.6
	kab	65.8	41.2	55.4	44.2	50.4	56.9	51.2	74.4	99.2	82.3	81.5	72.7	91.2	82.3
	msw	63.7	26.7	55.6	46.2	56.5	72.5	42.9	73.1	89.8	97.5	93.8	90.6	83.7	92.1
	reh	61.3	35.8	69.2	56.0	63.1	77.5	42.3	69.8	84.0	87.1	97.5	88.1	72.3	89.8
	rld	61.5	54.6	66.5	46.5	61.7	79.6	50.4	78.8	89.8	92.9	96.5	98.3	86.2	93.8
	tbs	32.7	37.3	26.5	18.1	32.3	42.9	33.3	86.5	87.3	78.7	81.9	80.2	98.3	83.3
	wmf	41.0	31.9	46.3	33.8	47.5	56.5	39.0	81.5	88.5	88.7	88.3	84.2	85.4	97.9

Taux de reconnaissance monolocuteur et interlocuteur sans adaptation

	F	H
F	85.1	56.8
H	47.3	83.8

moyenne du taux interlocuteur

	Moy
F	97.2
H	97.7

moyenne du taux monolocuteur

REF \ TEST		F							H						
		alk	cjp	dfg	gnl	jws	sas	sjn	grd	kab	msw	reh	rld	tbs	wmf
F	alk	---	82.7	78.5	88.7	86.9	85.8	70.6	80.8	89.0	75.0	87.1	80.2	86.9	86.0
	cjp	81.9	---	66.2	88.3	84.6	77.1	80.8	68.3	82.9	68.5	87.7	82.5	87.3	75.6
	dfg	89.2	70.6	---	74.6	83.1	81.9	78.1	73.1	90.6	84.8	73.3	90.8	78.3	75.6
	gnl	94.0	80.0	79.0	---	87.3	79.2	86.7	80.0	92.3	81.5	89.8	86.7	86.2	80.6
	jws	92.7	83.8	83.8	90.6	---	87.1	80.6	85.6	78.8	71.5	87.9	88.5	90.4	69.4
	sas	94.2	86.2	90.0	91.5	91.3	---	85.0	87.9	94.4	87.5	92.7	94.0	95.0	85.8
	sjn	87.9	75.0	79.0	91.0	84.8	87.9	---	78.3	88.1	76.3	87.9	88.1	88.7	70.0
H	grd	90.4	66.7	64.8	77.3	80.4	91.9	82.5	---	90.2	89.8	91.9	94.2	96.0	89.4
	kab	90.8	62.9	89.0	86.3	72.9	84.6	72.7	88.8	---	95.2	81.5	92.1	93.3	92.7
	msw	88.5	56.5	64.2	89.2	83.3	84.8	78.8	75.0	91.3	---	88.8	82.7	91.3	84.8
	reh	82.3	64.2	68.1	78.5	75.4	78.3	58.3	75.2	94.8	77.3	---	93.7	91.2	93.1
	rld	78.1	65.6	67.1	75.8	75.2	92.5	73.3	88.3	87.5	87.5	89.8	---	92.5	92.5
	tbs	79.6	76.5	63.1	90.8	76.9	88.1	78.8	94.0	91.0	88.7	94.6	95.6	---	89.0
	wmf	90.2	71.0	58.7	73.1	74.8	78.3	68.1	81.3	87.9	81.7	87.7	89.2	89.4	---

Taux de reconnaissance interlocuteur avec adaptation par regression lineaire multiple

	F	H
F	83.8	83.4
H	76.7	89.1

moyenne du taux interlocuteur

REF \ TEST		F							H						
		alk	cjp	dfg	gnl	jws	sas	sjn	grd	kab	msw	reh	rld	tbs	wmf
F	alk	---	83.7	79.4	87.5	83.8	90.2	83.3	83.8	89.6	82.9	88.5	83.3	88.3	89.4
	cjp	80.4	---	70.2	81.5	82.1	81.5	75.8	76.7	78.1	75.4	89.4	88.8	85.8	84.2
	dfg	88.8	71.9	---	76.5	79.0	84.8	77.5	74.2	91.2	87.9	79.0	89.6	84.0	78.3
	gnl	89.4	74.0	78.8	---	82.7	70.0	76.3	77.5	85.0	79.0	85.4	81.5	85.2	76.0
	jws	87.9	85.4	92.3	90.8	---	88.5	77.9	90.2	84.8	84.8	91.2	91.9	87.5	81.0
	sas	93.7	84.4	90.8	90.8	90.2	---	85.6	85.6	92.1	92.7	94.2	95.8	94.6	91.2
	sjn	89.2	76.0	80.6	91.7	85.0	88.3	---	81.9	87.7	79.0	87.7	89.2	85.4	81.3
H	grd	89.0	64.8	75.0	76.7	76.9	93.1	77.3	---	90.8	90.2	85.2	92.1	92.9	90.2
	kab	88.1	66.3	90.6	85.2	68.1	89.0	69.4	85.6	---	96.0	86.0	91.3	94.4	91.9
	msw	80.0	64.4	64.2	82.3	67.1	80.2	79.0	81.2	93.3	---	91.2	84.4	92.5	91.0
	reh	76.0	66.2	81.7	67.5	77.5	89.6	68.5	77.1	93.7	90.0	---	92.5	90.0	94.2
	rld	83.7	70.4	79.2	69.0	76.7	94.4	69.2	86.5	87.7	89.0	89.6	---	91.3	95.8
	tbs	81.7	70.4	76.9	89.2	75.4	94.2	82.3	93.3	91.7	90.4	95.2	96.7	---	90.2
	wmf	87.7	70.6	59.4	66.5	71.3	77.1	65.2	77.5	91.5	87.9	87.7	89.8	90.0	---

Taux de reconnaissance interlocuteur avec adaptation par perceptron multi-couches

	F	H
F	83.3	85.5
H	76.8	90.0

moyenne du taux interlocuteur

Dans ce cas d'adaptation homme-homme, l'amélioration est de l'ordre de 6 %, avec des variations (allant jusqu'à une légère dégradation pour quelques couples de locuteurs).

L'adaptation croisée homme-femme fournit les plus spectaculaires améliorations (presque + 30 % en moyenne). De façon un peu surprenante, dans le cas femme-femme, les performances tendent à baisser. Il s'agit très probablement d'une mauvaise représentation des voix à fondamental élevé par les coefficients MFCC utilisés pour l'analyse du signal.

Certains cas défavorables indiquent que les mots utilisés pour l'adaptation ne contiennent pas l'information convenable pour définir la transformation. L'utilisation du perceptron multicouches n'est alors évidemment pas susceptible d'améliorer le résultat.

L'étude détaillée de l'amélioration de la reconnaissance dans ces deux approches fait donc apparaître des comportements différents. Pour mieux les qualifier globalement, on utilise un rapport d'adaptation défini par

$$\rho = \frac{\tau_{sn}^{a} - \tau_{sn}}{\tau_{ss} - \tau_{sn}}$$

où τ_{sn} est le score de reconnaissance avec les références du locuteur S et les élocutions du locuteur N ; τ_{ss} est le score de reconnaissance monolocuteur (S et S) et τ_{sn}^{a} est le score de reconnaissance après adaptation.

La valeur de ρ est comprise entre 0 et 1, bornes incluses. Plus elle est proche de 1, meilleure est l'adaptation. Les valeurs correspondant à nos expériences sont :

— 50,3 % pour la régression linéaire multiple
— 52,7 % pour le perceptron multicouches

(ces chiffres mesurent la qualité de l'adaptation réalisée, par rapport aux 100 % que fournirait une adaptation idéale).

5. Conclusion, perspectives

Nous avons présenté une étude comparative de techniques permettant l'adaptation des références d'un système de reconnaissance monolocuteur pour un nouveau locuteur. On utilise des transformations d'espace spectral, linéaires et non linéaires, pour prendre en compte les caractéristiques propres du nouveau locuteur.

Dans un contexte de reconnaissance de mots isolés, l'utilisation d'un réseau neuromimétique (perceptron multicouches) se révèle légèrement plus efficace que la régression linéaire multiple. Il n'en reste pas moins vrai que la RLM, par son poids calculatoire nettement moindre, et le fait qu'elle peut se prêter mieux à une adaptation « en ligne », demeure une solution intéressante.

Ces méthodes peuvent être appliquées également à l'important problème de l'adaptation des systèmes de reconnaissance aux conditions de prise de son (changement de microphone, ...) et à l'environnement acoustique. Un travail est en cours dans le cadre du projet Esprit ARS (Adverse environment Recognition of Speech).

Ce type de transformation peut trouver une autre application dans le domaine de la synthèse de la parole (« changement de voix » d'un système de synthèse à partir du texte) (Abe *et al.* [15]).

Manuscrit reçu le 30 novembre 1989.

BIBLIOGRAPHIE

[1] Y. Grenier, Y., *Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonétique.* Thèse de Docteur Ingénieur, TÉLÉCOM Paris, 1977.

[2] K. Vintsuk, *Speech recognition by dynamic programming.* Kibernetika No. 1, 1968, pp. 81-88.

[3] Calliope, *La parole et son traitement automatique* (J. P. Tubach, éditeur principal), Collection technique et scientifique des télécommunications, Masson, Paris, 1989.

[4] L. Rabiner, S. Levinson, A. Rosenberg, *Speaker independent recognition of words using clustering techniques.* IEEE Transactions on ASSP, 1979, Vol. ASSP-27, No. 4, pp. 336-369.

[5] S. Levinson, *A unified theory of composite pattern analysis for automatic speech recognition*, in Computer Speech Processing (edited by F. Fallside and W. Woods), Prentice Hall International, 1983, pp. 243-292.

[6] B. H. Juang, *On the Hidden Markov Model and Dynamic Time Warping for Speech Recognition - A unified view.* AT & T Bell Laboratories Technical Journal. Vol. 63, No. 7, septembre 1984.

[7] D. Burton, J. Shore, J. Buck, *Isolated-Word speech recognition using multisection V.Q codebooks.* IEEE trans. on ASSP, 1986.

[8] K. N. Stevens, S. J. Kayser, H. Kawasaki, *Towards a phonetic and phonological theory of redundant features,* in Variability and invariance on speech processes (Perkell J. and Klatt D., editors). Erlbaum, Hillsdale, NJ, 1986.

[9] Shikano, Lee, Reddy, *Speaker adaptation through vector quantization.* Proc. IEEE-ICASSP 86 (Tokyo), pp. 2643-2646.

[10] J. K. Baker, *A second generation large vocabulary system.* Speech Technology. Vol. 4, No. 4, 1989, pp. 20-24.

[11] P. Duvaut, *Le filtrage de Wiener linéaire-quadratique à horizon fini. Application à la prédiction.* Traitement du Signal, Vol. 6, No. 3, 1989, pp. 152-161.

[12] T. Koh, E. J. Powers, *Second order Volterra filtering and its application to non linear systems identification.* IEEE Trans. on ASSP, Vol. ASSP 33, No. 6, décembre 1985.

[13] R. P. Lippman, *An introduction to computing with neural nets.* IEEE ASSP Magazine, Vol. 4, No. 2, April 1987, pp. 4-22.

[14] G. Chollet, K. Choukri, C. Montacié, *A test workstation for the evaluation of speech recognition algorithms, applications and data bases.* Proc. 7th FASE symposium (Speech'88), Edinburgh, août 1988, pp. 145-151.

[15] M. Abe, K. Shikano, H. Kuwabara, *Cross-Language Voice Conversion.* Proc. IEEE-ICASSP 90 (Albuquerque), pp. 345-348.