INVARIANCE DES SPECTRES DE PAROLE PAR ANALYSE DES CORRELATIONS CANONIQUES.
Adaptation d'un Système de Reconnaissance de Mots Isolés à de nouveaux locuteurs.

K.Choukri*,** ,G.Chollet** ,Y.Grenier**

* Laboratoires de Marcoussis, CRCGE. Route de NOZAY, 91460 Marcoussis.
** ENST-SYC, CNRS UA 820, 46 rue Barrault, 75013 Paris, France.

RESUME:

Cet article décrit une technique d'adaptation d'un dictionnaire de formes de référence à de nouveaux locuteurs, dans le cadre d'un Système de Reconnaissance Automatique de la Parole (SRAP). Elle se base sur l'hypothèse d'une corrélation maximale entre les espaces spectraux du locuteur standard et du nouveau locuteur pour déterminer un espace commun où les spectres respectifs sont invariants. Une application à la reconnaissance des dix chiffres montre les améliorations qu'elle apporte.

ABSTRACT:

Various speaker normalization and adaptation techniques of a knowledge data base or reference templates to new speakers in automatic speech recognition (ASR) have been studied during last years. This paper focusses on a technique for learning spectral transformations, based on a statistical analysis tool (Canonical correlation analysis), to adapt a standard dictionary to arbitrary speakers which does not require prior knowledge about them. The proposed method permits to improve speaker independance in Large vocabulary ASR. Application to an isolated digit recognizer improved a 70% correct score to 87%.

1. Introduction:

La représentation mathématique du signal de parole est déduite de l'onde acoustique acquise dans différents environnements (microphone, bruit ambiant, ...). La production de la parole (vibrations des cordes vocales et transmission par le conduit vocal) dépend fondamentalement des caractéristiques physiologiques et articulatoires des locuteurs, de l'influence des contraintes sémantiques, syntaxiques et lexicales (compétence et aptitude linguistiques), de l'état physique du locuteur (fatigue, émotion, ...) ainsi que d'autres facteurs paralinguistiques.

Ces différences expliquent la variabilité inter-locuteur observée dans le signal de parole. On observe aussi une variabilité intra-locuteur, mais beaucoup moins importante, ce qui explique les meilleures performances des systèmes dépendents du locuteur par rapport aux systèmes indépendents des locuteurs. Cela explique aussi le biais introduit dans les mesures de distance spectrale.

Pour réaliser des systèmes de reconnaissance indépendents du locuteur, plusieurs axes de recherche sont actuellement explorés. On distingue trois grandes directions. La première tente d'atténuer l'influence de la variabilité inter-locuteur en augmentant le nombre d'archétypes associés à chaque son dans le dictionnaire de référence, de telle sorte que tous les locuteurs représentatifs de la population d'utilisateurs fassent partie des locuteurs d'apprentissage. Pour y parvenir on utilise différents artifices tels que chaines de Markov, analyse discriminante, Clustering,...

La seconde méthode cherche des traits invariants aussi bien au niveau articulatoire, acoustique que perceptuel et ne garde que ces paramètres pour la représentation de la parole.

La première technique est telle que l'acquisition, la sélection et le codage des références deviennent vite une longue et coûteuse procédure. En outre le dictionnaire de références résultant occupe une place mémoire substantielle et on ne fait pas appel à des données spécifiques à la parole. La seconde technique, quoique plus attrayante, a encore besoin de quelques années de recherche avant d'être operationnelle, en élaborant un modèle de l'influence des caractéristiques du locuteur et de ses habitudes articulatoires sur le signal observé.

Ce papier concerne la troisième technique qui est **l'adaptation d'un SRAP de base à chaque nouvel utilisateur.**

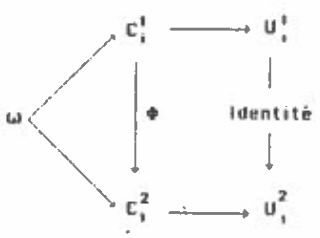
Chaque individu écoutant pour la première fois la parole d'un locuteur inconnu a souvent besoin de s'adapter à la nouvelle voix (ou d'adapter son apparail de perception), et les premiers mots d'un dialogue n'apportent guère d'informations que celles nécessaires à cette adaptation. D'une façon similaire on peut envisager une adaptation de SRAP basé sur le dictionnaire spécifique à un locuteur, à d'autres locuteurs sans acquérir leur dictionnaire spécifique respectif. Ceci permettra d'utiliser des algorithmes qui ont fait leurs preuves et dont on connait les performances. Par ailleurs cette procédure peut être accomplie d'une façon dynamique (Choukri et al.,1986), c'est à dire incorporée dans un système en configuration réelle d'exploitation ou d'utilisation.

2. Principe de l'adaptation de SRAP au locuteur:

Beaucoup d'auteurs qui s'intéressent aux problèmes dûs à la variabilité du signal de parole cherchent à normaliser des paramètres utilisés dans sa représentation. Il tiennent compte de paramètres articulatoires tels que la longueur du conduit vocal ou d'autres paramètres tels que les positions relatives des formants.

Le principe de la méthode est basé sur le constat qu'un même "son", produit par différents locuteurs, est interprété de manière identique par les personnes qui l'entendent malgré la variabilité inter-locuteur. On peut donc envisager un espace où des sons phonétiquement identiques seront représentés par des modèles identiques (Choukri et al.,1986).

Si on considère des cepstres sur une échelle Mel (MFCC) comme paramètres représentant la manifestation acoustique de chaque mot, l'espace associé à chaque locuteur est donc, dans un premier temps, un espace cepstral où la variabilité inter-locuteur s'exprime pleinement. Si on considère un son ω (mot, syllable, ...), on peut schématiser ces constats par la figure suivante où $\{C_i^j\}$, représente une succession de vecteurs cepstraux associée au locuteur j (Grenier,1980), (Grenier et al.,1981):



Production/Perception de la parole

Le problème de l'adaptation sera résolu si on arrive à déterminer les références $\{C_i^2\}$, associés au nouveau locuteur (2) à partir de celles associées à un locuteur standard (1). Il va de soi que nous ne connaitrons jamais - à moins de refaire un apprentissage sur le locuteur 2 - les références exactes mais uniquement une estimation de celles-ci.

Au lieu de chercher des transformations directes $C_2^i = \Phi(C_1^i)$, on se propose de chercher des transformations qui permettent de définir l'espace commun U. Pour cela on va partir d'un échantillon représentatif des espaces paramètriques C_1 et C_2, par exemple une phrase code ou un nombre limité de mots. Ensuite on va déterminer les projecteurs P_N et P_S tels que les projections soient identiques.

On se contentera dans un premier temps de transformations linéaires qui donnent des spectres projetés aussi proches que possible au sens d'un critère d'erreur. Si on choisit le critère des moindres carrés l'erreur de projection se traduit par l'équation (1):

$$J = \sum_i (u_i^1 - u_i^2)^T (u_i^1 - u_i^2) \qquad (1)$$

Il est facile de montrer à partir de cette équation qu'on peut minimiser l'écart entre les spectres projetés si et seulement si la corrélation entre les spectres associés est maximale, ce que réalise l'Analyse des Corrélations Canoniques (Golub,1970), en fournissant les projecteurs P_N et P_S en question (Choukri et al.,1986).

L'analyse canonique a pour but d'étudier la position relative d'un nuage de points par rapport à un autre (dans notre cas chaque nuage représentera l'échantillon d'un espace spectral d'un locuteur). Elle recherche des couples de variables, formés d'une combinaison des variables du premier nuage et d'une combinaison du second, les plus corrélés possible. Elle permet ainsi de définir un espace paramètrique où les projections de ces nuages coincident au mieux (au sens d'un critère d'erreur), qui sera alors une sorte d'espace "typologique" des deux locuteurs. On parle alors d'invariance des spectres par analyse des corrélations canoniques.

3. Procédure d'adaptation:

Pour valider notre propos on se propose d'appliquer cette méthode dans le cadre d'un système de reconnaissance de mots isolés avec un vocabulaire des dix chiffres.

Le spectre est paramétrisé avec 6 coefficients MFCC par trame. Durant la phase d'apprentissage chaque chiffre est prononcé une fois par un locuteur standard pour obtenir le dictionnaire de référence. La reconnaissance se fera grâce à des algorithmes de comparaison dynamique classiques, la détection de début et fin de mot est réalisée manuellement pour éviter toute erreur de détection pendant l'évaluation de cette méthode.

La première phase de la procédure d'adaptation consiste à acquérir et à aligner temporellement un échantillon représentatif de l'espace spectral associé à chacun des deux locuteurs. Il se pose alors le problème du choix de cet échantillon: que doit-on faire prononcer au nouveau locuteur comme "phrase code"?.

Dans une évaluation préliminaire cet échantillon sera réduit à un mot (le dixième du vocabulaire). les meilleurs mots semblent ceux qui reflètent le mieux la structure de l'espace phonétique (meilleure distribution dans le plan des premiers axes canoniques). Un logiciel d'analyse des corrélations canoniques permet alors de définir le nouvel espace de projection.

Grâce à ce logiciel on détermine la base génératrice du nouvel espace sur laquelle on projette le dictionnaire associé au locuteur standard pour obtenir le nouveau dictionnaire. On se retrouve dans le cas d'un "système monolocuteur" et on reprendra les algorithmes du système de base.

4. Evaluation:

Pour l'évaluation de cette méthode on dispose d'un corpus de 130 mots (comprenant les dix chiffres) prononcés par 100 locuteurs une seule fois. On cherche à évaluer la méthode dans le cadre d'un système monoréférence en insistant sur la variabilité inter-locuteur.

Des tests préliminaires ont pour but d'évaluer le système non-adapté en mono-locuteur croisé: le dictionnaire est obtenu grâce à un locuteur standard et on le teste sur des locuteurs choisis parmi les autres. Ensuite avec les mêmes données on évalue le système après adaptation.

Les taux de reconnaissance sont présentés en donnant les "bons" candidats qui sont reconnus en première position ou dans les deux premières positions avec l'intervalle de confiance correspondant à une probabilité d'erreur de 5%). Le taux de reconnaissance d'un système multi-référence utilisant les techniques de clustering (Syril) est de 93% en première position (Flocon et al.,1984).

Taux de reconnaissance et intervalle de confiance		
	première position	deux premières positions
non adapté	70% (68,73)	84% (81,86)
adapté	87% (84,89)	92% (91,94)

5. Conclusion:

Ce papier montre une adaptation de dictionnaires de formes à de nouveaux locuteurs. Une application à des systèmes mono-référence montre que les taux de reconnaissance sont améliorés de quelque 17%. Ce résultat reste à confirmer dans le cadre des systèmes Multi-références et de vocabulaire plus grands (130 mots).

6. Références:

Choukri, K. , Chollet, G. & Grenier, Y. (1986), Spectral transformations through canonical correlation analysis for speaker adaptation. in Proc. ICASSP, Tokyo (to be published).

Flocon, B. and Briant, N. (1984), SYRIL: système temps réel de reconnaissance de mots isolés indépendant du locuteur, 4ème congrés AFCET RFIA, Paris.

Golub, G.H. (1970), Matrix decomposition and statistical calculations. in Statistical computation, Edited by Milton, R.C. & Nelder, Y.A. (Academic press), PP. 365-397.

Grenier, Y. (1980), Speaker adaptation through canonical correlation analysis. in Proc. ICASSP, Denver, pp.888-891.

Grenier, Y., Miclet, L., Maurin, J.C. & Michel, H. (1981), Speaker adaptation for phoneme recognition. in Proc. ICASSP, Atlanta, pp.1273-1275.