Mise en correspondance d'espaces de représentation de la parole en vue de la fusion de données

Jérôme Piard et Harouna Kabré

Institut de la Communication Parlée
CNRS UA 368, INPG/ENSERG, Université Stendhal, 46 Av. Félix Viallet,
38031 Grenoble Cedex 1, France
piard kabre@icp.grenet.fr

RESUME

L'objectif de cette étude est la mise en correspondance d'espaces de représentation différents du signal de parole. Une méthode est proposée pour projeter ces différentes représentations dans un espace commun où la topologie des espaces initiaux sera au mieux préservée. Cette méthode s'appuie sur des techniques d'analyse multidimensionnelle utilisant le gradient pour minimiser une erreur quadratique. Le résultat de cette projection est analysé dans un cadre de fusion de données acoustiques et de données articulatoires fournies par un modèle anthropomorphique. Un essai de classification est proposé à partir des données acoustiques et articulatoires analysées par cette nouvelle technique.

ABSTRACT

The objective of this study is to take advantage of the coherence which can exist between different representations of the speech signal. First, a method is proposed to project some representations into a common space which conserves the topology of the initial spaces. This method relies both on the Multidimensional Data Analysis Technics and on the minimization of the root square error. Second, the results of this projection is analysed in the perspective of data fusion obtained with an anthropomorphic model of speech production and with acoustic data analysis methods. Some priliminary classification results are given with acoustic and articulatory data.

I INTRODUCTION

La mise en correspondance d'espaces de représentation est un problème crucial dans la résolution de problèmes complexes (vision, parole par exemple) nécessitant la fusion d'informations provenant de différentes sources [*11*]. Ces différentes sources d'informations donnent lieu à différentes représentations (par exemple le cepstre, le modèle d'oreille, etc. dans le cas de la parole) et leur mise en correspondance passe par leur caractérisation en vue de définir les complémentarités et les conflits qui pourraient exister entre elles.

Géneralement ces espaces présentent une trop grande dimension pour qu'il soit aisé d'y trouver une corrélation simple. Pour résoudre ce problème de dimensionalité, nous avons choisi de projeter ces espaces vers un cadre commun de dimension réduite où la comparaison sera plus simple à effectuer.

Il existe de nombreuses méthodes connues de réduction de dimension d'espaces d'observation. Les méthodes de projection linéaires en sont les représentantes les plus connues. La plus connue de ces transformations est l'analyse en composantes principales ou transformation de Karhunen-Loeve [*4*] souvent utilisée pour l'analyse de la parole. L'analyse discriminante et l'analyse en coordonnées principales [*13*] sont deux autres techniques bien connues. Toutes ces méthodes si elles sont pratiques pour des cas simples, s'avèrent cependant incapables de rendre compte de structures de données complexes. Il faut alors envisager d'utiliser une transformation non-linéaire. Historiquement les premières méthodes de ce type ont été développées à partir de techniques d'analyse multidimensionnelle statistique. On peut citer le non-metric MDS (Multi Dimensional Scaling) de Kruskal et Shepard [*8*] et le NLM (Non-Linear mapping) de Sammon [*12*]. Des méthodes récentes à base de réseaux de neurones tendent aujourd'hui à les supplanter, on peut par exemple citer les cartes tonotopiques de Kohonen [*6*], ou les réseaux VQP [*2*].

Cet article décrit une nouvelle méthode inspirée par ces travaux et qui est évaluée sur des données acoustiques et articulatoires.

II METHODOLOGIE

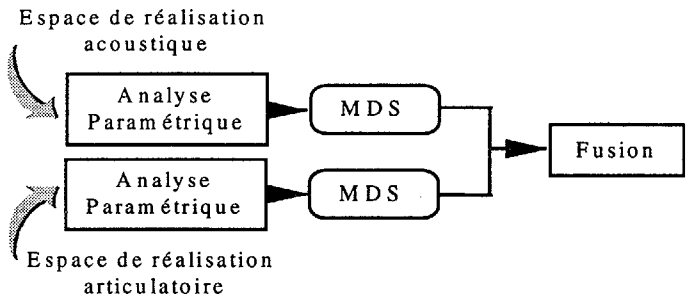

Figure 1 *Schema de principe*

La caractérisation passe par deux étapes, tout d'abord une paramétrisation qui va réduire la dimension de l'espace d'observation, puis une technique de projection MDS.
La méthode de réduction de dimension utilisée dans cet article s'inspire de travaux de Shepard et Kruskal [*7*]. Elle constiste à effectuer une projection non-linéaire d'un espace statistique de dimension élevée dans un espace de dimension réduite par conservation des distances mesurées dans l'espace initial.
On dispose de N objets notés $\{O_1, O_2, \ldots, O_N\}$ entre lesquels existent des mesures de dissimilarité $\delta_{ij,ij=1,..,N}$.
Le seul renseignement que l'on possède sur ces objets est cet ensemble de mesures de dissimilarité. Ces mesures sont supposées imparfaites et peuvent être bruitées ou distordues de manière quelconque et inconnue. On désire construire dans un espace, dit de *modélisation*, de faible dimension t (visualisable, c'est à dire t = 2 ou 3) une distribution de points rendant compte des dissimilarités. A chaque objet O_i est associé un point X_i de l'espace de modélisation.
Pour préserver la topologie de l'espace d'entrée tout en étant robuste aux différents bruits de mesure, la contrainte que l'on va essayer de maintenir est une conservation de la relation d'ordre existant sur les dissimilarités avec les distances que l'on peut calculer dans l'espace de modélisation.
On considère une configuration de points $\{X_1, X_2, \ldots, X_n\}$ dans l'espace de modélisation avec des distances d_{ij} entre chacun de ces points. Ces d_{ij} sont des distances euclidiennes calculées directement à partir des coordonnées des points X_i. Pour mesurer la qualité du placement de ces points vis à vis des dissimilarités, on définit une fonction associée à une configuration de points

$$STRESS = \frac{\sqrt{\sum_{(i,j)\in\Omega}\left(d_{ij} - \hat{d}_{ij}\right)^2}}{\sqrt{\sum_{(i,j)\in\Omega} d_{ij}^{\,2}}} \qquad (2.1)$$

où $\Omega = [1, N]x[1, N]$

Les $\hat{d}_{ij}$ sont appelées distances cibles ou pseudo-distances, elles servent à faire le lien entre les d_{ij} et les δ_{ij} qui sont supposées bruitées. La minimisation du STRESS se fait itérativement par la méthode du gradient. Le minimum du STRESS correspond à la configuration que l'on souhaite obtenir dans l'espace de modélisation.

III EXPERIMENTATIONS ET DISCUSSIONS

1. Corpus

Les expérimentations ont été mises en oeuvre sur un corpus audio-visuel. Les données dont nous disposons sont d'une part 100 réalisations pour chacune des 10 voyelles du français [i, e, ɛ, a, y, ɔ, œ, u, o, ø], consistant en des extraits sonores de longueur 64 ms, et d'autre part des paramètres articulatoires extraits automatiquement des mouvements des lèvres [*9*]. Nous avons également visualisé une représentation de l'espace articulatoire des voyelles du français en utilisant un modèle de production à 7 degrés de liberté, le modèle de MAEDA [*10*]. Dix prototypes de voyelles cardinales ont alors été utilisés pour obtenir l'espace articulatoire réduit.
A l'entrée de MDS nous avons construit une matrice de dissimilarité 10x10 symétrique représentant les distances entre chacun des 10 voyelles du français. Pour chacune des paramétrisations nous avons ainsi défini une distance afin de déterminer cette matrice de dissimilarité. Nous avons adopté une distance euclidienne pondérée définie par :

$$D^2 = \sum_{k=1}^{P} w(k)\left(X_k - Y_k\right)^2 \qquad (3.1)$$

où P est la dimension de l'espace d'observation,
$w(k)$ les coefficients de pondération,
X_k et Y_k $k = 1 \ldots n$ les paramètres de deux prototypes X et Y.

2. Expériences

a. *Les paramétrisations*

Nous avons utilisé trois méthodes d'analyse pour extraire des paramètres pertinents des signaux acoustiques. Tout d'abord les *4 premiers formants* exprimés sur une échelle Bark car ils ont une correspondance perceptive connue. Dans un deuxième temps les coefficients cepstraux MFCC exprimés sur une echelle Mel [*1*]. Nous avons utilisé *12 coefficients MFCC* issus d'une série de 40 filtres triangulaires appliqués sur une fenêtre de 32ms. Finalement *5 coefficients PLP* [*5*] avec une analyse d'ordre 5 sur des fenêtres de 32ms.
Des images du visage parlant ont été extraits *3 paramètres de lèvres* modélisant leur forme: l'écartement, la largeur, et l'aire [*9*].

b. *Protocole expérimental*

Nous avons effectué plusieurs essais afin de déterminer la distance la plus adaptée à partir de la formule (3.1) pour obtenir une représentation réduite la plus proche d'un espace réduit commun.
Nous avons appliqué une distance euclidienne ($w(k) = 1$) sur les formants pour obtenir l'espace MDS associé.
Pour les autres méthodes d'analyse acoustique, les pondérations ont été ajustées pour que les distances calculées soient les plus proches possible de celles dans l'espace formantique. Les tableaux 1, 2 et 3 donnent les pondérations que nous avons utilisées.

Coef	***1***	***2***	***3***	***4***	***5***	***6***
Valeur	0	0,42	0,27	0,22	0,26	0,99
Coef	***7***	***8***	***9***	***10***	***11***	***12***
Valeur	0	0	0	1,32	0,53	0,8

Tableau 1 *Pondération MFCC*

Coef	*1*	*2*	*3*	*4*	*5*
Valeur	5	0,5	2	1	0

Tableau 2 *Pondération PLP*

LipP	*LipH*	*Body*	*Jaw*	*Apex*	*Drsm*	*Lx*
0,3	1,4	1,3	1,5	0,6	0,8	0,2

Tableau 3 *Pondération articulatoire MAEDA*

La pondération des paramètres de Maeda correspond ici à l'influence perceptive lors de la variation de chacun des paramètres (*Tableau 3*)[***15***].
Contrairement au travail sur les paramètres acoustiques, il n'a pas été possible de réduire l'espace articulatoire visuel à un espace bidimensionnel proche du triangle vocalique. La projection se réduisant à regrouper les voyelles en trois ensembles : [i e ɛ a], [ɔ œ] et [y u o ø] des lèvres les plus ouvertes aux plus fermées.
De l'ensemble des réalisations a été extrait par chaque analyse, un jeu de paramètres « moyen » pour chaque phonème afin d'obtenir un espace réduit de référence.
Puis, pour juger la validité du choix de la distance adaptée ainsi que la robustesse de la projection, nous avons réalisé quelques expérimentations utilisant trois séries de paramètres prises au hasard dans les 100 réalisations associées à chaque voyelle. Nous avons calculé la corrélation entre les distances dans chaque espace réduit issu des 3 jeux de test et celles de l'espace formantique de référence.
Nous avons également procédé à des tests de classification des phonèmes sur les résultats fournis par MDS avec des réseaux de neurones à retro-propagation du gradient ***[14]***. Dans ces tests, l'apprentissage est fait sur les 10 premiers phonèmes de chaque série et le test est fait ensuite sur les 20 phonèmes suivants.

3. Discussion

Les figures 2, 3, 4 et 5 présentent une visualisation des espaces réduits par MDS. Le résultat obtenu avec les formants montre le triangle vocalique des voyelles. Il traduit une représentation perceptive des voyelles du français. Nous considérons donc cet espace comme l'espace de référence. Nous allons voir s'il est possible de projeter d'autres paramétrisations acoustiques ou articulatoires dans un espace MDS « perceptif » similaire à celui-çi. Les PLP déforment légèrement cette vision du triangle vocalique (*Figure 3*) ce qui est en accord avec les observations faites par Hermansky qui démontrent une bonne vérification de la théorie du F'2 à partir du spectre obtenu par les PLP [***5***].
Les coefficients de corrélation par rapport à l'espace de référence (*Tableau 4*) montrent que les PLP sont plus « perceptifs » que les MFCC. On peut aussi noter une forte corrélation entre les résultats obtenus à partir des formants et des données articulatoires de MAEDA. (*Tableau 5*). Dans une perspective d'interprétation phonétique, les lèvres ne semblent pas être pertinents au vu de nos résultats (Corrélation = 0.35).

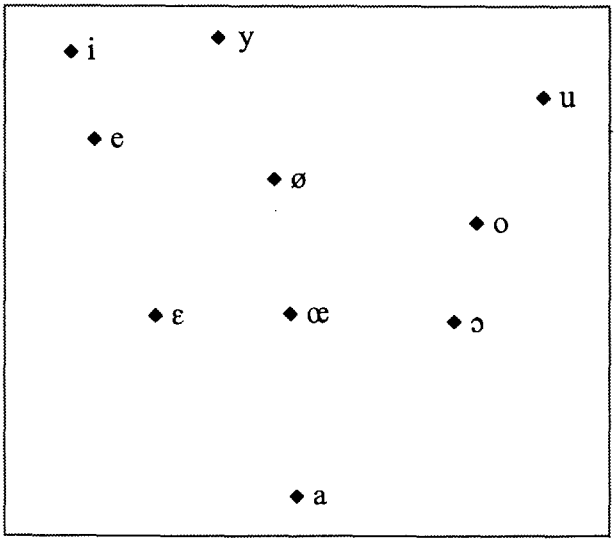

Figure 2 *Espace formantique réduit*

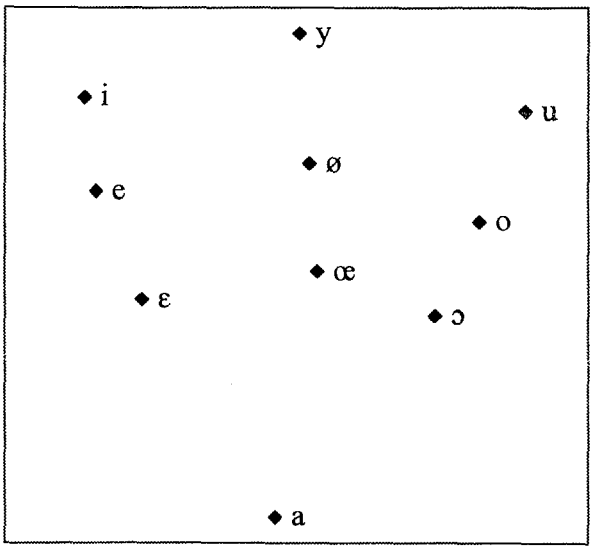

Figure 3 *Espace PLP réduit*

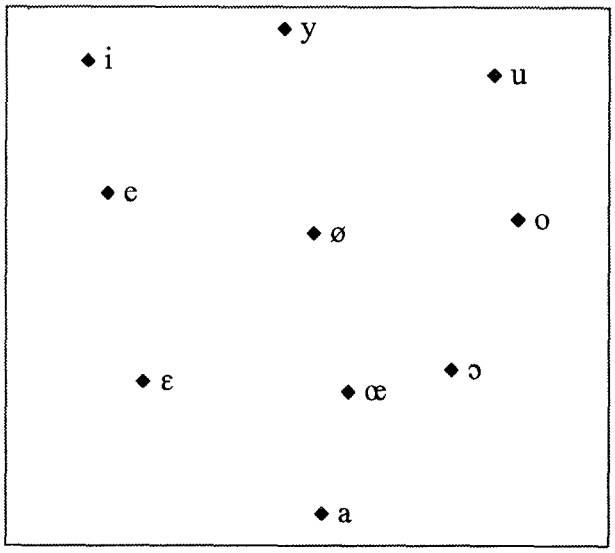

Figure 4 *Espace MFCC réduit*

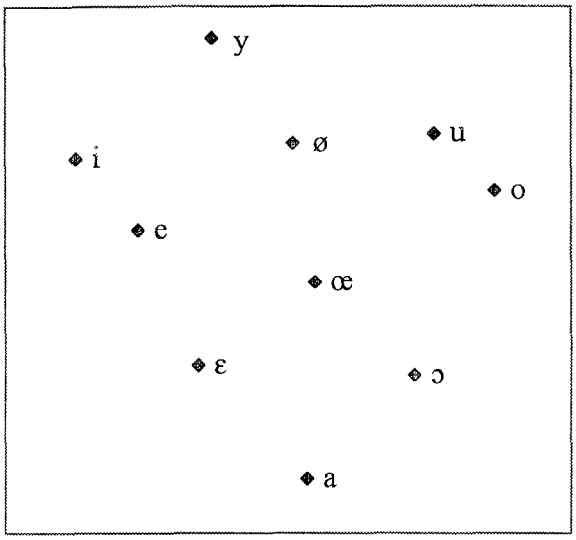

Figure 5 *Espace articulatoire associé à MAEDA*

Dans une perspective de classification (*Tableau 6*), les MFCC semblent plus performants que les formants qui ont une meilleure corrélation à l'espace articulatoire. Les PLP constituent alors, une solution intermédiaire entre la qualité perceptive et celle de discrimination.

	Formants	*PLP*	*MFCC*
Référence	1,00	0,94	0,92
Test1	0,96	0,91	0,83
Test2	0,92	0,83	0,74
Test3	0,97	0,76	0,91

Tableau 4 *Coefficients de corrélation vis à vis de l'espace formantique de référence*

	Maeda	Levres
Référence	0,96	0,35

Tableau 5 *Coefficients de corrélation articulatoires*

	Train	**Test**
FRM	98	92,5
PLP	95	89,5
MFCC	100	91,5

Tableau 6 *Résultats de Classification*

IV CONCLUSION

Nous avons pu constater lors de nos expérimentations que la caractérisation d'espaces d'observation était largement dépendante du but recherché. Certaines représentations nous ont donné une visualisation de l'espace MDS assez distordue par rapport au triangle vocalique alors qu'elles présentaient une très bonne capacité de discrimination (c'est le cas des MFCC). En revanche d'autres comme les PLP se sont révélées être une solution intermédiaire. Notre étude montre que d'une part, la distance utilisée et d'autre part, les pondérations adoptées, peuvent ainsi contribuer à une caractérisation d'espaces en vue d'une analyse discriminante ou perceptive dans un cadre de fusion d'informations. Même si le choix des pondérations reste en général empirique, ceux ayant trait aux paramètres de MAEDA et à la représentation formantique caractérisent une prépondérance perceptive de F1 et F2 qui est bien connue.

Dans une perspective de fusion de données, les informations fines et cohérentes extraites des différentes représentations acoustiques et celles des représentations articulatoires qui ne présentent pas les mêmes caractéristiques pourront être combinées pour accroître la robustesse des systèmes de reconnaissance automatique de la parole [*14*].

V. REFERENCES

[1] DAVIS S.B. & al. (1980) *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* IEEE Transactions on ASSP Vol 28(4) pp 357-365 Aug. 1980

[2] DEMARTINES P. & al. (1994) *Apprentissage de structures de données par auto-organisation*

[3] DROUET d'AUBIGNY G. (1989), *L'analyse multidimensionnelle des données de dissimilarité.* Thèse de mathématiques - Grenoble.

[4] FUKUNAGA K. & al. (1970) *Application of the Karhunen-Loève expansion to feature selection and ordering.* IEEE Transactions on Computers, C-19(4) pp 311-318, Apr 1970.

[5] HERMANSKI H. (1990) *Perceptual linear predictive (PLP) analysis of speech.* JASA Vol 87(4) pp 1738-1752 Oct 1990.

[6] KOHONEN T. (1990) *The self-organising map.* Proceedings of IEEE 78(9) pp 1464-1480, 1990.

[7] KRUSKAL J.B. (1964), *Nonmetric multidimensional scaling: a numerical method.* Psychometrika, 29 pp. 115-129.

[8] KRUSKAL J.B. (1979), *Multidimensional Scaling and other methods for discovering structure.* in Statistical method for digital computers. Enslein K. & Ralston A. eds

[9] ROBERT-RIBES J. (1994), *Models of Audio-Visual Integration*, Doctoral Thesis, INPG, Grenoble, France

[10] MAEDA S. (1989), *Compensatory Articulation during Speech : Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model,* Speech Production and Modelling pp 131-149.

[11] PASCAL D. (1990) *Mise en correspondance d'espaces acoustique et perceptif par l'analyse factorielle multiple* 18èmes JEP - Montréal Mai 1990.

[12] SAMMON J.W. (1969) *A non-linear mapping for data stucture analysis.* IEEE Transactions on Computers, C-18(5) pp 402-409, May 1969.

[13] TORGERSON W.S. (1952) *Mustidimensional Scaling I, Theory ans method.* Psychometrika 17, pp 401-418.

[14] KABRE H., *A Probabilistic Model for Multi-sensor Data Fusion*, J. of Speech Communication, 1995 (à paraître).

[15] BOË L.J. & al. (1992) *Une prédiction de l'« audibilité » des gestes de la parole à partir d'une modélisation articulatoire.* 19èmes JEP pp 151-157.