Analyse en composantes principales temps-fréquence : application à la reconnaissance de la langue

Michel Dutat[1][2], Ivan Magrin-Chagnolleau[3], Frédéric Bimbot[3]

[1] LSCP / CNRS UMR 8554, 54 Boulevard Raspail, 75270 Paris cedex
[2] ENST - Dépt Signal, CNRS URA 820, 46 Rue Barrault, 75634 Paris cedex 13
[3] IRISA / CNRS & INRIA Rennes, Campus universitaire de Beaulieu, 35042 Rennes cedex
dutat@lscp.ehess.fr - ivan@ieee.org - bimbot@irisa.fr

Abstract

In this paper, we use a new speech parameterization based on a principal component analysis applied to feature parameters augmented by their context. This new parameterization is called time-frequency principal component (TFPC) analysis. We apply the new parameterization in the framework of automatic language recognition. This new approach allows us to improve the identification rate compared to the use of the classical cepstral coefficients augmented by their Δ coefficients.

1. Introduction

Une grande variété d'analyses paramétriques du signal de parole a été utilisée en reconnaissance de la langue par modélisation acoustique. Les meilleurs résultats sont généralement obtenus en incorporant l'information dynamique du signal par le biais d'approximations de la dérivée et de la dérivée seconde (les coefficients Δ et $\Delta\Delta$). Dans cette étude, nous abordons une nouvelle paramétrisation qui prend également en compte l'aspect dynamique du signal. Cependant, cette méthode opére sur le signal une sélection à la fois temporelle et fréquentielle du matériel acoustique et ceci en fonction de la langue. Cette sélection est réalisée par des filtres temps-fréquences dont les coefficients sont calculés à partir du corpus d'apprentissage de la langue considérée. On fait l'hypothèse qu'un énoncé ainsi filtré est mieux représenté lorsque le filtre qui lui est appliqué est celui correspondant à sa langue. Cette approche rend plus optimal la paramétrisation et améliore ainsi le taux de reconnaissance par rapport à celui obtenu avec une paramétrisation classique.

2. Présentation de la méthode

2.1. Approche classique

Lors d'une procédure de reconnaissance de la langue par une modélisation acoustique, on tente de capter la structure sonore propre à cette langue en utilisant un grand nombre d'enregistrements prenant en compte le plus de locuteurs possible. Vient ensuite la paramétrisation des enregistrements dans une forme acceptable pour les modèles. Ceux-ci sont en général constitués de réseaux de Markov cachés à structure ergodique. La figure 1 présente la procédure classique d'une phase d'apprentissage. L'ensemble des énoncés d'apprentissage $E_{app}^{(i)}$ de la langue i est transformé par une analyse acoustique en une séquence de vecteurs de paramètres $\{\mathbf{x}_t\}^{(i)}$. Cette séquence est utilisée afin d'estimer les paramètres $\kappa^{(i)}$ du modèle de la langue i.

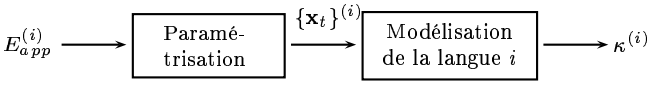

Figure 1 – *Apprentissage classique.*

Une fois que tous les paramètres des modèles ont été calculés, on obtient n modèles, chacun représentant une langue différente. La phase de test consiste à paramétriser l'énoncé à tester, puis à calculer une mesure de confiance par les n modèles du système. Un algorithme de décision combine alors les n scores obtenus, avec éventuellement d'autres sources d'informations, afin de sélectionner la langue la plus probable.

2.2. Approche par filtrage vectoriel

La méthode que nous présentons a été initialement proposée dans [4, 5]. Elle diffère de l'approche classique à la fois en ce qui concerne l'apprentissage des modèles et en ce qui concerne la procédure de test. Son but est de rendre l'énoncé plus facilement identifiable dès l'étape de paramétrisation. Les énoncés, après une analyse acoustique, subissent un filtrage vectoriel temps-fréquence dont les caractéristiques sont dépendantes de la langue. Ainsi un énoncé filtré avec le filtre de sa langue est mieux représenté que s'il l'est par un tout autre filtre. On obtient ainsi une paramétrisation qui doit être plus optimale. La phase d'apprentissage comporte deux étapes. La première permet l'obtention des coefficients du filtre et la seconde permet le calcul des paramètres $\lambda^{(i)}$ du nouveau modèle. La figure 2 illustre l'apprentissage du modèle de la langue i. L'ensemble $E_{app}^{(i)}$ des énoncés d'apprentissage est transformé par une analyse acoustique[1] en une séquence $\{\mathbf{x}_t\}^{(i)}$ de vecteurs de dimension p. Celle-ci permet le calcul des coefficients de la matrice de filtrage, puis une fois les caractéristiques du filtre $\mathbf{H}^{(i)}$ obtenues, le filtrage de la séquence $\{\mathbf{x}_t\}^{(i)}$ en une nouvelle séquence $\{\mathbf{f}_t\}^{(i)}$ de vecteurs de dimension r permet le calcul des paramètres $\lambda^{(i)}$ du modèle de la langue i.

1. qui peut être spectrale, cepstrale, par prédiction linéaire, etc.

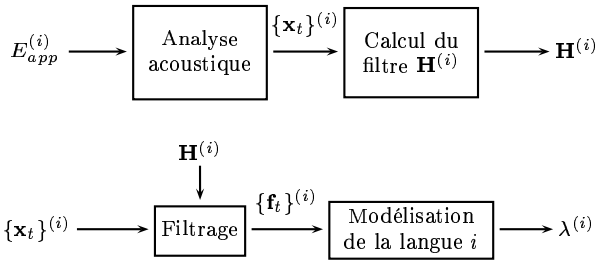

Figure 2 – *Apprentissage avec filtrage vectoriel.*

2.3. Fabrication du filtre

Les caractéristiques du filtre sont capitales pour obtenir des vecteurs bien répartis dans l'espace de représentation des langues. Ainsi, pour une langue donnée, et à partir de l'analyse acoustique de ses données d'apprentissage, on recherche une représentation des données qui maximise l'inertie, dans un sous espace qui sera caractéristique de la langue. Pour ce faire, nous utilisons une analyse en composantes principales de la séquence des vecteurs d'apprentissage. Le but est de rechercher les liaisons qui existent entre les différents coefficients des vecteurs paramètres et qui peuvent caractériser la langue en question. De plus, pour capter l'information dynamique résidant dans la suite de vecteurs consécutifs, on va également prendre en compte un contexte temporel autour de chaque trame analysée. On obtient en fin de compte un filtrage vectoriel à base de composantes principales temps-fréquences (que l'on nommera par le sigle TFPC *Time Frequency Principal Components*)[4, 5].

Les différentes étapes pour l'élaboration d'un filtre sont les suivantes : soit la séquence de vecteurs $\{\mathbf{x}_t\}$, de dimension p, issue d'une analyse acoustique des énoncés d'apprentissage, avec t variant de 1 à T, on construit les matrices de covariances décalées V_q avec $q = 0, 1, 2, \cdots$:

$$\mathsf{V}_q = \frac{1}{T}\sum_{t=q+1}^{T}(\mathbf{x}_t - \overline{\mathbf{x}})(\mathbf{x}_{t-q} - \overline{\mathbf{x}})^{\mathsf{T}} \quad \text{avec} \quad \overline{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$$

Puis on construit la matrice de covariance contextuelle $\mathbf{V}_{2q+1}$ à partir des matrices de covariances décalées de telle sorte que l'on obtienne une matrice de structure bloc-Toeplitz de dimension $(2q+1)p \times (2q+1)p$. On appellera ordre de l'analyse le nombre q qui prend en compte $2q+1$ trames temporelles.

$$\mathbf{V}_{2q+1} = \begin{bmatrix} \mathsf{V}_0 & \mathsf{V}_1 & \cdots & \mathsf{V}_{2q} \\ \mathsf{V}_1^{\mathsf{T}} & \mathsf{V}_0 & \cdots & \mathsf{V}_{2q-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{V}_{2q}^{\mathsf{T}} & \mathsf{V}_{2q-1}^{\mathsf{T}} & \cdots & \mathsf{V}_0 \end{bmatrix}$$

L'analyse en composantes principales des données augmentées de leur contexte temporel revient à rechercher les vecteurs propres $\mathbf{v}_i$ de la matrice $\mathbf{V}_{2q+1}$. L'espace de représentation composé des vecteurs propres correspondant aux plus grandes valeurs propres est celui d'inertie maximale. Il est à noter que toutes les directions de cet espace sont orthogonales entre elles. De plus, les composantes principales de moindres variances peuvent être assimilés à du bruit. Il est ainsi possible de diminuer l'espace de représentation en ne conservant que certains vecteurs propres représentant aux mieux les données. Il existe d'ailleurs plusieurs stratégies pour réduire la dimension de l'espace de représentation [3]. À partir des composantes principales retenues, on élabore une matrice de filtrage en juxtaposant ces vecteurs et en la transposant. Ainsi, si l'on veut les cinq premières composantes suivies de la 10e à la 12e, on construit la matrice de filtrage comme suit :

$$\mathbf{H} = [\mathbf{v}_1 \cdots \mathbf{v}_5\ \mathbf{v}_{10} \cdots \mathbf{v}_{12}]^{\mathsf{T}}$$

Le filtrage d'un énoncé consiste en un produit de convolution entre sa représentation paramétrique et la matrice $\mathbf{H}$. Soit la séquence de vecteurs $\{\mathbf{x}_t\}_{1\leq t\leq T}$ de taille p issue de l'analyse acoustique de l'énoncé. On définit la séquence de vecteurs centrés $\mathbf{x}_t^*$ entre le temps $t-q$ et le temps $t+q$:

$$\mathbf{X}_{t-q}^{t+q} = \begin{bmatrix} \mathbf{x}_{t+q}^* \\ \vdots \\ \mathbf{x}_t^* \\ \vdots \\ \mathbf{x}_{t-q}^* \end{bmatrix} \qquad \text{avec } \mathbf{x}_t^* = (\mathbf{x}_t - \overline{\mathbf{x}})$$

La dimension du vecteur $\mathbf{X}_{t-q}^{t+q}$ est $(2q+1)p \times 1$. En supposant que nous ayons choisi r composantes principales, la matrice de filtrage, de dimension $r \times (2q+1)p$, peut s'écrire en faisant apparaître sa structure temporelle :

$$\mathbf{H} = [\mathsf{H}_{-q} \cdots \mathsf{H}_0 \cdots \mathsf{H}_q]$$

Les vecteurs filtrés $\mathbf{f}_t$ sont obtenus par le produit de convolution suivant avec $1 \leq t \leq T$:

$$\begin{aligned} \mathbf{f}_t &= \mathbf{H} \cdot \mathbf{X}_{t-q}^{t+q} \\ &= \sum_{k=-q}^{+q} \mathsf{H}_k \cdot \mathbf{x}_{t-k}^* \end{aligned}$$

2.4. La phase de test

Une fois la phase d'apprentissage terminée, à chaque langue est associée un modèle mais également un filtre. Tester l'origine linguistique d'un énoncé revient à faire calculer par chaque modèle une mesure de vraisemblance. Mais ici la paramétrisation de l'énoncé subit, pour chaque modèle, le filtrage correspondant. La figure 3 illustre la procédure de test employée avec la méthode TFPC. Une analyse acoustique traduit l'énoncé de test E_{test} de langue inconnue en une séquence de vecteurs $\{\mathbf{x}_t\}$. Pour chacune des n langues du système, cette séquence est filtrée. Puis une estimation de la probabilité d'observation conditionnelle $P\left(\{\mathbf{f}_t\}|\lambda^{(i)}\right)$ est calculée. L'étape de décision revient à sélectionner la langue $L^{(i)}$ de plus forte probabilité.

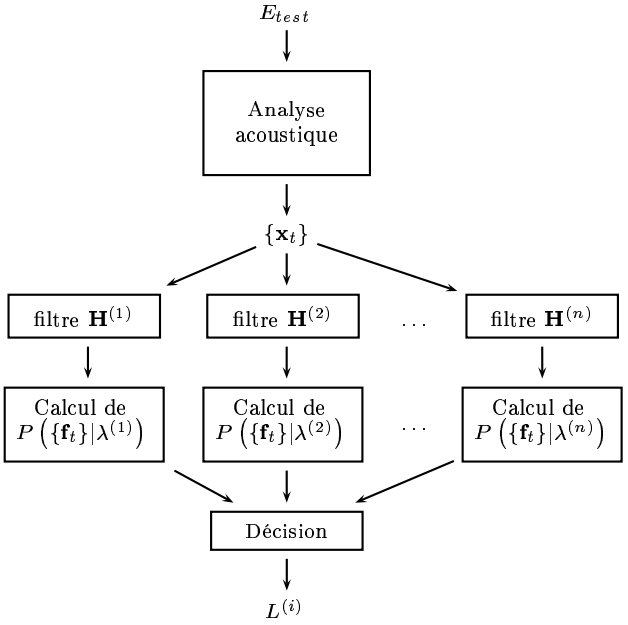

Figure 3 – *Procédure de test en utilisant la méthode* TFPC.

Cette méthode offre un large éventail de possibilités quant à la paramétrisation finale des énoncés. Ces choix portent sur les points suivants :

- Analyse acoustique des énoncés. (bancs de filtres, coefficients lpc, coefficients cepstraux, etc.)
- Choix de l'ordre de l'analyse TFPC, c'est-à-dire le nombre de trames temporelles prises en compte.
- Choix des vecteurs de la matrice des composantes principales à faire intervenir pour l'élaboration du filtre.

Dans les sections suivantes, on présente des résultats obtenus en faisant varier certains de ces paramètres. Une expérience classique faisant intervenir une paramétrisation cepstrale des énoncés nous permet d'évaluer l'efficacité de cette nouvelle approche.

3. Résultats

3.1. Objectif

Nous désirons étudier l'influence de cette nouvelle paramétrisation sur les taux de succès d'un système de reconnaissance de la langue. Pour ce faire, on construit deux systèmes de reconnaissance qui ne diffèrent que par leur étape de paramétrisation des énoncés, l'un appelé modèle classique, l'autre appelé modèle TFPC. Nos systèmes doivent attribuer à l'énoncé testé une langue parmi les langues apprises.

3.2. Corpus

Nous avons constitué nos corpus d'apprentissage et de test en choisissant des enregistrements dans la base de données OGI MLTS (Oregon Graduate Institute Multilanguage Telephone Speech) qui a été spécialement conçue pour des travaux de reconnaissance de la langue [6]. Pour cette étude, nous avons choisi quatre langues : l'anglais, l'espagnol, le français et le japonais. Pour chaque langue, nous avons utilisé pour l'apprentissage (resp. pour le test) 80 énoncés (resp. 54) prononcés par 20 locuteurs (resp. 12), ce qui représente une durée totale de 1240 secondes (resp. 860).

3.3. Caractéristiques du modèle classique

Ce modèle de contrôle nous permet d'analyser l'apport de cette nouvelle paramétrisation. Il consiste en une modélisation de la structure acoustique de la langue à partir d'un apprentissage non supervisé. Il est constitué d'un modèle de Markov caché par langue dont les caractéristiques sont données dans la table 1.

Table 1 – *Caractéristiques du modèle de Markov caché:*

Structure	ergodique
Nombre d'états	24
Nombre de gaussiennes par état	2
Matrices de covariances	diagonales

Pour le choix de la paramétrisation du signal de parole, nous avons écarté l'analyse par bancs de filtres, des études antérieures ayant montré que les résultats d'un système de reconnaissance de la langue utilisant cette paramétrisation sont moins bons que ceux utilisant une analyse cepstrale [1]. On a donc choisi cette dernière pour la mise en forme des énoncés du modèle de contrôle. Les caractéristiques de cette analyse sont résumées dans la table ci-dessous :

Table 2 – *Caractéristiques de l'analyse cepstrale :*

Type de la paramétrisation	MFCC [a]
Répartition fréquentielle	échelle Mel
Nombre de coefficients	12 ou 24
Largeur de la fenêtre d'analyse	30 ms
Période d'analyse	10 ms
Fenêtre d'analyse	Hamming
Soustraction cepstrale	oui

[a] Mel Frequency Cepstral Coefficient

Comme certaines expériences avec le modèle TFPC font intervenir le contexte temporel dans la paramétrisation, il faut le faire également intervenir dans le modèle classique. Ceci est réalisé par l'apport des coefficients différentiels Δc_t^i qui sont calculés à partir des coefficients cepstraux c_t^i par la formule [2] :

$$\Delta c_t^i = \frac{\sum_{k=1}^{N} k(c_{t+k}^i - c_{t-k}^i)}{2 \cdot \sum_{k=1}^{N} k^2} \qquad \text{avec} \quad i = 1, 2 \cdots, p$$

La variable N définit l'horizon temporel à prendre en compte pour le calcul des coefficients différentiels. Suivant l'ordre de l'expérience TFPC, on aura l'équivalence suivante :

Ordre	N	Nombre de c^i	Nombre de Δc^i
0	-	24	0
1	1	12	12
2	2	12	12
3	3	12	12
4	4	12	12

3.4. Expériences TFPC

Pour comparer les résultats, il faut que la différence entre les deux systèmes ne porte que sur le filtrage des

énoncés. Ainsi, la configuration du modèle de Markov caché pour le modèle TFPC est identique à celle du modèle de contrôle (cf. Table 1). Le modèle classique utilisant une paramétrisation cepstrale, nous gardons le même type d'analyse acoustique pour les expériences TFPC (cf. Table 2). Afin d'analyser l'apport de l'information dynamique sur les taux de succès de reconnaissance, on fait varier l'ordre de l'analyse TFPC en prenant soin de modifier en conséquence la paramétrisation de l'expérience classique de telle sorte que l'horizon temporel pris en compte soit identique pour les deux expériences. On s'intéresse également au nombre de composantes principales retenues pour l'élaboration des filtres TFPC. Ainsi, selon les expériences, on sélectionne de 16 à 32 composantes prises parmi les premiers vecteurs propres de la matrice $\mathbf{V}_{2q+1}$.

Ordre 0 L'analyse TFPC d'ordre 0 n'utilise pas de contexte temporel. La matrice de covariance contextuelle $\mathbf{V}_{2q+1}$ est donc la matrice de covariance V_0 de dimension 24×24. La matrice de filtrage sera au maximum de même dimension et les vecteurs filtrés auront donc aux maximum 24 coefficients. Par conséquent, l'expérience classique doit également avoir un vecteur de paramètres composé de 24 coefficients. Le tableau ci-dessous présente les différents taux de succès moyen sur les quatre langues considérées obtenus avec l'expérience TFPC en fonction du nombre de composantes retenues pour fabriquer le filtre. La dernière ligne est composée des gains relatifs par rapport à l'approche classique qui dans cette configuration a obtenu un taux de succès moyen de 53,10%. On constate une amélioration des résultats et celle-ci d'autant plus forte que l'on diminue le nombre de composantes du filtre. Il semble que l'effet du filtrage TFPC ait bien permis de retenir les informations dépendantes de la langue, mais si les premières composantes principales sont efficaces, les dernières semblent avoir des caractéristiques partagées par plusieurs langues, ce qui expliquerait l'accroissement des confusions lorsque l'on augmente la taille de l'espace de représentation.

Composantes	16	20	24
Taux de Succès	58,02	55,78	53,47
Gain relatif	9,27	5,05	0,70

Ordre 1 Avec un ordre $q = 1$, le contexte temporel s'étend sur trois trames du signal. La matrice de covariance contextuelle $\mathbf{V}_3$ est de dimension 72×72. L'expérience de contrôle fait intervenir l'estimation de la dérivée première dans sa paramétrisation du signal, avec un horizon temporel de 3 trames ($N = 1$). Son taux de succès moyen sur quatre langues est de 59,94%. Le tableau ci-dessous résume les résultats obtenus. Comme précédemment, une amélioration est à mettre au compte de la méthode TFPC, mais l'effet dû à la diminution de l'espace de représentation est moins marqué. Il semble que l'information pertinente soit plus confusément répartie à travers les composantes principales et une étude plus poussée du choix de celles-ci semble nécessaire.

Composantes	16	20	24	28	32
Taux de Succès	62,71	60,81	57,67	59,99	61,65
Gain relatif	4,62	1,45	-3,79	0,08	2,85

Ordre 4 L'ordre $q = 4$ fait intervenir 9 trames temporelles. La matrice de covariance contextuelle $\mathbf{V}_9$ a pour dimension 216×216. La paramétrisation de l'expérience classique comporte les coefficients ΔMFCC calculés avec $N = 4$. Le taux de succès moyen pour l'expérience de contrôle est de 60,34%. Le tableau des résultats confirme l'effet de confusion constaté précédemment. L'information utile pour la discrimination des langues ne peut être exploitée en n'utilisant qu'une trentaine de composantes. Cependant avec 32, mais surtout 24 composantes retenues, on obtient une amélioration par rapport à l'expérience de contrôle.

Composantes	16	20	24	28	32
Taux de Succès	53,91	58,55	63,49	53,59	60,75
Gain relatif	-10,66	-2,97	5,22	-11,19	0,68

4. Conclusions

Nous avons abordé une nouvelle paramétrisation qui, à partir d'une analyse acoustique classique, applique sur les vecteurs de paramètres un filtrage vectoriel dépendant de la langue. Les caractéristiques des filtres sont obtenues par une analyse en composantes principales des données d'apprentissage. Par ce procédé, on utilise des espaces de représentations optimisés pour chaque langue. Les expériences en reconnaissance automatique de la langue que nous avons réalisées montrent une légère augmentation des taux de succès lorsqu'on applique cette méthode. L'un des avantages de cette approche est la grande liberté que nous avons quant aux choix concernant la paramétrisation initiale, la taille du contexte temporel à prendre en compte, la sélection des vecteurs de la matrice des composantes principales pour l'élaboration du filtre. Cependant les expériences impliquant un contexte temporel important nous ont montré la nécessité d'un choix plus subtil des composantes à intégrer aux matrices de filtrage. On s'intéressera également à l'influence d'une telle paramétrisation sur la régularité des séquences d'états parcourus dans les modèles de Markov cachés. Ceci est particulièrement important dans l'optique de travaux sur la localisation de suites de séquences d'états typiques propres aux langues.

Références

[1] Michel Dutat. Reconnaissance automatique de la langue parlée. Rapport d'avancement de thèse, Ecole National Supérieure des Télécommunications, 1997.

[2] Sadaoki Furui. Comparison of speaker recognition methods using static features and dynamic features. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):342–350, June 1981.

[3] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[4] Ivan Magrin-Chagnolleau. *Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte*. PhD thesis, Ecole National Supérieure des Télécommunications, 1997.

[5] Ivan Magrin-Chagnolleau, Geoffrey Durou, and Frédéric Bimbot. Application of time-frequency principal component analysis to text-independent speaker identification. Submitted to IEEE Transactions on Speech and Audio Processing.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. Technical report, Center for Spoken Language Understanding Oregon Graduate Institute of Science and Technology., Portland, 1993.