ⴰⵙⵉⵏⴰⴳ-*Asinag,* 9, 2014, p. 133-145

# Vers un système de reconnaissance automatique de la parole en amazighe basé sur les transformations orthogonales paramétrables

Abenaou Abdenbi (1), Ataa Allah Fadoua (2) & Nsiri Benayad (3)

(1) ENSA, Université Ibn Zohr,
(2) CEISIC, IRCAM,
(3) FSAC, Université Hassan II

*In this paper, we propose a new approach for Amazigh isolated word recognition, based on relevant speech signal parameters' extraction algorithm. In general, the approach consists on the application of adaptive orthogonal transforms that are characterized by a linear operator constituted of configurable functions, which allows the transform adaptation to the initial data and the reduction of feature vector dimension, that improve the isolated word recognition rate.*

## 1. Introduction

Durant cette dernière décennie, l'évolution permanente des technologies de l'information et de la communication a été marquée par des progrès majeurs dans le déploiement du traitement du langage humain, notamment la reconnaissance automatique de la parole, pour la promotion et le développement des langues peu dotées.

De nos jours, en effet, la reconnaissance automatique de la parole est introduite dans de nombreuses applications ; à savoir les systèmes d'apprentissage des langues pour améliorer la prononciation des apprenants (Bahi, 2008), les applications téléphoniques du type serveur vocal pour l'accès aux services (Barnard et *al.*, 2009) ou l'accès à l'information à travers la recherche dans des bases de données vocales particulièrement pour les personnes à besoins spécifiques et les analphabètes surtout dans les régions rurales (Barnard et *al.*, 2010 ; Patel et *al.*, 2010 ; Kumar et *al.*, 2011), ainsi que les applications de transcription automatique des documents radio et télédiffusés.

Cependant, les technologies de la parole ne sont pas suffisamment exploitées pour la langue amazighe. Afin de profiter des avantages de ces technologies, nous avons consacré cette étude à la réalisation d'un premier système de reconnaissance de mots isolés amazighes à la base des transformations orthogonales paramétrables.

Généralement, en traitement du signal vocal, la résolution des problèmes de reconnaissance passe nécessairement par une étape d'extraction des caractéristiques informatives des signaux avant d'entamer la phase d'analyse. Parmi les travaux de recherche traitant l'extraction des caractéristiques à partir des mots isolés, nous distinguons deux principaux types d'approches : les méthodes à base des théories statistiques (Bourlard et Morgan, 1993 ; Doddington, 1985 ; Cappe, 1995) et les méthodes déterministes à base des transformations orthogonales classiques (Walsh, Haar, Fourier, …) (Kekre et *al.*, 2010 ; Ahmed et Rao, 1975). Néanmoins, les méthodes statistiques, tel que le modèle de Markov caché, ont atteint leurs limites dans l'amélioration des systèmes de la reconnaissance automatique des signaux vocaux, malgré la disposition de corpus suffisamment représentatifs. Tandis que les méthodes spectrales ont émergé dans plusieurs applications du traitement du signal vocal grâce à la richesse de leurs propriétés et la rapidité du calcul de leur algorithme (Bello et *al.*, 2004 ; Doets et Lagendijk, 2004).

Le principe fondamental de ces méthodes, particulièrement celles liées à un système de fonction de base orthogonale (non paramétrable) comme la transformée de Fourier ou la transformée en ondelettes, est d'obtenir le vecteur spectral des caractéristiques informatives.

Cependant, le spectre obtenu par ces méthodes est généralement trop large, vu que le signal vocal est un processus non stationnaire. Ce qui complique souvent la procédure de reconnaissance des signaux et conduit, dans certains cas, à des résultats insatisfaisants. D'où la nécessité d'une méthode de détermination des caractéristiques informatives du signal vocal dont le coût de calcul est optimal.

Dans cet article, nous proposons une solution au problème en utilisant les transformations orthogonales adaptables pour l'extraction des caractéristiques informatives du signal vocal, tout en visant la réalisation d'un système de reconnaissance de la parole amazighe dédié à l'apprentissage de la prononciation. L'utilisation de ces transformations (Abenaou et Sadik, 2011a, 2011b, 2011c) est favorisée par la possibilité d'adaptation de la forme de leurs fonctions de base en fonction du caractère du vecteur étalon. Ce dernier est formé par les différents signaux vocaux de chaque mot. Autrement dit, à chaque classe de mots est associé un système de fonctions de base paramétrables pour la projection des signaux. En outre, ces fonctions répondent au critère de la complétude du système, qui assure les transformations des signaux sans perte de leur contenu informatif. Le système de fonctions de base formé s'exprime sous forme d'un opérateur matriciel orthonormé factorisable, ce qui permet une transformation à la base d'un algorithme à calcul rapide.

## 2. Méthode et algorithme de synthèse de l'opérateur de la transformée orthogonale adaptable

En traitement numérique, la transformée linéaire orthogonale d'un signal X peut être représentée par l'équation matricielle (1):

$$Y = \frac{1}{N} HX \qquad (1)$$

où:

- $\boldsymbol{X} = [x_1, x_2, \ldots, x_N]^T$ est le signal initial à transformer, dont la taille $\boldsymbol{N} = 2^n$ ;
- $\boldsymbol{Y} = [y_1, y_2, \ldots, y_N]^T$ est le vecteur des coefficients spectraux, calculé par l'opérateur spectral orthogonal $\boldsymbol{H}$ de dimension $\boldsymbol{N}$ x $\boldsymbol{N}$.

La factorisation de Good (Good, 1960) a montré la possibilité de représenter l'opérateur matriciel $\boldsymbol{H}$ sous forme de produit de matrices creuses $\boldsymbol{G_i}$ (2) avec une proportion plus élevée des zéros ce qui permet la construction des algorithmes de transformation rapide de Walsh, de Haar et de Fourier. Les matrices $\boldsymbol{G_i}$ ($i = 1,\ldots, n$) sont construites par des blocs de matrices $\boldsymbol{V_{i,j}}$ de dimension minimale qui s'appellent noyaux spectraux (Abenaou et Sadik, 2011a).

$$G_i = \left[\begin{array}{cccc:cccc}
\begin{bmatrix}\alpha_{i1}\\ \beta_{i1}\end{bmatrix} & 0 & \cdots & 0 & \begin{bmatrix}\gamma_{i1}\\ \delta_{i1}\end{bmatrix} & 0 & \cdots & 0 \\
0 & \begin{bmatrix}\alpha_{i2}\\ \beta_{i2}\end{bmatrix} & \cdots & 0 & 0 & \begin{bmatrix}\gamma_{i2}\\ \delta_{i2}\end{bmatrix} & \cdots & 0 \\
\vdots & \ddots & & \vdots & \vdots & \ddots & & \vdots \\
0 & \cdots & 0 & \begin{bmatrix}\alpha_{iN/2}\\ \beta_{iN/2}\end{bmatrix} & 0 & \cdots & 0 & \begin{bmatrix}\gamma_{iN/2}\\ \delta_{iN/2}\end{bmatrix}
\end{array}\right] \qquad (2)$$

où :

$$V_{i,j} = \begin{bmatrix}\alpha_{ij} & \cdots & \gamma_{ij}\\ \beta_{ij} & \cdots & \delta_{ij}\end{bmatrix} = \begin{bmatrix}\cos(\varphi_{i,j}) & \cdots & w_{i,j}\sin(\varphi_{i,j})\\ \sin(\varphi_{i,j}) & \cdots & -w_{i,j}\cos(\varphi_{i,j})\end{bmatrix},$$

$$w_{i,j} = \exp(j\theta_{i,j}), \quad \varphi \in [0, 2\pi], \quad \theta \in [0, 2\pi].$$

D'où la relation (1) peut s'écrire comme suit :

$$Y = \frac{1}{N} HX = \frac{1}{N} G_1 G_2 \ldots G_n X = \frac{1}{N} \prod_{i=1}^{n} G_i X \qquad (3)$$

En définissant les paramètres angulaires $\boldsymbol{\varphi}_{i,j}$ et $\boldsymbol{\theta}_{i,j}$, les opérateurs de transformations orthogonales $\boldsymbol{H}$ peuvent être formés avec des fonctions de base complexes, ou avec des fonctions réelles lorsque $\boldsymbol{\theta}_{i,j} = 0$. Le calcul des paramètres $\boldsymbol{\varphi}_{i,j}$ dépend du choix des structures des noyaux spectraux $\boldsymbol{V_{i,j}}$ (Abenaou et Sadik, 2011c). Ce qui permet de générer un système de fonctions de base adaptable à une classe de signaux donnée.

Or, dans la perspective d'assurer un calcul rapide, dans ce travail, les noyaux spectraux dans les matrices $\boldsymbol{G_i}$ sont constitués de telle sorte qu'ils contiennent une proportion plus importante de zéros, tel qu'il est expliqué ci-dessous.

L'adaptation de l'opérateur $\boldsymbol{H}$ (1) est assurée par la condition :

$$\frac{1}{N} H_a Z_{et} = Y_c = [y_{c,1}, 0, 0, \ldots 0]^T, \; y_{c,1} \neq 0 \qquad (4)$$

où :

- $\boldsymbol{Y}_c$ est le vecteur cible qui construit le critère d'adaptation de l'opérateur $\boldsymbol{H_a}$ ;
- $\boldsymbol{Z}_{et}$ représente le vecteur étalon d'une classe calculé par la moyenne des estimations statistiques des enregistrements de plusieurs signaux vocaux, d'un même mot, prononcés par divers locuteurs ;
- $\boldsymbol{H_a}$ est l'opérateur adaptable à synthétiser.

La synthèse de l'opérateur adaptable $\boldsymbol{H_a}$ à l'étalon $\boldsymbol{Z}_{et}$, pour une classe donnée, consiste à calculer les paramètres angulaires $\boldsymbol{\varphi}_{i,j}$ des matrices $\boldsymbol{G_i}$ selon la condition (4). La procédure du calcul des paramètres est illustrée par la figure 1 dont le principe est basé sur l'algorithme itératif introduit par la figure 2, qui permet le calcul du vecteur cible $\boldsymbol{Y_c}$ selon la relation :

$$Y_i = G_i Y_{i-1}$$

Le calcul du vecteur $\boldsymbol{Y_c}$ permet l'obtention de l'opérateur adapté $\boldsymbol{H_a}$. Pour la reconnaissance des signaux, nous devons disposer de deux ensembles d'enregistrements de signaux vocaux pour chaque mot. Le premier sert à calculer l'étalon $\boldsymbol{Z_{et,i}}$ du mot $i$ (classe $i$) et permet de générer la synthèse de l'opérateur. Tandis que le deuxième ensemble sert à former l'étalon spectral $\boldsymbol{Y_{et,i}}$ du mot $i$, qui est obtenu par la projection des enregistrements du deuxième ensemble dans les bases adaptables $\boldsymbol{H_a}$.

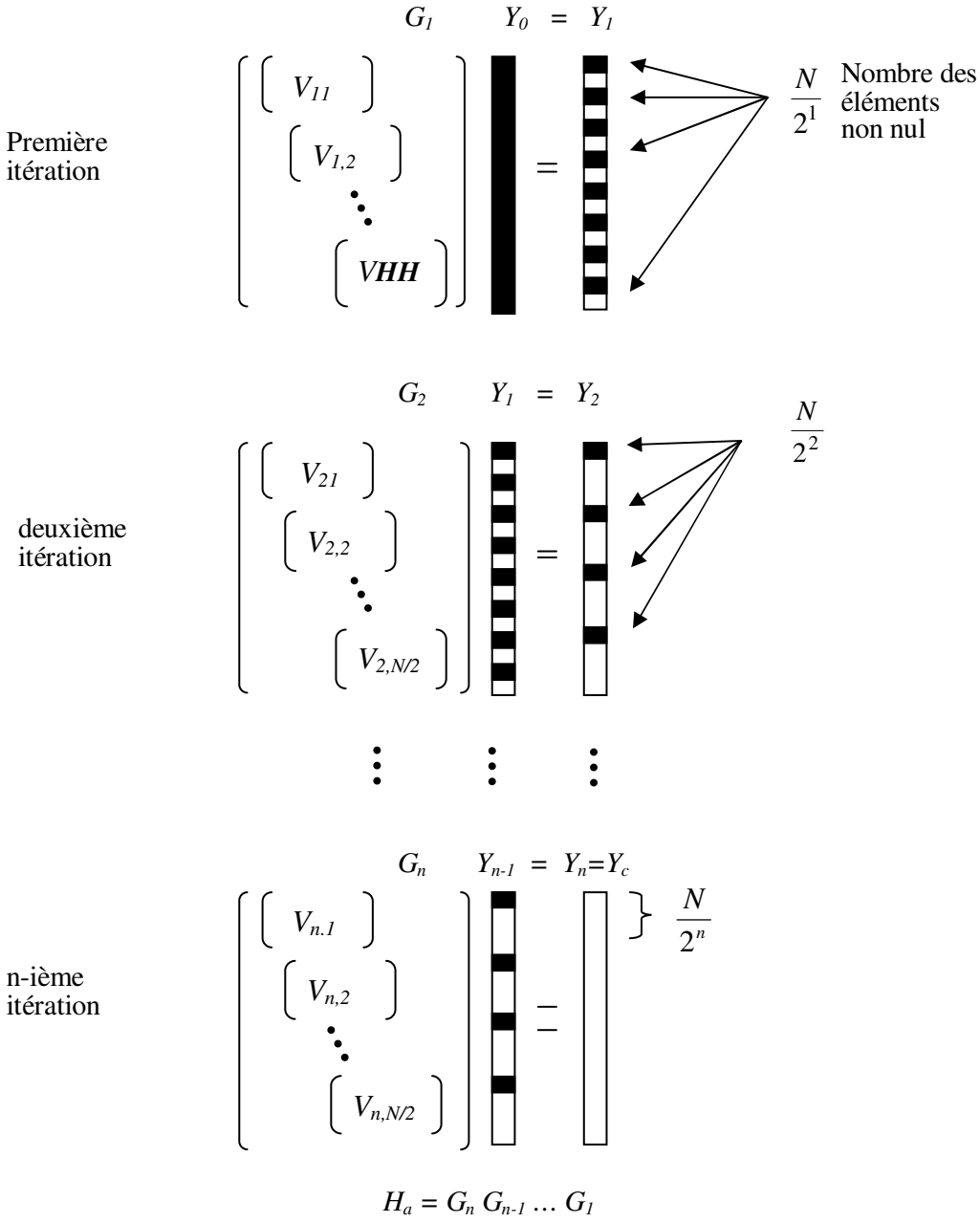


*Figure 1 : Schéma illustratif de la procédure de synthèse de l'opérateur de la transformée adaptable*

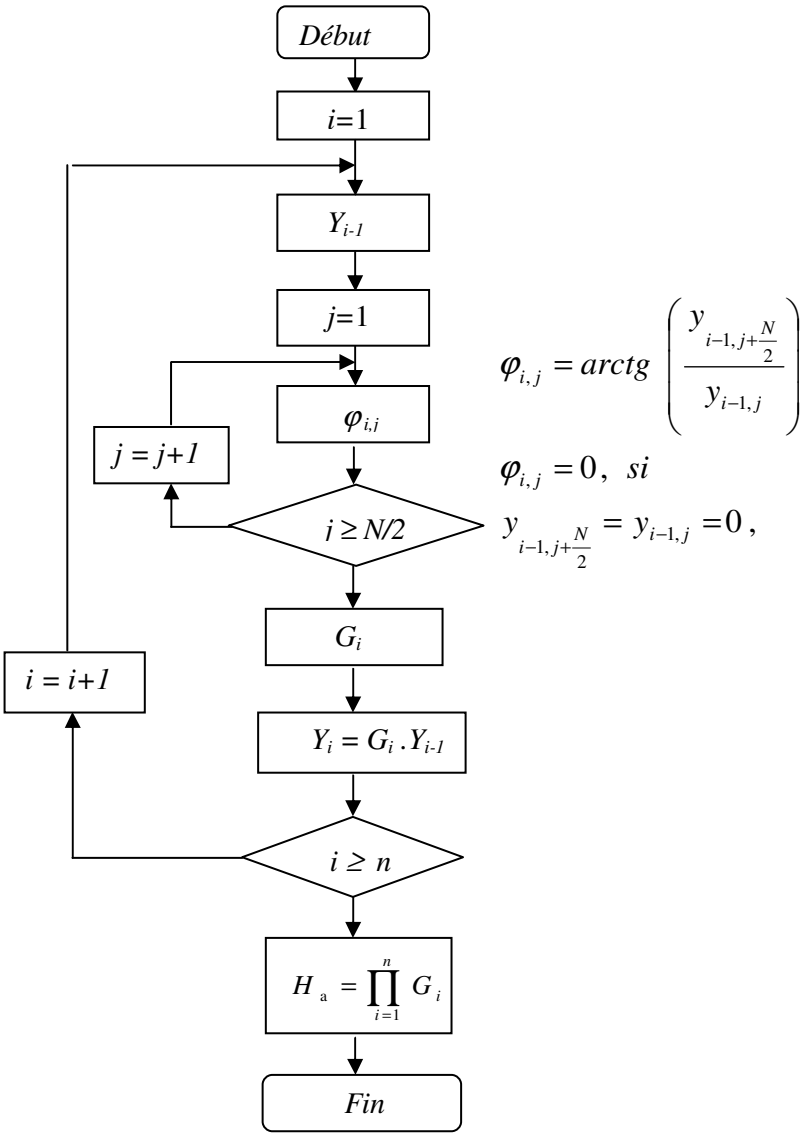


*Figure 2 : Schéma de l'algorithme de synthèse de l'opérateur de la transformée adaptable*

La reconnaissance d'un vecteur $\boldsymbol{Z}$ consiste à calculer son spectre $\boldsymbol{Y_i}$ dans chaque base $\boldsymbol{H_{a,i}}$. Pour définir le mot correspondant au vecteur $\boldsymbol{Y_i}$ des caractéristiques informatives, nous nous appuyons sur une règle de décision formée par une combinaison de deux critères :

- la distance euclidienne $\delta_i = \| \boldsymbol{Y_i} - \boldsymbol{Y_{et,i}} \|$ et
- la différence de l'énergie concentrée dans leurs premiers coefficients de la décomposition $\varepsilon_i = \left| y_{1,i}^2 - y_{1,et,i}^2 \right|$.

Ainsi, le vecteur $\boldsymbol{Y_i}$ correspondra au mot $i$ si $\delta_i = \min(\delta_{k=1..M})$ et $\varepsilon_i = \min(\varepsilon_{k=1..M})$, avec $M$ est le nombre de classes. Cette procédure de reconnaissance est illustrée par la figure 3.

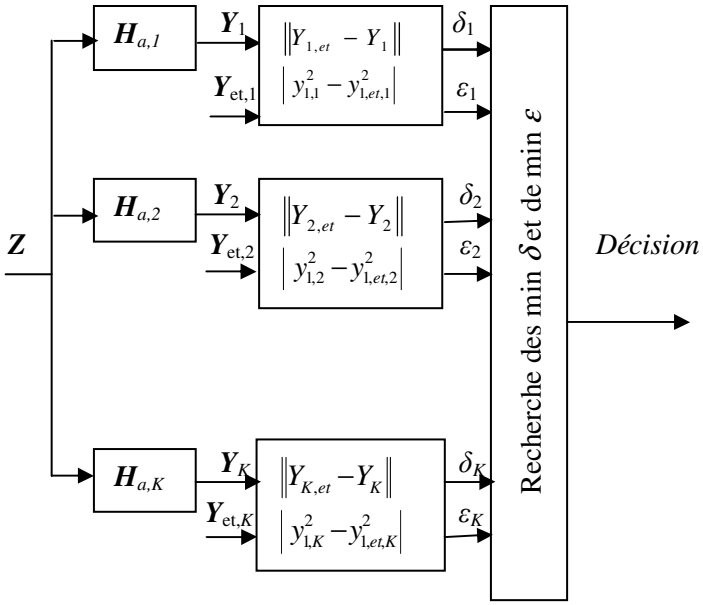


*Figure 3 : Procédure de reconnaissance*

## 3. La reconnaissance de la parole amazighe

Malgré l'avènement des technologies de la reconnaissance de la parole pour l'anglais, le français et l'arabe, des recherches approfondies au profit de la langue amazighe semblent insuffisantes et la mise en œuvre de ses applications est presque inexistante. D'où l'intérêt de la réalisation d'un système de reconnaissance de la parole en amazighe, en particulier un système qui pourra être dédié à l'apprentissage de la prononciation. Néanmoins, dans la perspective d'atteindre cet objectif, nous avons recours à un corpus qui caractérise la langue parlée.

### 3.1. Corpus

Vu la rareté et la non disponibilité des ressources électroniques en langue amazighe, particulièrement les corpus audio, nous avons recueilli, pour une première paramétrisation de notre système de reconnaissance de mots amazighes isolés, un

corpus de données vocalees multi-locuteurs. Ce dernier est constitué de 140 enregistrements des chiffres de 1 à 10, réalisés par trois locuteurs de différentes variétés régionales (Tarifit, Tamazight, Tachelhit).

En outre, ce corpus est regroupé en trois ensembles : le premier servira à calculer l'étalon de chaque mot pour générer la synthèse de l'opérateur ; le deuxième, à former les étalons spectraux des mots tandis que le troisième sera utilisé pour évaluer et analyser les performances de l'approche proposée.

### 3.2. Mesures d'évaluation

Afin d'évaluer la performance de notre système de reconnaissance, nous utilisons la mesure de taux d'exactitude par mot (Word Accuracy, WA), définies par la formule suivante (Sopheap, 2010) :

$$Taux\ d'exactitude = j/h * 100 \quad (5),$$

où $j$ correspond au nombre de mots justes et $h$ est le nombre total de mots.

### 3.3. Résultats expérimentaux

Pendant l'expérience, nous avons utilisé des enregistrements de signaux vocaux des mots amazighes isolés prononcés par différents locuteurs de diverses régions, ce qui a induit à un chevauchement assez considérable entre les classes des mots. Pour évaluer l'efficacité du système proposé, un test a été effectué pour la reconnaissance d'un même mot amazighe prononcé par divers locuteurs de diverses régions. La figure 4 illustre la projection du signal vocal du mot « ⵙⵉⵏ (*sin*) » (deux) dans les bases classiques (Walsh, Haar et Fourier). D'après cette figure, nous constatons que les spectres calculés du mot sont trop larges. Cependant, en utilisant la méthode proposée, nous remarquons une convergence rapide du spectre obtenu à l'aide des fonctions de base paramétrables.

Par ailleurs, grâce à l'application des transformations orthogonales paramétrables aux bases synthétisées, nous constatons que :

- l'énergie de la projection du signal dans la base adéquate est concentrée dans les premiers composants du spectre (figure 4) ; et
- la projection du signal d'un mot donné, qui caractérise une classe, dans d'autres classes (représentant d'autres mots amazighes) permet l'obtention de spectres assez larges dont l'énergie est dispersée sur plusieurs coefficients (figure 5).

Ce qui nous permet de reconnaître le mot prononcé avec une grande certitude.

En effet, les résultats de l'étude expérimentale de la méthode élaborée pour la reconnaissance des mots amazighes isolés indiquent, selon les courbes de la figure 6 qui présente les taux de certitude de la reconnaissance des signaux, une efficacité considérable par rapport aux autres méthodes qui sont basées sur l'application des transformations spectrales dans les bases traditionnelles (Walsh, Haar et Fourier).

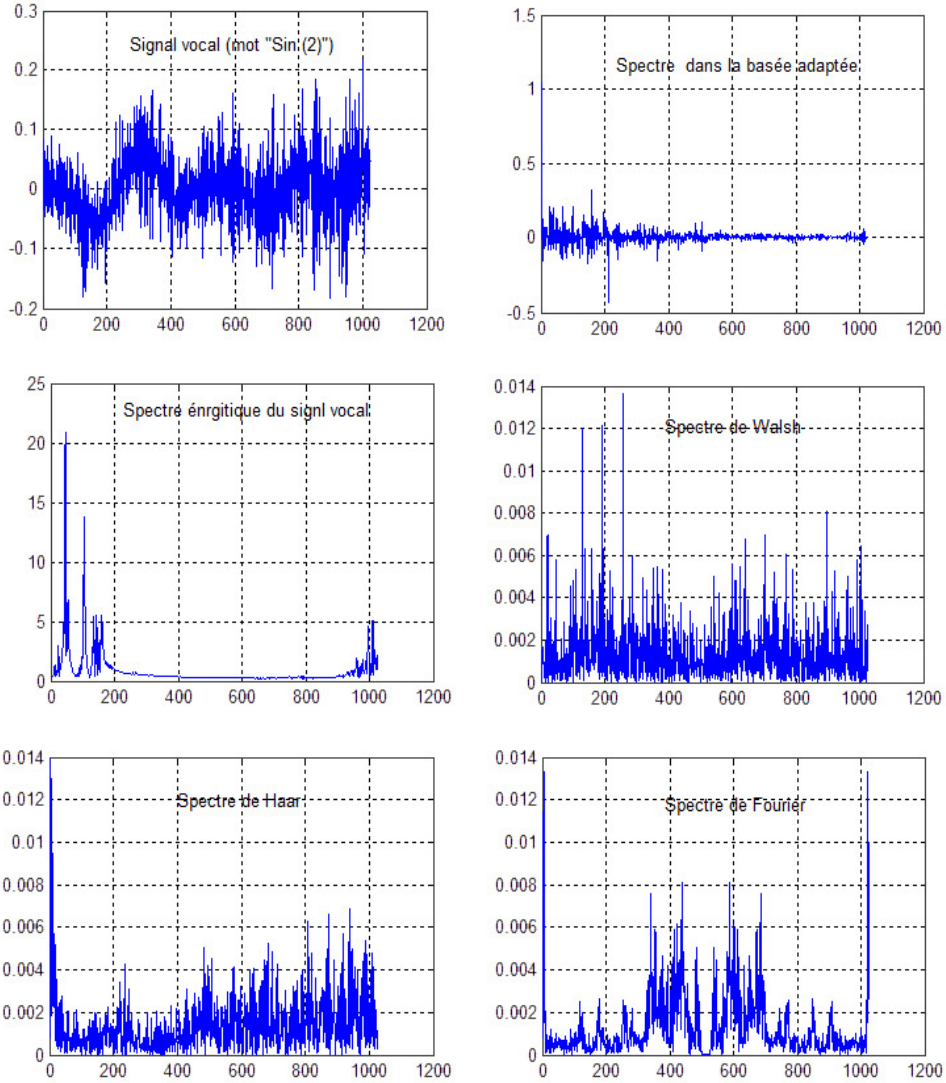


*Figure 4 : Projection du signal vocal (fragment du mot ⵙⵉ / « Sin » - 2)*

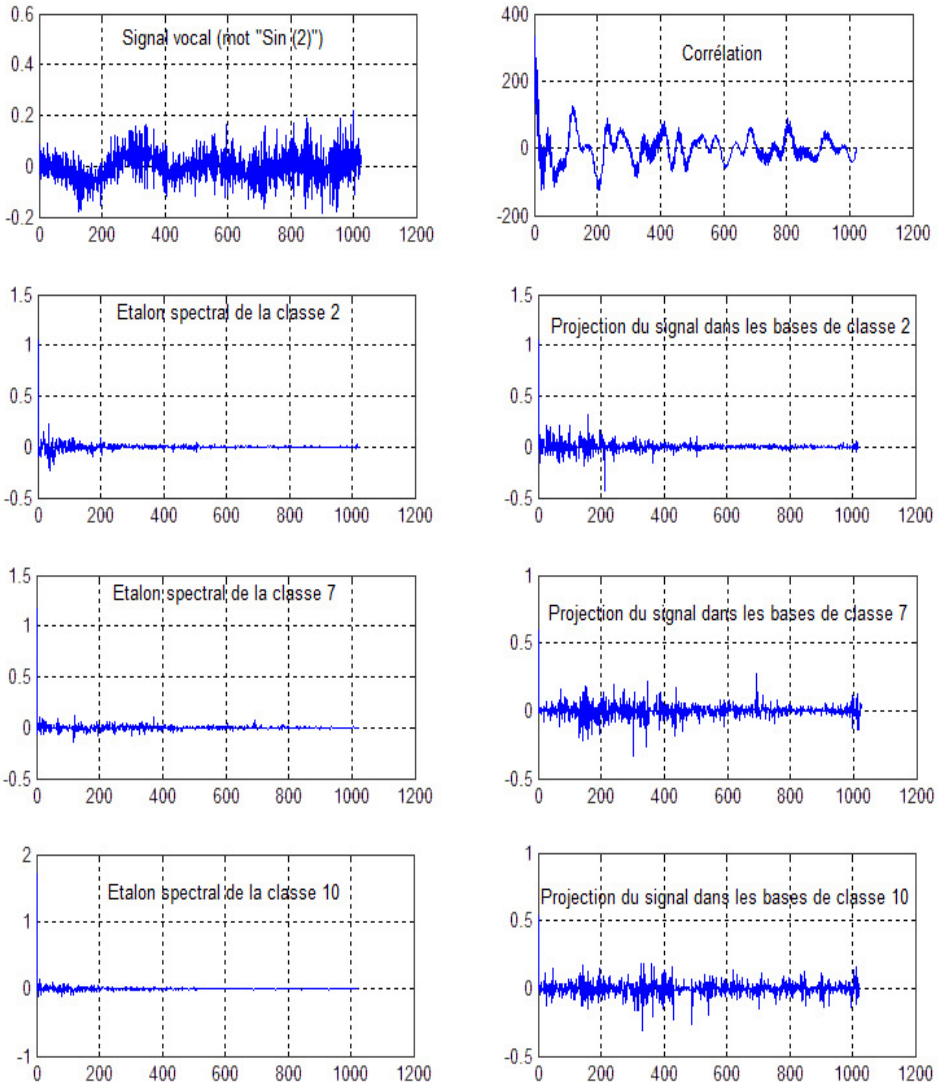


*Figure 5 : Calcul du spectre du signal vocal à l'aide des fonctions de bases paramétrables*

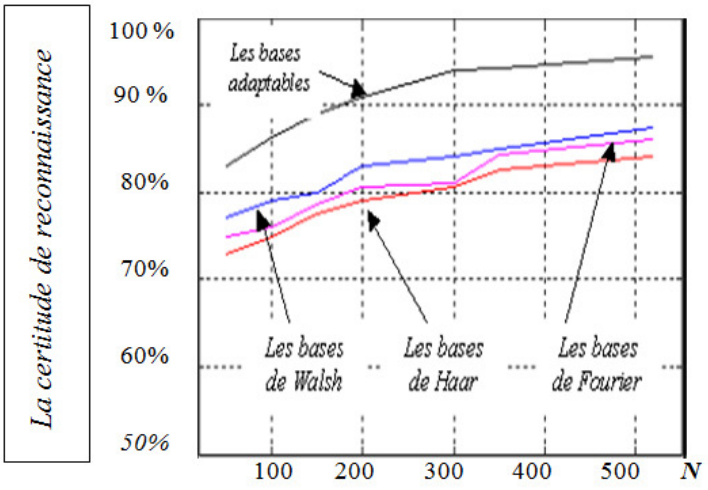


*Figure 6 : Résultat de la reconnaissance lors de l'application de divers systèmes de fonctions de base*

A partir de ces courbes, nous pouvons constater que dans le cas de l'utilisation des bases traditionnelles, la certitude de reconnaissance des signaux des mots amazighes ne dépasse pas 87%. Tandis que dans le cas où nous utilisons les fonctions de bases adaptables, le taux de reconnaissance des signaux s'élève à 96% lorsque la taille de l'intervalle de l'analyse est égale à 512.

Ce qui peut être expliqué par le fait que :

- la méthode proposée est basée sur l'utilisation des fonctions de bases adaptables selon le caractère du signal vocal du mot prononcé par les divers locuteurs des différentes régions ; et
- la propriété de sélectivité des fonctions de base synthétisées simplifie la distinction des signaux dans l'espace des caractéristiques informatives.

## Conclusion

Dans la présente contribution, nous proposons un système de reconnaissance automatique des mots isolés de la langue amazighe basée sur les transformations orthogonales paramétrables. Ce dernier est composé de deux sous-systèmes : un sous-système d'apprentissage et un sous-système de reconnaissance. Le sous-système d'apprentissage est conçu à la base d'un corpus multi-locuteurs de la parole amazighe de différentes régions afin de prendre en considération la diversité de la prononciation d'un même mot et de contribuer à la stabilité des caractéristiques statistiques dans le calcul de l'étalon de chaque mot. En outre, ce sous-système nous offre la possibilité de synthétiser des fonctions de base adaptables de chaque mot avec une propriété de sélectivité plus importante, ce qui

nous a permis d'extraire les caractéristiques les plus informatives de chaque mot prononcé indépendamment du locuteur.

Suite à l'étude comparative réalisée sur le système à base des transformations orthogonales paramétrables et sur les transformations orthogonales de Walsh, Haar et Fourier, nous avons pu atteindre un taux de reconnaissance plus élevé qui tend vers les 96%. Cependant, nous considérons que ce travail est une première initiative pour la réalisation d'un système de reconnaissance de la parole amazighe assurant l'apprentissage de la prononciation, qui suscite l'intérêt de recueillir un corpus oral riche et varié composé de mots amazighes dédiés à l'apprentissage de la langue.

## Références bibliographiques

Abenaou A. et Sadik M. (2011), « Elaboration d'une méthode de compression des signaux aléatoires à base d'une transformation orthogonale paramétrable avec algorithme rapide », *The Fourth Workshop on Information Technologies and Communication (WOTIC'11), ENSEM*, Casablanca.

Abenaou A. et Sadik M. (2011), « Méthode et algorithme de formation d'un système de fonctions de base adaptables pour le diagnostic des signaux biologiques », *Colloque International des Telecom'2011 & 7èmes Journées Franco-Maghrébines des Micro-ondes et leurs Applications*, Tanger.

Abenaou A. et Sadik M. (2012), « Méthode et algorithme d'identification des signaux vocaux à base des transformations orthogonales adaptables », *Network Security and Systems*, p. 41-45.

Ahmed N. et Rao K.R., (1975), *Orthogonal transforms for digital signal processing*, Springer Ber.

Bello J. et *al.* (2004), "*On the use of phase and energy for musical onset detection in the complex domain", IEEE Signal Procesing letters*, 11(6) : 553–556.

Bahi H. (2008), "Hybrid ASR system for teaching pronunciation". ICL Conference, 24 -26 Septembre, Villach, Austria.

Barnard E. et *al.* (2009), "Asr corpus design for resource-scarce languages". In *Interspeech*.

Barnard E. et *al.* (2010). Voice search for development. In Interspeech.

Bourlard H. et Morgan N. (1993), "Continuous speech recognition by connectionist statistical methods", *IEEE Transaction on Neural Networks,* Vol. 4, n° 6, pp. 893-909.

Cappe O. (1995), « Etat actuel de la recherché en reconnaissance du locuteur et des application en criminalistique », Rapport interne, Ecole Nationale des Télécommunications, Département du Signal, Paris.

Doddington G.R. (1985), "Speaker reconnaissance-identification people by their voices", Proc. IEEE; vol 73; no.11; p. 1651

Doets P. et Lagendijk R. (2004), "Theoretical modeling of a robust audio fingerprinting system", In *IEEE Benelux Signal Processing Symposium.*

Good I.J. (1960), The interaction algorithm and practical Fourier analysis, J. Roy. Statist. Soc. Ser. B, B-20, 361-372, 1958, B-22, 372-375.

Kekre H. B. et *al.*, "Performance Comparison of Speaker Identification Using DCT, Walsh, Haar On Full And Row Mean Of Spectrogram", *International Journal of Computer Applications*, August 2010 Edition, in press.

Kumar A. et *al.* (2011), Rethinking speech recognition on mobile devices. In IUI4DR. ACM.

Patel N. et *al.* (2010), "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural India", In *CHI*, pages 733–742, ACM.

Sopheap S. (2010), *Vers une modélisation statistique multi-niveau du langage, application aux langues peu dotées*. Thèse de doctorat, Université de Grenoble..