Synthèse de la parole en Arabe Standard Par Réseaux de Neurones Artificiels

Salim Djeghiour, Mhania Guerti

الملخّـص

إنّ طرق التركيب الآليّ للكلام كثيرة؛ منها ما يتمّ تركيب الكلام فيه باستعمال قواعد يجمعها ويستنبطها لغويّون متمكّنون، ومنها ما يُعتمد على وحدات صوتيّة مزدوجة مسجّلة مسبقا لتكوين الجمل المطلوبة. الطريقة الثانية هي المعتمدة في دراستنا هذه، بتطبيق الشّبكة العصبونيّة الاصطناعيّة متعدّدة الطبقات، وقد اختيرت اللّغة العربيّة الفصحى، حيث جرى تحويل جمل مكتوبة -أخذت في سياقات مختلفة- إلى سلسلة من الفونيمات.

في المرحلة التّالية، استعملنا مولّدا للأصوات، ينطق الفونيمات الصّادرة عن الشبكة العصبونيّة، وهذا بطريقة تركيب الأصوات المزدوجة، هذه الوحدات الصوتية سجّلت مسبقا في قاموس معدّ لذلك.

الكلمات المفاتيح : تركيب الكلام، الشّبكة العصبونيّة الاصطناعيّة، الأصوات المزدوجة، اللّغة العربيّة الفصحى.

Synthèse de la parole en Arabe Standard Par Réseaux de Neurones Artificiels

Salim Djeghiour[1], Mhania Guerti[2]

[1,2]Laboratoire Signal et Communications
Ecole Nationale Polytechnique, Alger, Algérie.
d.salim3@yahoo.fr, mhania.guerti@enp.edu.dz

Résumé

Les méthodes de synthèse de la parole sont nombreuses, parmi les principales, nous pouvons citer la synthèse par règles et la synthèse par concaténation d'unités phonétiques.

Dans le cadre de ce travail, nous optons pour cette dernière méthode en lui appliquant des Réseaux de Neurones Artificiels (RNA). Le choix de la langue sera l'Arabe Standard (AS). Pour cela, nous utilisons un RN qui réalise la transformation d'une phrase écrite en une suite de phonèmes. Les graphèmes à transformer sont pris dans un corpus constitué par des mots et des phrases en AS. Ensuite, nous employons un générateur de sons qui réalise la prononciation des phonèmes fournis par le réseau et ceci à l'aide de la synthèse par concaténation de diphones qui sont préenregistrés dans un dictionnaire.

Mot clés: Synthèse de la Parole par diphones, Réseaux de Neurones Artificiels, Arabe Standard.

1. Introduction

La parole étant le moyen de communication le plus naturel chez l'Homme, celui-ci a très vite cherché à l'intégrer dans les interfaces Homme-Machine. Cela a été rendu possible grâce aux efforts consentis en reconnaissance et en synthèse de la parole, alors que la première permet à la machine de traiter des informations fournies oralement par un utilisateur humain, la seconde est un ensemble des procédés utilisés pour passer d'un message textuel à son correspondant sonore.

Malgré les avancées réalisées ces dernières années dans ces domaines, des progrès restent à faire pour accroître le confort d'utilisation des systèmes actuels.

Les systèmes de synthèse ne sont réellement sortis des laboratoires pour des applications commerciales que depuis environ une trentaine d'années.

Avec la diffusion des potentialités de l'ordinateur, la popularisation d'Internet et l'émergence de la société de l'information, la communication Homme-Machine voit croître la part de la parole. En particulier, de nouvelles technologies donnent un regain d'intérêt à la synthèse de la parole à partir du texte, pour répondre aux besoins des applications embarquées (automobile, traducteurs automatiques de parole, etc.), des télécommunications (services de consultation de courrier électronique par téléphone, serveurs vocaux interactifs, livres et journaux parlants) et du multimédia (jeux informatiques, aide aux handicapés, etc.).

2. La synthèse de la parole

Un système de synthèse à partir du texte est une machine capable de lire a priori n'importe quel texte à voix haute, que ce texte ait été directement introduit par un opérateur sur un clavier alphanumérique, qu'il ait été scanné et reconnu par un système de reconnaissance optique des caractères (OCR : Optical Character Recognition), ou qu'il ait été produit automatiquement par un système de dialogue Homme-Machine [1].

Les principales méthodes de synthèse de parole sont :

- la synthèse par concaténation : cette approche consiste à synthétiser le signal par concaténation d'unités sonores naturelles ;
- la synthèse par règles : les synthétiseurs par règles ont principalement la faveur des phonéticiens et des phonologistes. Ils permettent une approche cognitive, générative du mécanisme de la phonation. Ils sont basés sur l'idée que, si un phonéticien expérimenté est capable de lire un spectrogramme, il doit lui être possible de produire des règles qui permettent de créer un spectrogramme artificiel pour une suite de phonèmes donnés.

3. Réseau de Neurones Artificiels (RNA)

Les Réseaux de Neurones Artificiels consistent en un ensemble d'outils et de méthodes de calcul, pouvant être appliqués dans divers domaines, tels que le traitement d'information, la classification de données, la statistique, le traitement de signal (image, parole), la prédiction de séries temporelles ou encore le contrôle, les nombreuses applications industrielles.

Un Réseau de Neurones Artificiels est formé d'un grand nombre de cellules élémentaires simples, fortement interconnectées.

La cellule élémentaire est appelée "neurone" car son fonctionnement est fondé sur celui d'un automate proposé comme une approximation du fonctionnement du neurone biologique [2-3]. La sortie de cette cellule est une fonction non linéaire de la somme pondérée de ses entrées. Une forme analytique très courante pour la décision est la fonction sigmoïde, mais d'autres fonctions peuvent également être utilisées, la réponse finale est calculée selon la formule suivante :

$$S_i = f\left(\sum_{j=1}^{n} w_{ij} x_j\right) \qquad (1)$$

Le neurone formel peut être représenté comme suit dans Figure 1.

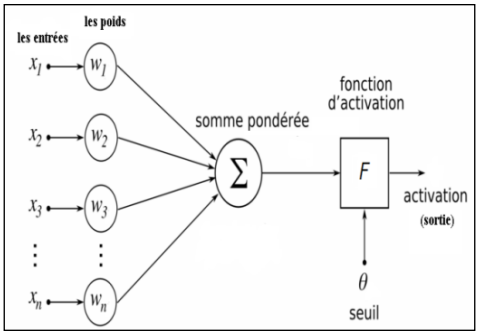

Figure 1: Principe d'un neurone artificiel

4. Mise en œuvre des RN dans la Synthèse de la Parole

Le développement d'un système de Synthèse de la Parole (SP) utilisant les Réseaux de Neurones (RN) de type Multi Layer Perceptron (MLP), implique nécessairement un ajustement des paramètres tels que l'architecture du réseau, l'initialisation des poids synaptiques, le nombre d'itérations et les constantes d'apprentissage, etc.

L'ajustement de ces paramètres se fait plus d'une manière empirique que par une

méthode basée sur un fondement théorique [2, 4-7]. Ces paramètres ajustés nous permettent d'optimiser certains critères dont nous citons les plus courants :

- l'amélioration du pouvoir de généralisation (les performances du système neuronal) ;
- la minimisation de l'architecture (réduction de la quantité de calculs en phase d'utilisation) ;
- la minimisation du temps d'apprentissage.

4.1. Analyse des données

Dans la synthèse de la parole, il est nécessaire d'effectuer une analyse des données de manière à déterminer les caractéristiques discriminantes pour détecter ou différentier ces données. Ces caractéristiques constituent l'entrée du réseau de neurones.

4.2. Initialisation des poids synaptiques

L'initialisation des poids avant l'application de l'algorithme d'apprentissage par rétropropagation du Gradient dans un système de SP est une tâche importante, car elle influe sur la vitesse de convergence du réseau [8]. Plus les poids initiaux sont proches de leur valeur finale et plus la convergence est rapide. En effet, quand ces poids sont trop faibles ceci entraîne un apprentissage très long. On peut distinguer deux méthodes d'initialisation : les méthodes d'initialisation aléatoires dans un intervalle choisi de manière adéquate et celles basées sur des techniques non aléatoires. En ce qui concerne notre travail, nous nous sommes limités aux algorithmes basés sur la rétropropagation du gradient.

4.3. Architecture du réseau

Nous avons utilisé le Perceptron Multi Couches (MLP), connu comme étant le type de réseau le plus répandu et le plus utilisé, vues la simplicité de sa structure et la rapidité de son apprentissage. L'architecture de réseau doit correspondre au problème qu'il est conçu pour résoudre [9]. Trouver une architecture adéquate d'un RN à un problème donné, n'est pas une tâche simple, car le nombre optimal de couches cachées ainsi que le nombre de neurones dans chaque couche et leurs connexions se fait plus de manière empirique que par une méthode basée sur un fondement théorique. Une méthode appropriée consiste à utiliser un réseau très petit (exemple, un neurone dans la couche cachée) puis ajouter des neurones jusqu'à l'obtention de bonnes performances [10].
Le MLP que nous avons utilisé contient une seule couche cachée avec 100 neurones. La couche d'entrée forme une fenêtre glissante sur le jet de données à l'entrée. Elle contient trois groupes de neurones. Chaque groupe contient 22 neurones (Figure 2).
Le phonème associé à un graphème dépend du contexte de ce dernier, la donnée fournie au réseau est constituée de trois-graphèmes, le graphème à traiter, sa précédente et sa suivante. Chacun des contextes est représenté par un vecteur de caractéristiques articulatoires, associé à un phonème. Une suite de trois graphèmes suffit à déterminer le phonème associé au graphème central. La couche de sortie contient 34 neurones, correspondant aux 34 phonèmes de l'AS, 28 consonnes et 6 voyelles, comme indiqué ci-dessous :

[ء, ب, ت, ث, ج, خ, ح, د, ذ, ر, ز, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, خ, ل, م, ن, ه, و, ي].

On fait la correspondance selon la transcription phonétique de l'auteur [11] :

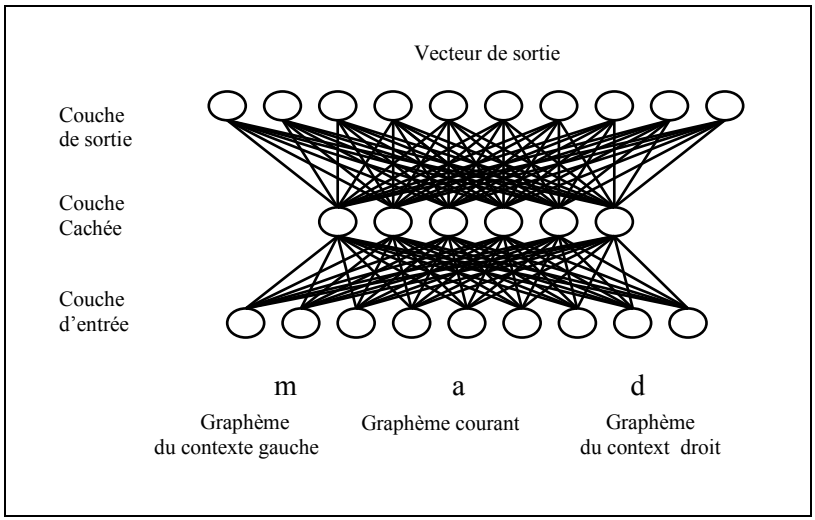

Figure 2: Architecture du module Orthographique Phonétique.

[?, b, t, ç, G, K, H, d, D, r, z, S, c, $, µ, T, £, §, E, f, q, k, l, m, n, h, w, y] en plus les 3 voyelles courtes et 3 voyelles longues (a, A, i, I, u, U).

Une cellule de la couche de sortie doit être active (état >0.7) correspondant au phonème désiré).

5. Elaboration du Corpus

L'élaboration du corpus d'apprentissage et de test, le choix de ce dernier n'est pas arbitraire, et afin de tester les performances de notre système de synthèse, nous avons pris un corpus constitué des mots et des phrases contenant des graphèmes de l'AS pris dans des différentes positions (Initiale, Médiane et Finale) (Table .1). Nous avons choisi 40 mots et phrases pour la phase d'apprentissage.

1 ان الوضوح نقيض الغموض	21 صار
2 جمع مالا	22 الولد
3 ظلم الظالمين	23 صار الولد
4 القمر جميل	24 طاف
5 غمام كثيف	25 رجل
6 غداء لذيذ	26 ظفر # بي
7 ظل ظلالا	27 سار رجل
8 حاد يمينا	28 صار #رجل
9 جمع غفير	29 جعله حطاما
10 جموع غفيرة	30 بهره جماله
11 غابة كثيفة	31 نهر كبير
12 نزل المطر	32 طليق حر
13 سكت سكوتا	33 ألم #أبي
14 أخذ قلما	34 خبر #مصدر
15 طبع المطبوعات	35 محيط #سقط
16 المسجد العظيم	36 نضيف #مريض #ديك
17 غسل	37 مجيب #لبيب
18 شمال	38 قام كمال من مجلسه
19 جمال	39 أمر فظيع
20 قال	40 قادر

Table 1. Corpus d'apprentissage.

1 ثمن القسمة	14 غلاف
2 أكلت التفاح	15 سار
4 ذهب الى المدرسة	16 جلس زيد
5 نظر في وجهه	17 مدينة وهران
6 كتاب مفيد	18 ثمن بخس
7 جاء عمر	19 طبيب
8 قرأ رسالة	20 صيدلي
9 ضربه طويلا	21 طماطم
10 سنم من الحياة	22 صلصال
11 جمع	23 شعب
12 العمل عبادة	24 الولد ذاهب
13 ظمأ العطش	25 قضم التفاحة

Table 2. Corpus de test.

Concernant la phase de test nous avons choisi 24 mots et phrases (Table 2).

6. Elaboration du dictionnaire de diphones

Le diphone est, par définition, "un élément sonore caractéristique de la transition entre deux phonèmes s'étendant de la partie stable d'un phonème à la partie stable du phonème suivant " [12]. Pour élaborer le dictionnaire de diphones, il faut d'abord enregistrer des mots en utilisant le logiciel Praat, qui nous donne la représentation temps-fréquence et énergie du mot ou de la phrase. Après avoir enregistré les sons, nous procédons à la segmentation. La segmentation a été effectuée manuellement, cette phase joue un rôle très important, de bonnes conditions de ladite phase induit de bonnes performances de système de synthèse de la parole. Chaque segment sonore représente un diphone, à la fin de l'opération, nous disposons d'un certain nombre de diphones dans le dictionnaire (*.wav), contient environ 68 diphones indiqués ci-après:

[#r, #S, $aa, £a, 0$, 0£, 0b, 0E, 0G, 0m, 0q, 0S, 0T, a0, a#, aaf, aal, aam, aar, aG, ak, al, am, aq, ar, aT, bI, Ei, f0, fa, fi, Ga, Gl, Gu, hi, i0, ih, il, in, i0, ir, iS, IT, ka,

l0, laa, li, lu, ma, maa, mi, mu, n0, n#, nm, qa, qaa, ra, Sa, Saa, Si, T#, T#, Ta, Taa, uH, ul, un].

7. Apprentissage et test de généralisation

L'apprentissage est la propriété la plus intéressante des RN. L'apprentissage est une phase du développement d'un réseau de neurones durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré.

La synthèse de la parole par les RN de type MLP nécessite un apprentissage supervisé, en utilisant l'algorithme de rétropopagation du gradient comme méthode d'apprentissage. Si un système neuronal possède un nombre trop grand de neurones de la couche cachée par rapport à celui des exemples de la base d'apprentissage, tous les exemples seront parfaitement appris: on parle d'apprentissage par cœur ou surparamétrisation [2]. En ce qui concerne notre travail, l'apprentissage est réalisé en présentant au réseau une suite de triplets de graphèmes, correspondant à une suite de phonèmes. Exemple 2 : pour la phrase "$Ara#raGulun", nous présentons la suite :

P$A, $Ar, Ara, ra#, a#r, #ra, raG, aGu, Gul, ulu, lun, unP. (P représente un espace)

Associée à la suite de phonèmes :

$, A, r, a, #, r, a, G, u, l, u, n. (# représente une virgule).

Exemple 2 : "maGd", on présente la suite :

Pma, maG, aGd, GdP. (P représente un espace).

Associée à la suite de phonèmes :

m, a, G, d.

	Voisé/non voisé	Occlusive	Fricative	Nasale	Vibrante	Latérale	Afriquée	Semi – voyelle	Longue/courte	Antérieure/Postérieure	Fermée/Ouverte	Bilabiale	Labio-dentale	Dentale	Alvéo-dentale	Palatale	Vélaire	Uvélaire	Pharyngale	Laryngale	Emphatique	Voyelle
M	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
A	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1
G	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
D	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

Table 3. Exemple de codage d'un graphème.

Par exemple : le tableau 3 donne les codes des graphèmes [m], [a], [G] et [d] selon les caractéristiques articulatoires associées à leurs phonèmes correspondants.

Plusieurs stratégies d'apprentissage ont été testées :

- apprentissage progressif : on apprend à phrase prononcer une phrase, puis on ajoute une autre et on recommence l'apprentissage, et ainsi de suite ;
- constitution d'un corpus d'apprentissage, en sélectionnant dans un ensemble de phrases les différents triplets, sans répétition. Ces triplets peuvent être ensuite présentés dans l'ordre d'apparition dans les phrases ou dans un ordre aléatoire, variable à chaque étape ;
- apprentissage global : on donne directement toutes les phrases et on apprend en parcourant à chaque étape l'ensemble des phrases.

Parmi ces méthodes, nous choisissons la dernière, car elle est rapide.

7.1. Choix des paramètres d'apprentissage

Le choix des paramètres d'un algorithme d'apprentissage influe beaucoup sur la rapidité de calculs.

Dans le cas de l'algorithme de rétropropagation, le calcul du gradient consiste à définir, la direction dans laquelle doit s'effectuer la modification des poids.

Après l'étape de l'initialisation des poids synaptiques par des valeurs comprises entre -0.5 et 0.5, un apprentissage est effectué sur tout le corpus d'apprentissage avec un pas de 1, un taux d'erreur égal 0.01 et un nombre d'itérations égal 800.

8. Description du Générateur de Sons

Le matériel utilisé est une carte audio adaptable sur un micro-ordinateur. Cette carte permet d'échantillonner des sons, de les stocker sous cette forme et de les restituer.

La fréquence d'échantillonnage est fixée à la valeur :

$$f_e = 11025Hz \qquad (2)$$

Nous donnons quelques exemples des résultats de Transcription Orthographique Phonétique (TOP), en utilisant notre réseau de neurones (Table 4).

TO (1)	**TP** (2) Désirée	**TP** Obtenue
EilAf	EilAf	EilAf
Sara	Sara	Sara
$Ara	$Ara	$Ara
$Ara lwaladu	$Ara lwaladu	$Ara lwala0u
TAfa	TAfa	TAfa
raGulun	raGulun	raGulun
£afira # bi	£afira # bi	£afira # bi
Ga§alahu huTaman	Ga§alahu huTaman	Ga§alahu 0uTaman
nahrun kabIrun	nahrun kabIrun	nahrun kabIrun

(1) **TO** : Transcription Orthographique
(2) **TP** : Transcription Phonétique

Table 4. Résultats de la TOP par le RN.

9. Conclusion

Avec le corpus de test, nous avons testé les capacités du réseau sur un texte comprenant 200 triplets ou des mots qui ne figurent pas dans le corpus d'apprentissage. Le réseau associe la bonne réponse dans 96% des cas (Tab 4), ceci nous permet d'avoir un taux de synthèse de 71,42%, ce qui prouve les bonnes capacités de généralisation d'une telle solution et permet de ne pas exiger l'apprentissage sur un corpus complet. D'autre part, si nous traçons la courbe du nombre de triplets inconnus en fonction du nombre de caractères d'un texte nouveau, nous constatons que le pourcentage de triplets inconnus va décroître, ce qui signifie que le nombre de triplets nouveaux à apprendre est limité. Les phrases ou les sons prononcés par notre système sont intelligibles, dans la mesure où nous comprenons bien ce qui est énoncé, car nous avons pris en considération lors de la phase de segmentation les informations importantes transportées par les transitions entre phonèmes.

Nous suggérons des perspectives à notre travail telles que :

- l'augmentation du corpus d'apprentissage ;
- l'augmentation du nombre de diphones échantillonnés.

10. Références

[1] Dutoit, T., "Introduction au Traitement Automatique de la Parole3, Notes de cours, Faculté polytechnique de Mons, Belgique, 2000.

[2] Jadouin, J. F., "Les réseaux de neurones principes et définitions", Ed. Hermes, 1994.

[3] Héraut, J. and Jutten,C., "Réseaux neuronaux et traitement de signal", Ed. Hermes, 1994.

[4] Bunet, L., "Traitement Automatique de la Parole en milieu bruité : Etude de modèles connexionnistes statique et dynamique", Thèse de Doctorat, Université Henri Poincaré-Nancy1, Spécialité informatique, France, 1997.

[5] Dutoit, T., "Introduction au Traitement Automatique de la Parole", Notes de cours, Faculté polytechnique de Mons, Belgique, 2000.

[6] Jadouin, J. F., "Les réseaux de neurones principes et définitions", Ed. Hermes, 1994.

[7] Héraut, J. and Jutten, C., "Réseaux neuronaux et traitement de signal", Ed. Hermes, 1994.

[8] Bunet, L., "Traitement Automatique de la Parole en milieu bruité : Etude de modèles connexionnistes statique et dynamique", Thèse de Doctorat, Université Henri Poincaré-Nancy1, Spécialité informatique, France, février 1997.

[9] Boutou, L., "Reconnaissance de la Parole par réseaux multi-couches3, International Workshop on Neural Networks and Their Application", 197-217, 1988.

[10] Selouani, S., ”Reconnaissance automatique de la parole par des techniques multi-agents, connexionnistes et hybrides : application à la langue Arabe”, Thèse de Doctorat d’Etat, USTHB, Alger, Algérie, 2000.

[11] Boutou, L., ”Une approche théorique de l’apprentissage connexionniste, Application à la reconnaissance de la Parole”, Thèse de Doctorat, Paris Sud, France, 1991.

[12] Moutard, F., ”Introduction aux Réseaux de Neurones”, Ecoles des Mines de Paris, France, 2003.

[13] Vainio, M., ”Artificiel neural network based prosody models for Finnish Text-To-Speech Synthesis”, Univesity of Helsinki, Departement of Phonetics, Finland, 2001.

[14] Nguyen, L. and Widrow, B., ”Approving the learning speed of Tow – Layer Neural Network by chosing initial values of the adaptation weights”, International Joint Conference Occidentale, 1991.

[15] Hamdani-Droua, G., ”Prédiction des Durées des Phonèmes de l’Arabe Standard”, thèse de Magister, Ecole Nationale Supérieure des Sciences Humaines (ENSSH), Alger, Algérie, 2004.

[16] Emerard, F., ”Synthèse par diphones et traitement de la prosodie”, thèse de Doctorat, Université de Grenoble III, France, 1997.