Un Modèle Prédictif de la Durée Segmentale pour la Synthèse de la Parole Arabe à Partir du Texte

A. Zaki [1,2], A. Rajouani [2], M. Najim [1]

[1] Equipe Signal et Image-LAP UMR 5131, ENSERB. B.P 99, F-33 402 Talence Cedex, France
[2] LEESA, Faculté des Sciences. B.P 1014 - Rabat, Maroc
Tél.: ++33 556 84 61 85 - Fax: ++33 556 84 84 06
e-mail: {zaki,najim}@tsi.u-bordeaux.fr, arajouani@yahoo.fr

ABSTRACT

This paper deals with a neural-network based model of segmental duration for a TTS Arabic system. Given a set of factors influencing phoneme duration, a Multi-Layer Perceptron (MLP) is used to predict phoneme duration. Different linguistic features are extracted automatically from the text and coded for networks with binary and analog input nodes. The correlation coefficient measured of the generalization test database is 0.882. This coefficient corresponds to 14.3 ms as a mean absolute prediction error of segmental duration.

1. INTRODUCTION

Le développement de l'utilisation de la synthèse de la parole dans des services qui nécessitent une interaction conviviale PERSONNE-MACHINE requiert plusieurs étapes de traitement. L'amélioration du naturel de la parole de synthèse (fluidité, prosodie) figure comme un traitement prioritaire qui fait l'unanimité aussi bien des industriels que des utilisateurs. C'est un traitement primordial situé au niveau du traitement linguistico-prosodique pour tout système de synthèse à partir du texte. Du point de vue phonétique, il s'agit du traitement des paramètres prosodiques définis par : la fréquence fondamentale (F0), la durée segmentale et l'intensité. La modélisation de ces paramètres a fait l'objet de plusieurs travaux portant essentiellement sur la fréquence fondamentale et, dans une moindre mesure, sur la durée. Par contre le paramètre intensité a été peu étudié pour les recherches en prosodie [Lac99].

Le manque de fluidité et, par conséquent, de naturel de la parole synthétique, est dû pour une grande partie à un traitement inadéquat du rythme et de la durée segmentale. Le contrôle de l'organisation temporelle de l'énoncé nécessite la mise en jeu d'un modèle prédictif pour différents aspects temporels tels que le débit, la durée des pauses et la vitesse d'articulation. On s'intéresse dans cette communication à la prédiction de la durée segmentale. Sa plus grande difficulté de mise en œuvre, est due, indépendamment de la langue étudiée, à l'interaction complexe d'une multitude de facteurs. En effet, les variations temporelles sont régies par de multiples paramètres qui correspondent à des niveaux d'analyse différents (paralinguistique, intrinsèque et co-intrinsèque, linguistique) et qui font de la durée un paramètre très difficile à interpréter [Ros81].

Dans le contexte de la synthèse de la parole à partir du texte et dans la mesure où les domaines d'application visés correspondent essentiellement à des situations de dialogue contrôlé exemptes de toute improvisation, les corpus utilisés dans ce cas sont des corpus lus. La construction des corpus dépend aussi de plusieurs facteurs que l'on peut contrôler pour la modélisation de la durée segmentale. Dans la synthèse de la parole, l'intérêt primordial d'un modèle de durée réside dans sa capacité de prédire des durées relativement proches des durées optimales à partir de toutes les combinaisons de facteurs linguistiques possibles.

Différentes méthodes ont été appliquées pour la modélisation de la durée segmentale pour la synthèse de la parole à partir du texte. On peut distinguer deux tendances de modélisation : les modèles basés sur un système par règles [Kla79] et les méthodes statistiques, telles que celles fondées sur les réseaux de neurones [Rie95] ou les méthodes de régression [Rie97].

L'objectif de cette étude est le développement d'un modèle de la durée segmentale basé sur les réseaux de neurones qui peut être intégré, avec le modèle de génération des variations de F0 [Zak01], au niveau du bloc de traitement automatique linguistico-prosodique.

L'approche neuronale est basée sur l'apprentissage automatique qui consiste à faire le lien entre les informations linguistiques reflétées par le texte et la durée segmentale.

Les réseaux de neurones ont été utilisés avec succès pour la modélisation de plusieurs systèmes et en particulier ceux dédiés au traitement acoustico-linguistique : prononciation (grapheme-to-phoneme) [Sej87], génération de F0 [Sco89]. L'utilisation des réseaux de neurones pour la modélisation de la durée syllabique a été proposée par [Cam90].

2. TRAITEMENT DE LA DUREE SEGMENTALE

Le modèle présenté dans cette communication consiste à prédire la durée segmentale en utilisant des facteurs qui affectent la durée. La figure 2 illustre l'organigramme du modèle prédictif.

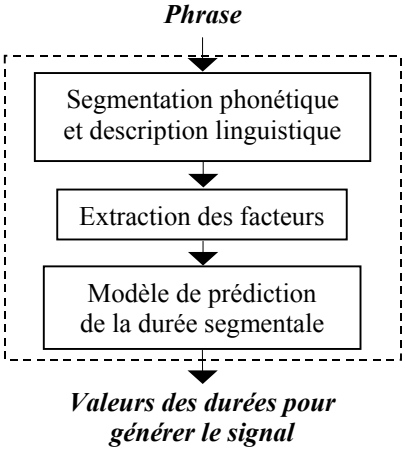

Figure 1 : schéma général du modèle de traitement automatique de la durée segmentale

Le modèle de prédiction de la durée est composé de deux blocs distincts. Le premier est dédié au traitement linguistique, le deuxième permet de transformer les informations issues du traitement linguistique en données quantitatives. Le bloc de

traitement linguistique permet une description du texte acquis. Cette description fournit des informations sur la nature de la phrase, syllabe, accentuation, type de phonèmes, frontière de mot etc. Ces informations feront l'objet des facteurs qui influencent la durée segmentale. Les facteurs sont choisis à partir d'une analyse de la durée segmentale.

2.1. Analyse de la durée segmentale

L'analyse de la durée segmentale consiste à identifier quelques effets du contexte immédiat sur la durée des phonèmes. Il s'agit, essentiellement de l'effet du contexte consonantique sur la durée vocalique. En effet, les syllabes dans la langue arabe sont basées sur des éléments contrastés situés à l'intérieur de la frontière de la syllabe. Chaque syllabe a une partie principale saillante. Cette partie est connue par *le noyau de la syllabe* qui est la *voyelle*. Les éléments restants sont appelés *les facteurs marginaux* et sont représentés par les *consonnes*. Une syllabe commence toujours par une seule consonne et se termine soit par une consonne soit par deux consonnes soit par aucune consonne. Cette définition de la syllabe de l'arabe met en évidence le rôle principal de la voyelle au sein de cette unité phonologique. Ce qui explique la priorité de l'étude des variations temporelles des voyelles par rapport aux consonnes.

Une étude statistique a été réalisée par [Raj89] sur un corpus composé de mots prononcés dans une phrase porteuse. Le corpus a été lu par deux locuteurs. Le premier locuteur ayant un débit normal et le deuxième ayant un débit relativement lent. Ce choix permet de vérifier l'influence du débit sur l'analyse de la durée segmentale. Les conclusions dégagées à partir des expériences réalisées pour mesurer l'effet du contexte consonantique sur la durée vocalique et l'effet des facteurs morphologiques et phonologiques sont :

- une voyelle précédée d'une consonne géminée est relativement plus longue que celle précédée d'une consonne simple ;
- une voyelle suivie d'une consonne géminée est plus brève que celle suivie d'une consonne simple ;
- la pharyngalisation n'affecte pas la durée vocalique ;
- les voyelles sont plus longues en syllabe ouverte ;
- la voyelle est plus longue lorsqu'elle est précédée d'une consonne sonore ;
- la durée de la voyelle brève n'est pas affectée par le voisinage d'une voyelle longue que la consonne intervocalique soit simple ou géminée ;
- la durée d'une consonne géminée est plus longue lorsqu'elle est suivie d'une voyelle longue ;
- la durée d'une consonne géminée est plus brève lorsqu'elle est précédée d'une voyelle longue ;
- l'influence du rythme est plutôt quantitative que qualitative ;
- l'accentuation affecte la durée de la voyelle selon le type de syllabe ouverte/fermée. La durée de la voyelle d'une syllabe ouverte et accentuée est plus longue que la durée d'une syllabe fermée ;
- une voyelle située avant une pause est plus longue que dans les autres positions ;
- le nombre de syllabes dans le mot affecte la durée des voyelles. La durée décroît avec une moyenne de 10 ms pour chaque voyelle du début à la fin du mot ;
- la durée des voyelles des mots grammaticaux est inférieure à celle des voyelles des mots lexicaux.

Cette analyse a permis la construction d'un ensemble de règles pour la modélisation de la durée segmentale. Le modèle réalisé a été dédié à un système de synthèse par règles [Raj89]. Dans cette communication, on propose d'exploiter cette analyse pour choisir l'ensemble des facteurs qui affectent la durée segmentale, ainsi que l'ordre des informations contextuelles dont il faut tenir compte pour prédire la durée d'un phonème cible.

A partir des conclusions soulignées auparavant, on peut extraire les facteurs qui influencent la durée segmentale : accent lexical, gémination, nature phonétique des consonnes, type de syllabe, position par rapport à la pause, nombre de syllabes du mot, type de mot. A cet ensemble de facteurs, on ajoute : la classe de chaque phonème, l'effet de la liaison phonologique, et d'autres facteurs d'ordre phonotactique. Il est clair que pour prédire la durée d'un phonème au niveau d'une phrase, il faut tenir compte au moins du phonème précédent et de celui subséquent.

3. APPROCHE NEURONALE

Dans cette approche, on utilise un réseau de neurones standard en l'occurrence le Perceptron multicouche (PMC). C'est un réseau de neurones artificiel à couches cachées. Le réseau neuronal nécessite une étape d'apprentissage supervisé en se basant sur des données segmentées de la parole naturelle. Chaque vecteur caractérisant la durée du phonème et son contexte phonétique est fourni au réseau. En même temps, la durée du phonème cible est présentée à la sortie du réseau. L'algorithme d'apprentissage utilisé est celui de retro-propagation de l'erreur [Hay94].

3.1. Base de données

Le corpus utilisé dans cette application est similaire à celui que nous avons utilisé précédemment pour l'étude de l'intonation de la langue arabe standard [Zak01]. Il s'agit d'un corpus composé de 112 phrases déclaratives de taille variant de 1 à 10 mots. Le corpus contient : 395 mots (10% de mots grammaticaux et 90% lexicaux), 1013 syllabes (524 CV, 129 CVV, 356 CVC, 4 CVVC) et 3575 phonèmes.

3.2. Codage des paramètres

Selon la catégorie de chaque facteur caractérisant la durée segmentale, nous avons utilisé différentes méthodes de codage des paramètres présentés au réseau neuronal.

- **Codage binaire** : c'est un codage standard pour les paramètres vrais/faux ;
- **One-of-n** : on utilise *n* neurones et un seul parmi eux sera activé, celui-là correspond à une classe ou à une catégorie ;
- **Transformation en pourcentage** : cette méthode consiste à diviser la valeur en cours par la valeur maximale pour obtenir un pourcentage. Il s'agit des valeurs en virgule flottante en entrée.

3.3. Evaluation du réseau

Pour évaluer et comparer l'ensemble des réseaux considérés durant les simulations, nous avons utilisé le coefficient de corrélation calculé entre les durées segmentales prédites et optimales.

Le coefficient de corrélation est défini par :

$$\rho_{x,y} = \frac{Cov\ (x,y)}{\sigma_x \sigma_y}$$

σ_x et σ_y sont les écarts types.

où : $-1 \le \rho_{xy} \le 1$

et $Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$

3.4. Paramètres d'entrées

Les paramètres d'entrées sont extraits automatiquement du texte en se basant sur une hiérarchie de modules : syllabation, classification de types de syllabes, accentuation, classification phonétique, acoustique et articulatoire, ainsi que la détermination des paramètres de position.

Nous avons classé les paramètres d'entrée en trois catégories, correspondant aux types de codage :

- La première catégorie qui est de type binaire comprend les facteurs suivant :
  - la gémination (consonne géminée ou simple),
  - la liaison phonologique (existence d'une liaison entre deux mots juxtaposés ou non, /daxa**lal**waladu/),
  - la pause (phonème situé avant une pause ou non),
  - type de mot (grammatical ou lexical).
- La deuxième catégorie comprend :
  - la classe des phonèmes (Voyelle longue/courte, classe phonétique des consonnes : occlusives, fricatives, nasales, etc. avec la précision du voisement ou non). Ce facteur est codé sur 10 bits,
  - le type de syllabe (CV, CVV, CVC, CVVC, CVCC). Ce facteur est codé sur 5 bits,
  - le niveau d'accent (Accent principal, secondaire et tertiaire). Ce facteur est codé sur 3 bits,
  - la position du phonème dans la syllabe (de 1 à 4). Ce facteur est codé sur 4 bits.
- La troisième catégorie comprend les autres paramètres de position :
  - la position du phonème dans le mot (dépend du nombre de phonèmes dans chaque mot),
  - la position du phonème dans la phrase (dépend du nombre de phonèmes total dans la phrase),
  - la position de la syllabe dans le mot (dépend du nombre de syllabes dans le mot),
  - la position de la syllabe dans la phrase (dépend du nombre total de syllabes dans la phrase),
  - la position du mot dans la phrase (dépend du nombre de mots dans la phrase).

Pour cette dernière catégorie, les facteurs sont codés avec des valeurs analogiques résultant de la transformation en pourcentage.

3.5. Paramètres de sortie

Nous avons choisi un codage linéaire de la durée observée à la sortie du réseau neuronal. Ce type de codage transforme linéairement les durées segmentales pour qu'elles soient comprises dans l'intervalle [0,1]. L'utilisation de la fonction sigmoïde[1] permet cette adaptation d'intervalle.

[1] $f(x) = \frac{1}{1+\exp(-\alpha x)}$

3.6. Architecture et mise en œuvre

Comme nous l'avons souligné auparavant, le réseau utilisé est le PMC. Pour notre application, nous avons déduit de tests de performances qu'une seule couche cachée est suffisante. Ce choix est bien commenté dans [Zak01]. En tenant compte de tous les facteurs présentés en 3.4, chaque phonème sera caractérisé par 13 paramètres. L'influence d'une voyelle voisine, quelle que soit la consonne intervocalique, suggérée le choix d'une fenêtre qui inclut les informations contextuelles de trois phonèmes : le phonème cible, le phonème précédent et celui subséquent. Par conséquent la fenêtre d'entrée est composée de 13*3 facteurs qui seront codés sur 32*3 bits. Cela nécessite 96 neurones au niveau de la couche d'entrée. La couche de sortie du réseau est composée d'un seul neurone qui correspond à la durée du phonème cible de la fenêtre d'entrée. Le choix du nombre de neurones de la couche cachée est effectué d'une manière empirique. Nous avons testé plusieurs nombres de neurones variant de 5 à 50 pour identifier la meilleure architecture. Le choix du meilleur réseau est fait selon les coefficients de corrélation mesurés sur les données d'apprentissage et celles de test. Le réseau présente les meilleures performances pour un choix de 30 neurones dans la couche cachée.

3.7. Expériences et résultats

L'algorithme d'apprentissage dépend de plusieurs paramètres qui ont un caractère aléatoire, tel que l'initialisation des poids de connexions, le pas de l'algorithme et le moment[2]. Nous avons expérimenté plusieurs paramètres sur un réseau à une seule couche cachée composée de 30 neurones, une couche d'entrée à 96 neurones et une couche de sortie à un neurone. Durant la phase d'apprentissage effectuée sur 75% de la base de données totale, on mesure pour chaque cycle les coefficients de corrélation sur les données d'apprentissage et sur les données de test qui représentent 25% de la base de données totale. Pour éviter le phénomène de sur-apprentissage « over learning », on suit l'évolution des deux coefficients de corrélation mesurés. L'algorithme doit être arrêté à partir de l'itération où l'on remarque une dégradation du coefficient de corrélation mesuré sur les données de test[3]. La courbe de la figure 2 illustre l'évolution des coefficients de corrélation pour 300 itérations. On remarque qu'au bout de la 100ème itération, il y a apparition du phénomène de sur-apprentissage comme on peut le voir sur la figure 2. Dans cette zone, le réseau atteint ses meilleures performances avec un coefficient de corrélation d'apprentissage de l'ordre de 0.889. Celui mesuré sur des données de test est de l'ordre de 0.882. L'erreur de prédiction absolue mesurée dans ce cas est de l'ordre de 14.3 ms.

Il faut noter que le modèle neuronal peut être utilisé aussi bien pour l'analyse que pour la synthèse de la durée segmentale. L'étape d'analyse consiste à évaluer la contribution de chaque facteur à la prédiction de la durée segmentale. Pour réaliser cette démarche on utilise la méthode du variant simple proposée par [Cam90]. La méthode consiste à désactiver alternativement les neurones qui correspondent aux facteurs à tester et calculer le coefficient de corrélation. Les résultats de ce test sont présentés dans le tableau 1.

[2] Le moment η est un coefficient introduit dans l'équation d'adaptation des poids pour accélérer l'apprentissage du réseau neuronal et pallier le problème d'instabilité. $\Delta w_{ij}(n) = \alpha \Delta w_{ij}(n-1) + \eta \delta_j(n) y_i(n)$.

[3] Il s'agit des données qui ne figurent pas dans la base d'apprentissage.

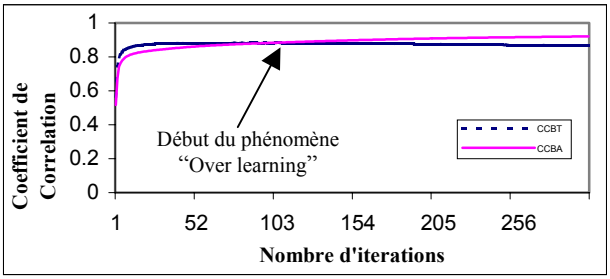

Figure 2 : évolution des coefficients de corrélation au cours de la phase d'apprentissage en fonction du nombre d'itérations. CCBA et CCBT représentent respectivement les coefficients de corrélation mesurés sur les données d'apprentissage et de test.

Table 1 : résultat du test d'analyse des facteurs.

Facteurs désactivés	Coef_Cor_Test
Classe des phonèmes	0.3022
Position du phonème, syllabe et mot dans la phrase	0.3521
Position du phonème dans la syllabe	0.4574
Position du phonème dans le mot	0.5845
Type de la syllabe	0.6054
Position de la syllabe dans la phrase	0.6104
Type de mot	0.6125
Position de la syllabe dans le mot	0.6145
Position du phonème dans la phrase	0.6158
Position du mot dans la phrase	0.6246
La gémination	0.6451
Niveau d'accent	0.7965
La pause	0.8064
La liaison phonologique	0.8101

Le tableau 1 représente des résultats préliminaires d'une analyse par synthèse de la durée segmentale. La contribution des facteurs est présentée dans l'ordre croissant du coefficient de corrélation du test. On remarque que la classe des phonèmes est un facteur déterminant. Ce résultat est tout à fait évident car le facteur en question est l'identificateur de chaque segment. Les facteurs de position dans la phrase figurent en deuxième position. Nous avons désactivé en même temps les trois facteurs de position concernant le placement du phonème, syllabe et mot dans la phrase. On remarque que les performances du réseau dans ce cas se détériorent complètement. En désactivant un seul des trois paramètres liés à la position dans la phrase, les performances du réseau s'améliore : le coefficient de corrélation passe de 0.3521 à plus de 0.6. La position du phonème dans la syllabe est un paramètre qui contribue considérablement aux performances du réseau neuronal. Tel est également le cas, pour le type de mot et de syllabe. En ce qui concerne le niveau d'accent on remarque qu'il n'influence pas beaucoup le modèle. Cela peut s'expliquer par son effet limité à la voyelle uniquement. Pour ce qui est de la pause et la liaison phonologique, leur absence n'influence pas beaucoup les performances du réseau, mais ces facteurs restent nécessaires pour assurer de bons résultats.

4. CONCLUSION

Nous avons présenté dans cette communication, les différentes étapes pour la réalisation d'un modèle prédictif de la durée segmentale de la langue arabe standard, pour la synthèse à partir du texte. Nous nous sommes basés sur une analyse statistique pour identifier les paramètres qui affectent la durée segmentale. Nous avons utilisé l'approche neuronale pour la synthèse de la durée segmentale à partir des facteurs qui affectent la durée des phonèmes. L'analyse par synthèse nous a permis de classer la contribution des paramètres utilisés. Cette analyse préliminaire est en accord avec certains résultats obtenus par analyse statistique. Les paramètres de position contribuent fortement aux performances du modèle prédictif. Les résultats sont encourageants. On estime que le modèle actuel peut être simplifié davantage si on arrive à analyser la contribution de chaque facteur en tenant compte de toutes les combinaisons possibles. Pour obtenir un bon modèle prédictif on propose d'analyser les performances du modèle en tenant compte de différentes méthodes de codages des paramètres d'entrée et de sortie. Le modèle de prédiction de la durée segmentale proposé est testé par un synthétiseur de la parole arabe basé sur la technique TD-PSOLA. Des exemples qui illustrent des résultats du modèle sont disponibles sur le site de l'Equipe Signal et Image[4]

5. BIBLIOGRAPHIE

[Cam90] W. Campbell "Analog I/O nets for Syllable Timing", in speech communication, vol 9, pp 57-61, North-Holland, 1990.

[Hay94] S. Haykin. *Neural Networks. A Comprehensive Foundation*. IEEE Computer Society Press. 1994.

[Kla79] D. Klatt. "Synthesis by Rule of Segmental Durations in English Sentences". In Frontiers of Speech Communication Research, edited by B. Lindblom and S. Öhman (Academic, London), pp. 287-301, 1979.

[Lac99] A. Lacheret-Dujour et F. Beaugendre. *La prosodie du français*. CNRS Editions. Paris 1999.

[Raj89] A. Rajouani. *Contribution à la synthèse de la parole Arabe par Règles*. Thèse de Doctorat d'Etat, Université Mohamed V, Faculté des Sciences Rabat, Maroc 1989.

[Rie95] M. Riedi "A Neural-Network-Based Model of Segmental Duration for Speech Synthesis". Proceedings of Eurospeech'95 conference, pp. 599-602, Madrid, 1995.

[Rie97] M. Riedi "Modelling Duration With Multivariate adaptive Regression Splines". Proceedings of Eurospeech'97 Conference. Volume 5, pp.2627 – 2630. Rhodes 1997.

[Ros81] M. Rossi et Al. *L'intonation de l'Acoustique à la Sémantique*. KLINCKSIECK. Paris 1981.

[Sco89] M. S. Scordilis and J. N. Gowdy "Neural Network Based Generation of Fundamental Frequency Contours". Proc. IEEE-ICASSP, Vol. 1, pp. 219-222, Glasgow, 1989.

[Sej87] T. J. Sejnowski and C. R. Rosenberg "Parallel Networks that Learn to Pronounce English Text". Complex Systems, Vol. 1, pp. 145-168, 1987.

[Zak01] A. Zaki, A. Rajouani, M. Najim, "Synthesizing Intonation of Standard Arabic Language Using Neural Networks", Proc. of Eurospeech'01 Conference. Volume 1, pp. 541-544, September, Aalborg, 2001.

[4] http://www.tsi.u-bordeaux.fr/zaki/arabic-synthesis-demo.htm