Discrimination des langues suivant leur classe rythmique par un réseau récurrent temporel

Jean-Marc Blanc, Peter Ford Dominey

Institut des Sciences Cognitives
UMR 5015 CNRS-Université Claude Bernard Lyon 167, boulevard Pinel 69675 BRON cedex
Mél: blanc@isc.cnrs.fr, dominey@isc.cnrs.fr

ABSTRACT

Rhythmic properties of speech could play an important role in language acquisition. Indeed, newborns acquire rhythmic structure of speech, so that they can discriminate languages from different rhythmic classes [16]. Dominey and Ramus [7] have shown that a Temporal Recurrent Network (TRN) inspired by neurophysiological data can deal with succession of consonants and vowels to simulate this discrimination. However these phonemes were segmented by hand. We suggested here that this network could be employed to process directly speech sounds via a time frequency representation of low frequencies corresponding to speech prosody.

1. INTRODUCTION

L'amorçage prosodique est souvent proposé comme base de l'acquisition du langage et de la syntaxe. Sous cette hypothèse, les indices prosodiques (intonation, rythme) permettent de mettre en valeur des éléments clefs du langage et guideraient les nourrissons lors de l'apprentissage du langage. La sensibilité des nouveau-nés pour la prosodie a déjà été démontrée en particulier lors de la discrimination des langues [16]. Notre objectif est de montrer qu'un réseau récurrent temporel (TRN) distinguent les langues (Anglais, Japonais et Néerlandais) suivant leur classe rythmique uniquement à partir du signal de parole.

2. LE RYTHME

2.1 Perception du rythme

Le rythme est induit par la perception d'événements réguliers comme les voyelles. La perception des nourrissons pourrait être focalisée sur la voyelle, car ils se montrent plus attentifs aux voyelles qu'aux consonnes [1]. Ils représenteraient alors la parole sous la forme d'une succession de voyelles, entrecoupées de bruits non analysés, les consonnes.

Le rythme se définit également par une succession d'unités faibles ou fortes. Dans le discours, ces unités se traduisent par des accents, résultats d'une combinaison complexe de l'intonation, l'intensité et la durée. Les nourrissons différencient des stimuli multisyllabiques qui se distinguent uniquement par la position de l'accent (rythmes trochaïque et ïambique) [12]. Quel système leur permet-il d'analyser des structures temporelles ?

2.2 Traitement des séquences temporelles

Historiquement, les réseaux de neurones ont été utilisés pour des problèmes de classification ou d'approximation de fonctions sur des données statiques. Pour modéliser de façon réaliste les réseaux de neurones biologiques, l'aspect dynamique du problème doit être pris en compte, soit de façon **explicite**: le temps est réduit à une succession d'évènements discrets sans durée [9]; soit **implicite**: le temps est échantillonné avec un pas très inférieur aux unités étudiées [14, 7].

Dans ces deux cas, l'état du réseau dépend non de l'entrée courante, et des états précédents. Seulement, les méthodes classiques de descente du gradient sont limitées par la taille de l'historique à prendre compte. Ainsi, l'apprentissage devient impossible pour de longues périodes, puisque le contexte se retrouve mal représenté. En augmentant la taille de la fenêtre d'analyse, les détails de la description du signal pourrait être perdues.

Les recherches en neurophysiologie ont permis le développement de modèles neuromimétiques dédiés au traitement implicite de l'information temporelle. Deux hypothèses [13] prédominent pour expliquer le codage de l'information par les impulsions nerveuses: le **codage par fréquence** privilégie la fréquence moyenne d'émission de décharges [4, 7]; le **codage temporel** emploie la structure temporelle de ces impulsions [13, 14]. Ces modèles ont principalement été testés avec des séquences temporelles abstraites [4, 8]; ou sur des fragments de discours (phonèmes [4] ou mots isolés [13, 14]). Quelle tâche pourrait être traitée afin de démontrer leurs capacités pour l'analyse des structures rythmiques de la parole ?

3. DISCRIMINATION DES LANGUES

Les structures rythmiques de la parole se définissent en fonction de la langue employée. Trois classes rythmiques sont communément décrites : langues accentuelles (Anglais, Arabe, Néerlandais), syllabiques

(Espagnol, Français, Italien) et moraïques (Japonais, Tamoul) [6]. Dans les langues accentuelles, le rythme est perçu par le biais des syllabes accentuées, alors qu'il se manifeste dans chaque syllabe pour les langues syllabiques. Le rythme des langues moraïques se défini à partir des mores, unité caractéristique du Japonais.

3.1. Discrimination perceptuelle

Dans les expériences perceptuelles d'identification des langues, la prosodie est souvent citée par les sujets comme étant caractéristique des langues : l'intonation pour le Français [15], l'Italien [20] ou pour les langues asiatiques [15] et le rythme pour l'Espagnol [15]. Les propriétés phonétiques et phonotactiques sont rendues inexploitables de façon à ne laisser passer que les informations prosodiques (Filtre passe-bas). Sous ces conditions, les nourrissons distinguent l'Anglais du Japonais, mais pas l'Anglais du Néerlandais [17]. Ainsi, la discrimination est possible quand les langues appartiennent à des classes rythmiques distinctes. Comment le rythme de la parole peut-il être traité pour retrouver les différences rythmiques entre les langues ?

3.2. Discrimination automatique

Étiquetage manuel du signal

Les propriétés rythmiques des langues ont été étudiées au travers de statistiques basées sur la durée des consonnes et des voyelles, obtenues par un découpage manuel du signal de parole. Deux de ces variables statistiques sont le pourcentage d'intervalles vocaliques %V qui correspond à la durée des intervalles vocaliques divisée par la durée totale de la phrase, et l'écart type des durées d'intervalles consonantiques au sein de la phrase C. Elles permettent de regrouper les langues selon les trois grandes classes rythmiques. Une régression logistique à partir de %V est utilisée pour discriminer les paires de langues en fonction des classes rythmiques auxquelles elles appartiennent [19].

Étiquetage automatique du signal

Galvès et coll. [11] se basent sur l'idée que les nourrissons utilisent des catégories grossières pour représenter les consonnes et les voyelles. Effectivement, le filtrage passe-bas simulant la prosodie ne permet pas de retrouver très précisément les voyelles. Ils introduisent donc une fonction de sonorité, qui reconnaît les motifs réguliers du signal acoustique à partir de l'entropie locale. Ils obtiennent alors la classification rythmique des langues sans faire appel à une segmentation manuelle. Pellegrinno et coll. [17, 10] définissent également la voyelle à partir des fréquences basses. Le signal est d'abord segmenté à partir de ruptures acoustiques. Les voyelles sont alors identifiées à partir de l'énergie contenue dans les basses fréquences d'un spectrogramme. Cette technique de segmentation a été appliquée à l'identification de 5 langues européennes [17, 10].

4. MATÉRIEL ET MÉTHODES

4.1 Corpus

Trois langues ont été extraites du corpus LSCP [15]: le Japonais (langue moraïques), l'Anglais, et le Néerlandais (langues accentuelles). Chaque langue fait appel à quatre locuteurs différents : deux de ces locuteurs composent le corpus d'apprentissage, les deux autres constituent le corpus de validation.

4.2 Le Réseau Récurrent Temporel

Ce modèle est initialement basé sur le système frontostriatale du primate, pour apprendre des ordres sensorimoteurs et pour simuler l'activité neurophysiologique pendant une tâche accomplie par des primates non humains [8]. Le contexte des événements passés est représenté grâce à des boucles récurrentes qui permettent à l'information présentée au moment t d'influencer la représentation des nouvelles informations au moment $t+1$.

L'utilisation d'un tel réseau récurrent temporel pour l'apprentissage de séquences n'est pas nouvelle [9]. Cependant, avec les méthodes de type récurrent couramment utilisées, la structure temporelle (et donc prosodique) ne peut pas être traitée indépendamment de la structure sérielle. Chaque neurone est un intégrateur à fuite et les connexions impliquées entre les couches d'entrées et les couches récurrentes sont fixes et caractérisées par une constante de temps [8]. Un tirage aléatoire du poids de ces connexions est effectué pour déterminer un individu-réseau.

A chaque fin de phrase, le motif d'activation formé par les deux couches cachées constituées de 25 neurones chacune est enregistré. L'apprentissage correspond à la création d'un prototype issu de la moyenne des vecteurs d'une même langue. Pendant la validation, une phrase est reconnue comme appartenant à la catégorie dont le prototype est le plus proche de l'activation qu'elle a engendrée dans la couche *State*. Les performances sont indiquées pour une population de 10 réseaux.

4.3 Obtention de la fréquence fondamentale

Trois méthodes d'extraction de la fréquence fondamentale (F0) ont été testées avec le réseau récurrent temporel. Chacune de ces méthodes est effectuée à l'aide du logiciel PRAAT. Elles permettent d'assigner l'activité d'un neurone d'entrée à l'énergie contenue dans une bande de fréquences, définies différemment pour chaque méthode. En outre, cette activité est normalisée entre les valeurs 0 et 100. Il s'agit d'une représentation temps-fréquence des fréquences basses (< 400 Hz) obtenue à partir 1) de l'algorithme melfilter (PRAAT) basé sur une échelle perceptuelle des fréquences (mel) (fenêtre d'analyse de 60ms, résolution de 10 mels ce qui conduit à 40

neurones d'entrées), 2) d'un spectrogramme à bande étroite (fenêtre d'analyse de 80ms, résolution de 12,5 Hz soit 145 neurones), 3) de l'autocorrélation du signal (< 400 Hz). Dans ce dernier cas, F0 est alors représentée par une population de 60 neurones dont l'activité est calculée d'après une courbe de Gauss [2]. Cette représentation inclue la distinction entre partie voisée et non voisée du signal, car la F0 n'est pas interpolée. Les deux premières méthodes ont un pas de 5 ms et de 10 ms pour l'autocorrélation. En outre, les deux premières représentations intègrent des informations sur l'intensité du signal pour chaque bande de fréquence ce que ne fournit pas la dernière méthode. Un seuil a été appliqué aux valeurs d'énergie obtenues par la méthode melfilter, de façon à ne laisser que l'énergie supérieure à 50 %. Ce filtrage isole la fréquence fondamentale du bruit de fond lié aux conditions d'enregistrement.

5. SIMULATION

La figure 1 présente la moyenne des performances obtenues par 10 réseaux récurrents sur le corpus de validation pour identifier les paires de langues Anglais/Japonais et Anglais/Néerlandais. La première représentation utilisée donne de bonnes performances quelles que soient les classes rythmiques. De surcroît, les performances sont supérieures à 80 % dès la première trame de signal. Elle inclut donc des informations supplémentaires. Cette représentation étant fortement bruitée, un seuil est appliquée de manière à ne tenir compte que des valeurs qui dépassent ce seuil. Lorsque ce seuil est appliqué nous retrouvons les performances rendant compte des classes rythmiques. Pour les méthodes utilisées, la discrimination avec l'Anglais est plus aisée pour le Japonais que pour le Néerlandais.

Pour les représentations melfilter avec un seuil et le spectrogramme, les performances dépendent des classes rythmiques des langues, comme pour les nourrissons [16]. (ANOVA : 1) b. $p=0.001$, $F=14.7$; 2) $p<0.001$, $F=25.8$). Lorsque la fréquence fondamentale est transmise seule, les réseaux ne parviennent pas à distinguer les langues quelque soit leur classe rythmique (ANOVA : 3) $p=0.77$, $F=0.09$) Les performances sont faibles pour la simulation (62 % et 63 %, le seuil du hasard est à 59 %). Lorsque l'intensité est disponible (spectrogramme à bande étroite), les performances augmentent (78 %), mais uniquement dans le cas ou les langues appartiennent à des classes rythmiques distinctes. Seules les représentations tenant compte de l'intensité permettent de distinguer l'Anglais du Japonais, qui appartiennent à deux classes rythmiques distinctes.

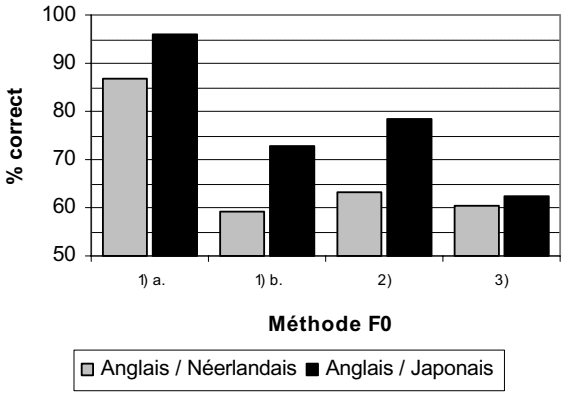

Figure 1 : Performance de discrimination en fonction des représentations prosodiques

6. DISCUSSION

Nous avons montré qu'un réseau récurrent temporel pouvait distinguer les langues avec un profil similaire aux performances des nourrissons [16]. Les valeurs les plus faibles de la représentation spectro-temporelle de la prosodie devaient être éliminées pour que les signaux contenant les phrases anglaises ne soient pas distingués des signaux contenant les phrases néerlandaises. Dans le cas contraire, la discrimination des langues se base sur des propriétés spectrales liées aux conditions d'enregistrement, mais la nature rythmique des langues influence quand même les résultats, puisque la paire Anglais/Japonais est mieux discriminée dans toutes les conditions.

Lorsque seule l'intonation est transmise aux réseaux récurrents temporels, les performances (62 %) approchent le seuil du hasard (59 %). Les phonèmes sont difficilement identifiables, et ne peuvent permettre de qualifier le rythme de la parole. En outre, les adultes ne parviennent pas à distinguer l'Anglais du Japonais à partir de l'intonation seule, alors qu'ils les distinguent lorsque le rythme est isolé [18]. Les représentations de la prosodie transmises aux réseaux peuvent contenir d'autres informations que le rythme, cependant le réseau récurrent temporel produit également des performances fonction des classes rythmiques des langues, avec des stimuli ne contenant que le rythme [7]. Pour affirmer que le rythme permet de distinguer les langues avec le réseau récurrent temporel, les stimuli devraient synthétisés en consonnes et voyelles [19], avant d'être transmis au réseau.

Lorsque ces classes rythmiques sont retrouvées, une segmentation du signal en consonnes / voyelles est effectuée soit à la main [19, 7], soit à partir d'une segmentation automatique [11, 17]. Le réseau récurrent temporel ne nécessite pas de segmentation explicite du signal en unité proche des phonèmes, ce qui est intéressant dans le cas du traitement du rythme de la parole, qui ne contient pas de marqueur évident.

Cependant, si les performances dépendent des classes rythmiques, le mécanisme que nous proposons peut être différent de celui présent chez le nourrisson [16], même si ce modèle est initialement inspiré de l'architecture frontostriatale du singe et simule l'apprentissage de séquences [8]. Les primates non humains distinguent aussi les langues de différentes classes rythmiques, mais il est possible qu'ils s'appuient sur d'autres indices.

Un système de discrimination des classes rythmiques pourrait s'avérer utile en Identification Automatique des Langues en effectuant une préclassification des langues en fonction de leur classe rythmique [10]. Le réseau récurrent temporel a déjà été employé pour l'identification des attitudes prosodiques [2] et la discrimination des mots de fonction et de contenu [3]. Le rythme pourrait être relié à trois problèmes du traitement de la parole : la segmentation pré-lexicale du signal acoustique [5] ; l'adaptation au discours compressé, la structure syllabique. Les réponses apportées par le TRN à ces tâches devraient être fonction de la classe rythmique des langues avec lesquelles il a été entraînées.

BIBLIOGRAPHIE

[1] J. Mehler, J. Bertoncini. Syllables as units in infant perception. *Infant Behavior and Development,* 4:271-284, 1981.

[2] J.-M. Blanc, P.F. Dominey. Identification of prosodic attitudes by a temporal recurrent network, *Cognitive Brain Research*, 17(3):693-699, 2003.

[3] J.-M. Blanc, C. Dodane, P.F. Dominey. Temporal Processing for Syntax Acquisition: A simulation study, *Proceedings of the 25th Annual Conference of the Cognitive Science Society,* 2003.

[4] D.V. Buonomano. Decoding Temporal Information: A Model Based on Short-Term Synaptic Plasticity, *Journal of Neuroscience* 20(3):1129-1141, 2000.

[5] A. Cutler. Prosody and the word boundary problem. In *J. L. Morgan & K. Demuth (Eds.), Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition. Mahwah, NJ:Lawrence Erlbaum Associates,* 1996.

[6] R.M. Dauer. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics,* 11:51-62, 1983.

[7] P.F. Dominey, F. Ramus. Neural Network Processing of Natural Language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, 15(1):87-127, 2000.

[8] P.F. Dominey. Complex sensory-motor sequence learning based on recurrent state representation and reinforcement learning. *Biological Cybernetics,* 265-274, 1995.

[9] J.L Elman. Finding structure in time. *Cognitive Science*, 14, 179-211, 1990.

[10] J. Farinas. Une modélisation automatique du rythme pour l'identification des langues. Thèse d'informatique de l'université Toulouse III, 2003.

[11] A. Galvès, J.E. Garcia, D. Duarte, C. Galves. Sonority as a Basis for Rhythmic Class Discrimination, *Proceedings of Speech Prosody* 2002.

[12] P.W. Jusczyk, A.D. Friederici, J. Wessels, V.Y. Svenkerud, and A.M. Jusczyk. Infants' sensitivity to the sound patterns of native language words. In Journal of Memory and Language, 32:402-420, 1993.

[13] J.S. Liaw and T.W. Berger. Dynamic synapse: Harnessing the computing power of synaptic dynamics. *Neurocomputing*, 26-27:199-206, 1999.

[14] W. Maass, T. Natschläger and H. Markram. Computational models for generic cortical microcircuits. In *J. Feng, editor, Computational Neuroscience: A Comprehensive Approach. CRC-Press, 2003.*

[15] Y.K. Muthusamy, N. Jain and R.A. Cole. Perceptual Benchmarks for Automatic Language Identification. In *Proceedings. of the IEEE International Conference on Acoustics, Speech, and Signal Processing* 1994.

[16] T. Nazzi, J. Bertoncini and J. Mehler. Language discrimination by newborns: Towards an underst,ing of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3), 1-11, 1996.

[17] F. Pellegrino, J.-L. Chauchat, R. Rakotomalalla, and J. Farinas. Can Automatically Extracted Rhythmic Units Discriminate Among Languages? In *Proceedings of Speech Prosody*, 2002.

[18] F. Ramus and J. Mehler. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105(1) :512-521, 1998.

[19] F. Ramus, M. Nespor and J. Mehler. Correlates of linguistic rhythm in the speech signal. Cognition, 73 : 265-292, 1999.

[20] I. Vasilescu, F. Pellegrino, J.-M. Hombert. Perceptual Features for the Identification of Romance Languages. In *Proceedings Of ICSLP'2000, Beijing, Chine,* 2000.