Macro-classification de signal de parole suivant les catégories phonétiques

Pierre-Sylvain Luquet, Serge Mauger

GREYC - UMR 6072 - Université de Caen
Campus II, F14032 Caen Cedex, France
psluquet@info.unicaen.fr, mauger@iutc3.unicaen.fr

ABSTRACT

In this paper we expose a method allowing to study the phonetic features of a speech signal through time. More specifically, we focus on five binary features : nasal/oral, voiced/unvoiced, strident/mellow, compact/diffuse and flat/plain. For exemple, for the nasal feature we try to consider the signal as [+nasal] or [-nasal] (wich is equivalent to [+oral]) at any given time, and we do the same with other features. For each feature, we proceed with a classifier system based on a multilayer perceptron neural net. We traine our nets on MFCC vectors with different number of coefficient extracted from a single speaker corpus. We do not select any phone subset (as vocalic/non-vocalic) in the constitution of our training and testing corpus.

INTRODUCTION

Dans cet article nous nous intéressons à la macro-classification de signal de parole et nous décrivons une expérience visant à montrer que cette macro-classification est possible à partir de coefficients cepstraux. Nous nous focaliserons sur les caractéristiques de voisement, de nasalisation, de compacité, de stridence et de bémolisation des phonèmes. Nous utilisons pour cela un ensemble de réseaux de neurones de type perceptron multi-couches. L'entrée reçoit les échantillons de signal de parole prétraités, la couche de sortie indique si l'échantillon considéré présente ou ne présente pas la caractéristique recherchée, par exemple si le signal est de type nasal ou oral.
Nous préciserons dans un premier temps quels sont les fondements théoriques et les modèles sur lesquels nous nous appuyons, ainsi que les caractéristiques ou traits d'un signal de parole.
Dans un second temps, nous préciserons la méthodologie suivie dans notre expérience, la manière dont nous avons constitué notre corpus et les outils utilisés. Nous ferons enfin une présentation critique des résultats obtenus.

1. TRAITEMENTS STOCHASTIQUES *VS* APPROCHE SYMBOLIQUE

Les traitements automatiques de la parole, et tout particulièrement ceux qui permettent sa reconnaissance, ont, grâce aux techniques Markoviennes, fait un bond qualitatif important ces dernières années. Les décodeurs acoustiques, tels que ceux de Lamel et Gauvain [10], atteignent des taux de reconnaissance proches des 75%. Cependant, les limitations restent nombreuses et la critique la plus largement formulée vis à vis de ce type de système est la quasi absence de connaissances linguistiques dans les modèles sous-jacents [15]. Les travaux actuels s'articulent autour de deux axes. Le premier s'intéresse à l'amélioration des techniques de description du signal, citons entre autres Chetouani [4], Lee [11] et Linares [12]. Le second est orienté vers la production : Vaxelaire [20] s'intéresse à l'acquisition de connaissances concernant les gestes articulatoires des locuteurs, tandis que Montagu [14] étudie leurs influences sur le signal. Hawkins [8] s'attarde sur les processus cognitifs mis en jeux. Soquet [19], Roweis [16] et Wrench [1] visent à l'intégration de ces éléments dans les systèmes de reconnaissance automatique de la parole.
Nous adoptons ci-après une démarche sensiblement différente, fondée sur une technique de décodage acoustico-phonétique différentiel. Nos travaux se raprochent donc de ceux développés par Zarader, Gas et Chetouani [2] et [3] qui développent un décodeur acoustique basé sur une architecture coopérative entre codeurs neuro-prédictifs. Néanmoins, si nous retenons également une architecture coopérative entre extracteurs de caractéristiques, nous travaillons à l'aide de coefficients cepstraux.
Ce type d'approche a pour vocation de combler le vide entre les modèles stochastiques de décodage acoustique et les connaissances que les phonologues et phonéticiens ont pu acquérir sur le langage. Nous procédons donc par macro-classification et chaque macro-classe correspond à une catégorie phonétique. A chacune d'elle est associé un classifieur (basé sur un réseau de neurones type perceptron multi-couches) et deux corpus : un pour l'apprentissage, le second pour le test. Cette modélisation du décodage acoustique permet d'obtenir un vecteur phonétique de la tranche de signal analysée. La différence avec les systèmes classiques est fondamentale car elle permet de « casser » la vision boîte noire du décodage acoustique. Le vecteur phonétique délivré par le décodeur offre autant d'informations symboliques potentiellement exploitables dans la phase de décodage lexicale. Le nombre de classifieurs dans notre architecture est dépendant de la finesse du modèle phonétique sous-jacent. Dans l'ensemble de notre travail nous avons retenu 7 catégories, mais nous ne présenterons ici les résultats que sur 5 d'entre-elles : la bémolisation, la stridence, le voisement, la compacité et la nasalité.

2. APPUIS THÉORIQUES

Saussure [17] affirme que « *dans la langue, il n'y a que des différences [...] sans termes positifs* » . Coursil [5][1]

[1]Les travaux sur la phonologie de Coursil s'inscrivent dans un projet global dénommé *ANADIA*. On lira dans Mauger [13] une extension de ce

exploite cette notion en disant que : « *Pour tout phonème* x*, il existe un phonème* y *tel que* y = x *à une et une seule différence catégorique près* ». C'est à partir de cette dernière affirmation qu'il construit la *topique des phonèmes du français contemporain* dont nous décrivons le principe dans les lignes qui suivent et qui est à la base de nos travaux. Notons enfin que la classification automatique d'un signal de parole suivant un trait phonétique donné suppose que le phonème est une *substance*, hypothèse validée par la « Dispersion-Focalisation Theory » publiée par Schwartz [18].

Phonologie et topique des phonèmes du français contemporain. La topique[2] des phonèmes du français contemporain décrite par Coursil dans [6] et [5] peut être vue comme un graphe dans lequel chaque noeud est un phonème et où seuls les phonèmes distincts par une et une seule caractéristique sont reliés. Chaque lien porte une étiquette, dont le label est le nom d'un trait, et plus précisément le nom du trait dont la valeur diffère entre deux phonèmes liés.

La topique forme donc un espace à 7 dimensions, chaque dimension étant une catégorie phonologique : *laxité*, *hauteur*, *compacité*, *voisement*, *stridence*, *nasalisation* et *bémolisation*[3]. Les phonèmes sont positionnés dans cet espace, et leurs coordonnées sont définies par les valeurs associées à chacune de leur catégorie. Il devient alors possible de mesurer la distance qui les sépares. Nous donnons à titre d'exemple dans le tableau 1 la définition extensive en termes de catégories[4] de l'ensemble des phonèmes nasalisés.

TAB. 1: Code phonétiques des phonèmes nasalisés

Phonème Exemple	L	H	C	V	S	N	B
[ɛ̃] d**aim**	5	0	0	1	0	1	0
[õ] **on**	5	0	0	1	0	1	1
[œ̃] **un**	5	0	1	1	0	1	1
[ɑ̃] r**ang**	6	1	1	1	0	1	0
[m] **m**ain	1	0	0	1	0	1	0
[n] **n**ain	1	1	0	1	0	1	0
[ɲ] **gn**on	1	1	1	1	0	1	0

3. LES CATÉGORIES ÉTUDIÉES

Nous rappelons ici quelques éléments sur les catégories étudiées dans les résultats présentés ci-après. Les définitions ci-dessous sont reprises dans [6] (voire section 2).

Compacité La compacité se manifeste par une concentration de l'énergie dans la région des fréquences moyennes du spectre, au contraire des phonèmes diffus pour lesquels cette énergie se répartit sur une zone de fréquence plus large.

Voisement Pour Saussure le voisement est le son qui résulte de l'appareil vocal selon que le résonateur pharingal est en position étroite ou large. Quand il est étroit le son est non voisé (ou sourd). Quand il est large la résonnance est effective et le son est voisé.

Stridence Pour Jakobson les stridentes sont caractérisées par une turbulence au point d'articulation. Il oppose les phonèmes stridents aux phonèmes mates. Les labiodentales, les sifflantes, les chuintantes et les uvulaires sont stridents alors que les bilabiales, les dentales, les palatales non sifflantes et les vélaires sont catégorisés mates.

Nasalisation La nasalisation est décrite comme une connexion du conduit vocal avec le conduit nasal par le biais de l'abaissement du vélum. Feng et Kotenkoff [7] ont constaté que l'abaissement de ce dernier a deux effets distincts : pour le conduit vocal le rétrécissement engendre le rapprochement des formants F3 et F4, et pour le conduit nasal sa connexion entraîne un rayonnement au niveau des narines caractérisé par une concentration dans les basses fréquences et aux alentours de 3000 Hz.

Bémolisation La bémolisation est décrite chez Jakobson par rapport au dièsement : il s'agit d'un déplacement vers le bas de la zone de fréquence des formants, le spectre conservant sa forme générale.

4. CORPUS ET MÉTHODOLOGIE

4.1. Corpus

Les résultats présentés ici sont obtenus à partir d'un corpus de parole spontanée monolocuteur. Il s'agit d'un extrait de $6min.$ du corpus C-ORAL-ROM. Durant ces six minutes un homme est interviewé (l'intervieweur n'intervient que très peu) et raconte une anecdote. On distingue très clairement dans l'enregistrement différents états émotionnels retranscrivant la joie, la surprise ou la douleur influant grandement sur la prosodie du locuteur. Aucun filtrage n'a été réalisé relativement à ces variations. Ce corpus a été ségmenté à la manuellement et 3300 phonèmes ont été labellisés et extraits. Pour chaque catégorie des jeux d'apprentissage et de test ont automatiquement et aléatoirement été générés (sans recouvrement entre les jeux d'apprentissage et les jeux de test). Un jeux est composé de 300 phonèmes (son monophonique échantilloné à $22KHz$) également répartis entre les deux étiquettes de la catégorie. Chaque jeu est normalisé en puissance à $70dB$. Les tableaux 2 et 3 donnent les phonèmes constitutifs de chacune des 5 catégories étudiées.

4.2. Paramètres

Corpus. Concernant l'analyse de nos corpus nous avons choisi de faire varier plusieurs paramètres : la taille des fenêtres ($30ms$ ou $15ms$ avec respectivement $10ms$ et $5ms$ de décalage entre fenêtres) et le nombre de coefficients cepstraux extraits : 6, 12, 18 et 24 coefficients. Pour chaque catégorie retenue nous obtenons donc 8 taux de classification par jeu d'expériences. En outre nous avons mené une série d'expériences sans normaliser la puissance de nos entrées. Le corpus étant monolocuteur les résultats étaient sensiblement meilleurs que ceux publiés ici.

Classifieur. Nos classifieurs sont générés par l'application PRAAT. Nous utilisons des réseaux de neurones type perceptron à une couche cachée. L'entrée du réseau comporte autant de cellules que nous avons de valeurs par vecteur de signal, autrement dit, la taille du réseau est dépendante

projet.

[2]La topique est définie ici par analogie aux travaux de Greimas : « l'espace topique est le lieu où se manifeste syntaxiquement une transformation, eu égard à un programme narratif donné, défini comme une transformation entre deux états narratifs stables », la transformation étant ici de nature phonétique, entre deux états phonologiques stables.

[3]Cette description en traits s'inspire en grande partie des études de caractérisation des phonèmes de Jakobson [9] qui lui même s'appuie sur les écrits de Delattre, et Fant.

[4]Code des catégories : L ≡ Laxité, H ≡ Hauteur, C ≡ Compacité, V ≡ Voisement, S ≡ Stridence, N ≡ Nasalisation et B ≡ Bémolisation

TAB. 2: Catégorisation des consonnes

	C	V	S	N	B
/p/	-	-	-	-	-
/f/	-	-	+	-	-
/v/	-	+	+	-	-
/b/	-	+	-	-	-
/m/	-	+	-	+	-
/n/	-	+	-	+	-
/d/	-	+	-	-	-
/t/	-	-	-	-	-
/k/	+	-	-	-	-
/g/	+	+	-	-	-
/ɲ/	+	+	-	+	-
/z/	-	+	+	-	-
/s/	-	-	+	-	-
/ʃ/	+	-	+	-	-
/ʒ/	+	+	+	-	-
/ʁ/	-	+	+	-	-
/l/	+	+	+	-	-
/w/	-	+	-	-	+
/ɥ/	-	+	-	-	+
/j/	-	+	-	-	-

TAB. 3: Catégorisation des voyelles

	C	V	S	N	B
/u/	-	+	-	-	+
/y/	-	+	-	-	+
/i/	-	+	-	-	-
/ɛ/	-	+	-	-	-
/e/	-	+	-	-	-
/ɛ̃/	-	+	-	+	-
/ɔ/	-	+	-	-	+
/o/	-	+	-	-	+
/õ/	-	+	-	+	+
/œ/	+	+	-	-	+
/ø/	+	+	-	-	+
/œ̃/	+	+	-	+	+
/ə/	-	+	-	-	-
/a/	+	+	-	-	-
/ɑ/	+	+	-	-	-
/ɑ̃/	+	+	-	+	-

du nombre de coefficients extraits (6, 12, 18 ou 24). La couche cachée est composée de moitié moins de cellules que la couche d'entrée. Lors des phases d'apprentissage la méthode d'évaluation d'erreur (et d'arrêt du cycle d'apprentissage avec un maximum de 400 cycles) est calculée suivant la méthode *minimum squared error* avec une tolérance de $1e^{-7}$ où le coût de l'erreur est calculé comme suit : $\textit{coût} = \Sigma_{tous\ les\ partons}\Sigma_{toutes\ les\ sorties}(o_k - d_k)^2$ avec : o_k la sortie courante de la cellule k et d_k la sortie désirée de la cellule k. Chaque classifieur possède deux cellules de sortie codant les deux états possibles de la catégorie.

5. RÉSULTATS

Nous présentons ici les résultats synthétiques d'une série de classification. Pour chaque catégorie étudiée nous avons réalisé 8 apprentissages distincts par deux facteurs : la longueur des fenêtres ($30ms$ et $10ms$) et le nombre de coefficients cepstraux extraits : 6, 12, 18, 24. La figure 1 donne les résultats de l'une de ces séries d'apprentissage selon la catégorie de bémolisation. La synthèse de l'ensemble de ces expériences se trouve dans le tableau 4.
D'une manière générale, nous avons pu observer l'influence de la longueur des fenêtres sur la qualité de la classification : pour toutes les séries de classification expérimentées, le meilleur taux a toujours été obtenu avec des fenêtres de $30ms$ (cf. tableau 4) et nous retiendrons cette valeur dans nos futures expériences. Le nombre de coefficients cepstraux extraits joue un rôle tout aussi important. Comme nous pouvons le voir sur la figure 1 on constate que la variation globale est de plus de 5% entre les taux maximaux et minimaux toutes longueurs confondues et qu'elle reste de $4,2\%$ pour les échantillons fenêtrés à $30ms$. Nous constatons également qu'il n'existe pas de relation linéaire entre le nombre de coefficients extraits et la qualité de la classification. Dans le cas de la bémolisation 6 coefficients permettent une meilleure classification qu'avec 12, mais une vectorisation avec 18 et 24 coefficients permettent d'obtenir des résultats meilleurs encore. Dans le tableau 4 on observe que 12 coefficients permettent d'obtenir la meilleure classification pour la catégorie de nasalmisation, que 18 sont nécessaires pour la compacité et le voisement et qu'il en faut 24 pour les catégories de bémolisation et de stridence. Notons enfin que, sans pouvoir encore l'évaluer précisement, nous avons noté une forte concentration des erreurs de classification en attaque et relache de phonème.

TAB. 4: Classification par catégorie

	Fenêtres	Coeff.	Taux (%)
Bémolisation	$30ms$	24	$81,69$
Compacité	$30ms$	18	$76,06$
Stridence	$30ms$	24	$75,15$
Nasalisation	$30ms$	12	$80,00$
Voisement	$30ms$	18	$90,40$

6. CONCLUSION ET PERSPECTIVES

Ainsi, dans ce qui précède, nous avons classifié nos phonèmes selon 5 catégories, or suivant le modèle retenu 7 sont nécessaires pour décrire un phonème. Par conséquent une telle classification ne nous permet pas encore de différencier un [i] d'un [ɛ] ou d'un [e] ni un [œ] d'un [ø] (cf. tableau 3). Si les résultats présentés sont encourageants, il ne nous permettent pas encore de tirer de conclusions quant aux performances d'un système de décodage acoustique basé sur cette technique. Il faut donc que nous rajoutions deux classifieurs : l'un concernant la hauteur des phonèmes et l'autre concernant la laxité. Comme pour les cas présentés ici, la hauteur est binaire (phonème grave ou aigue) alors que la laxité peut prendre 6 valeurs différentes. Par conséquent l'architecture de ce dernier classifieur sera différente des 6 autres. Une fois cette étape franchie et via l'integration dans un système tel que SIROCCO, nous pourrons comparer notre décodeur acoustique à d'autre plateformes et nous pourrons envisager la participation à différentes campagnes d'évaluations.

RÉFÉRENCES

[1] A. Wrench A. and K. Richmond. Continuous speech recognition using articulatory data. In *International Conference on Spoken Language Processing*, 2000.

[2] M. Chetouani, B. Gas, and J-L. Zarader. Coopéra-

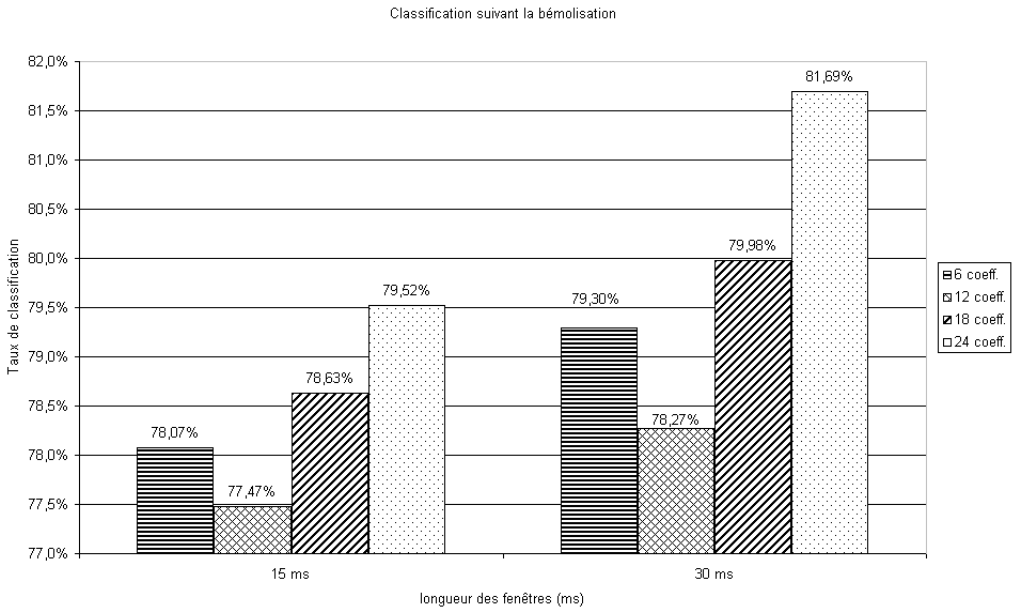

FIG. 1: Classification selon la catégorie de bémlisation

tion entre codeurs neuro-prédictifs pour l'extraction de caractéristiques en reconnaissance de phonèmes. In *Reconnaissances des formes et intelligence artificielle*, 2002.

[3] M. Chetouani, B. Gas, and J-L. Zarader. Une architecture modulaire pour l'extraction de caractéristiques en reconnaissance de phonèmes. In *GRETSI'03*, 2003.

[4] M. Chetouani, B. Gas, and J-L. Zarader. Classifieur à prototypes et codage neuro-prédictif pour l'extraction non linéaire de caractéristiques en classification de phonèmes. In *Journées d'Etude sur la Parole*, 2004.

[5] J. Coursil. *Essai d'intelligence artificielle et de linguistique générale*. PhD thesis, Université de Caen, 1992.

[6] J. Coursil, J.-C. Hilaire, D. Montlouis-Calixte, and C. Remi. Projet anadia, 2000.

[7] Feng and Kotenkoff. Vers un nouveau modèle acoustique des nasales basé sur l'enregistrement bouche - nez séparé. In *Journées d'Etude sur la Parole*, 2004.

[8] S. Hawkins. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 2003.

[9] R. Jakobson. *La charpente phonique du langage*. Editions de Minuit, Paris, 1980.

[10] L. Lamel and J. Gauvain. High performance speaker-independent phone recognition using cdhmm. In *European Conference on Speech Communication and Technology*, 1993.

[11] J-H. Lee, H-Y. Jung, T-W. Lee, and S-Y. Lee. Speech feature extraction using independent component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2000.

[12] G. Linares, P. Nocera, and H. Meloni. Model breaking detection using independent component classifier. In *International Conference on Artificial Neural Networks*, 1997.

[13] S. Mauger. *L'Interprétation des Messages Énigmatiques. Essai de Sémantique et de Traitement Automatique des Langues*. PhD thesis, Université de Caen, 1999.

[14] J. Montagu. L'articulation labiale des voyelles nasales postérieures du français : comparaison entre locuteurs français et anglo-américains. In *Journées d'Étude sur la Parole*, 2002.

[15] D.C. Plaut and C.T. Kello. *The Emergence of Language*, chapter The Emergence of Phonology from the Interplay of Speech Comprehension and Production : A Distributed Connectionist. Lawrence Erlbaum Assoc, Mahwah, 1999.

[16] S. Roweis and A. Alwan. Towards articulatory speech recognition : learning smooth maps to recover articulator information. In *European Conference on Speech Communication and Technology*, 1997.

[17] F. Saussure. *Cours de linguistique générale*. Mauro Payot, Paris, 1986.

[18] J-L. Schwartz, L-J. Boë, N. Vallée, and C. Abry. The dispersion-focalization theory of vowel systems. *Journal of Phonetics*, 1997.

[19] A. Soquet, M. Saerens, and V. Lecuit. Complementary cues for speech recognition. In *International Congress of Phonetic Sciences*, 1999.

[20] B. Vaxelaire, V. Ferbach-Hecker, and R. Sock. La perception auditive de gestes vocaliques anticipatoires. In *Journées d'Etude sur la Parole*, 2002.