Une méthode pour la classification de signal de parole sur la caractéristique de nasalisation

Luquet Pierre-Sylvain
GREYC - CNRS UMR 6072 - Université de Caen
Bd Maréchal Juin - F14032 Caen Cedex
psluquet@info.unicaen.fr
Date de soutenance prévue : décembre 2005

Mots-clefs – Keywords

Phonologie, phonétique, classifieur, réseaux de neurones
Phonology, phonetic, classifier, neural nets

Résumé - Abstract

Nous exposons ici une méthode permettant d'étudier la nature d'un signal de parole dans le temps. Plus précisément, nous nous intéressons à la caractéristique de nasalisation du signal. Ainsi nous cherchons à savoir si à un instant *t* le signal est nasalisé ou oralisé. Nous procédons par classification à l'aide d'un réseau de neurones type perceptron multi-couches, après une phase d'apprentissage supervisée. La classification, après segmentation du signal en fenêtres, nous permet d'associer à chaque fenêtre de signal une étiquette renseignant sur la nature du signal.

In this paper we expose a method that allows the study of the phonetic features of a speech signal through time. More specifically, we focus on the nasal features of the signal. We try to consider the signal as [+nasal] or [-nasal] at any given time. We proceed with a classifier system based on a multilayer perceptron neural net. The classifier is trained on a hand tagged corpus. The signal is tokenized into $30ms$ hamming windows. The classification process lets us tag each window with information concerning the properties of its content.

1 Appuis théoriques

La reconnaissance de la parole a, grâce aux techniques Markoviennes, fait un bond qualitatif énorme ces dernières années. Les décodeurs acoustiques, tels que ceux développés au LIMSI (Lamel & Gauvain, 1993), 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 sur le langage dans les modèles sous-jacents (Plaut & Kello, 1999). Les travaux actuels s'articulent autour de deux axes. Le premier s'intéresse à l'amélioration des techniques de description du signal (Chetouani *et al.*, 2002). Le second est orienté vers la production : acquisition de connaissances concernant les gestes articulatoires des locuteurs (Vaxelaire *et al.*, 2002), leurs influences sur le signal (Montagu, 2004), et les processus cognitifs mis en jeu (Hawkins, 2003). Ces connaissances font l'objet de différentes études visant à leur intégration dans les systèmes de reconnaissance automatique de la parole (Wrench & Richmond, 2000).

Nous décrivons dans ces lignes une approche sensiblement différente. Nous cherchons à appuyer une technique de décodage acoustico-phonétique sur la *notion de différence*. Saussure affirme que « *dans la langue, il n'y a que des différences […] sans termes positifs* » (Saussure, 1986). Coursil reprend à son compte cette notion dans (Coursil, 1992)[1] et affirme à son tour « *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*[2]. Le but de la classification est de mettre en évidence cette différence dans le signal. 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 dans (Schwartz *et al.*, 1997).

Le trait de nasalité. On distingue dans le français contemporain les phonèmes nasaux des phonèmes oraux, le tableau 1 en présente la partition[3]. Du point de vue de la mécanique articulatoire, la nasalité est décrite comme une connexion du conduit vocal avec le conduit nasal par le biais de l'abaissement du vélum. Les répercussions acoustiques de ce phénomène, sont décrites par Jakobson dans (Jakobson, 1980) en s'appuyant sur Fant et Delattre. Pour Fant, les consonnes nasales sont « caractérisées par un spectre où F2 est faible ou bien absent »[4] ; Delattre affine la description en précisant que pour les voyelles nasales, comparées aux orales, F1 perd une bonne part de son intensité au profit de F2. Plus récemment, Feng et Kotenkoff (Feng & Kotenkoff, 2004) ont mené à l'ICP[5] des observations basées sur une technique d'enregistrement du locuteur en différenciant les prises de son en provenance du conduit vocal et du conduit nasal. Ils ont constaté que l'abaissement du vélum 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.

[1] Les travaux sur la phonologie de Coursil s'inscrivent dans un projet global dénommé *ANADIA*. On lira dans (Mauger, 1999) l'une des extensions de ce projet.

[2] Le format dans lequel cet article est accepté ne me permet pas d'expliquer plus avant cette notion de topique. Néanmoins, je mets à disposition de tout lecteur en faisant la demande une version étendue décrivant plus finement celle-ci.

[3] La notation employée ici pour désigner les phonèmes est le codage SAMPA (Speech Assessment Methods Phonetic Alphabet). Pour plus d'informations se reporter au site de l'UCL : `http://www.phon.ucl.ac.uk/home/sampa/home.htm`

[4] F1 et F2 désignent respectivement le premier et le second formants. Les formants sont des fréquences de résonance maximum de l'enveloppe spectrale du signal de la parole à un instant donné.

[5] Institut de la Communication Parlée - Grenoble

	nasal	oral
voyelles	/e∼, o∼, 9∼, a∼/	/u, y, i, E, e, O, o, 9, 2, @, a, A/
consonnes	/m, n, J/	/p, f, v, b, d, t, k, g, z, s, S, Z, R, l, w, H, j/

TAB. 1 – Nasal vs. oral

2 Corpus

2.1 Constitution

Nous faisons ici l'hypothèse que le corpus présentant le moins de difficultés pour réaliser la partition oralisé *vs.* nasalisé est constitué de paires minimales oral *vs.* nasal. De plus, nous nous sommes concentrés sur les phonèmes dont la production pouvait être maintenue. Nous avons retenu dans notre corpus les quatre paires oppositives suivantes : /o∼/ - /o/, /e∼/ - /E/, /9∼/ - /9/ et /a∼/ - /A/. Ces phonèmes sont associés aux mots prototypiques du tableau 2.

Phonèmes	/o∼/	/o/	/e∼/	/E/	/9∼/	/9/	/a∼/	/A/
Prototype	tronc	trot	bain	baie	un	neuf	pente	pâte

TAB. 2 – Phonèmes et mots prototypes

Nous disposons aujourd'hui des résultats sur 3 corpus[6] de test monolocuteur (voir le tableau 3). Le premier corpus, $C1$ est constitué d'une seule paire minimale (/o∼/ & /o/), dont la seule variation est le trait de nasalisation (N). Les corpus $C2$ et $C3$ sont plus complexes : sur les 7 caractéristiques mises en jeu 5 varient. Pour les orales (/o/ & /a/) les variations portent sur la laxité (L), la compacité (C) et la bémolisation (B) ; la hauteur (H) intervient en plus pour les nasales (/o∼/ & /a∼/)[7].

	Phonèmes	Nb. Phonèmes	Nb. fenêtres	Maintenus	Variations
$C1$	/o∼/-/o/	22	2250	oui	N
$C2$	/o∼, a∼/-/o, a/	20	2000	oui	N, L, C, B, H
$C3$	/o∼, a∼/-/o, a/	24	470	non	N, L, C, B, H

TAB. 3 – Corpus

2.2 Paramètres

L'outil principalement utilisé dans cette expérience est le logiciel d'analyse acoustique PRAAT[8]. **Corpus.** Nous travaillons en utilisant la technique classique de fenêtrage du signal. Chaque signal de parole à analyser est segmenté en tranches de $30ms$ avec un décalage de $10ms$. A chaque

[6]Pour chacun de ces phonèmes, il a été demandé au locuteur de le prononcer dans un mot prototypique, puis de le répéter de façon isolée. Cette façon de procéder permet à l'utilisateur de « calibrer » le phonème qu'il doit ensuite prononcer isolément. Nous demandons au locuteur de répéter une dizaine de fois ce processus par couple prototype/phonème.

[7]Les quantités du tableau 3 (nombre de phonèmes et nombre de fenêtres) données ici correspondent pour chaque corpus aux versions de test. Les versions d'apprentissage sont du même ordre de grandeur.

[8]Pour plus d'informations voir la page web `http://www.fon.hum.uva.nl/praat/`

fenêtre est appliquée une fonction de $Hamming$. Chaque tranche de signal fenêtré constitue un $vecteur$ dont le nombre d'éléments est dépendant de la fréquence d'échantillonnage du signal (662 échantillons par tranche pour du signal échantillonné à $22kHz$). Chaque vecteur est labellisé par sa caractéristique acoustique ("nasal" ou "oral"). Ces vecteurs sont concaténés en matrices, qui selon le corpus servira soit à l'apprentissage, soit à la phase de test[9].
Classifieur. 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, soit 662 cellules. La sortie est composée de deux cellules correspondant aux classes activables. La couche cachée est composée de 331 cellules. Lors des phases d'apprentissage l'évaluation de l'erreur est calculée suivant la méthode *minimum squared error*.

3 Résultats

3.1 Variation restreinte

Les résultats du tableau 3 concernent le corpus $C1$ et ont été obtenus au terme d'un apprentissage de 400 cycles. Les phonèmes sont maintenus et la seule variation phonologique mise en jeu est la nasalisation. Nous voyons ici que sur une quantité restreinte de corpus il est possible de classifier le signal avec de bons résultats. En effet nous obtenons un taux d'erreurs faible ($2,8\%$), mais nous voyons surtout que le nombre de fenêtres continues incorrectement classifiées est très faible (8) en regard du nombre de fenêtres par phonème (102). Le risque de mal classifier un phonème est donc minime.

fenêtres	2252	fenêtres par phonème	102
erreurs	63	groupes d'erreur	17
taux	**2,8%**	erreurs par groupe	3,71
erreurs consécutives maximum			**8**

TAB. 4 – Résultats $C1$

3.2 Augmentation de la dissemblance

Les résultats donnés ici concernent le corpus $C2$. Le tableau 5 donne les résultats obtenus pour 400 cycles d'apprentissage, tandis que le tableau 6 nous donne les résultats au bout de 600 cycles. Comme précédemment les phonèmes sont maintenus mais plusieurs variations phonologiques sont ici mises en jeu (voire 2.1). La focalisation du classifieur sur la caractéristique de nasalisation est donc rendue plus complexe en raison du bruit apporté par les autres variations. Cependant, les résultats obtenus montrent qu'une classification est toujours possible. Avec 400 cycles (tableau 5), nous obtenons un taux d'erreurs qui reste faible ($5,3\%$). Le nombre de fenêtres contiguës incorrectement classifiées l'est aussi (12 fenêtres mal classifiées). Néanmoins, si nous augmentons d'un tiers le nombre de cycles (tableau 6), le taux d'erreurs retombe à $2,3\%$.

[9]Dans les deux cas, les valeurs des échantillons sont décalées et mises à l'échelle pour être dans le domaine de définition de notre classifieur. Les valeurs d'origine varient dans l'intervalle $[-1, 1]$. Nous les réduisons d'un facteur $1/2$ puis les décalons de 1 pour qu'elles soient comprises dans l'intervalle d'entrée du classifieur : $[0, 1]$.

fenêtres	1996	fenêtres par phonème	100
erreurs	106	groupes d'erreur	47
taux	**5,3%**	erreurs par groupe	2,3
erreurs consécutives maximum			**12**

TAB. 5 – Résultats $C2$ - Apprentissage : 400 cycles

fenêtres	1996	fenêtres par phonème	100
erreurs	47	groupes d'erreur	16
taux	**2,3%**	erreurs par groupe	2,9
erreurs consécutives maximum			**9**

TAB. 6 – Résultats $C2$ - Apprentissage : 600 cycles

3.3 Phonèmes non maintenus

L'expérience menée sur le corpus $C3$ est similaire à l'expérience précédente, mais concerne des phonèmes non maintenus. Les résultats obtenus (tableau 7 et 8) sont nettement en retrait, mais restent néanmoins très intéressants. Au terme d'un apprentissage de 300 cycles, nous observons un taux d'erreur de 20% que nous pouvons réduire à $15,8\%$ au terme de 600 cycles d'apprentissage (soit une réduction de ce taux de $21,5\%$). En revanche, le doublement du nombre de cycles d'apprentissage n'apporte rien ici en terme de réduction du nombre d'erreurs contiguës (7 fenêtres mal classifiées[10]). Néanmoins ce nombre reste acceptable, dans le cas d'une stratégie de classification *winner-takes-all* dans la mesure où un phonème compte en moyenne 20 fenêtres. Notons que la taille de notre corpus d'apprentissage (412 fenêtres) pose ici un problème ; le nombre de patrons étiquetés limite la capacité de classification. Enfin, un dernier cycle long d'apprentissage (2000 cycles) ne nous a pas permis d'améliorer sensiblement le taux d'erreurs et a également confirmé qu'au delà de 600 cycles, la réduction de l'erreur est faible pour un coût très élevé ; dans notre cas le nombre de cycles a été plus que triplé pour un gain de 2 erreurs seulement sur le corpus de test.

fenêtres	469	fenêtres par phonème	20
erreurs	94	groupes d'erreur	37
taux	**20,0%**	erreurs par groupe	2,5
erreurs consécutives maximum			**7**

TAB. 7 – Résultats $C3$ - Apprentissage : 300 cycles

4 Perspectives et conclusion

Les résultats présentés dans cet article sont prometteurs, cependant certains aspects sont à approfondir. D'autre types de descripteurs sont envisagés : techniques d'extraction de type MFCC (Mel Frequency Cepstral Coefficients), LPC (Linear Predictive Coding) ou plus encore PLP (Perceptual Linear Predictive coding). Par ailleurs, la limite en terme de fréquence d'échantillonnage en deçà de laquelle l'apprentissage n'est plus réalisable n'est pas connue. Qu'en

[10]Le nombre donné ici correspond au nombre maximal de fenêtres contiguës mal classifiées dans un phonème.

fenêtres	469	fenêtres par phonème	20
erreurs	74	groupes d'erreur	30
taux	**15,8%**	erreurs par groupe	2, 5
erreurs consécutives maximum			**7**

TAB. 8 – Résultats $C3$ - Apprentissage : 600 cycles

est-il d'un signal de qualité téléphonique échantillonné à $8kHz$?
Nous envisageons également d'augmenter la complexité du corpus : nombre de locuteurs et nombre de phonèmes présents. L'augmentation du nombre de locuteurs a pour but de tester l'indépendance de l'apprentissage du classifieur. Pour valider notre méthode sur du signal de parole continue, une nouvelle série d'expériences est envisagée. L'augmentation du nombre de phonèmes doit permettre de multiplier les caractéristiques prises en considération.
En outre, nous faisons l'hypothèse que le croisement de résultats issus de plusieurs classifieurs (avec un apprentissage sur des catégories phonétiques différentes) permettra de situer le signal dans l'espace topique et de déterminer ainsi la classe phonétique à laquelle il appartient.

Références

CHETOUANI M., GAS B. & ZARADER J. (2002). Coopération entre codeurs neuro-prédictifs pour l'extraction de caractéristiques en reconnaissance de phonèmes. In *Reconnaissances des formes et intelligence artificielle*.

COURSIL J. (1992). *Essai d'intelligence artificielle et de linguistique générale*. PhD thesis, Université de Caen.

FENG & KOTENKOFF (2004). Vers un nouveau modèle acoustique des nasales basé sur l'enregistrement bouche - nez séparé. In *Journées d'Etude sur la Parole*.

HAWKINS S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*.

JAKOBSON R. (1980). *La charpente phonique du langage*. Paris : Editions de Minuit.

LAMEL L. & GAUVAIN J. (1993). High performance speaker-independent phone recognition using cdhmm. In *European Conference on Speech Communication and Technology*.

MAUGER S. (1999). *L'Interprétation des Messages Énigmatiques. Essai de Sémantique et de Traitement Automatique des Langues*. PhD thesis, Université de Caen.

MONTAGU J. (2004). Les sons sous-jacents aux voyelles nasales en français parisien : indices perceptifs des changements. In *Journées d'Étude sur la Parole*, p. 385–388.

PLAUT D. C. & KELLO C. T. (1999). *The Emergence of Language*, chapter The Emergence of Phonology from the Interplay of Speech Comprehension and Production : A Distributed Connectionist. Lawrence Erlbaum Assoc : Mahwah.

SAUSSURE F. (1986). *Cours de linguistique générale*. Paris : Mauro Payot.

SCHWARTZ J.-L., BOË L.-J., VALLÉE N. & ABRY C. (1997). The dispersion-focalization theory of vowel systems. *Journal of Phonetics*.

VAXELAIRE B., FERBACH-HECKER V. & SOCK R. (2002). La perception auditive de gestes vocaliques anticipatoires. In *Journées d'Etude sur la Parole*.

WRENCH A. A. & RICHMOND K. (2000). Continuous speech recognition using articulatory data. In *International Conference on Spoken Language Processing*.