Représentation de connaissances indépendantes du locuteur pour la reconnaissance de mots acoustiquement proches

H. Méloni, P. Gilles

Laboratoire d'Informatique - Faculté des Sciences
33 rue Louis Pasteur - 84000 Avignon - France

RESUME

Nous proposons une méthodologie pour la discrimination descendante entre des mots phonétiquement proches d'une cohorte. Les connaissances utilisées ne dépendent que de quelques caractéristiques très limitées du locuteur (position des formants pour les voyelles) et décrivent les traces acoustiques de phénomènes articulatoires dans un contexte connu. Ces techniques sont appliquées à l'identification des occlusives sourdes dans des logatomes constitués des consonnes /p/, /t/ et /k/ suivies d'une des voyelles du français.

1. PRESENTATION DU PROBLEME

Le Décodage Acoustico-Phonétique de la parole est rendu difficile notamment à cause des variations inter-locuteurs et des effets de la coarticulation des phonèmes. Le premier type de variabilité, de nature statique (cibles différentes), peut être traité partiellement de manière ascendante par l'utilisation de quelques caractéristiques d'un locuteur (modèle spectral des parties stables des unités phonétiques). L'acquisition, la mémorisation et le traitement de ces connaissances sont aisément effectués et permettent de mettre en œuvre une première phase efficace du DAP [2], [4], [5]. Les résultats d'un tel processus sont constitués par un treillis de phonèmes valués comportant toutes les hypothèses vraisemblables d'occurrence d'une unité. Ces éléments déterminent des ensembles de mots qui sont susceptibles de coïncider de manière optimale - au sens de critères de proximité et de densité de recouvrement - avec une zone du treillis.
Les mots proposés dans la phase ascendante sont acoustiquement proches et les scores de reconnaissance qui leurs sont associés ont été calculés au moyen de distances par rapport à des références idéales non altérées par le contexte. Il convient donc, dans une étape descendante du processus de décodage, de classer plus précisément ces hypothèses.
Les phénomènes de coarticulation ont pour conséquence la modification des cibles phonétiques et apparaissent sur l'évolution temporelle des paramètres acoustiques et phonétiques (formants par exemple). La phase descendante du DAP consiste à localiser et évaluer les traces acoustiques de phénomènes articulatoires distincts sur les zones appropriées du signal. Cette opération est effectuée en utilisant les connaissances disponibles sur le contexte phonémique.
Les travaux présentés ici décrivent la méthodologie utilisée et les résultats obtenus pour la discrimination des occlusives sourdes dans le cas où les mots sont des logatomes constitués d'une consonne suivie de l'une quelconque des voyelles du français. Nous examinerons plus particulièrement le processus d'identification du lieu d'articulation.

2. METHODOLOGIE

L'identification du lieu d'articulation des occlusives sourdes peut être effectuée au moyen d'informations diverses (spectrales et temporelles) qui apparaissent sur l'explosion et dans la transition vers la voyelle adjacente [2], [3], [7]. Nous n'envisagerons que les traces acoustiques détectées sur les paramètres spectraux.

2.1. paramétrisation du signal

Le signal de parole est numérisé sur 16 bits à une fréquence de 12,8 kHz puis préaccentué et caractérisé chaque 10 ms par son énergie globale, la densité des passages par zéro et les énergies spectrales dans 24 canaux répartis suivant une échelle de Mel. Les spectres sont obtenus par prédiction linéaire et cette représentation est suffisamment efficace pour représenter la plupart des connaissances. Il est cependant parfois indispensable de disposer de paramètres plus précis, notamment pour suivre les trajectoires formantiques. Dans ce cas, nous disposons d'une caractérisation plus fine des spectres LPC (figure 1).
Un ensemble d'outils permet de définir et de calculer dynamiquement de nombreux paramètres auxiliaires obtenus par combinaisons des attributs initiaux [5]. Les informations les plus utilisées mesurent et comparent les densités d'énergie dans certaines bandes spectrales. L'évolution temporelle de ces paramètres est modélisée au moyen de formes élémentaires.

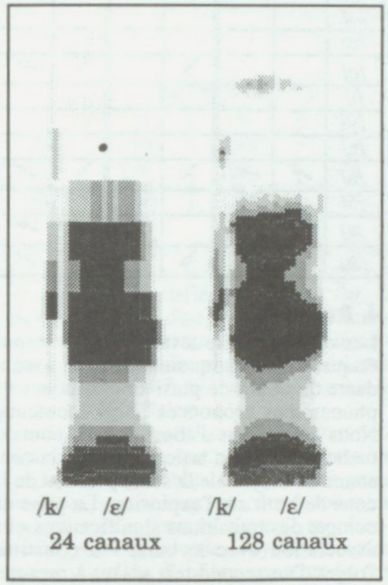

Figure 1 - La représentation spectrale au moyen de 128 canaux est nécessaire pour permettre le suivi précis des formants.

2.2. Identification sur l'explosion

Dans la phase de DAP ascendant, la position de l'explosion a été repérée au moyen de paramètres calculés en fonction du phonème. Nous disposons par ailleurs des valeurs des formants de chacune des voyelles pour un locuteur donné.
Pour les occlusives /p/, /t/ et /k/ des règles définissent et calculent les paramètres caractérisant l'énergie spectrale de la composante principale du bruit d'explosion en fonction de la position des formants de la voyelle adjacente. Si nous notons $E(p,v)$ la densité d'énergie dans la zone désignée pour la consonne p dans le contexte de la voyelle v, nous pouvons calculer la fonction :

$$f(p1,v) = 2*E(p1,v) - E(p2,v) - E(p3,v)$$

qui définit la valuation de l'hypothèse correspondant à la consonne $p1$. La valeur de la fonction est d'autant plus grande que la position spectrale du bruit coïncide avec celle définie pour cette situation.

Table 1 - Position du bruit d'explosion pour les occlusives sourdes en fonction de la position des formants de la voyelle.

	/p/	/t/	/k/
/a/	≤ F2-1	≥ F3+2	[F2, F2+3]
/ɔ/	[F2+3, F3-1]	≥ F3	[F2, F2+2]
/ɛ/	≤ F2-1	≥ F3+2	[F2, F2+3]
/œ/	≤ F2-1	≥ F3+2	[F2, F2+3]
/o/	[F2+3, F3-1]	≥ F3	[F2, F2+2]
/e/	≤ F2-1	≥ F3+1	[F2, F3]
/ø/	≤ F2-1	≥ F3+1	[F2+2, F3+1]
/i/	≤ F2+1	≥ F3+2	[F2, F3]
/y/	≤ F2-1	≥ F2+2	[F2, F2+1]
/u/	[F2+3, F3-1]	≥ F3	[F2, F2+2]
/ã/	≤ F2-1	≥ F3+1	[F2, F3]
/ɛ̃/	≤ F2-1	≥ F3+1	[F2, F3]
/ɔ̃/	≤ F2-1	≥ F3+1	[F2, F3]
/œ̃/	≤ F2-1	≥ F3+1	[F2, F3]

Les calculs de $E(p,v)$ sont effectués à partir des valeurs de la table 1. L'énergie correspond à celle du canal dont l'amplitude est maximale dans la zone spectrale indiquée par la règle associée à la situation donnée.

2.3. Identification sur la transition

Pour modéliser les informations relatives à l'évolution spectrale de l'énergie autour des formants, il est nécessaire d'utiliser une représentation des spectres au moyen de 128 valeurs (figure 1). Toutefois, la caractérisation au moyen des 24 canaux permet de mesurer les évolutions temporelles des formants dans le cas où les

pôles significatifs sont suffisamment séparés.
La direction de la transition des formants est évaluée sur la portion de la voyelle située entre le début d'apparition des pics spectraux et la trame de plus grande stabilité. Le calcul des valeurs de la pente du formant (repéré par le canal *i* au maximum de stabilité) est effectué à partir de l'évolution de l'énergie dans les canaux adjacents (canaux *i-1* et *i+1*). La différence de densité d'énergie entre la zone stable et le début d'apparition des formants dans les canaux *i-1* et *i+1* constitue le paramètre essentiel permettant d'apprécier le sens de l'évolution d'un formant au contact de la consonne (figure 2).

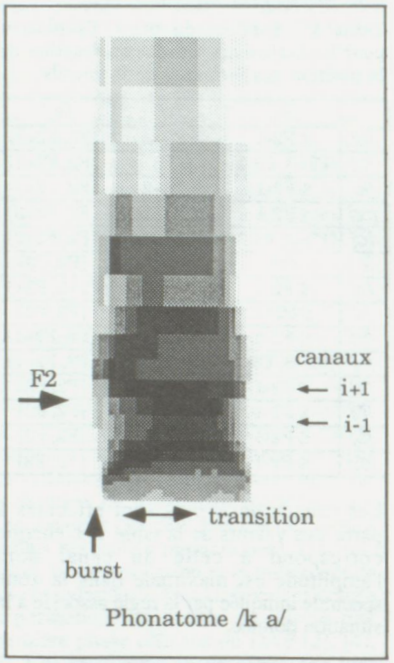

Figure 2 - Les canaux i-1 et i+1 sont utilisés pour mesurer l'évolution temporelle de l'énergie autour du formant (canal i).

Les informations concernant les transitions sont utilisées pour compléter celles qui sont évaluées sur l'explosion. Nous avons limité ces connaissances aux seules situations qui sont pertinentes pour de nombreux locuteurs et qui peuvent être traitées à partir de la représentation paramétrique sur 24 canaux. Les formes des transitions de référence utilisées sont données dans la table 2. Il s'agit d'une tendance générale plus ou moins marquée suivant le contexte et le locuteur. Ces indices acoustiques traduisent l'influence du lieu articulatoire de la consonne sur la cible de la voyelle.

Table 2 - Formes des transitions des formants F2 et F3 pour les voyelles précédées des occlusives sourdes. Seules les formes utilisées dans notre système pour l'identification du lieu articulatoire sont présentées dans cette table.

	/p/		/t/		/k/	
	F2	F3	F2	F3	F2	F3
/a/	↗	→	↘		↘	↗
/ɔ/						
/ɛ/	↗	→				
/œ/	↗	→	↘		↘	→
/o/						
/e/						
/ø/			↘		↘	→
/i/						
/y/						
/u/						
/ã/	↗	→	↘		↘	↗
/ɔ̃/						
/ɛ̃/	↗	→			↘	→
/œ̃/	↗	→	↘		↘	↗

3. RESULTATS

Les règles ont été testées sur un corpus étiqueté automatiquement (étape ascendante du DAP) de plusieurs centaines de phonatomes prononcés par 4 locuteurs. Nous avons tout d'abord évalué contextuellement le lieu articulatoire de la consonne au moyen de la seule position de la zone de bruit sur l'explosion. La prise en compte des transitions significatives - en association avec le burst - a constitué l'objet d'un second test visant à mesurer si ces deux types de connaissances étaient complémentaires.

3.1. Résultats sur l'explosion

Les résultats obtenus avec les règles caractérisant le lieu d'articulation sur l'explosion de la consonne sont donnés par la matrice de confusion de la table 3. Les performances sont intéressantes pour /t/ et /k/ mais demeurent insuffisantes pour /p/. Les confusions pour la consonne bilabiale résultent d'une absence fréquente du burst et de la diffusion de l'énergie dans le spectre.

Table 3 - Matrice de confusion pour l'identification du lieu articulatoire des occlusives sourdes à partir de l'explosion.

	consonne reconnue		
	/p/	/t/	/k/
/p/	**70%**	14%	16%
/t/	3%	**89%**	8%
/k/	7%	3%	**90%**

3.2. Résultats avec les transitions

Les résultats obtenus si l'on ajoute les règles caractérisant le lieu d'articulation sur les transitions de la voyelle sont donnés par la matrice de confusion de la table 4.

Table 4 - Matrice de confusion pour l'identification du lieu articulatoire des occlusives sourdes à partir de l'explosion et des transitions.

	consonne reconnue		
	/p/	/t/	/k/
/p/	**75%**	13%	12%
/t/	2%	**90%**	8%
/k/	6%	3%	**91%**

L'amélioration des résultats n'est sensible que dans le cas de la consonne /p/ qui est moins bien identifiée que /t/ et /k/. Il semble difficile d'augmenter significativement les performances de reconnaissance sans prendre en compte d'autres informations (diffusion de l'énergie sur le burst de /p/, VOT, etc.).
Les transitions utilisées (table 2) font nettement apparaître que quelques contextes sont plus favorables que d'autres pour l'évaluation des mouvements de certains formants dans notre système de représentation paramétrique (les voyelles fermées et les voyelles d'arrière constituent des environnements peu favorables). Une paramétrisation au moyen de 128 valeurs spectrales permet de mieux apprécier les transitions formantiques, mais ces informations varient parfois considérablement et sont rarement complémentaires de celles mesurées sur l'explosion [1].

4. CONCLUSION

L'identification descendante (contexte phonétique connu) des consonnes occlusives sourdes en reconnaissance de la parole est une opération qui peut être effectuée avec de bonnes performances en utilisant des systèmes de représentation des connaissances. Ces techniques ont produit des résultats intéressants dans d'autres circonstances [6] et sont opérationnelles pour la caractérisation multilocuteur et la discrimination d'autres phonèmes dans des contextes connus ou hypothétiques.
La modélisation par auto-organisation des informations de ce type avec un processus d'apprentissage implique la prise en compte d'une grande quantité d'exemples pour de nombreux locuteurs. Nous envisageons, pour comparer les performances de notre méthode, de réaliser un système utilisant des techniques connexionnistes qui serait supervisé par des règles de manière à fournir des entrées prétraitées aux organes effectuant l'apprentissage et la reconnaissance et limiter ainsi le nombre des exemples nécessaires.

5. REFERENCES

[1] BLUMSTEIN S.E., STEVENS K.N. (1979), *"Acoustic invariance in speech production : evidence from measurement of the spectral characteristics of stops consonants"*, JASA 66

[2] CALLIOPE (1989), *La parole et son traitement automatique*, Collection technique et scientifique, Masson, Paris

[3] DURAND P. (1982), *"Etude acoustique des consonnes occlusives du français commun"*, Doctorat de 3ème cycle, Université de Provence, Aix-Marseille

[4] HATON J.P. et Col. (1990), *"Décodage Acoustico-Phonétique : problèmes et éléments de solution"*, Traitement du Signal, volume 7 n°4, pp. 293-313

[5] MELONI H., GILLES P. (1991), *"Décodage Acoustico-Phonétique ascendant"*, Traitement du Signal, (à paraître)

[6] MELONI H., GILLES P., BETARI A. (1991), *"Representation of acoustic and phonetic knowledge for speaker-independent recognition of small vocabularies"*, Speech Communication, volume 10 n°2

[7] VAISSIERE J. (1987), *"Effect of phonetic context and timing on the F-pattern on the vowels in the continuous speech"*, 11ème ICPS, Tallinn, Estonia, URSS