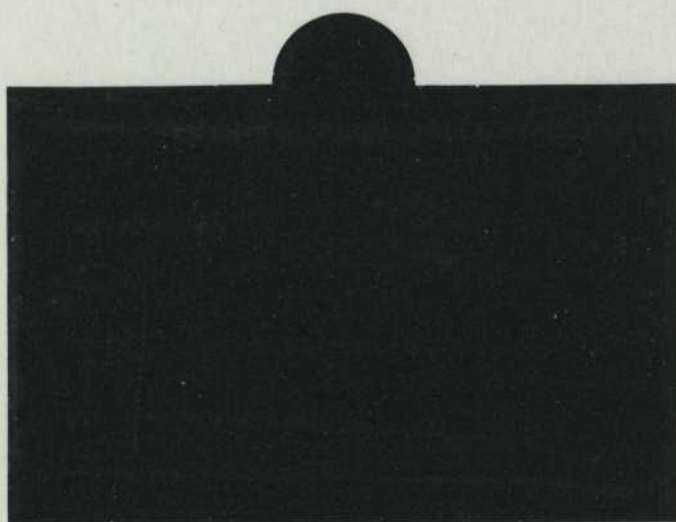


006.454
P96423
1989

Notes de recherches

11

57^E CONGRÈS DE L'ACFAS
PRODUCTION ET PERCEPTION
DE LA PAROLE: SYNTHÈSE ET
RECONNAISSANCE AUTOMATIQUES



11

57E CONGRÈS DE L'ACFAS
PRODUCTION ET PERCEPTION
DE LA PAROLE: SYNTHÈSE ET
RECONNAISSANCE AUTOMATIQUES

Association canadienne-française pour
l'avancement des sciences

COLLOQUE

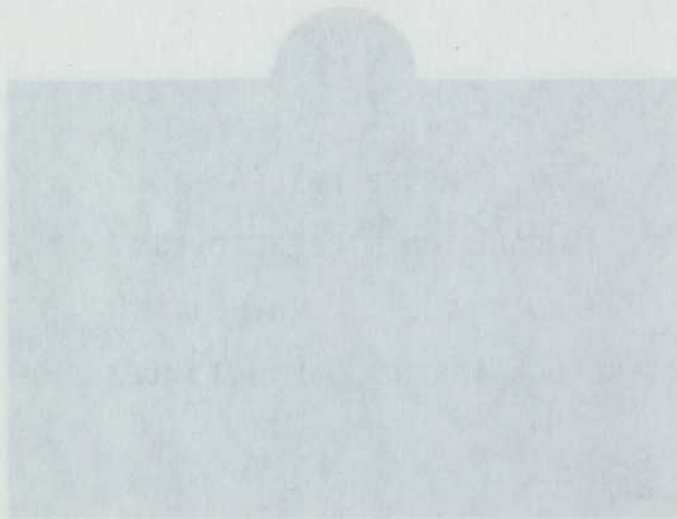
Raymond Descout de Communications Canada - Centre
canadien de recherche sur l'informatisation du travail

Danièle Archambault de l'Université de Montréal
département de linguistique et Philologie

Montréal

17 mai 1989

D8904521



TK
7882
T73 P76
1989

Association canadienne-française pour
l'avancement des sciences

COLLOQUE

Raymond Desautels de l'Université du Canada - Centre
canadien de recherche en linguistique et en sciences du travail

Danièle Archambault de l'Université de Montréal
Département de Linguistique et de Psychologie

08904521

Introduction

Les activités de Recherche & Développement reliées aux "Industries de la Langue" présentent un intérêt croissant tant dans le monde industriel que dans les milieux universitaires. En effet, la Connaissance d'une Langue va de pair avec son utilisation effective dans les systèmes d'information qui deviennent chaque jour de plus en plus puissants et sophistiqués. Dans les champs bien connus des systèmes de gestion et d'interrogation de bases de données, de la traduction assistée, des lexiques et thésaurus, cette observation s'applique également aux interfaces vocales mettant en oeuvre la synthèse et la reconnaissance automatique de la parole.

Dans ce domaine des "Sciences de la Parole" (production et perception humaines, technologies vocales), les dix dernières années ont vu se développer de nouvelles équipes de recherche. Les sujets spécifiques propres à une langue donnée se sont considérablement accrues, tout particulièrement pour le français.

Pour sa part, dans une situation bilingue exceptionnelle, le Canada compte malheureusement plusieurs équipes de recherche oeuvrant dans le domaine de la Communication Parlée. C'est la raison pour laquelle il nous a semblé opportun de profiter de la tenue du 57^{ème} Congrès de l'ACFAS pour faire le point sur l'état actuel des recherches menées en traitement de la parole au Canada, et plus particulièrement au Québec. Un colloque spécialisé d'une journée a donc été consacré à ce thème, réunissant en un même lieu des représentants des mondes universitaire et industriel.

Plus de 40 personnes ont participé à cette rencontre au cours de laquelle douze exposés ont donné un aperçu rétrospectif des travaux effectués dans ces domaines. Dix présentations venaient de chercheurs canadiens et deux conférenciers européens avaient été invités pour nous faire partager leur expertise sur deux sujets: François Néeil du CNRS-LIMSI, France, en reconnaissance automatique et Francis Charpentier du CNET, France, en synthèse de la parole. Étant donné le qualité de l'ensemble de ces présentations, il nous a paru important d'en assurer la diffusion par la publication du présent recueil.

Pour conclure, après une table ronde consacrée au thème: "Transferts technologiques et formation des chercheurs", l'ensemble des participants a émis le souhait de maintenir ce lieu qui nous avait réuni pour un jour, par la création d'une nouvelle association.

Introduction

Les activités de Recherche & Développement reliées aux "*Industries de la Langue*" présentent un intérêt croissant tant dans le monde industriel que dans les milieux universitaires. En effet, la Connaissance d'une Langue va de pair avec son utilisation effective dans les systèmes d'information qui deviennent chaque jour de plus en plus puissants et sophistiqués. Outre les champs bien connus des systèmes de gestion et d'interrogation de bases de données, de la traduction assistée, des lexiques et thésaurus, cette observation s'applique également aux interfaces vocales mettant en oeuvre la synthèse et la reconnaissance automatiques de la parole.

Dans ce domaine des "Sciences de la Parole" (production et perception humaines, technologies vocales), les dix dernières années ont vu se développer de nouvelles équipes de recherche. Les études spécifiques propres à une langue donnée se sont considérablement accrues, tout particulièrement pour le français.

Pour sa part, dans une situation bilingue exceptionnelle, le Canada compte maintenant plusieurs centres de recherche oeuvrant dans le domaine de la Communication Parlée. C'est la raison pour laquelle il nous a semblé opportun de profiter de la tenue du 57^{ème} Congrès de l'ACFAS pour faire le point sur l'état actuel des recherches menées en traitement de la parole au Canada, et plus particulièrement au Québec. Un colloque spécialisé d'une journée a donc été consacré à ce thème, réunissant en un même lieu des représentants des mondes universitaire et industriel.

Plus de 40 personnes ont participé à cette rencontre au cours de laquelle douze exposés ont donné un aperçu réaliste des travaux effectués dans ces domaines. Dix présentations venaient de chercheurs canadiens et deux conférenciers européens avaient été invités pour nous faire partager leur expertise sur deux sujets: Françoise Néel du CNRS-LIMSI, France, en reconnaissance automatique et Francis Charpentier du CNET, France, en synthèse de la parole. Etant donnée la qualité de l'ensemble de ces présentations, il nous a paru important d'en assurer la diffusion par la publication du présent recueil.

Pour conclure, après une table ronde consacrée au thème: "*Transferts technologiques et formation des chercheurs*", l'ensemble des participants a émis le souhait de maintenir ce lien, qui nous avait réuni pour un jour, par la création d'une nouvelle association.

Table des Matières

Indicateurs d'une synthèse de qualité	43
<i>Laurent SENTERRE, Université de Montréal</i>	
Contribution à l'analyse et à la reconnaissance des nasales, des liquides et des fricatives	1
<i>Jean ROUAT, Université du Québec à Chicoutimi</i>	
Système de reconnaissance automatique pour la détection et l'analyse des troubles neurologiques de la parole	7
<i>Eric KELLER, Université du Québec à Montréal</i>	
Utilisation de la méthode d'analyse discriminante avec différentes analyses acoustiques dans un système de reconnaissance de la parole	13
<i>Claude LEFEBVRE, Melvyn J. HUNT Centre National de la Recherche à Ottawa</i>	
Comparaison entre la segmentation implicite et explicite pour les modèles de Markov	19
<i>Yves NORMANDIN, Claude ACHCAR Centre de Recherche Informatique de Montréal</i>	
Analyse discriminatoire de l'alphabet français par anamorphose temporelle	25
<i>Jean-Pierre CORDEAU, INRS-Télécommunications</i>	
Les Technologies Interactives Vocales et leurs Applications	31
<i>Matthew LENNIG, Recherches Bell-Northern & INRS-Télécommunications</i>	
Reconnaissance vocale et Application en Aéronautique	37
<i>Françoise NEEL, K. MATROUF, J.L. GAUVAIN, J. MARIANI CNRS - LIMSI (France)</i>	

Intérêts d'une synthèse de qualité	43
Laurent SANTERRE, <i>Université de Montréal</i>	
Analyse syntaxique du texte pour la synthèse automatique de la parole	51
Douglas O'SHAUGNESSY, <i>INRS-Télécommunications</i>	
L'aspect temporel de la synthèse de parole par règles	57
Danièle ARCHAMBAULT, <i>Université de Montréal</i>	
Synthèse de la parole dans le domaine temporel par superposition-addition pitch-synchrone	63
Francis CHARPENTIER, Eric MOULINES <i>CNET (France)</i>	
Un outil générateur d'applications de technologies vocales	69
Philippe MABILLEAU, <i>Université de Sherbrooke & Centre Canadien de Recherche sur l'Informatisation du Travail</i>	

CONTRIBUTION À L'ANALYSE ET À LA RECONNAISSANCE
DES NASALES, DES LIQUIDES ET DES FRICATIVES

Jean Rouat

Département des Sciences Appliquées

Université du Québec à Chicoutimi

Chicoutimi, Québec

Canada, G7H 2B1

Tél.: (418) 545-5042

Courrier électronique: P134@UQUEBEC.BITNET

RÉSUMÉ:

Cet article présente une partie des travaux de recherche effectués en vue de reconnaître certaines consonnes nasales, liquides et fricatives, en anglais, pour le vocabulaire composé des chiffres et des lettres de l'alphabet. Ce travail a été motivé par la réalisation d'un système dont les grandes lignes sont décrites dans [De Mori et al 88]. Avec ce système, il est possible de concevoir des tâches spécialisées dans l'analyse et/ou la reconnaissance de pseudo-syllabes spécifiques.

On se propose de montrer qu'il est possible d'obtenir de bonnes performances de reconnaissance sur certaines consonnes nasales, liquides et fricatives en utilisant des analyses acoustiques simples. Les analyses présentées ont été testées en utilisant une modélisation statistique du signal. L'apprentissage a été effectué avec seulement 40 locuteurs et la reconnaissance avec 40 autres locuteurs. Seuls 20 locuteurs sur les 40 de l'apprentissage parlaient couramment l'anglais et seuls 21 locuteurs sur les 40 de la reconnaissance étaient anglophones. Tous les autres locuteurs étaient d'origines très diverses, avec un accent très prononcé. Les résultats obtenus sont bons, compte tenu de la très grande variété des locuteurs. Les taux sont de 98% sur les fricatives et de 90.5% sur les nasales, liquides considérées. Il a été constaté qu'on peut améliorer les taux de reconnaissance de l'ordre de 2% à 4%, par rapport à l'analyse initiale, en jouant sur des paramètres aussi simples que la préaccentuation et sur une pondération dans le domaine spectral. Plusieurs paramètres acoustiques ont été comparés et nous présentons les résultats de reconnaissance obtenus pour différentes combinaisons.

1) INTRODUCTION

Après une trentaine d'années de recherche, il n'existe toujours pas de système de reconnaissance qui fonctionne sur de la parole continue indépendamment du locuteur et des accents régionaux. Toutefois, il existe différents petits systèmes de reconnaissance dont les performances sont relativement bonnes. Ils ont été réalisés en restreignant le problème, afin de diminuer la

complexité. En général, les voyelles sont assez facilement reconnues. Ce n'est pas le cas pour les consonnes en raison de leur courte durée, elles sont souvent plus instables et leurs paramètres articulatoires moins bien définis. Parfois, elles se confondent facilement avec les voyelles.

Les consonnes occlusives ont été assez bien étudiées [Chang 81] [Edwards 81] [Kewley - Port 83] et elles sont relativement bien reconnues [Demichelis 83] [Rossen 88]. En revanche, les fricatives, nasales et liquides, en reconnaissance, n'ont été étudiées que sur un nombre restreint de locuteurs. Pour réaliser un système qui soit robuste, il faut améliorer les résultats obtenus sur ces dernières.

Un des objectifs du présent article est d'effectuer la reconnaissance de ces consonnes pour un nombre relativement important de locuteurs. Par rapport à certaines études antérieures, les performances ont été évaluées avec des locuteurs d'origines ethniques très diverses. En effet, les accents régionaux et étrangers influent énormément sur les performances de reconnaissance.

Un autre objectif est de montrer qu'il est possible d'obtenir de bonnes performances, sur un vocabulaire difficile, en utilisant des techniques d'analyse relativement simples. Les analyses présentées ici reposent sur des bancs de filtres à échelle Mel. Toutefois, nous avons modifié certains paramètres: préaccentuation, pondération,..., afin de montrer qu'il est possible d'améliorer les taux de reconnaissance de l'ordre de 2 à 4%. Cette amélioration se compare très bien avec celles obtenues à l'aide d'analyses beaucoup plus complexes.

Enfin, nous testons différents paramètres acoustiques (coefficients cepstraux, centres de gravité, position et valeur du minimum dans le spectre,...) et présentons les résultats.

Le paragraphe 2) décrit les analyses, le paragraphe 3) souligne l'importance de l'interaction entre les méthodes d'analyse et de reconnaissance. Le paragraphe 4) décrit les paramètres acoustiques, le paragraphe 5) décrit les expériences et résultats.

2) ANALYSE

a) Nasales, liquides:

La sortie d'un banc de filtres par FFT est calculée toutes les 10 ms sur une fenêtre de Hamming de 17,5 ms. Ce banc comprend 14 filtres ayant chacun une bande passante de 1 Bark (100 Mels). Il couvre le spectre de 200 Hz à 2300 Hz. Pour chaque fenêtre d'analyse n (17,5 ms de parole), 14 réels $Y_{i,n}$ sont calculés ($i=1, \dots, 14$) avec:

Ces paramètres sont calculés avec ou sans préaccentuation, et avec ou sans pondération des $Y_{i,n}$.

$$Y_{i,n} = \log \frac{E_{i,n}}{TE_n}$$

où $E_{i,n}$ est l'énergie en sortie du filtre i pour la fenêtre n . TE_n est l'énergie totale dans la bande de 200 Hz à 2300 Hz pour la fenêtre n . Chaque $E_{i,n}$ est divisée par l'énergie totale, afin d'éviter une trop grande dynamique des $E_{i,n}$ et pour tenir compte, de façon approximative, du fait que l'oreille est plus sensible aux valeurs relatives de l'énergie qu'aux valeurs absolues de celle-ci. Le logarithme est utilisé pour respecter, ici aussi, une des propriétés de la perception.

b) fricatives:

Cette fois-ci, on utilise un banc de 8 filtres répartis entre 2kHz et 7,7 kHz selon l'échelle Bark.

3) INTERACTION ANALYSE - RECONNAISSANCE

Afin de pouvoir comparer les performances de l'analyse, une modélisation de Markov est utilisée à la reconnaissance. On ne peut ignorer le choix des algorithmes de reconnaissance pour concevoir une analyse, et vice-versa. Ainsi, si la reconnaissance est basée sur des algorithmes statistiques, (c'est notre cas) l'analyse doit extraire des paramètres qui sont statistiquement significatifs, sinon les performances ne seront pas bonnes, quelque soit la qualité et la complexité de l'analyse. Afin de tenir compte de ceci, il a été proposé [J. Rouat 88] de multiplier la sortie des filtres, pour plusieurs fenêtres d'analyses successives, par une matrice de transformation F qui permet de modifier les lois de probabilité des paramètres $Y_{i,n}$. Dans le cadre de la présente étude, nous nous contentons de multiplier les $Y_{i,n}$ par des coefficients de pondération, notés b_i .

4) PARAMÈTRES ACOUSTIQUES

a) Nasales, liquides:

- Position et valeur du minimum spectral cherché entre 200 Hz et 1430 Hz ou entre 200 Hz et 2300 Hz.
- Coefficients de type cepstral.

Ces paramètres sont calculés avec ou sans préaccentuation ($a = 0.95$), avec ou sans pondération des $Y_{i,n}$.

b) Fricatives:

- Centre de gravité de la sortie du banc, position du maximum.
- Coefficients de type cepstral.

Ces paramètres sont aussi calculés avec ou sans préaccentuation, et avec ou sans pondération.

5) EXPÉRIENCES ET RÉSULTATS

a) Corpus:

Les expériences ont porté sur la reconnaissance de /l/, /m/, /n/, /f/, /s/, /ʃ/ dans le contexte des chiffres et lettres de l'alphabet anglais prononcés de façon isolée. Les consonnes choisies sont celles pour lesquelles il n'est pas possible d'effectuer la reconnaissance en se basant sur les voyelles ou diphtongues adjacentes, compte tenu du vocabulaire considéré. Le tableau 1 présente les locuteurs, leur nombre et leur langue maternelle. Il est important de noter qu'au moins dix locuteurs de la reconnaissance n'ont pas de représentants de leur ethnie ou nationalité parmi les locuteurs d'apprentissage. De plus, il existe un déséquilibre volontaire entre la proportion d'hommes et de femmes à l'apprentissage et à la reconnaissance.

Locuteurs d'apprentissage		Locuteurs de tests	
16 femmes 24 hommes		25 femmes 15 hommes	
Langue maternelle ou accent	Nombre	Langue maternelle ou accent	Nombre
Français de France	3	Français de France	3
Français du Québec	6	Français du Québec	3
Américain	2	Américain	4
Anglais du Canada	16	Anglais du Canada	16
Anglais d'Angleterre	2	Anglais d'Angleterre	1
Arabe	1	Arabe	2
Chinois de Canton	1	Akan (Ghana)	1
Chinois	1	Australien	1
Créole de Mauritanie	1	Tchécoslovaque	1
Grec	2	Allemand	3
Hébreu	2	Italien	1
Islandais	1	Jamaïcain	1
Perse	1	Polonais	1
Espagnol	1	Vietnamien	2

Tableau 1 Liste des locuteurs de la base de données.

b) Reconnaissance:

Chaque /l/, /m/, /n/, /f/, /s/ et /ʃ/ est représenté par un modèle de Markov. Une chaîne est utilisée pour les femmes, une autre pour les hommes. Les chaînes sont simples: 4 états pour les nasales/liquide, 3 états pour les fricatives. L'apprentissage a été effectué en utilisant

une seule prononciation par locuteur et par lettre "l", "m", "n", "s", "f", "h" et "x". La reconnaissance s'est faite sur les 40 autres locuteurs, à raison de 2 prononciations par locuteur et par lettre à reconnaître.

La figure 1 présente les résultats obtenus pour les nasales/liquide. Sans préaccentuation et sans pondération (npré, npond), les meilleures performances sont observées pour 5 coefficients cepstraux. Le taux est de 88.2%. Dans ce contexte, l'utilisation du minimum ne donne rien d'intéressant. Lorsque la préaccentuation est utilisée, le taux de reconnaissance reste stationnaire à 88.7% pour les 5 coefficients cepstraux, mais passe à 90.5% dans le cas où le minimum est utilisé (pré, npond 1.43 et pré, npond 2.3). La préaccentuation améliore nettement les performances. Ceci s'explique par le fait que la distribution statistique du minimum est alors modifiée et convient mieux à une modélisation de Markov. Par contre, la pondération choisie, notée #1, améliore assez peu les résultats (npré, #1, 1.43).

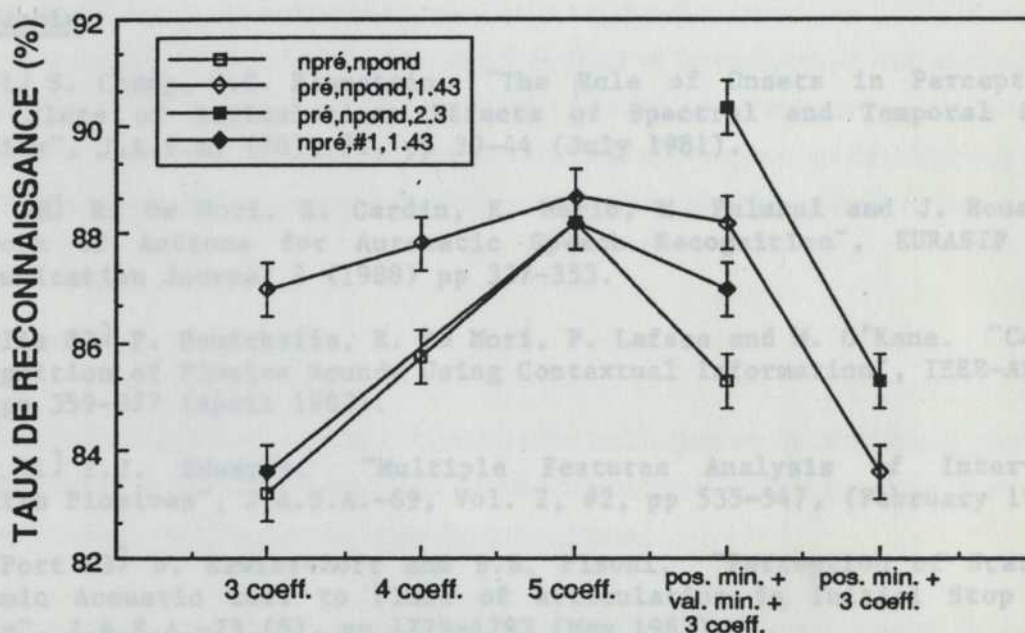


Fig. 1: "l", "m" ou "n" parmi "l", "m" et "n".

En effectuant le même type d'expériences avec les fricatives, il a été trouvé que, contrairement aux nasales, la préaccentuation dégrade légèrement les performances. Par contre, l'utilisation d'une pondération, inspirée des courbes d'isosonie de l'oreille, permet de passer d'un taux initial de 97% (sans préaccentuation, ni pondération) à un taux de 98.5%. Compte tenu du corpus utilisé, on ne peut guère espérer mieux faire.

6) CONCLUSION

Les améliorations obtenues se comparent très bien avec celles observées par d'autres auteurs utilisant des analyses plus sophistiquées. Il est probable qu'il est possible de trouver une pondération qui améliore nettement les résultats dans le cas des nasales. Le but de ces expériences n'était pas de quantifier de façon précise les améliorations, mais de vérifier s'il était possible d'obtenir des changements significatifs.

En conclusion, il est possible d'obtenir de bonnes performances de reconnaissance sur un corpus difficile en utilisant des analyses simples. Elles ne sont utilisables que si le système de reconnaissance est capable de décomposer le problème global en sous-problèmes simples. Nous les avons intégrées dans un système complet de reconnaissance et avons observé s'il y avait des changements sur les taux. Ceux-ci n'ont pas été trop modifiés pour les fricatives (97%) mais la reconnaissance de /l/ /m/ /n/ a chuté jusqu'à 88% en raison de confusions avec /ε/ et /i/.

Bibliographie

- [Chang 81] S. Chang, S.E. Blumstein. "The Role of Onsets in Perception of Stop Place of Articulation: Effects of Spectral and Temporal Discontinuity", J.A.S.A. (70), #1, pp 39-44 (July 1981).
- [De Mori 88] R. De Mori, R. Cardin, E. Merlo, M. Palakal and J. Rouat. "A Network of Actions for Automatic Speech Recognition", EURASIP Speech Communication Journal 7 (1988) pp 337-353.
- [Demichelis 83] P. Demichelis, R. De Mori, P. Laface and M. O'Kane. "Computer Recognition of Plosive Sounds Using Contextual Information", IEEE-ASSP-31, #2, pp 359-377 (April 1983).
- [Edwards 81] T.J. Edwards. "Multiple Features Analysis of Intervocalic English Plosives", J.A.S.A.-69, Vol. 2, #2, pp 535-547, (February 1981).
- [Kewley-Port 83] D. Kewley-Port and B.B. Pisoni. "Perception of Static and Dynamic Acoustic cues to Place of Articulation in Initial Stop consonants", J.A.S.A.-73 (5), pp 1779-1793 (May 1983).
- [Rossen 88] M.L. Rossen et al. "A Connectionnist Model for Consonant - Vowel Syllable Recognition", IEEE-ICASSP 1988, pp 59-62.
- [Rouat 88] Jean Rouat. "Analyse et reconnaissance des consonnes nasales, liquides et fricatives indépendamment du locuteur pour intégration dans un système de reconnaissance automatique de parole alliant approches cognitive et statistique", Thèse de Ph.D., Université de Sherbrooke, Juin 1988.

Conception d'un système de reconnaissance automatique pour la détection et l'analyse des troubles neurologiques de la parole

Eric Keller

Département de linguistique, Université du Québec à Montréal
et Centre de recherche, Centre hospitalier Côte-des-Neiges, Montréal

Résumé

En comparaison au manque de distinctivité qui caractérise les jugements perceptuels des troubles neurologiques de la parole, les analyses instrumentales (et particulièrement les analyses acoustiques) permettent une quantification et une évaluation plus détaillée et plus fiable. Cependant, la quantité de données à évaluer et la complexité des analyses à effectuer, imposent un recours au traitement semi-automatique ou automatique du matériel recueilli. Nous décrivons brièvement l'approche informatisée que nous développons afin d'effectuer le repérage et la quantification des perturbations neurologiques de la parole.

I. Fondements

L'objectif de ces explorations est de circonscrire des techniques de reconnaissance pouvant potentiellement poser un diagnostic semi-automatique ou automatique pour les différentes formes de la dysarthrie et de l'aphasie motrice.

On comprend par dysarthrie, un ensemble de difficultés de la production de la parole impliquant des distortions de la voix et de l'articulation, mais excluant les difficultés de structuration grammaticale ou de compréhension verbale. Le terme aphasie motrice dénote, par contre, une perturbation caractérisée, en sus des distortions de parole, par une lenteur excessive et par des difficultés plus ou moins graves de la structuration, de la recherche lexicale et de la compréhension verbale. Les différents types de dysarthrie se manifestent typiquement suite à des lésions sous corticales, tandis que les cas d'aphasie motrice impliquent fréquemment des lésions ayant des foyers à la fois corticales et sous corticales.

L'oreille humaine ne possède que des capacités limitées pour la différenciation et l'identification des perturbations phonétiques de ce type. Par exemple, nos publications antérieures ont documenté deux cas de parkinsonisme d'évolution semblable (cinq ans après le premier diagnostique), auprès de deux hommes du même âge (65 ans) ayant tous deux une élocution classée "distortionnée" et "dysarthrique" (Keller, 1987a, b). Pourtant, les mouvements du dos de la langue des deux patients étaient forts différents. L'un montrait des mouvements réguliers mais déphasés par rapport au voisement, tandis que l'autre avait des mouvements forts irréguliers mais bien coordonnés sur le plan interarticulatoire. Nous avons également pu documenter, auprès de patients avec dysarthrie cérébelleuse, des dévoisements transitoires mais pathologiques de voyelles qui n'étaient pas ou à peine perceptibles pour l'auditeur entraîné. Il est peu surprenant, par conséquent, qu'une tentative élaborée de poser un diagnostic par unique référence aux indices perceptuelles (Darley, Aronson & Brown, 1969; 1975) aît connu un échec notable.

II. Avantages d'un système de reconnaissance automatique

L'analyse acoustique de la parole pilotée par des algorithmes (semi-)automatisés offrent deux avantages distincts pour l'obtention du diagnostic des perturbations neurologiques de la parole. Premièrement, l'analyse acoustique instrumentale permet une résolution temporelle plus fine, et peut par conséquent identifier un grand nombre de perturbations difficiles à identifier à l'oreille. A l'opposé des systèmes de reconnaissance automatique de la parole normale, un tel système ne cherchera donc pas à modéliser l'opération de l'oreille humaine, mais utilisera la précision temporelle offerte par la numérisation du signal (p.ex., ± 10 kHz).

Deuxièmement, la conception d'algorithmes spécifiques pour la reconnaissance des perturbations neurologiques permettra de simplifier et de rendre plus accessible des analyses qui dans le passé ont été jugées trop lourdes ou trop spécialisées à effectuer dans le contexte clinique de tous les jours. De plus, un système de type hardware/software intégré permettrait l'établissement d'un diagnostique en absence d'un technicien spécialisé.

III. Algorithmes de reconnaissance de troubles neurologiques de la parole

Après avoir exploré pendant plusieurs années les mesures ultrasoniques du dos de la langue pour les fins de diagnostic des perturbations de la parole, nous avons décidé de favoriser plutôt des enregistrements acoustiques; ceux-ci combinent la possibilité de recueillir une grande richesse d'observations avec une simplicité maximale et une intrusion minimale lors du recueil des échantillons. L'analyse est basée sur un examen de la parole spécifiquement développé à ces fins. Nous poursuivons actuellement trois approches dans le développement d'algorithmes de dépistage.

a. Approche traditionnelle

Parmi les épreuves traditionnelles, on compte celle des voyelles et fricatives prolongées qui, chez les patients dysarthriques, met fréquemment en évidence des variations excessives de la fréquence fondamentale et de l'amplitude de voix, ainsi que des cas de voix bitonale ou d'autres irrégularités de la voix.

Par un tableau histographique des sorties d'un banc de filtres ajustés à la fréquence fondamentale, nous préconisons le repérage de variations excessives de la voix et des cas de voix atonale ou bitonale. De plus, en comparant les résultats de filtrages par bande étroite à ceux obtenus par bande large, nous pourrions établir si la pathologie affecte la voix, le conduit vocal supérieur, ou les deux en même temps.

b. Approche par réseaux neuronaux

S'il s'avère vrai que les différents types de pathologie ont des effets variés mais prévisibles sur une réponse donnée, les systèmes d'apprentissage de type "réseau neuronal" devraient être en mesure de les identifier et de les distinguer. Ainsi, notre protocole prévoit 54 répétitions de [pi], [pa] et [pu], et ceci en ordre aléatoire, que nous soumettrons à une analyse par réseau neuronal. Comme entrée à l'algorithme, nous utiliserons les amplitudes des signaux secondaires suivants: l'enveloppe d'amplitude, les passages par zéro, l'extraction de F0 et les sorties du banc de filtres. Le choix précis du type de réseau neuronal à utiliser fera objet d'une étude particulière qui sera abordée sous peu.

c. Approches exploratoires

Finalement, nous planifions trois évaluations exploratoires ayant trait à des perturbations connues, qui ne font cependant pas objet d'une analyse traditionnelle.

(1) **Variations de durées.** En comparant les degrés de variabilité des durées de la voyelle, de la consonne, et du délai intersyllabique dans [papa], [tata], [kaka], [bama] et [ʔaʔa], nous tentons d'identifier les régions dans le conduit vocal qui sont particulièrement affectées par la perturbation neurogénétique. Spécifiquement, nous déduisons des enveloppes d'amplitude du signal acoustique pour chaque suite de stimuli, et nous mesurons les intersections entre l'enveloppe et un seuil donné. Nos recherches pilotes sur la dysarthrie cérébelleuse indiquent que différents patients peuvent souffrir d'une affectation inégale des différentes valves du conduit vocal, ce qui s'exprime par une augmentation substantielle des variations de durée, particulièrement en ce qui concerne la durée de la période silencieuse caractérisant la consonne intervocalique.

(2) **Examen des groupes de consonnes intra- et interlexicaux.** Identifié par Alexandre Luria comme difficulté caractéristique d'un de ses deux types d'aphasie motrice (Luria, 1966), les transitions consonne-consonne forment une difficulté notable pour un grand nombre de patients avec perturbation neurologique de la parole. Cependant, l'examen systématique de ces transitions n'est que rarement entrepris. Le protocole actuel intègre cinq cas de [kt] et de [st] intralexical (p.ex. "acteur"), ainsi que cinq cas de [kt], [st] et de [pt] interlexical (p.ex. "le duc te regarde"). Nous prévoyons l'analyse des durées des différentes consonnes, et nous nous attendons à des prolongations excessives dans le cas de certains tableaux cliniques tel que l'aphasie motrice.

(3) **Analyses du rythme.** Le rythme régulier sera mesuré, (semi-) automatiquement et séparément pour les débits conversationnel et rapide, sur les stimuli du type [papa] et [tata]. (Nous prenons comme indice de rythme, la variation sur le délai intersyllabique). De plus, le protocole intègre cinq conditions de rythme irrégulier (p.ex. [pápápá] ou [pápápá]), impliquant des mesures similaires. Nous nous attendons à ce que certaines pathologies, telle que les dysarthries parkinsonnienne ou cérébelleuse, soient caractérisées d'une part, par des variations excessives et d'autre part, par une incapacité d'effectuer des

différences systématiques entre la production conversationnelle et la production rapide, ou entre les syllabes accentuées et les syllabes non accentuées.

VI. Implantation soft- et hardware

Nous développons actuellement une station de travail à bas prix et facile d'utilisation (user-friendly) incorporant la présentation des stimuli, le recueil des réponses, leur analyse interactive sur écran et la présentation sommaire des résultats. Au niveau de l'équipement, nous préconisons soit une unité de Macintosh Plus ($\pm 2500\$$) dotée d'un système de numérisation de qualité raisonnable mais peu couteux («MacRecorder», 8 bits, 22kHz, $\pm 200\$$), soit une unité de Macintosh SE-30 ($\pm 5500\$$, fonctionnement fort accéléré) incorporant une carte de numérisation et de resynthèse de haute qualité («Digidesign», 16 bits, 43 kHz, $\pm 2600\$$).

Au niveau du logiciel, nous prévoyons l'adaptation de notre logiciel de traitement et d'affichage du signal ("Signalize"), afin de permettre la présentation sériée des stimuli, l'acquisition des données, leur traitement interactif sur écran et l'affichage des résultats obtenus.

Références

- Darley, F.L., Aronson, A.E., & Brown, J.R. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, 12, 246-269.
- Darley, F.L., Aronson, A.E., & Brown, J.R. (1975). *Motor Speech Disorders*. Philadelphia: W.B. Saunders.
- Keller, E. (1987a). Mesures ultrasoniques des mouvements du dos de la langue en production de la parole: Aspects cliniques. *Folia Phoniatrica*, 39, 52-62.
- Keller, E. (1987b). Ultrasound measurement of tongue dorsum movements in articulatory speech impairments. In J.H. Ryalls (Ed.), *Phonetic Approaches to Speech Production in Aphasia and Related Disorders* (pp. 93-112). San Diego, CA: College-Hill Press.
- Luria, A. R. (1966). *Higher Cortical Functions in Man*. New York: Basic Books.

*Present address: Manual Speech & Information Systems,
Alpex-Labs, The Airport, Portsmouth, Hants, PO3 8EX, U.K.

Utilisation de la Méthode d'Analyse Discriminante avec Différentes Analyses Acoustiques dans un Système de Reconnaissance de la Parole.

*Claude Lefèbvre et Melvyn J. Hunt**

Conseil National de Recherche du Canada
Edifice U61, Chemin de Montréal
Ottawa, K1A 0R6, Canada

Introduction

Nous avons considéré dans le passé que les analyses acoustiques basées sur le spectre d'énergie à court terme ou le cepstre dans les systèmes de reconnaissance de la parole ne permettaient pas d'obtenir de bons résultats à cause de leur sensibilité au bruit environnemental, au changement de balance spectrale comme quand une personne élève la voix et à la ressemblance des mots utilisés. Pour contrer ces problèmes, nous avons pensé que l'analyse acoustique devrait simuler l'analyse de la parole faite par le système auditif humain. Nous avons donc construit et testé un modèle de la cochlée. Ce modèle a la particularité de produire deux représentations acoustiques pour des signaux de parole. Afin d'optimiser la combinaison des représentations, nous avons fait une analyse discriminante de toutes les représentations acoustiques des mots utilisés. Les résultats obtenus pour des tests de reconnaissance de la parole utilisant le modèle de la cochlée avec l'analyse discriminante ont été comparés à ceux obtenus avec l'utilisation d'une analyse acoustique cepstrale [1]. Pour tous les tests performés, nous avons toujours obtenu les plus bas taux d'erreurs avec les représentations provenant du modèle de la cochlée.

Dans des tests de reconnaissance de la parole que nous avons faits plus récemment avec les représentations acoustiques obtenues d'une banque de filtres triangulaires, nous avons utilisé une transformation matricielle discriminante au lieu d'une transformation cosinuidale. Dans ce communiqué, nous comparons les résultats des tests de reconnaissance pour ces deux transformations. Nous présentons aussi les résultats de l'utilisation de pondérations sur les coefficients cepstraux ainsi que de l'utilisation de l'analyse discriminante pour combiner plusieurs représentations obtenues d'une banque de filtres

*Present address: Marconi Speech & Information Systems,
Airspeed Lane, The Airport, Portsmouth, Hants, PO3 5RE, U.K.

triangulaires. Tout dernièrement, nous avons fait des tests de reconnaissance avec un groupe de logatomes dont les résultats sont discutés dans ce texte.

Tests de reconnaissance de la parole

Pour tous les tests de reconnaissance des chiffres, nous avons utilisé des séries de trois chiffres enchaînés prononcés en anglais par 9 hommes et 6 femmes. Cinq des 9 hommes et toutes les femmes ont l'anglais nord-est américain comme langue primaire. Trois hommes ont le français et un homme a l'anglais avec un accent sud-africain. Les chiffres de référence sont obtenus de la façon suivante: la moyenne pour chaque paire d'exemplaires est calculée, les moyennes résultantes sont recombinaées en paires et le même calcul est répété jusqu'à ce qu'il reste un exemplaire. Les bordures de départ et d'arrivée de toutes les chiffres ont été trouvées manuellement. Pour tester la résistance des analyses acoustiques aux dégradations, nous avons ajouté aux signaux de parole du bruit blanc et nous avons changé la balance spectrale de 6 dB/octave.

Des tests de reconnaissance ont été faits dépendemment et indépendemment du locuteur. Les représentations acoustiques utilisées dans ces tests sont des tranches spectrales ou cepstrales de 19.2 msec. La dissimilarité entre les tranches spectrales ou cepstrales est calculée avec une distance Euclidienne. Les résultats, présentés ailleurs [2], des tests avec un algorithme reconnaissant des mots enchaînés sont consistents avec ceux d'un algorithme conçu pour des mots isolés présentés dans ce communiqué. Tous les deux algorithmes utilisent la programmation dynamique dans l'alignement des mots comparés.

Analyses acoustiques cepstrales

Les représentations cepstrales utilisées dans les tests de reconnaissance sont calculées en appliquant au signal de la parole une transformée de Fourier rapide de 256 points à toutes les 6.4 msec. Une tranche spectrale est obtenue en calculant la moyenne pour chaque trois TFR, donc à un taux de 19.2 msec, et en multipliant la réponse en fréquence de 20 filtres triangulaires espacés par une échelle Mel au spectre à court terme. Les coefficients cepstraux sont obtenus en appliquant une transformation cosinuidale sur la sortie de la banque de filtres triangulaires [3]. Dans tous les tests de reconnaissance que nous avons faits, nous avons exclus C_0 (voir 1,2). Les nombres entre parenthèses dans le texte réfèrent aux résultats dans le tableau 1.

Il a été montré que les taux d'erreurs obtenus avec une analyse cepstrale peuvent être réduits en pondérant les valeurs relatives entre les coefficients cepstraux [4]. La plus

simple façon est d'augmenter linéairement la valeur des coefficients - i.e. en multipliant les coefficients par leur indice [5]. En expérimentant avec différentes pondérations il a aussi été montré que les taux de reconnaissance sont optimisés lorsque les poids augmentent jusqu'au cinquième et sont constants par la suite [6].

Nous avons fait plusieurs tests de reconnaissance de mots isolés avec différents poids appliqués sur les coefficients cepstraux. Parmi ceux-ci (voir fig. 1), il y avait un test pour lesquels les poids ont été augmentés proportionnellement à leurs indices (3). Dans un autre test, les poids appliqués sur les coefficients cepstraux ont été calculés à partir des variances des tranches spectrales pour toutes les classes (4). Dans un troisième, les poids ont été obtenus à partir des variances des tranches spectrales au sein d'une même classe. Les résultats des tests dans le tableau 1, indiquent que les poids appliqués sur les coefficients cepstraux sont optimisés lorsqu'ils sont dérivés par cette dernière méthode [7].

Il a aussi été montré que les taux d'erreurs peuvent être réduits en ajoutant aux représentations cepstrales leur changement en temps [8]. Comme il est montré dans le tableau 1, nous avons confirmé ces résultats (6).

Transformations matricielles discriminantes

L'année passée, nous avons utilisé une transformation matricielle calculée à partir de l'analyse discriminante de toutes les tranches spectrales de mots permis dans un système de reconnaissance pour optimiser la combinaison des deux représentations générées par un modèle de la cochlée. Nous avons aussi appliqué cette même analyse discriminante aux représentations obtenues à la sortie d'une banque de filtres triangulaires [9].

Une transformation matricielle discriminante est calculée à partir des matrices de covariance au sein d'une même classe et de toutes les classes. Assumant que les séquences de tranches spectrales sont multivariées normalement distribuées autour de leurs moyennes, nous avons défini que toutes les tranches spectrales des mots de références représenteraient toutes les classes. Et assumant que les distributions autour de leurs moyennes des tranches spectrales pour chaque mot sont semblables, la matrice de covariance au sein d'une même classe est calculée en alignant tous les exemplaires des mots enregistrés avec leur mot de référence et en estimant la covariance de la différence obtenue entre les tranches spectrales. Les trois premières fonctions discriminantes appliquées sur les tranches spectrales provenant d'une banque de filtres ainsi que les trois premières fonctions cosinoidales, excluant C_0 , sont montrées dans la figure 2.

Nous avons aussi utilisé la méthode d'analyse discriminante pour pouvoir combiner deux et même trois représentations acoustiques obtenues d'une banque de filtres triangulaires. Les résultats présentés dans le tableau 1 correspondent à la combinaison de la sortie d'une banque de filtres et de la différence en temps de cette sortie (10). Nous avons aussi fait des tests avec trois combinaisons, soit les deux précédentes et une troisième qui est la sortie de 20 filtres coupe-bandes (11). Ces résultats très encourageants, nous permettent de réaliser l'efficacité d'une analyse discriminante à combiner plusieurs représentations.

Modèle de la cochlée

Dans le passé, nous avons fait plusieurs tests de reconnaissance de la parole avec des représentations acoustiques obtenues d'un modèle de la cochlée [10]. Le modèle consiste en une banque de filtres de 32 canaux. La résolution fréquentielle de chaque canal varie selon l'échelle de Bark. La sortie de chaque canal est envoyée dans deux fonctions non-linéaires: un détecteur de périodicité [11], et un système de compression non-linéaire. Le premier détecteur est surtout sensible aux trois premiers formants contenus dans la parole. Le deuxième détecteur est surtout sensible à la variation d'énergie contenue dans le signal. L'information contenue dans les représentations acoustiques des deux fonctions non-linéaires sont pratiquement non-corrélées et par conséquent elles sont tous les deux utiles dans la reconnaissance des mots. Nous avons combiné les deux représentations avec une transformation matricielle discriminante. Les résultats des tests de reconnaissance pour des mots isolés sont montrés dans le tableau 1 (7).

Tests avec des logatomes

Quoiqu'une transformation discriminante permet d'obtenir de meilleurs résultats qu'une transformation cosinuidale dans la reconnaissance des chiffres, il reste à savoir si elle peut être utilisée indifféremment des mots permis par le système de reconnaissance. Récemment, nous avons fait quelques tests de discrimination entre plosives initiales dans un contexte CVC utilisant la transformation discriminante calculée à partir des chiffres. Les tests montrent que la transformation discriminante appliquée seulement sur la sortie d'une banque de filtres donnent des taux d'erreurs inférieurs à ceux d'une transformation cosinuidale, mais ils sont supérieurs lorsqu'on utilise deux ou trois combinaisons de la sortie de la banque de filtres. Nous avons aussi calculé une transformation discriminante à partir des logatomes. Comme pour les chiffres, les taux d'erreurs sont les plus bas pour une transformation discriminante appliquée sur trois représentations.

Discussion

Les résultats des tests de reconnaissance de la parole montrent qu'une transformation matricielle calculée à partir d'une analyse discriminante permet d'obtenir des taux de reconnaissance assez élevés pour de la parole dégradée comparé à l'utilisation d'autres transformations comme la transformation cosinuidale. L'utilisation de la méthode d'analyse discriminante avec des représentations obtenues d'une transformée de Fourier rapide permet d'obtenir d'aussi bons taux de reconnaissance, sinon meilleurs que ceux des représentations d'un modèle de la cochlée. La combinaison de trois représentations est celle qui donne les plus bas taux d'erreurs pour toutes les conditions de signaux de parole testées. Les représentations provenant d'une transformée de Fourier rapide ont l'avantage de demander beaucoup moins de temps de calcul. Par contre les représentations obtenues du modèle de la cochlée montrent que le premier formant est détecté sans qu'il soit affecté par les harmoniques du fondamental [12].

Pour tous les tests qui ont été faits, nous avons utilisé le même corpus de parole pour le calcul de la transformation discriminante et les tests de reconnaissance. Dans un système réel de reconnaissance, on ne pourrait pas procéder de cette façon, et par conséquent on pourrait avoir donné un certain avantage à la transformation discriminante. Afin de vérifier que ce n'est pas le cas, nous avons divisé en deux groupes les chiffres enregistrés par chaque locuteur de sexe masculin. Dans un premier test (12), nous avons utilisé une moitié pour le calcul des mots de référence et de la transformation discriminante, l'autre moitié pour les tests de reconnaissance. Le même test a été fait avec une transformation cosinuidale (13). La différence entre les taux d'erreurs est légèrement plus petite que ce que nous avons obtenu précédemment (2,11). Dans un deuxième test (14), nous avons utilisé une transformation discriminante calculée à partir du groupe test. La comparaison des taux d'erreurs obtenus pour ces tests suggère que l'insertion du groupe test dans le calcul de la transformation discriminante ne biaise pas les résultats de façon considérable.

Les tests de reconnaissance avec différents vocabulaires de mots montrent que les transformations matricielles discriminantes sont plus performantes pour le vocabulaire sur lequel elles ont été calculés. Il serait intéressant de calculer et de tester une transformation matricielle discriminante sur des logatomes contenant toute la gamme des sons d'une langue.

Comparaison entre la segmentation implicite et
explicite pour les modèles de Markov

Références

1. M.J. Hunt & C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model," *Proc. IEEE Int. Conf. Acoustics, Speech & Sig. Proc., ICASSP-88*, New York, April 1988, pp. 215-218.
2. M.J. Hunt & C. Lefebvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undergraded Speech," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-89*, Glasgow, Scotland, May 1989.
3. S. B. Davis & P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. on Acoust., Speech & Sig. Proc.*, vol. ASSP-28, pp.357-366, Aug. 1980.
4. B.H. Juang, L.R. Rabiner & J.G. Wilpon, "On the Use of Bandpass Filtering in Speech Recognition," *Proc. IEEE Int. Conf. Acoustics, Speech & Sig. Proc., ICASSP-88*, Tokyo, April 1988, pp. 765-768.
5. K. Pallwal, "On the Performance of the Quefrency-Weighted Cepstral Coefficients in Vowel Recognition," *Speech Communication*, 1982, Vol. 1, pp. 151-154.
6. Y. Tohkura, "A Weighted Cepstral Distance Measure for Speech Recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-88*, Tokyo, April 1988, pp. 761-764.
7. M.J. Hunt, "A Statistical Approach to Metrics for Word and Syllable Recognition," Fall 79 Meeting, Acoust. Soc. America, Salt Lake City, abstract published in *J. Acoust. Soc. America*, 1979, Vol. 86, pp. S636-S638; full original text in the appendix of [9].
8. K. Aikawa & S. Furui, "Spectral Movement Function and its Application to Speech Recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-88*, New York, April 1988, pp. 223-226.
9. M.J. Hunt & C. Lefebvre, "Distance Measures for Speech Recognition," *National Aeronautical Establishment Aeronautical Note AN-57*, Ottawa, March 1989.
10. M.J. Hunt & C. Lefebvre, "Speech Recognition with the NRC Auditory Model," *Proc. Canadian Conference on Electrical and Computer Engineering*, Vancouver, Nov. 1988, pp. 135-138.
11. S. Seneff, "Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model," *Proc. IEEE Int. Conf. Acoustics, Speech & Sig. Proc., ICASSP-84*, San Diego, March 1984, pp. 36.2.1-36.2.4.
12. M.J. Hunt & C. Lefebvre, "Speech Recognition using an Auditory Model with Pitch-Synchronous Analysis," *Proc. IEEE Int. Conf. Acoustics, Speech & Sig. Proc., ICASSP-87*, Dallas, April 1987, pp. 813-816.

Locuteurs de sexe masculin							
# Representation	# de coefs	DL (% d'erreurs)			IL (% d'erreurs)		
		C	B	F	C	B	F
1 trf. cos.	8	1.2	20.6	72.3	6.3	30.9	75.8
2 trf. cos.	12	1.0	18.3	71.2	5.3	28.8	76.4
3 trf. cos. ech. de quefr.	12	0.2	7.6	51.8	6.2	19.8	84.8
4 trf. cos. ech. var. tot.	12	0.2	5.3	54.3	4.2	15.4	85.8
5 trf. cos.-var. d'une cl.	12	0.2	2.9	40.0	3.2	9.4	77.5
6 trf. cos. + δ -cos.	24	0.6	9.7	68.9	3.6	19.9	75.2
7 modèle de la cochlée	8	0.3	2.2	5.3	1.3	6.1	12.6
8 trf. discr. 1	8	0.4	1.1	0.5	1.8	5.3	4.5
9 trf. discr. 1	12	0.2	1.2	0.4	1.5	5.6	3.7
10 trf. discr. 2	12	0.2	0.9	0.4	0.8	3.0	1.7
11 trf. discr. 3	12	0.1	0.5	0.0	0.3	1.9	0.4
12 trf. discr. 3 groupe 1	12	0.0	1.9	0.0	1.0	3.4	1.6
13 trf. discr. 3 groupe 2	12	0.0	1.0	0.0	0.4	3.7	2.1
14 tr. cos. groupe 1	12	1.0	18.4	73.4	4.8	26.9	73.6
Locuteurs de sexe féminin							
1' trf. cos.	8	1.0	23.3	57.4	5.4	29.3	65.2
2' trf. cos.	12	0.8	20.0	55.2	5.3	27.9	65.2
3' trf. cos. ech. de quefr.	12	1.3	20.6	18.6	22.7	46.1	53.2
4' trf. cos. ech. var. tot.	12	0.6	13.8	20.8	15.2	35.0	53.4
5' trf. cos.-var. d'une cl.	12	0.3	9.0	13.1	11.3	29.7	46.7
6' trf. cos. + δ -cos.	24	0.3	8.8	48.0	4.8	13.3	60.3
7' modèle de la cochlée	8	0.1	4.8	4.1	3.3	10.1	9.9
8' trf. discr. 1	8	0.1	2.6	0.6	3.7	8.6	9.1
9' trf. discr. 1	12	0.1	2.8	0.1	6.7	8.8	13.3
11' trf. discr. 3	12	0.1	3.8	0.1	1.1	5.3	1.3

Table 1. Taux d'erreurs pour les tests de reconnaissance de mots isolés avec 1346 et 900 chiffres prononcés par des locuteurs de sexe masculin et féminin respectivement. Les tests ont été faits dépendamment (DL) et indépendamment (IL) du locuteur, avec des signaux de parole enregistrés dans un milieu calme (C), avec du bruit blanc additionné (B) et filtrés (F) de façon à modifier la balance spectrale de 6 dB/octave.

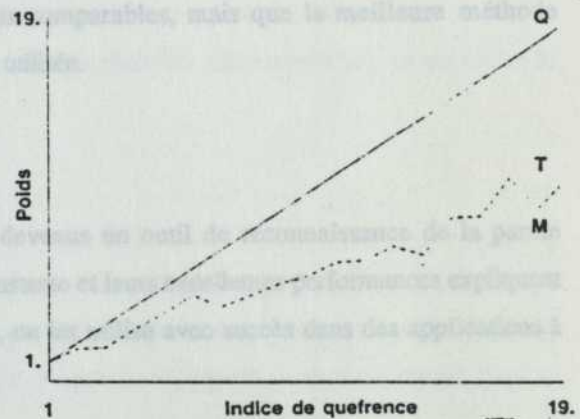


Figure 1. Poids appliqués aux coefficients cepstraux, basé sur la matrice de covariance de toutes les tranches spectrales (T), sur la matrice de covariance au sein d'une même classe de tranches spectrales (M) et sur l'indice de quefrency (Q).

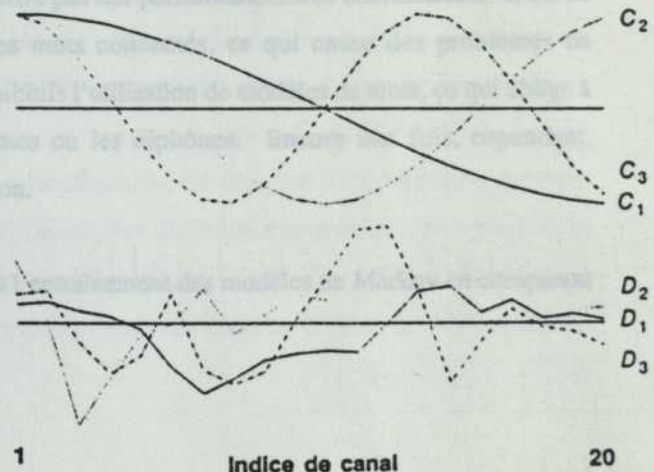


Figure 2. Les trois premières transformations matricielles discriminantes (D) et cosinoidales (C) appliquées sur la sortie des filtres triangulaires.

Comparaison entre la segmentation implicite et explicite pour les modèles de Markov

Yves Normandin et Claude Achcar

Centre de Recherche Informatique de Montréal

1550, boul. de Maisonneuve O., bureau 1000

Montréal (Québec) H3G 1N2

Tel: (514) 848-3990

Résumé:

Cette étude compare trois différentes façons de faire l'apprentissage de Modèles de Markov dans le cas où une segmentation de l'ensemble d'apprentissage s'avère nécessaire. Deux de ces méthodes sont basées sur une segmentation explicite automatique, alors que la troisième accomplit la segmentation implicitement. Les résultats obtenus montrent que les trois méthodes donnent des résultats comparables, mais que la meilleure méthode d'entraînement dépend en général de la méthode de reconnaissance utilisée.

1. Introduction

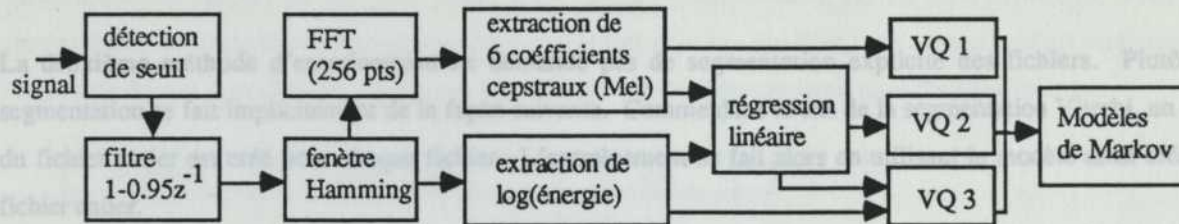
Au cours des dernières années, les modèles de Markov sont devenus un outil de reconnaissance de la parole extrêmement populaire. Leur simplicité, leur flexibilité, leur robustesse et leurs excellentes performances expliquent en grande partie cette popularité. Depuis assez longtemps déjà, on les utilise avec succès dans des applications à petits vocabulaires de reconnaissance de mots isolés.

De plus en plus, cependant, il devient important de pouvoir reconnaître des mots connectés et des vocabulaires toujours plus grands. Cela cause des problèmes au niveau de l'entraînement. En effet, la reconnaissance de mots connectés à partir de modèles entraînés sur des mots isolés n'offre pas des performances très intéressantes. Il est de loin préférable d'entraîner les modèles directement sur des mots connectés, ce qui cause des problèmes de segmentation. De même, des grands vocabulaires rendent prohibitifs l'utilisation de modèles de mots, ce qui oblige à utiliser des modèles d'unités plus petites, telles les phonèmes ou les diphônes. Encore une fois, cependant, l'entraînement de ces modèles pose un problème de segmentation.

Le présent article adresse ce problème de la segmentation dans l'entraînement des modèles de Markov en comparant trois méthodes d'entraînement différentes.

2. Description du système

Le système utilisé dans le cadre de ce travail se voulait aussi conventionnel que possible. La figure suivante illustre comment les paramètres utilisés par les modèles de Markov sont extraits des fichiers de parole. Dans la figure, les blocs "VQ" désignent chacun des livres de code utilisés dans l'étape de quantification vectorielle [1]. La régression linéaire sert à évaluer le taux de changement des 7 paramètres utilisés, générant ainsi 7 paramètres additionnels.



Les distributions de probabilités des modèles de Markov sont associées aux transitions [2]. Cependant toutes les transitions arrivant dans un même état partagent la même distribution de probabilité (elles sont liées), ce qui réduit de beaucoup le nombre de paramètres à évaluer.

3. Méthodes d'entraînement

Tel que mentionné dans l'introduction, le problème considéré ici est celui d'entraîner des modèles de Markov en utilisant des fichiers de parole qui ne sont pas explicitement segmentés. Dans le cas présent, nos efforts ont été concentrés sur les problèmes d'entraînement de modèles de chiffres à partir de fichiers de chiffres connectés et le problème d'entraînement de diphtonges à partir de chiffres isolés.

Deux méthodes ont été essayées. Dans la première, chaque itération d'apprentissage se fait en deux étapes. D'abord, le fichier est segmenté de façon automatique, puis chaque segment est utilisé individuellement pour entraîner le modèle de Markov correspondant. La segmentation automatique se fait en utilisant les modèles provenant de l'itération précédente. Pour la première itération, on utilise des modèles entraînés à partir de mots isolés dans le cas de modèles de chiffres, ou bien à partir d'une segmentation manuelle faite sur un petit nombre de fichiers dans le cas de modèles de diphtonges. Deux techniques de segmentation automatique ont été essayées. La première, basée sur l'algorithme de Viterbi, se fait de la façon suivante. Un modèle de Markov du fichier complet est créé à partir de la concaténation des modèles appropriés. Par exemple, si un fichier représente la suite de chiffres 3-5-3, alors le modèle de Markov du fichier est simplement la concaténation des modèles pour les chiffres 3, 5 et 3. En appliquant

l'algorithme de Viterbi avec ce modèle, on peut, en retraçant le chemin, segmenter le fichier en ses divers chiffres constitutants. Cette segmentation sera d'autant meilleure que les modèles utilisés sont bons. C'est pourquoi il importe de faire un apprentissage initial avant d'appliquer cet algorithme. La deuxième technique de segmentation utilise l'algorithme de "Level-Building" [3, 4] au lieu de Viterbi. Cet algorithme a une ressemblance assez étroite avec l'algorithme de Viterbi, mais il ne nécessite pas de concaténation préalable des modèles de Markov; l'algorithme s'occupe explicitement de combiner les modèles.

La deuxième méthode d'entraînement ne demande pas de segmentation explicite des fichiers. Plutôt, cette segmentation se fait implicitement de la façon suivante. Comme dans le cas de la segmentation Viterbi, un modèle du fichier entier est créé pour chaque fichier. L'entraînement se fait alors en utilisant le modèle ainsi créé sur le fichier entier.

4. Expériences

4.1 Bases de données

Toutes les expériences ont été faites avec les locuteurs adultes mâles de la base de données des chiffres connectés de Texas Instruments, échantillonnée à 10 kHz. Dans cette base de données, chaque fichier contient une répétition d'une séquence de 1 à 7 chiffres (1-9, 'oh' et 'zero'). Les chiffres isolés ont été répétés deux fois par chaque locuteur, alors que les séquences de plusieurs chiffres varient d'un locuteur à un autre. La base de données a été divisée en deux catégories: chiffres isolés et suites de plus de 1 chiffre. Pour les chiffres isolés, nous avons utilisé 40 locuteurs pour l'ensemble d'apprentissage et 43 locuteurs pour l'ensemble de tests. Pour les suites de chiffres, nous avons utilisé 616 suites dans l'ensemble d'apprentissage et 330 dans l'ensemble de test. Nous avons intentionnellement évité les suites de sept chiffres (numéro de téléphone) pour éviter les complications dues à la pause habituellement faite après le premier groupe de 3 chiffres. Pour les besoins de l'expérience, il n'était pas vraiment important de les utiliser.

4.2 Méthode de reconnaissance

Dans presque tous les cas, la reconnaissance a été effectuée en utilisant l'algorithme Viterbi avec un modèle de Markov de type bouclé, fait de tous les modèles de chiffre en parallèle. Comme le but de l'expérience était de comparer trois différentes méthodes d'apprentissage, aucun effort spécial n'a été fait pour optimiser les résultats de reconnaissance. Par exemple, l'utilisation d'un modèle de silence et un traitement a posteriori sur les durées auraient pu grandement diminuer le nombre d'insertions, qui, en général comptent pour plus de la moitié des erreurs. Les résultats de reconnaissance donnés sont le taux d'erreur par mot, qui est le nombre total d'erreurs (incluant insertions,

omissions et substitutions) divisé par le nombre total de mots, et le taux d'erreur par séquence, qui est le nombre de séquence avec erreur(s), divisé par le nombre total de séquences.

4.3 Tests avec modèles de chiffres

Pour les tests avec les modèles de chiffres, les modèles de Markov utilisés avaient tous 8 états. Un entraînement initial a d'abord été fait sur des mots isolés avant de faire l'entraînement sur les mots connectés. Pour fins de comparaison, des tests de reconnaissance ont été effectués avec ces modèles sur l'ensemble d'apprentissage et sur l'ensemble de test. Les résultats (donnés en pourcentage d'erreur) sont montrés dans le tableau suivant.

ensemble	apprentissage	test
mot	8.4%	4.5%
séquence	24.5%	14.5%

Ces modèles ont été utilisés dans les trois cas comme modèles initiaux pour l'entraînement. Dans chaque cas, quatre itérations d'apprentissage ont été faites et des tests de reconnaissance ont été faits sur l'ensemble d'apprentissage et l'ensemble de test avec les modèles obtenus après 2 et 4 itérations. Le tableau suivant montre les résultats.

ensemble	apprentissage				test			
	2		4		2		4	
erreur	chiffre	séquence	chiffre	séquence	chiffre	séquence	chiffre	séquence
Viterbi	1.37	4.61	1.18	4.18	1.71	5.35	1.80	5.66
Level-building	1.56	5.27	1.41	4.61	2.07	6.40	2.52	8.18
implicite	1.46	5.11	1.41	4.95	2.52	8.18	2.25	6.92

4.4 Tests avec modèles de diphônes

Le tableau qui suit montre la transcription en diphônes de chacun des 11 chiffres.

1	WAH	+	AHN	+	N				
2	T	+	TUW	+	UW				
3	THR	+	RIY	+	IY				
4	F	+	FAO	+	AOR	+	R		
5	FAA	+	AA	+	AAYV				
6	S	+	SIH	+	IHK	+	KS	+	S
7	S	+	SEH	+	EHV	+	VAX	+	AXN
8	EY	+	EYYT	+	T				

9	NAA	+	AA	+	AAY	+	YN
OH	AO	+	OW				
ZERO	Z	+	ZIY	+	IYR	+	RAO + OW

Les tests avec modèles de diphônes consistaient à entraîner un total de 35 modèles de diphônes (à 4 états) avec différentes méthodes et d'utiliser ces modèles dans des tests de reconnaissance. Tous les tests dans cette section ont été faits avec des mots isolés. De façon à obtenir des modèles de Markov initiaux (nécessaires pour la segmentation automatique), 14 répétitions de chaque chiffre ont été segmentées manuellement. Pour la reconnaissance, les modèles d'unités obtenus ont été utilisés de deux façons différentes. Premièrement, dans le but de faire de la reconnaissance de chiffres, les modèles d'unités ont été combinés de façon à créer un modèle pour chacun des 11 chiffres. Les modèles d'unités ont aussi été utilisés tels quels pour tenter de reconnaître la séquence de diphônes apparaissant dans les chiffres présentés. Pour ce dernier test, on donne aussi le pourcentage de chiffres dont la suite de diphônes a été reconnue sans erreur. Les résultats sont donnés dans le tableau suivant.

ensemble	apprentissage					test				
	CH	UN (LC)		UN (LI)		CH	UN (LC)		UN (LI)	
test	chiffres	unités	chiffres	unités	chiffres	chiffres	unités	chiffres	unités	chiffres
reconnaissance										
MAN	5.23	67.8	-	-	-	4.02	-	-	-	-
MAN + IMP	0.00	7.79	20.4	12.5	35.3	0.21	11.9	29.8	22.5	53.0
IMP	0.23	7.51	21.0	13.1	36.5	0.21	12.5	33.7	24.4	54.4
MAN + VIT	0.00	7.23	19.2	12.0	33.7	0.21	11.9	31.7	22.3	52.7
MAN + LB	0.00	6.66	17.9	12.9	36.0	0.21	11.7	31.6	22.9	53.4

Colonne 1: tests avec modèles obtenus après:

MAN: segmentation manuelle seulement
 MAN + IMP: segmentation manuelle + segmentation implicite
 IMP: segmentation implicite seulement
 MAN + VIT: segmentation manuelle + segmentation automatique Viterbi
 MAN + LB: segmentation manuelle + segmentation automatique Level-Building

Ligne 2: expériences de reconnaissance

CH: reconnaissance de chiffres à partir de modèles de chiffres dérivés de modèles d'unité
 UN (LC): reconnaissance des unités (longueur connue; utilise Level-Building)
 UN (LI): reconnaissance des unités (longueur inconnue; utilise modèle de Markov bouclé)

5. Discussion

En observant les résultats obtenus dans la section 4.3, on constate que la méthode basée sur l'algorithme de Viterbi donne systématiquement les meilleurs résultats. De plus, l'écart entre Viterbi et les autres méthodes est

généralement assez important. Dans le cas des modèles de diphônes, cependant, il est beaucoup plus difficile de déterminer quelle est la meilleure méthode. On peut toutefois constater que lors des tests de reconnaissance d'unités, les meilleurs résultats ont été obtenus lorsque la méthode d'entraînement utilisée était cohérente avec la méthode de reconnaissance (longueur connue utilise Level-Building; longueur inconnue utilise Viterbi), ce qui semble tout-à-fait raisonnable. Cela peut aussi expliquer les résultats obtenus avec les modèles de chiffres (qui étaient tous du type longueur inconnue). La segmentation implicite, quant à elle, se situe en général entre les deux autres méthodes.

On constate aussi que les modèles obtenus suite à une segmentation manuelle seulement donnent des résultats assez médiocres. Cela s'explique par le petit nombre de fichiers utilisés et aussi par le fait que nous ne désirions en fait qu'une segmentation approximative, capable de conduire à une première segmentation automatique raisonnable. Il est intéressant aussi de constater que les résultats obtenus avec segmentation implicite sans segmentation manuelle préalable sont assez proches de ceux obtenus avec les autres méthodes. Cela nous permet de penser qu'il serait possible de faire l'apprentissage d'un groupe de modèles d'unités sans utiliser de segmentation manuelle, en n'utilisant que les contraintes inhérentes à la transcription en unités de l'ensemble d'apprentissage.

6. Conclusion

Nous avons étudié trois méthodes d'entraînement de modèles de Markov pour des situations dans lesquelles une segmentation de l'ensemble d'apprentissage semble nécessaire. Nous avons constaté qu'en général, les meilleurs résultats sont obtenus lorsque la méthode de reconnaissance est cohérente avec la méthode d'apprentissage.

7. Bibliographie

- [1] K.-F. Lee et H.-W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", *Proc. ICASSP 1988*, article S3.7, pp.123-126, New-York, NY, avril 1988
- [2] L. R. Bahl, F. Jelinek et R. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-5, no. 2, mars 1983
- [3] L. R. Rabiner, J. G. Wilpon et F. K. Soong, "High Performance Connected Digit Recognition, Using Hidden Markov Models", *Proc. ICASSP 1988*, article S3.6, pp. 119-122, New-York, NY, avril 1988
- [4] C. S. Myers et L. R. Rabiner, "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-29, no. 2, avril 1981

ANALYSE DISCRIMINATOIRE DE L'ALPHABET FRANÇAIS PAR ANAMORPHOSE TEMPORELLE

J-P. Cordeau
INRS - Télécommunications
3, place du Commerce
Ile-des-Soeurs, Verdun
H3E 1H6

RÉSUMÉ

La reconnaissance automatique de mots isolés par anamorphose temporelle (ou DTW) est un outil adéquat lorsque les mots dictés se distinguent facilement les uns des autres. Par contre, lorsque ces mots sont tirés d'une base de données à haut niveau de difficulté, où les mots sont acoustiquement semblables, un algorithme spécifique doit être employé pour faciliter la tâche. L'alphabet français est un vocabulaire qui se compose de lettres similaires que l'on peut regrouper en 6 classes distinctes. Les erreurs obtenues par les systèmes de reconnaissance travaillant sur cette base de données sont uniquement dues à des fautes intra-classes. Pour pallier à cette carence, un système à discrimination mixte (cepstrale et temporelle) avec alignement des fichiers de parole est introduit. Cette discrimination pondère davantage les coefficients utiles à la dissociation (les régions temporelle et cepstrale de grande variance) en modifiant légèrement la notion de distance entre la lettre test et la lettre de référence.

I - INTRODUCTION

Le problème théorique de la reconnaissance automatique de mots isolés (RMI) est d'exposer les invariances de la parole. Par exemple, lorsqu'un même mot est prononcé à plusieurs reprises, des différences temporelles et fréquentielles se manifestent entre les sons, et ce sont ces différences qu'il faut éliminer. Un système de reconnaissance doit compenser ces anomalies en se concentrant sur la phonétique des mots qui reste constant, tout en négligeant les autres niveaux d'information, qu'ils soient prosodique ou esthétique.

Cet article traite la reconnaissance des lettres épelées de l'alphabet français par anamorphose temporelle (DTW). D'une part, un tel vocabulaire est essentiel si l'on veut qu'un système de reconnaissance soit assez souple pour accepter l'épellation des mots non reconnu, et d'autre part, il est possible d'envisager la création d'un système de consultation automatique de répertoire téléphonique, où les noms doivent être épelés pour que l'appel soit acheminé.

L'alphabet parlé est un vocabulaire particulièrement difficile pour les systèmes RMI. Il est constitué de mots uni-syllabiques (sauf pour *W* et *Y*) composés souvent d'une consonne et d'une voyelle. Ces mots sont souvent similaires, distincts par la voyelle et/ou la consonne qui les forment. La difficulté provient donc du peu d'information de discrimination disponible et des durées identiques des lettres qui dans presque tous les cas sont inférieures à 600 msec.

L'alphabet considéré dans ce rapport peut être trié en 14 classes, dont 8 sont formées d'un seul élément et les autres de 2 à 8 éléments:

$\Phi(1) = \{ /a/, /ka/ \}$; $\Phi(3) = \{ /i/, /ʒi/ \}$; $\Phi(4) = \{ /y/, /ky/ \}$; $\Phi(5) = \{ /ɛf/, /ɛs/ \}$
 $\Phi(2) = \{ /e/, /be/, /se/, /de/, /ʒe/, /pe/, /te/, /ve/ \}$; $\Phi(6) = \{ /ɛm/, /ɛn/ \}$

La représentation phonétique des classes nous mène à deux constatations: 1) les différences inter-classes sont d'ordre vocalique (sauf $\Phi(5)$ et $\Phi(6)$) et 2) les différences intra-classes proviennent d'un ajout ou d'une modification consonantique.

La classe $\Phi(2)$, composée de 8 lettres à terminaison /e/, est l'un des ensembles posant le plus de problèmes en reconnaissance de la parole. Les composantes sont formées d'une plosive suivie de la voyelle /e/ pour les lettres B, D, P, T; d'une fricative suivie de la même voyelle pour C, G, et V; ou simplement de la voyelle /e/ (pour la lettre é). Dans le contexte alphabétique, la durée de la voyelle est similaire pour chacune des lettres, et ne peut être utilisée à des fins de discrimination. Les régions distinctives se situent plutôt en début de lettre: du silence à la partie stable du noyau vocalique.

Du point de vue articulatoire, les plosives proviennent d'une accumulation de pression à l'intérieur de la cavité orale, puis au relâchement soudain de l'énergie suite à l'ouverture de la bouche. Leur dissociation provient du point d'articulation et de la sorte de modulation comprise dans la rétention de la pression antérieure à la constriction. Dans le cas des explosives, les organes phonateurs ont trois points de contact, menant alors à 6 occlusives.

Une étude réalisée sur les plosives françaises suggère de séparer l'occlusive en quatre: 1) Un silence correspondant à l'obstruction complète du conduit vocal; 2) Une barre d'explosion correspondant à l'ouverture soudaine du conduit oral; 3) Un bruit de friction qui correspond au bruit d'écoulement aux alentours de la région de resserrement et 4) Les transitions des formants qui apparaissent après le bruit de friction et correspondent à l'évolution du conduit vocal au phonème qui suit.

• Conclusions d'ordre temporel:

La durée de l'explosion est supérieure dans le cas des occlusives sourdes (puisque la pression buccale l'est aussi); qu'elle est d'autant plus longue que l'inertie de l'articulatoire est faible (les labiales ont un point d'articulation situé au niveau des lèvres qui se meuvent plus facilement que le dos de la langue); que l'aspiration est présente pour /t/ et /p/ mais absente pour /b/ et /d/.

• Conclusions d'ordre fréquentiel:

L'explosion des occlusives labiales est très faible et l'estimation de la fréquence explosive est par ce fait même très difficile. Pour les occlusives alvéolaires, il est tout de même possible d'affirmer que généralement l'explosion du /t/ est supérieure en fréquence à celle du /d/. D'une manière générale, il est possible d'affirmer que la reconnaissance des plosives est un procédé qui incorpore

à la fois des phénomènes statiques et dynamiques, et que la transition de l'explosion à la voyelle subséquente possède une information discriminatoire considérable [1].

Les systèmes par anamorphose temporelle ne peuvent utiliser ces données dissociatives d'une façon simple car la notion de distance intègre toute l'information fréquentielle d'une façon uniforme par le biais de la distance locale (dans notre cas euclidienne). Une approche permettant d'utiliser cette information est l'analyse discriminatoire.

Le système de référence sur lequel sera transposée l'analyse est celui élaboré par Sakoe et Chiba [2], légèrement modifié pour y inclure une détection de début et fin de mot, en utilisant une fenêtre d'ajustement de 40 msec. Le module de pré-traitement est l'analyse cepstrale à tous les 6.4 msec, avec une fenêtre de 25.6 msec, en utilisant une série de 20 filtres à l'échelle Mel et production de 8 coefficients cepstraux (dont les $N_{cep} = 7$ derniers seront utilisés). La parole est de bonne qualité, quantifiée à 16 KHz et enregistrée dans une chambre anéchoïque.

II - LES FONCTIONS DE DISCRIMINATION

Nous avons mentionné qu'une des raisons pour expliquer la reconnaissance erronée des lettres intra-classes est le peu d'information de discrimination disponible, souvent concentrée dans le temps (les plosives sont temporellement inférieures aux voyelles). Ce phénomène peut être observé en traçant les courbes des distances ponctuelles en fonction des fenêtres d'analyse pour les différents cas qui peuvent se présenter. La figure 1 représente les distances ponctuelles $d_e(T(j(k)), R(i(k)))$ en fonction de l'indice temporel déformé k . Si les lettres T et R sont identiques (courbe pleine), la fonction est distribuée plus ou moins uniformément aux alentours d'une moyenne. Si les lettres sont de classes différentes (courbe tiretée), la fonction est supérieure à la moyenne pour toutes les fenêtres d'analyse. Par contre, pour des lettres T et R de même classe, la courbe démontre un maximum dans la région de dissimilitude (courbe pointillée). Pour les lettres de la classe $\Phi(2)$, d_e n'aura que deux aspects différents: pour des lettres identiques il existera aussi un maximum, attribué à la grande variabilité individuelle des momentanées lors de l'explosion. C'est pour cette raison que les courbes ne peuvent être employées comme fonctions de pondération. Par contre, dans tous les cas, la région du maximum est une région de discrimination, et une fonction qui suit l'allure générale des courbes aiderait la reconnaissance en plaçant l'emphasis de la comparaison dans une région temporelle bien déterminée.

Une expression mathématique de la fonction de pondération peut être obtenue en considérant que les courbes de distance locale suivent une distribution gaussienne ($D_1 \sim N(\mu_1, \sigma_1)$ lorsque les lettres sont identiques et $D_2 \sim N(\mu_2, \sigma_2)$ lorsque les lettres sont différentes). La probabilité de bonne reconnaissance nous est alors donnée par:

$$P(C) = \frac{1}{2} + \frac{1}{2} \operatorname{erf} \left(\frac{\mu_2 - \mu_1}{\sqrt{2\sqrt{\sigma_1^2 - \sigma_2^2}}} \right). \quad (1)$$

Si $\mu_1 \ll \mu_2$, i.e. la différence entre les lettres non identiques est supérieure à la différence des lettres identiques, $P(C)$ s'approche de 1; tandis que lorsque $\mu_2 \simeq \mu_1$, $P(C) \simeq 0.5$ signifiant que la métrique utilisée n'est pas adéquate pour séparer les lettres dans cette région.

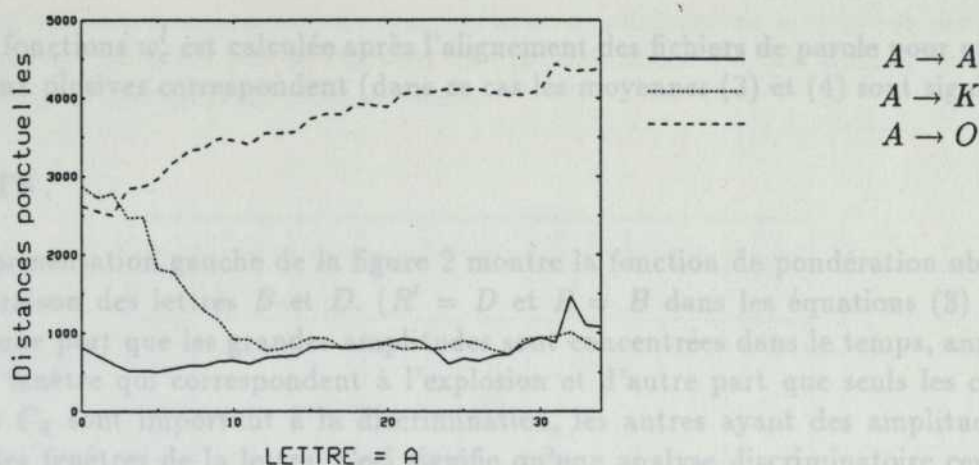


Fig. 1 Distance ponctuelle pour la comparaison des lettres A, K et O.
A → A (pleine), A → K (pointillée) et A → O (tiretée).

Comme nous l'avons mentionné, les indices temporels sont insuffisants pour dissocier toutes les lettres des six classes. En particulier, pour la classe $\Phi(2)$, μ_1 et μ_2 sont presque identiques pour toutes les fenêtres de la comparaison. Pour pallier à cet effet, les fonctions de pondération doivent non seulement dépendre du temps, mais aussi du domaine cepstral. Dans la présente étude la fonction

$$w_l^c(k) = \frac{|\hat{d}_2^{RR'}(k, l) - \hat{d}_1^{RR}(k, l)|}{\sqrt{\sigma_{d_1}^2 + \sigma_{d_2}^2}} \quad (2)$$

est utilisée. $\hat{d}_2$ et $\hat{d}_1$ correspondent alors aux moyennes μ_2 et μ_1 :

$$\hat{d}_1^{RR}(k, l) = \frac{1}{N_{d1}^2} \sum_R \sum_{R'=R} (C_l^{R(k)} - C_l^{R'(k)})^2 \quad (3)$$

$$\hat{d}_2^{RR'}(k, l) = \frac{1}{N_{d1}N_{d2}} \sum_R \sum_{R' \neq R} (C_l^{R(k)} - C_l^{R'(k)})^2 \quad (4)$$

La fonction w_l^c peut alors être utilisée dans la métrique pour mettre une certaine emphase aux régions temporelle et cepstrale d'intérêt.

$$d_e^c(T, R) = \sqrt{\sum_{l=1}^{N_{cep}} w_l^c(k) (C_l^{T(k)} - C_l^{R(k)})^2} \quad (5)$$

Les fonctions w_l^c dépendent de deux groupes distincts de lettres R et R' (N_{d1} lettres R et N_{d2} lettres R'). Elles représentent les fonctions de discrimination mixtes (temporelle et cepstrale) des lettres R et R' . Une grande amplitude pour un certain coefficient à une certaine fenêtre indique une région de l'espace cepstral-temps ayant un grand pouvoir de discrimination.

Chacune des fonctions w_c^l est calculée après l'alignement des fichiers de parole pour que l'explosion entre deux plosives correspondent (dans ce cas les moyennes (3) et (4) sont significatives).

RÉSULTATS

La représentation gauche de la figure 2 montre la fonction de pondération obtenue lors d'une comparaison des lettres *B* et *D*. ($R' = D$ et $R = B$ dans les équations (3) et (4)). Il faut noter d'une part que les grandes amplitudes sont concentrées dans le temps, aux environs de la 19^{ème} fenêtre qui correspondent à l'explosion et d'autre part que seuls les coefficients C_1, C_2, C_3 et C_4 sont important à la discrimination, les autres ayant des amplitudes faibles pour toutes les fenêtres de la lettre. Ceci signifie qu'une analyse discriminatoire cepstrale est nécessaire puisqu'une analyse uniquement temporelle aurait intégrée toute l'information des sept coefficients en un terme de pondération unique. Une représentation équivalente dans le domaine fréquentiel est donnée dans le schéma droit de la figure 2. Cette dernière montre que les régions de haute discrimination (noircies) sont localisées à l'explosion et aux transitions formantiques ($[F_2, F_3] \simeq [2.2, 3.0]$). Les régions noircies antérieures à l'explosion s'expliquent par la présence de la barre de voisement. Les coefficients cepstraux responsables pour cette région fréquentielle sont aussi responsables pour la région $[2.0, 3.0]$ et $[4.0, 4.5]$.

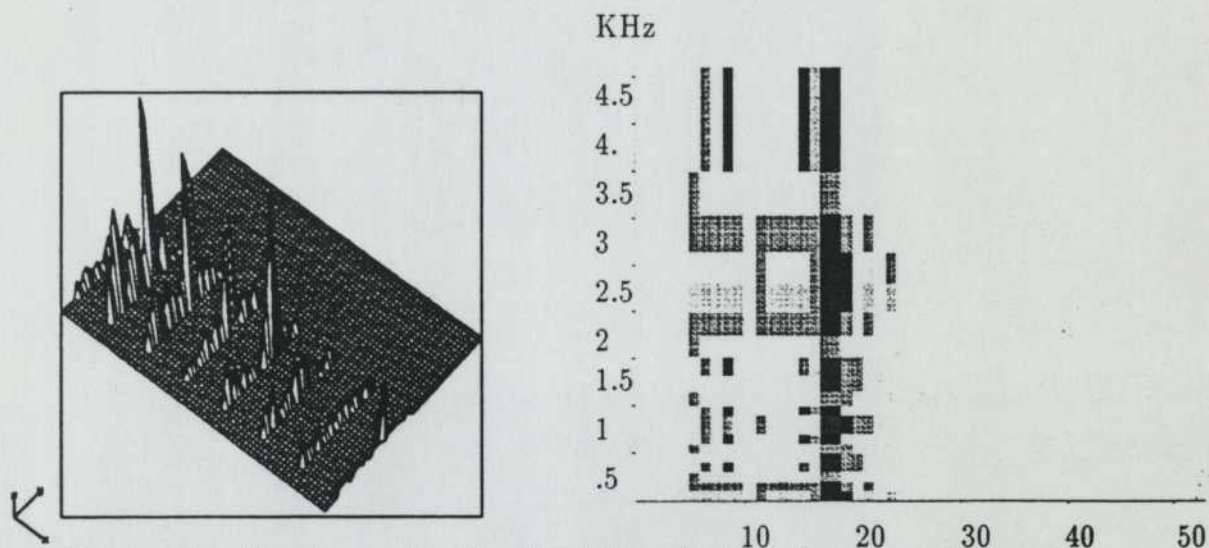


Fig. 2 Fonction de pondération des lettres *B* et *D*. Représentation cepstrale (à gauche) et fréquentielle (à droite).

La matrice de confusion avant et après la discrimination est indiquée à la table 1. Elle montre que la discrimination mixte permet une réduction de 47 % des erreurs de reconnaissance en ne considérant que les lettres des classes problématiques. Le taux de reconnaissance total passe de 88 % à 94 %. L'analyse discriminatoire non seulement aide la reconnaissance mais

regroupe aussi les lettres ensembles: la matrice avec discrimination a tous ses éléments près de la diagonale ce qui n'est pas le cas sans la discrimination. Ce fait diminue le nombre de possibilités dans le processus de reconnaissance.

	F	S	B	D	V	P	T	/	TO
F	6/6	-/-							6
S	1/-	5/6							6
B			6/5	-/1	-	-	-	-	6
D			5/3	1/2	-/1	-	-	-	6
V			1/-	1/-	4/6	-	-	-	6
P			-	-	-	1/2	4/4	1/-	6
T			1	-	-	2/1	1/5	2/-	6
/			-	-	-	1/-	-	5/-	6
ER	0/0	1/0	0/1	5/4	2/0	5/4	5/1	1/0	19/10

Tableau 1 Matrice de confusion sans la discrimination | avec discrimination mixte.

DISCUSSION

Nous avons montré qu'une analyse discriminatoire aide la RMI lorsque les lettres sont tirées d'un vocabulaire à haut niveau de difficulté. Quoique le système présenté est bénéfique, il semble que certaines erreurs ne peuvent être résolues en utilisant cette sorte d'analyse. En particulier, la reconnaissance de la lettre *P* est imparfaite: sur 6 lettres *P* prononcées, 4 sont prises pour un *T*, démontrant qu'une amélioration subséquente est nécessaire. Une procédure plus précise serait possible en changeant le module de pré-traitement (soit plus de paramètre cepstraux ou soit des paramètres fréquentiels plus fins), mais aux dépens d'une augmentation du temps de calcul et d'un accroissement de la complexité du système.

RÉFÉRENCES

- [1] D. Kewley-Port et D. Pisoni, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Am.*, vol. 73 (5), pp. 1779-1793, May 1983.
- [2] H. Sakoe et S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech and Signal Processing.*, vol. ASSP-26, No.1, pp. 43-49, Feb. 1978.

Les Technologies Interactives Vocales et leurs Applications

*Matthew Lennig
Recherches Bell-Northern et l'INRS-Télécommunications
3 Place du Commerce
Ile des Soeurs
Montréal, Québec H3E 1H6*

*tél.: (514) 765-7772
FAX: (514) 766-0732
e-mail: matt@inrs-telecom.quebec.ca*

Introduction

Depuis plus de dix ans l'INRS-Télécommunications et les Recherches Bell-Northern font des recherches dans le domaine des technologies interactives vocales. Celles-ci comprennent la reconnaissance de la parole, la vérification du locuteur, l'encodage et décodage de la parole et la synthèse de la parole. Prises ensemble, ces technologies fournissent une modalité vocale pour l'interaction entre l'humain et l'ordinateur. Cette modalité est particulièrement utile quand il s'agit d'une interaction vocale par téléphone parce qu'elle permet l'interaction du grand public directement avec l'ordinateur.

Le 5 mai 1989 à 6:59 du matin nous avons franchit une étape importante: l'automatisation du premier appel téléphonique à frais virés où les frais ont été acceptés en utilisant la reconnaissance de la parole. Ce système fut mis officiellement en service le 15 mai 1989 à Grand Rapids (Michigan). Il est capable de traiter jusqu'à 48 appels simultanés.

Dans cette communication je mentionnerai les différentes technologies interactives vocales utilisées dans ce système mais je mettrai l'emphase sur celle d'entre elles qui a représenté le plus grand défi pour l'équipe de développement, dont je faisais parti: la reconnaissance de la parole omnilocuteur à travers le réseau téléphonique.

L'automatisation des appels à frais virés et facturés à un tiers.

Dans l'application, il s'agit de l'automatisation des appels téléphoniques à frais virés et des appels facturés à un tiers. C'est une application qui a le potentiel d'épargner plus de \$260 millions *par an* pour les compagnies de téléphone nord-américaines.

Les technologies interactives vocales sont implantées sur un appareil spécialisé qu'on nomme TOPS VSN (Traffic Operator Position System Voice Service Node) qui est installé dans le central téléphonique. Cet appareil est lié au commutateur téléphonique DMS-200 par des voies de données (X.25) et des voies vocales (T-1). Quand un appel 0+ (c.à.d. où le numéro composé est précédé d'un zéro) arrive au commutateur, celui-ci envoie l'appel au VSN. Le VSN reproduit une tonalité de début suivie des instructions vocales expliquant au demandeur les trois options de facturation alternative: frais virés, facturé à un tiers numéro, facturé à une carte d'appel. Le demandeur doit en choisir une, en signalant son choix par les touches Touch-Tone: il signale *un un* pour un appel à frais virés, le numéro de téléphone en question pour facturation à un tiers ou le numéro de carte d'appel pour payer avec sa carte d'appel.

Si le client choisit un appel à frais virés ou facturé à un tiers, le VSN lui demande de dire son nom et enregistre numériquement sa réponse. Le VSN demande au client de bien vouloir attendre pendant qu'on vérifie l'acceptation des frais. Ensuite, le VSN demande au commutateur à travers le lien X.25 de composer le numéro du facturé (dans le cas des frais virés celui-ci est aussi le numéro demandé). Le texte suivant montre un exemple de ce qui arrive dans un appel à frais virés quand le demandé répond:

Demandé

Allo?

VSN

Ici Michigan Bell. Vous avez un appel à frais virés de Matthew Lennig.

Acceptez-vous les frais?

Demandé

Qu'est-ce que vous dites?

VSN

Ici Michigan Bell. Vous avez un appel à frais virés de Matthew Lennig. Veuillez répondre à la question suivante par oui ou non: acceptez-vous les frais?

Demandé

Ouais.

VSN

Merci. Allez-y.

Les technologies interactives vocales suivantes sont employées dans ce dialogue: la reconnaissance de la parole omnilocuteur à travers le réseau téléphonique, l'enregistrement numérique avec compression du débit (encodage), la reconstitution (décodage) de la parole encodée et la détection/réception des tonalités Touch-Tone. Toutes ces technologies sont implantées dans l'unité d'interface vocale qui fait partie du VSN.

Chaque unité d'interface vocale comporte quatre cartes: trois cartes vocales et une carte de processeur général qui fonctionne comme interface entre les applications (qui tourne sur d'autres processeurs) et les cartes vocales. Les cartes vocales contiennent un circuit intégré spécifique à la reconnaissance de la parole qui fait l'anamorphose temporelle. Cette puce, qui s'appelle le R32, est capable de faire environ mille anamorphoses temporelles par seconde pour comparer les gabarits du vocabulaire de reconnaissance avec le mot inconnu.

Précision du reconnaisseur sur le vocabulaire yes/no

Bien que le vocabulaire de reconnaissance soit *yes/no*, dans les conditions réelles d'utilisation les usagers emploient bien souvent d'autres termes pour répondre à la question posée ("Acceptez-vous les frais?"). Les réponses comprennent les exemples suivants: (traduction) *Oui Madame; oui j'accepte;*

ouais; ouah; quoi?; qui parle?; ne quittez pas; no Madame; pas du tout; non merci; etc.

Le reconnaisseur peut faire une des choses suivantes avec l'énoncé d'entrée: l'accepter comme *oui*, l'accepter comme *non* ou le rejeter comme étant en dehors du vocabulaire de reconnaissance. Quand il se produit un rejet, le système rappelle à l'abonné quels sont les choix valables et lui donne une deuxième occasion pour répondre (voir exemple).

Il y a quatre résultats possibles lorsque le reconnaisseur tente de reconnaître un énoncé d'entrée. Si le système accepte l'entrée (e.g., le reconnaît comme *non*) on considère le résultat comme une acceptation juste si la classification est correcte ou comme une fausse acceptation si elle est incorrecte (par exemple, si l'abonné a dit *Qui parle?*). Si le reconnaisseur rejette l'entrée, on considère avoir un rejet juste si l'entrée ne figure pas dans le vocabulaire de reconnaissance ou comme un faux rejet si l'entrée était bien *oui* ou *non*.

Il reste encore le problème des énoncés comme *Oui, Madame*, qui porte le sens de *oui* et contiennent même le mot *oui*. Nous avons créé une classe spéciale pour de tels énoncés; nous appelons cette classe équivalents-oui. De la même façon, nous nommons équivalents-non les énoncés comportant le sens de non et contenant une sous-séquence phonétique correspondant au mot *non*.

Les équivalents-oui et les équivalents-non peuvent être facultativement rejetés. Dans ce cas, ils sont considérés comme des rejets justes. Par contre s'ils sont acceptés dans la bonne classe, ils sont considérés comme des acceptations justes (i.e., équivalent-oui reconnu comme *oui*).

Certaines formes de *oui* et *non* (e.g., *ouais*) sont considérées suffisamment proches des mots en questions qu'elles doivent être obligatoirement reconnues. De telles formes sont considérées comme des faux rejets si elles sont rejetées.

Dans le laboratoire, nous avons testé le reconnaisseur sur un ensemble de 5,000 occurrences de *oui*, *non* et imposteurs (les mots ne figurant pas dans le vocabulaire). Le taux de faux rejet observé est moins de 5% et le taux de fausse

acceptation est de moins de 1%. Maintenant que le système est en service nous sommes en train de vérifier que la même performance est obtenue dans l'application réelle.

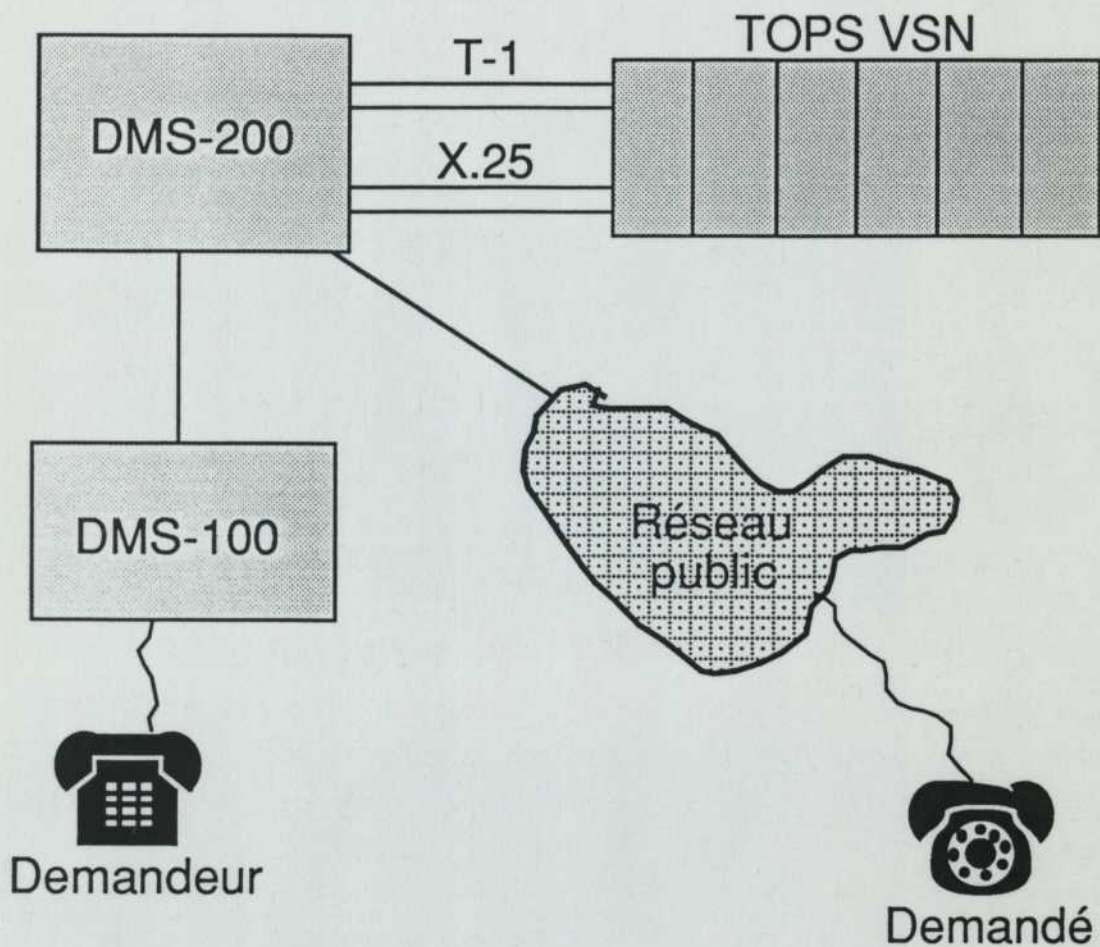


Schéma du système d'automatisation de facturation alternative.

Reconnaissance Vocale et Application en Aéronautique

F. Néel, K. Matrouf, J.L. Gauvain, J. Mariani

LIMSI-CNRS.

B.P. 30

91406-ORSAY CEDEX, FRANCE

tél: 33-1 69 85 80 64

Résumé:

Les recherches en reconnaissance de la parole et dialogue oral développées au LIMSI sont brièvement présentées ici. Une première version d'un système fonctionne avec comme objectif l'aide à la formation au Contrôle Aérien. L'accent est mis sur les performances des nouvelles cartes de reconnaissance, sur le choix d'une représentation des connaissances pour les niveaux supérieurs, compatible avec les exigences d'une application réelle, ainsi que sur les mécanismes de correction mis en oeuvre pour pallier les erreurs provenant du locuteur comme du système de reconnaissance.

I. Introduction:

Parmi les nombreux systèmes de reconnaissance commercialisés depuis plusieurs années, rares sont ceux réellement intégrés dans une application et utilisés en conditions opérationnelles; la plupart des évaluations résultent de tests effectués en laboratoire. En dépit de l'avantage évident de la parole (rapidité, naturel) pour nombre de tâches, les raisons qui expliquent les réticences des utilisateurs potentiels sont non moins évidentes: les contraintes d'élocution (pauses d'environ 200 ms pour certains systèmes), d'utilisation (phase d'apprentissage souvent nécessaire), de vocabulaire (taille limitée), de formulations (syntaxe rigide), d'ambiance sonore (environnement non bruité) sont souvent incompatibles avec le caractère intrinsèquement spontané de la parole, qui implique une grande variabilité à tous les niveaux de la compréhension.

Au LIMSI, l'objectif de nos recherches en traitement de la parole est de mieux comprendre les mécanismes de production et de perception orales. Ces études de nature fondamentale ne sont entreprises que dans la mesure où elles sont susceptibles de mener, à plus ou moins long terme, à la réalisation de prototypes de systèmes susceptibles d'être ensuite commercialisés. Ceci implique la prise en compte de l'utilisateur final et de l'environnement de l'application.

Nous présentons, tout d'abord, les derniers développements réalisés dans notre Laboratoire en matière de reconnaissance vocale, pour ensuite décrire brièvement un système de dialogue intégrant une entrée vocale dans le cadre d'une application à la formation au contrôle aérien. Nous insistons notamment sur ses caractéristiques de flexibilité et de robustesse. Cette étude a été effectuée en collaboration avec le CENA (Centre d'Etudes de Navigation Aérienne).

II. Reconnaissance de la parole

II.1 Le système AMADEUS:

Les premiers systèmes de reconnaissance réalisés au LIMSI en 1981 (MOISE-RMI88 pour les mots isolés) et en 1982 (MOZART-RME186 pour les mots enchaînés) (GAUVAIN, 1982) ont permis d'en cerner les limites et de mieux définir les spécifications d'une nouvelle génération de systèmes. Depuis trois ans, dans le cadre d'un contrat avec la DGT-DAII, nous avons développé en collaboration avec BULL-TRANSAC et VECSYS un processeur VLSI spécialisé pour le calcul des équations de programmation dynamique utilisées dans les algorithmes de reconnaissance de formes, en général, et de la parole, en particulier. Ce processeur μ PCD qui permet d'augmenter la puissance de calcul d'un facteur de 100 par rapport à un processeur standard Intel 80186, a été intégré sur une carte de reconnaissance au format IBM PC. La rapidité de calcul ainsi obtenue permet d'utiliser l'algorithme de comparaison dynamique optimal ou l'algorithme de Viterbi pour lequel les références sont représentés par des Modèles de Markov. Il est ainsi possible de reconnaître de 300 à 600 mots en parole continue et jusqu'à 5 000 mots prononcés en mode isolé.

Le système complet AMADEUS se compose, en plus de la carte de reconnaissance (RdF1000), d'une autre carte pour l'analyse du signal (TdS96/25), toujours au format IBM PC, et d'un boîtier analogique externe (ASA116) pour l'acquisition du signal. La carte de traitement de signal est construite autour du processeur TMS320C25 cadencé à 40 Mhz et contrôlée par la carte de reconnaissance qui comprend, outre le μ PCD cadencé à 20 Mhz, un processeur 80188 cadencé à 9 MHz. Le processeur du PC n'est utilisé que pendant la phase d'élaboration des modèles acoustiques et devient ensuite disponible dans le cadre d'une application pour les niveaux supérieurs de compréhension et de gestion du dialogue. Dans la phase d'analyse, le signal est échantillonné à 10 kHz et quantifié sur 16 bits. Après préaccentuation, on calcule, toutes les 12,8 ms, un jeu de 8 coefficients cepstraux sur des fenêtres glissantes de 25,6 ms.

II.2 L'unité de décision:

L'objectif de la reconnaissance de grands vocabulaires (supérieurs à 10000 mots) nous a conduits à mener plusieurs études en parallèle afin de comparer les performances respectives de différentes unités de décision, telles que le phone, la demisyllabe, le diphone, la syllabe... Les résultats ont montré l'importance de choisir une unité intégrant les parties transitoires du signal, là où la coarticulation est la plus marquée, et ont privilégié la dissyllabe qui se définit comme l'unité s'étendant de la partie stable d'une syllabe (noyau vocalique) ou d'un silence jusqu'à la partie stable de la syllabe suivante (noyau vocalique suivant) ou d'un silence, incluant une ou plusieurs consonnes (SINGER, 1988). Ce choix permet de contourner la difficulté qui existe pour segmenter des groupes consonantiques comprenant des liquides et nasales, et ne présentant souvent pas de zones stables. Pour un vocabulaire difficile tels que les nombres, 137 unités sont nécessaires; les tests montrent une amélioration de près de 6% (pour 4 locuteurs prononçant 400 nombres chacun: 0.87% d'erreurs, au lieu de

6,85% avec des mots comme références).

Le nombre des dissyllabes pour l'ensemble du français peut être estimé à plus de 10 000. L'approche dissyllabe est alors intéressante, dans un premier temps, pour un vocabulaire limité. Dans le cadre d'une application donnée telle que celle présentée ci-dessous, on constate l'occurrence de nombreuses erreurs d'omission ou d'ajout dans l'énoncé, notamment pour des mots monosyllabiques fréquents tels que les chiffres. On peut envisager une approche mixte: les références des mots longs sont stockées sous forme de mots, les mots courts plus difficilement reconnaissables sont stockées sous forme de dissyllabes. Cette approche est celle actuellement choisie pour l'application présentée ci-dessous.

Par ailleurs, AMADEUS prend en compte la syntaxe de l'application et utilise des modèles de signaux ne correspondant à aucun élément du vocabulaire (référence vide), ce qui permet d'accepter des prononciations hésitantes ou bruitées. Dans le cadre de l'application décrite ci-dessous, des règles de syntaxe binaires pondérées sont transmises, à chaque échange, par le système de dialogue à la carte de reconnaissance afin de restreindre l'espace de recherche et améliorer encore le taux de reconnaissance.

III. Système de dialogue oral dans le domaine du Contrôle Aérien:

III.1 Contexte de l'application:

Nous poursuivons depuis plus de cinq années une étude sur la réalisation d'un système de dialogue oral pour la formation des élèves-contrôleurs, dans le cadre d'une convention avec le CENA (Centre d'Etudes de Navigation Aérienne). Normalement est utilisé un simulateur d'entraînement au contrôle du trafic aérien: celui-ci nécessite l'intervention d'une ou quelquefois plusieurs personnes humaines jouant le rôle des pilotes et traduisant les ordres vers le simulateur. L'objectif est donc de réaliser un système susceptible de remplacer le pseudo-pilote et de communiquer directement le contenu des messages au simulateur chargé de mettre à jour l'écran radar.

Une telle application présente de nombreux intérêts:

- les contrôleurs sont habitués au port du casque muni d'un microphone et n'ont donc pas de réticences à utiliser les technologies vocales;
- les technologies vocales fonctionnant en temps réel les met dans des conditions réalistes analogues aux conditions opérationnelles;
- des études ergonomiques (FALZON, 1982) ont montré la taille limitée du vocabulaire qui est compatible avec les performances des systèmes actuels (pour le français: 132 mots-clefs + les chiffres + les lettres (alphabet international) + les noms propres de balises);
- le langage utilisé est de type opératif ("phraséologie"), c'est-à-dire un sous-ensemble du langage naturel, respectant normalement des règles strictes de composition: un message se compose généralement d'un indicatif suivi d'une question ou d'une instruction; la question ou l'instruction concerne des informations, en général, sur le maintien ou la modification de paramètres; chacun des messages peut être catégorisé par un ou plusieurs mots-clefs tels que: cap, niveau, transpondeur...

- l'enchaînement des messages est également déterminé: à une question du contrôleur, une réponse doit être fournie par le pilote; une instruction requiert normalement en retour un collationnement.

Ces constatations doivent cependant être tempérées par les remarques qui ne manquent pas de surgir à l'écoute d'enregistrements réels:

- les contrôleurs se permettent en fait une grande variabilité dans la structure de leurs énoncés qui présentent des omissions et des inversions fréquentes. Dans l'exemple suivant:

"Maintenez le cap" / "Gardez le cap actuel" / "Gardez votre cap" / "Le cap toujours",

ces différents énoncés ont tous la même signification de "maintien de cap". De même, l'indicatif qui est obligatoire dans chaque message admet des règles d'abréviation et de prononciation multiples: l'indicatif IJ0620Q peut s'énoncer sous la forme *"Air Touraine", "India Juliet", "zéro, six, deux, zéro"* ou encore *"zéro six cent vingt"* etc...

Cette constatation nous a amenés à développer un système de dialogue qui réponde à deux préoccupations majeures:

- l'acceptation de variabilité dans la formulation des messages;
- la correction des erreurs provenant tout autant du système de reconnaissance que du locuteur.

Le fonctionnement du système est plus amplement décrit dans (MATROUF, 1989). Nous insisterons ici plus particulièrement sur ces deux aspects: d'une part, le choix d'un mode de représentation des connaissances adéquat, et, d'autre part, les mécanismes de correction, notamment internes.

III.2 Représentation des Connaissances:

Il était exclu de représenter le langage utilisé par les contrôleurs sous forme d'ATN, ce qui aurait mené à une explosion combinatoire pour modéliser l'ensemble des variantes possibles. Le choix s'est porté sur une structure de schémas hiérarchisés (MINSKY, 1975). Un schéma permet de représenter un objet ou un concept (dans le cas présent, une catégorie de messages déterminée par des mots-clefs) par un ensemble de champs qui lui sont associés. A chaque champ, sont affectés des indicateurs servant de directives pour l'instanciation. L'intérêt de cette représentation est de fournir un formalisme unifié pour des informations de niveaux différents tels que syntaxique, sémantique et pragmatique. Dans l'exemple suivant:

(*NIVEAU

(indicatif	ISA IND)
(action	VAL (maintenez descendez...)
	SUC (sujet sens))
(quest	VAL (quel-est-votre...)
	SUC (sujet))
(sens	VAL (la-descente la-montee)
	SUC (sujet)
	PRD (action quest))

(sujet	VAL (niveau niveau-de-vol)
	SUC (param)
	PRD (action sens))
(param	ISA PNIVEAU)
(mexec	VAL (rapidement...))
(methode	INIVEAU))
(PNIVEAU	
(digit1	VAL (0 1 2 3 4 5 6))
(digit2	VAL (0 1 2 3 4 5 6 7 8 9))
(digit3	VAL (0 5)
(valeur	FONC VNBIV)
(sem	(valeur inf 600)
	(valeur sup 30)
	(valeur mul 5)))

les indicateurs VAL et SUC permettent de tester des règles de syntaxe locale: si le mot "maintenez" ou "descendez" a été détecté, il doit être suivi du sujet (niveau, niveau de vol) ou du sens (la descente, la montée). Dans le sous-schéma PNIVEAU, le champ "sem" (sémantique) permet de limiter le domaine de variation du paramètre associé au niveau. Les informations de type pragmatique sont également prises en compte, puisque ces valeurs, par exemple, peuvent être modifiées de façon dynamique au cours du dialogue, en fonction de la trajectoire de l'avion et de celles d'autres avions présents dans le même secteur aérien à un moment donné (connaissances qui sont consignés dans la mémoire dynamique de l'historique du dialogue et du contexte courant de la tâche).

Par ailleurs, il est possible d'utiliser la redondance du message: dans l'exemple "descendez au niveau 230", l'information apportée par descendez est incluse dans le paramètre 230, si l'on connaît le niveau actuel. Ainsi, l'instanciation des différents champs d'un schéma permet de tester la cohérence du résultat de la reconnaissance et de détecter les erreurs afin de les corriger.

III.3 Mécanismes de correction:

On constate que la représentation choisie, tout en permettant au locuteur une plus grande liberté d'expression, est à même de faciliter la correction d'un message incomplet ou entaché d'erreurs. Nous avons ainsi mis l'accent sur le développement de procédures de corrections interne et externe.

Ayant détecté des incohérences aux niveaux syntaxique, sémantique ou pragmatique, nous utilisons des informations acoustiques sous la forme de matrice de confusion de mots ou d'entités phonétiques. Dans le cas où cette procédure se révèle inopérante, nous avons la possibilité de modifier les pondérations des contraintes syntaxiques transmises au système de reconnaissance, avant de recommencer le processus de reconnaissance sur le même signal. Si ces deux tentatives se soldent par un échec, il est toujours possible d'entamer un dialogue de correction en coopération avec le locuteur.

La première des solutions a été la plus explorée jusqu'à présent. Des tests effectués avec 4 locuteurs, prononçant deux fois un corpus de 50 phrases ont montré que les erreurs les plus courantes (56,4%) étaient celles de confusion. Une matrice de confusion donne, pour chaque entité, une liste ordonnée selon la fréquence de confusion de toutes les entités susceptibles d'être confondues avec elle. La correction peut se faire en deux passes, l'entité corrigée devenant, à son tour, un îlot de confiance pour corriger une autre entité du message:

message prononcé "tournez gauche cap 2 3 0"

message reconnu "tournez gauche 4 2 droite 0"

première correction "tournez gauche cap 2 droite 0"

deuxième correction "tournez gauche cap 2 3 0"

Ce type de correction peut être personnalisé, chaque locuteur ayant une matrice de confusion qui lui est propre.

Des tests d'évaluation (sur un corpus de 200 phrases prononcées par 4 locuteurs) montrent que les mécanismes de détection d'incohérences et de corrections internes permettent d'améliorer les performances de reconnaissance de 7% (98,4% de succès obtenus au lieu de 91,4%).

IV. Conclusion:

Les systèmes de reconnaissance tels que ceux développés au LIMSI peuvent être utilisés dans de nombreuses applications de dialogue ainsi que pour des machines à écrire à entrée vocale. L'ensemble de ces cartes est actuellement commercialisé par la Société VECSYS (gamme DATAVOX). La synthèse à partir du texte utilisée est un système développé par le CNET. L'application présentée ici offre un intérêt immédiat pour la formation des élèves-contrôleurs et, à terme, peut être envisagée comme système de veille pour une meilleure sécurité des échanges entre pilotes et contrôleurs en situation opérationnelle. Le système de dialogue décrit brièvement ici fonctionne actuellement sur PC, en temps réel, en langage LE-LISP. Des évaluations de la première version du système ont été effectuées avec des contrôleurs (BAILLEUL, 1988). Plusieurs extensions sont actuellement étudiées, notamment à la langue anglaise, pour une modélisation plus complète des initiatives du pilote et enfin pour améliorer la formulation des réponses et la rendre plus proche du naturel. Ce projet fait l'objet d'une convention avec le CENA et des partenaires industriels STERIA, CROUZET-SA et VECSYS, et vise à réaliser un prototype industriel d'ici deux ans, avec des évaluations systématiques.

BIBLIOGRAPHIE

- GAUVAIN J.L., Reconnaissance de mots enchaînés et détection de mots dans la parole continue, Thèse de troisième cycle, Paris XI, juin 1982.
- SINGER H., GAUVAIN J.L., Connected speech recognition using dissyllable segmentation, Proceedings of the Acoustical Society of Japan Conference, October 1988.
- FALZON P., Les communications verbales en situation de travail: Analyse des restrictions du langage naturel, Rapport technique n° 19 INRIA, 1982.
- MATROUF K., NEEL F., et MARIANI J., Système de dialogue orienté par la tâche: une application en aéronautique, J. Acoustique 2, 1989, 85-93.
- MINSKY M., A framework for representing knowledge in P.H. Winston, ed. The Psychology of Computer Vision, Mc Graw Hill, 1975.
- BAILLEUL C., Evaluation des performances d'un système de reconnaissance vocale dans des tâches de contrôle aérien: poste vocal d'un simulateur de contrôle aérien, J. Acoustique 1, 1988, 237-240.

Intérêts d'une synthèse de qualité

Laurent Santerre

Département de linguistique

Université de Montréal

Introduction : Tout le monde connaît la synthèse par règles, à formants ou par concaténation de diphtonges. La synthèse dont je parle est très différente; elle est faite au moyen du synthétiseur de Klatt (1980) fonctionnant en parallèle dans lequel l'opérateur précise tous les dix ou cinq millisecondes les quinze ou vingt paramètres acoustiques dont la convolution et l'évolution dans le temps dessineront les syllabes de la phrase à imiter. Car il s'agit de synthèse par imitation, la plus fidèle possible; la fidélité de la synthèse repose sur l'analyse juste et précise de la phrase à reproduire (1).

L'intérêt de cette synthèse réside dans le fait qu'elle est un moyen de progresser dans la recherche fondamentale en vue de la synthèse par règles et de la reconnaissance automatique. Je ne connais pas de meilleur moyen d'analyser acoustiquement et psychoacoustiquement la parole réelle. J'ai souvent dit que la sonographie avait débouché les oreilles des phonéticiens en leur ouvrant les yeux. La synthèse fine le fait plus encore. Quelques décennies de lecture des sonagrammes et l'examen de plusieurs milliers d'images de radiofilms m'ont moins appris que quelques mois d'analyses vérifiées par une synthèse très fidèle. Inutile d'ajouter qu'il s'agit là de recherche fondamentale, mais applicable aux industries de la langue. Le nouveau défi que les pays industrialisés se sont donné les oblige à ne négliger aucun moyen qui peut faire ses preuves.

La synthèse fine m'amène à concevoir la recherche fondamentale en vue de la synthèse par règles dans les termes suivants: il faut d'abord connaître l'appareil de production, instrument dont le locuteur a la maîtrise, puis l'usage qu'il sait en faire, comme on distingue le piano et le pianiste. De plus, il y a intérêt à étudier en profondeur la production d'un seul et même locuteur à la fois. Enfin, je dirai un mot de la pratique de la synthèse fine.

L'instrument et le musicien

Les caractéristiques physiologiques d'un locuteur, sa voix, ses résonances formantiques et ses spectres de bruit, sont l'instrument stable et permanent dont il joue dans la production de la parole. Il y a lieu de bien définir les dimensions acoustiques de cet instrument et d'établir comment les divers paramètres interagissent les uns sur les autres au cours de leurs variations; par exemple, comment les spectres varient-ils en fonction des évolutions du F_0 et de l'intensité globale? Plus que le trapèze vocalique individuel, il y a lieu d'établir les résonances de toutes les positions articulatoires au moins sur trois formants, pour ne pas courir le risque de solliciter le résonateur en dehors de ses caractéristiques normales. Il y a des fréquences sur un formant qui ne sont pas naturelles pour un locuteur quand un autre formant est à telle position de sa course possible; ce qui est fréquent pour un locuteur peut ne pas être possible pour un autre. Les caractéristiques des fonctions de source et de transfert sont des contraintes instrumentales qu'il faut respecter à tout moment dans la synthèse.

L'autre aspect à définir consiste dans l'usage que le locuteur fait de cet instrument : ses habitudes articulatoires, ses débits rythmiques, ses écarts mélodiques et dynamiques, ses coarticulations, ses moyens de faire les diverses sortes d'accent; du musicien-locuteur relève la prise en charge de la syntaxe et de la sémantique au moyen de ses patrons prosodiques personnels, de ses groupes accentuels et de ses pauses. La parole synthétique du locuteur paraîtra naturelle seulement dans la mesure où sa parole réelle aura été étudiée et informatisée.

Il serait très long de montrer l'importance de la permanence des caractéristiques instrumentales du locuteur sur tous les points où elle se manifeste. Je ne retiens qu'un exemple : l'homogénéité du système de production des voyelles. La production de quelques mots implique toutes les caractéristiques de la voix et des résonateurs du locuteur; ce produit sonore est inséparable de l'instrument; de même pour la perception : une voyelle porte en elle un timbre phonologique précis, à condition de renseigner en même temps sur les caractéristiques de sa source (homme, femme, enfant, etc.). Le cadre de référence est primordial aussi bien pour la production que pour la perception (Kahn, 1978). Ce cadre de référence physique, c'est l'instrument de la parole. Si au cours de la production, la voix ou les

caractéristiques personnelles de production changent au-delà d'étroites limites, la communication est rompue, parce que la source a changé.

La description de l'appareil de production et la paramétrisation dynamique de ses états possibles pourraient être le travail des ingénieurs acousticiens et informaticiens; aux linguistes incombe surtout la tâche de décrire les règles de production d'une variabilité très complexe. La parole naturelle porte dans ses formes acoustiques toutes ces marques codées par le locuteur et perceptibles par l'auditeur.

Un tel programme peut paraître irréaliste, mais y a-t-il des moyens rapides et simples de faire une chose très complexe? Pourrions-nous contourner la recherche fondamentale dans la synthèse et la reconnaissance de la parole? Peu importent le temps et les moyens qu'il faudra mettre en oeuvre; ces considérations ont bien peu à voir en regard de la nature et de l'étendue des connaissances que l'homme est intéressé à percer.

La parole réelle est nécessairement celle d'un seul et même locuteur

Un ordinateur n'aura jamais à parler comme la moyenne de plusieurs locuteurs d'une langue; de même il n'aura pas à reconnaître la voix et le discours de plusieurs locuteurs parlant en même temps, même s'ils disent la même chose.

A la différence de la parole concrète, la langue est une entité notionnelle; elle n'existe que sous forme de connaissances dans la tête de ceux qui la parlent, l'étudient, l'analysent, la mettent dans des grammaires et des dictionnaires, etc. Par contre, chaque locuteur d'une langue a la compétence de coder du signal sonore de parole selon les conventions communes au groupe dont il fait partie, et il ne peut manquer de produire en même temps les caractéristiques de son dialecte, de ses sociolectes, de ses idiolectes, de ses styles de production, etc. Il ne lui est même pas toujours loisible de ne pas livrer son âge, son caractère, son instruction, son humeur. Le signal de parole est empreint de toutes ces marques individuelles du locuteur. Le champ de la variation est vaste, même pour un seul locuteur.

S'il y a lieu de prendre la moyenne et l'écart-type des mesures acoustiques du signal de la parole d'un locuteur, c'est dans la répétition très semblable et rapprochée des

mêmes mots ou des phrases brèves qu'il faut le faire. Assurément, il y a un intérêt immense à décrire et informatiser la production de deux, trois ou cent locuteurs, mais séparément; il faudra bien le faire avec le temps. Les comparaisons et même les statistiques contribuent à enrichir l'idée que les linguistes se font des langues. Ceci dit, il reste juste que la parole réelle se trouve toujours uniquement dans la production d'un locuteur à la fois, et c'est là qu'il faut d'abord chercher les systématiques.

On ne pourra pas toujours éluder ou minimiser la nécessité de décrire la parole dans ses formes les plus courantes ou systématiques. L'importance de la prosodie se fait de plus en plus sentir en linguistique et en synthèse et reconnaissance; elle est sûrement la partie la plus lacunaire du traitement de la parole. Voilà un domaine de recherche fondamentale qui implique toutes les disciplines linguistiques. Actuellement, les analyses syntaxiques et sémantiques ne visent qu'à permettre à l'ordinateur de choisir entre quelques patrons prosodiques stéréotypés et monotones. Le naturel de la synthèse repose pour une bonne part sur la prosodie (Sorin 1983). Parler de la synthèse de la parole réelle en ces termes paraît relever un peu de l'utopie; se cacher ces évidences ou espérer les contourner relèverait d'une utopie inverse.

La pratique de la synthèse fine (2)

Il faut avoir travaillé et réussi à faire de la synthèse de qualité pour savoir jusqu'à quel point elle oblige à raffiner l'analyse de la parole naturelle et pour comprendre ce qu'elle peut nous apprendre. Malheureusement, le synthétiseur en mode parallèle de Klatt, aussi précieux qu'il soit, devrait être repensé pour permettre de prévoir plus facilement la sortie quand on connaît les paramètres acoustiques d'entrée. Cette tâche relève des ingénieurs acousticiens. Nous avons fait peu de changements au plan du synthétiseur lui-même: nous pouvons introduire le bruit sur le premier formant et la sonorité sur le cinquième. (3)

Les analyses avec ILS et la sonographie analogique du type Kay ne m'ont pas permis de faire une synthèse fidèle au modèle naturel. C'est au moyen des sonagrammes numériques (Maceron 1982) que j'ai pu extraire des paramètres acoustiques assez précis, et là seulement où il fallait qu'ils le soient, pour faire de la synthèse fine. La fréquence exacte des formants (F), leur amplitude relative (A) et leur largeur de bande (B) peuvent être estimées

pour chaque trame, au moyen d'approximations polynomiales ou paraboliques du 2^{ème} degré dans les trois filtres qui contiennent chaque formant ((le sommet et les deux pentes) (4).

La comparaison des sonagrammes numériques du modèle naturel avec ceux des différentes synthèses aboutit, après de nombreux essais au début, et plus rapidement avec l'expérience, à une synthèse à peu près impossible à perfectionner. Les comparaisons visuelles ne sont pas à elles seules suffisantes; deux sonagrammes numériques peuvent paraître semblables, mais l'oreille entendra des différences assez importantes entre les deux stimuli qu'ils représentent. Par contre, des sonas différents à la vue en certains points peuvent provenir de stimuli impossibles à distinguer entre eux à l'oreille. Autrement dit, le signal de parole est très riche d'informations en certaines de ses dimensions acoustiques, et beaucoup moins en d'autres, même dans le bas du spectre.

L'analyse par synthèse plutôt rudimentaire (Delattre, 1951) a fait faire de grands progrès à la compréhension de la parole dans les années 50 (Stevens 1972); la synthèse de qualité et les moyens dont nous disposons aujourd'hui pourraient permettre d'autres progrès très appréciables. Ainsi, la synthèse fine m'a appris qu'il faut attacher autant d'importance à la précision temporelle et dynamique des attaques et des fins de voisement qu'aux explosions et aux V.O.T. Selon les locuteurs, le retard du F1 par rapport aux autres formants peut varier de quelques dizaines de ms. La fluctuation de la fréquence et de l'intensité de chaque formant est une condition importante de la qualité "naturelle".

La précision sur le F0 est à suivre au Hz près, et la période peut fluctuer d'un cycle à l'autre jusqu'à 10 pour cent. F1 et F2 doivent être précis à un ou deux pour cent, du moins globalement. Un écart de cinq pour cent sur 10 ou 20 ms peut passer inaperçu, mais il change le timbre sur un segment plus long. Les seuils de trois pour cent proposés par Flanagan (1957) sont à vérifier. Sur F3 et F4, la tolérance peut aller jusqu'à 10 pour cent et même plus.

Pour ce qui est de l'intensité relative des formants, elle est à surveiller à 1 dB près, surtout à l'attaque et à la fin; A1 doit être précis dans la mesure où F1 est bas. Au-dessus de 1500 Hz, des écarts allant jusqu'à 6 dB sur les A affectent peu le timbre. Quant aux largeurs de bande (B), sauf dans le cas difficile des nasales, elles peuvent être estimées grossièrement; il suffit de s'en tenir aux propositions de Klatt (Carlson 1979 et Klatt 1980).

Le point qui exige le plus d'ajustements par comparaison avec les sonagrammes numériques est celui qui tient à l'influence des variations du F_0 sur le spectre. Une variation de 15 à 20 pour cent sur la mélodie ne provoque pas trop de changements perceptibles dans le timbre, mais au-delà elle oblige à remodeler la pente de la fourniture laryngée et les intensités relatives des formants, et parfois même leurs fréquences. On peut trouver à peu près les mêmes constatations dans la thèse de Biétry (1986). Une des difficultés du synthétiseur en parallèle consiste à prévoir les intensités relatives des formants selon qu'ils se rapprochent ou s'éloignent les uns des autres.

Conclusion : la recherche est fondamentale

Quels que soient les moyens que nous employons, si les scientifiques veulent répondre aux attentes qu'ils ont suscitées dans le domaine des industries de la langue, il faudra faire avancer considérablement les applications de la technologie, mais surtout les connaissances sur la langue et le langage. Cette recherche est fondamentale et orientée; les résultats en sont presque immédiatement applicables; nous avons la possibilité d'examiner à loisir les aspects de la surface sonore, mais entre la surface et les profondeurs qui lui donnent ses infinies variations formelles, que de mystères nous sollicitent.

Même si une bonne synthèse par règles était possible dans quelque dix ans, combien seraient préparés à la mettre en oeuvre? En effet, il ne sera pas suffisant de taper du texte sur le clavier ou de le livrer au lecteur optique; il faudra encore introduire dans la chaîne une foule de marqueurs linguistiques qui seront autant de décisions de l'opérateur et qui supposeront des connaissances du langage beaucoup plus nombreuses, fines et approfondies que ce dont les linguistes et les scientifiques du langage disposent aujourd'hui. Pendant longtemps, les scientifiques de la parole seront, comme les scribes du Moyen-Age, les seuls capables de faire parler convenablement l'ordinateur. L'école devra enseigner dans l'avenir bien plus que l'orthographe et l'accord des participes. Pensons seulement aux fonctions très complexes de la prosodie dans la langue parlée et au peu qu'on en sait de concret aujourd'hui.

Victor Zue (1982) dit : "The speech scientists and acoustic phoneticians must specify the phonetic rules of a given language in quantitative detail. Parameters capturing

the relevant acoustic characteristics must be determined and extracted from the speech signal". La synthèse de qualité est non seulement un moyen de le faire, mais elle nous oblige à le faire de façon très instructive.

En science, les prévisions à long terme sont bien hasardeuses. Peu importe au fond que les ordinateurs parlent prochainement "comme du monde"; aujourd'hui, les sciences de la matière pure fascinent plus la société que celles de la matière grise. Mais au fond, est-ce que toute science n'est pas une science humaine dans la tête des hommes? C'est lui qui nous intéresse en fin de compte; saurons-nous s'il est capable à la longue de faire parler et entendre cette matière dont il est lui-même fait?

Notes

1. Ce travail est subventionné par le CRSHC (410-88-0164).
2. Cette synthèse a été mise au point avec le concours de Catherine Biétry du LIMSI, à la faveur d'une coopération France-Québec (1982-1985).
3. Grâce à Guy Basque des Services informatiques de l'U. de M.
4. Merci à Guy Basque.

Références

- Biétry, C. (1986).** Synthèse de haute qualité de la parole : étude multi-voix des caractéristiques individuelles de la parole. Thèse de 3e cycle. Paris III.
- Carlson, R., Branstrom, B. and Klatt, D. (1979).** "Vowel Perception: The relative perceptual Salience of selected acoustic Manifestation". STL-QPSR, 3 (4), 730-83.
- Delattre, P., Liberman, A.M., Cooper, F.S. and Gerstman, L. (1952).** "Speech Synthesis as a Research Technique", Proc. 7th Int. Congr. Ling. London, 555-561.
- Flanagan, J.L. (1957).** "Estimates of the Maximum Precision necessary in Quantifying certain Dimensions of Vowel Sounds". J.A.S.A. 29, 533-534.

Kahn, D. (1978). "On the Identification of isolated Vowels". UCLA Working papers in Phonetics 41. 26-31.

Klatt, D.H. (1980). "Software for a cascade/parallel Formant Synthesizer". J.A.S.A. 67(3), March 1980, 971-994.

Manceron, F. (1982). Contribution à l'analyse spectrotemporelle du signal de parole considéré comme une suite d'impulsions acoustiques. Thèse de doctorat Ing., Université de Paris XI.

Sorin, C. (1983). "Evaluation de la contribution de l'intonation à l'intelligibilité de phrases". Recherches acoustiques. C.N.E.T. VII, 141-155.

Stevens, K. N. (1972). "Segments, Features and Analysis by Synthesis". In J. F. Cavanaugh (ed.), Language by Eye and by Ear. Cambridge, Mass. MIT Press.

Zue, V.W. (1982). "Acoustic Phonetic Knowledge Representation : Implications from Spectrogram Reading Experiments". In Haton, J.-P., Automatic Speech Analysis and Recognition. Reidel.

Analyse syntaxique du texte pour la synthèse automatique de la parole

Douglas O'Shaughnessy

INRS-Télécommunications, Université du Québec

3, Place du Commerce

Ile des Soeurs, Québec H3E 1H6

Résumé:

Cette présentation porte sur la mise au point d'un système automatique qui détermine la structure syntaxique d'un texte français et sur les applications en synthèse de la parole à partir du texte. La prosodie de chaque phrase parlée est intimement liée à sa structure syntaxique; e.g.: on fait des pauses aux frontières syntaxiques majeures. Le système présenté ici se sert d'un ensemble de règles qui examinent le contexte de chaque mot (sous la forme de 40 catégories syntaxiques) pour estimer des structures syntaxiques de phrases. Afin de simplifier le programme et de minimiser le calcul, le système utilise un petit dictionnaire, comprenant moins de 200 mots communs comme base de l'analyse syntaxique. Comme complément au dictionnaire, 50 terminaisons de mot communes sont associées à leurs catégories syntaxiques (e.g., '-ment' → adverbe). Des contraintes aux possibilités syntaxiques imposées par des règles du texte français servent à aider à estimer la forme syntaxique des phrases. Par exemple, le système exploite l'accord trouvé entre le genre et le nombre du sujet dans la phrase et la conjugaison du verbe suivant. Les mots nouveaux, et même plusieurs fautes typographiques, ne causent aucun problème majeur avec cet algorithme. De plus, le système n'exige pas d'accent sur les lettres, dans le texte d'entrée. L'algorithme sera décrit et plusieurs exemples seront fournis.

1) Introduction

Un algorithme qui analyse la grammaire syntaxique d'un texte possède plusieurs applications. Pour des systèmes texte-à-parole, la prosodie (la hauteur et la durée des sons) dépend fortement de la structure syntaxique; les mots substantifs (e.g., les noms et les verbes) tendent à recevoir l'emphase, alors que ce n'est pas le cas pour les mots grammaticaux (e.g., les prépositions et les pronoms), et les pauses (ainsi que des phénomènes de hauteur qui leur sont reliés) tendent à se produire aux frontières syntaxiques majeures. Donc, une bonne analyse syntaxique automatique aiderait à produire une prosodie plus naturelle dans la parole synthétique. Une analyse syntaxique est aussi très utile pour aider à insérer les accents dans un texte d'entrée sans accent. Quant aux systèmes traitant les langues naturelles (e.g., la traduction automatique), l'analyse syntaxique est indispensable pour comprendre le texte d'entrée.

La plupart des systèmes d'analyse syntaxique utilise un grand dictionnaire pour identifier la catégorie syntaxique de chaque mot, présupposant qu'un texte d'entrée n'est pas restreint à son vocabulaire possible [1, 2, 3, 4, 5, 6]. Quelques fois, d'autres données utiles sont stockées avec chaque mot, e.g., information sémantique (pour un système qui essaie de comprendre une langue naturelle) ou phonétique (pour la prononciation dans un système texte-à-parole). Une analyse grammaticale normale, cependant, ne requiert pas un tel grand dictionnaire (e.g.,

pour hollandais, voir [7]; pour anglais, voir [8]). Aux cas où l'information supplémentaire d'un grand pourcentage des mots est facultative (e.g., pour insérer des accents en télétexte), il est plus efficace de n'utiliser qu'un petit dictionnaire au lieu de stocker et d'examiner une grande base de données.

Cette présentation décrit un système qui fait une analyse syntaxique du texte français sans accent (mais autrement normal). Il est couramment utilisé avec un synthétiseur texte-à-parole, et améliore la qualité de la parole quand le texte est sans accents (comme c'est le cas, par exemple, dans les textes en majuscules).

2) Applications de l'analyse syntaxique

Quelques systèmes texte-à-parole ignorent la structure syntaxique, mais une telle économie amène des pauses mal placées et une attribution aléatoire de la prosodie. Tandis que la hauteur et la durée dans la parole ont un rapport grammatical complexe aux variables linguistiques (e.g., la syntaxe, la sémantique, les émotions, la vitesse de parole), il est généralement reconnu que la structure syntaxique a un effet majeur sur la prosodie. De plus, du point de vue de la spécification automatique de la prosodie par règles, il est plus facile d'obtenir une analyse syntaxique (au moins une analyse partielle) d'un texte d'entrée que d'obtenir d'autre information linguistique (e.g., la sémantique). Alors qu'identifier l'analyse globale syntaxique complète d'une phrase serait idéal pour le texte-à-parole, une structure locale en termes de groupes nominaux et verbaux peut fournir la plupart de l'information syntaxique dont un synthétiseur a besoin afin d'attribuer une prosodie raisonnable [8, 9, 10].

Pour générer une parole de qualité adéquate, un synthétiseur doit connaître: 1) quels mots reçoivent l'emphase et 2) le point d'apparition des pauses et d'autres phénomènes de frontières qui leur sont reliés. Un locuteur met de l'emphase sur les mots qui sont importants pour l'idée qu'il essayait de communiquer; ces mots sont typiquement des mots substantifs et non pas des mots grammaticaux. Donc un algorithme capable d'identifier ces mots substantifs joue un rôle important. Le système décrit dans cette présentation remplit ce besoin. De plus, une analyse locale de la syntaxe pour chaque phrase d'entrée fournit au système une information suffisante afin de déterminer les frontières syntaxiques majeures, où les effets de prosodie reliés à la syntaxe se retrouvent.

3) Vue de l'analyse syntaxique

L'analyse syntaxique emploie souvent deux stratégies: une locale et l'autre globale. La stratégie locale s'applique d'abord normalement, et va de gauche à droite dans l'ordre d'entrée des mots dans le système. Parce qu'une stratégie globale peut exiger un examen de la phrase complète, parfois l'analyse syntaxique de la partie initiale de la phrase est révisée quand les mots suivants sont analysés. De plus, l'analyse globale devrait tenter une analyse gauche-à-droite (en temps réel).

Localement, notre analyse syntaxique regroupe des mots en unités de phrase. Ces groupes de mots sont plus petits que les unités traditionnelles globales tel que les ensembles nominaux et les ensembles verbaux: 1) groupe nominal (GN), qui est composé d'un nom et des mots immédiatement en avant (e.g., un article, une préposition) ainsi que des mots suivants qui le modifient (e.g., des adjectifs, un participe passé), 2) groupe verbal (GV), qui

est composé d'un verbe facultativement précédé des adverbes et des verbes auxiliaires. En parole, il est rare de trouver des pauses à l'intérieur des séquences de mots correspondant aux deux unités syntaxiques ci-constatées, mais elles apparaissent souvent dans les ensembles nominaux et verbaux. Pour aider à trouver ces ensembles de mots et ses frontières, l'analyse syntaxique exploite des contraintes à l'ordre de mots dans les GNs et les GVs; quand l'ordre normal semble être violé, il est probable qu'une frontière de phrase s'y trouve.

Le problème de segmentation des phrases est assisté par la ponctuation (e.g., des virgules), qui arrive souvent aux frontières d'une proposition. Les traits d'union et les virgules tendent à diviser la phrase en propositions. Cependant, beaucoup de phrases ont peu de ponctuation interne. Des deux-points et des point-virgules, bien sûr, séparent des propositions, mais des virgules peuvent arriver aux frontières syntaxiques plus mineures. Des ponctuations "gauches" (guillemets, parenthèses, etc.) débutent des ensembles, et des ponctuations "droites" terminent ces ensembles.

A) Le dictionnaire

Le système emploie un dictionnaire de moins de 200 mots, chacun identifié par catégorie syntaxique et par autres traits syntaxiques. Ce petit dictionnaire est très différent des autres systèmes d'analyse français dans la littérature, qui exigent plus de 250 000 mots [5, 6]. Les mots du dictionnaire appartiennent aux neuf catégories syntaxiques de base: noms, verbes, adjectifs, adverbes, prépositions, conjonctions, interrogatives, pronoms, et mots d'introduction. Les premières quatre classes comprennent les mots substantifs, et les autres classes sont constituées de mots grammaticaux. Le dictionnaire consiste en à peu près 100 mots grammaticaux et 60 mots substantifs. Les mots grammaticaux se trouvent souvent dans le texte, et fournissent de bons indices pour l'analyse syntaxique. Pour des raisons d'efficacité, seulement une trentaine de mots substantifs sont inclus dans le dictionnaire. Parmi ces mots, on retrouve les adverbes qui ne se terminent pas en *-ment*, quelques adjectifs qui tendent à apparaître devant le nom dans un GN, et une douzaine de verbes communs (la plupart viennent des conjugaisons d'*être*).

Chacune des neuf catégories syntaxiques élémentaires est sous-divisée en sous-catégories en fonction de ses traits syntaxiques, qui produit un quarantaine de classes syntaxiques. Le nombre de classes est moindre que celui trouvé dans les autres systèmes qui analysent le français (e.g., 92 [5] et 170 [6] – comme les autres ne listent pas leurs catégories, il est difficile de découvrir pourquoi ils emploient tellement de catégories).

Les prépositions sont divisées en sous-catégories selon leurs traits pour discriminer des mots suivants: 1) les prépositions qui peuvent précéder des substantifs verbaux et des infinitifs (e.g., *sans*), 2) ceux qui exigent d'être suivis par un mot singulier masculin (*au*, *du*), 3) ceux qui exigent un mot pluriel suivant (*aux*), 4) des cas spéciaux (*a*, qui peut alternativement être un verbe; *des*, qui peut facultativement avoir un accent grave), 5) des mots qui peuvent être soit des prépositions soit autre chose (e.g., *entre*), 6) autres prépositions (e.g., *sur*).

Les conjonctions présentent trois possibilités: 1) une classe générale dont les membres introduisent une proposition subordonnée (e.g., *mais*, *lorsque*), 2) *et*, qui est une conjonction coordonnée menant à une construction plurielle, et 3) *ou*, qui peut ou appartient à la classe générale (*où*) ou peut être une conjonction coordonnée

menant à une construction singulière (*ou*). Comme les conjonctions, presque tous les adverbes tombent en une classe générale (e.g., *aussi*, *enfin*), sauf quelques cas spéciaux: *peu*, *plus*, et *moins*, qui peuvent arriver dans des constructions syntaxiques spéciales.

Parmi les quelques adjectifs dans le dictionnaire, on peut retrouver ceux qui précèdent un mot singulier (e.g., *chaque*) et ceux qui précèdent un mot pluriel (e.g., *plusieurs*). Les mots interrogatifs, qui peuvent introduire une question ou une proposition subordonnée sont divisés en une classe générale (e.g., *comment*) et deux mots spéciaux (*qui*, *que*). *Qui* agit comme un sujet dans sa proposition, tandis que *que* peut agir soit comme un objet soit dans une construction comparative avec *plus* ou *moins*.

Les pronoms sont regroupés en plusieurs sous-catégories: 1) ceux qui agissent comme un sujet (e.g., *je*, *ils*), 2) ceux qui agissent comme un objet (e.g., *eux*, *toi*), 3) ceux qui peuvent être soit sujet soit objet (e.g., *nous*, *ceux*). De plus, les pronoms sont divisés selon leur nombre (e.g., *je* - singulier; *nous* - pluriel) et, dans quelques cas, selon leur personne (e.g., *tu* - deuxième personne) et leur genre (e.g., *elle* - féminin).

Une catégorisation semblable est faite pour la classe syntaxique appelée "mots d'introduction," qui inclut des articles, des pronoms adjectifs, des quantificateurs, et des adjectifs démonstratifs; tous ces mots débutent des GNs et agissent comme des mots d'introduction modifiants. Ils se divisent selon leur nombre (e.g., *notre* - singulier, *nos* - pluriel) et leur genre (e.g., *mon* - masculin, *ma* - féminin).

Quelques verbes communs, dont la plupart des formes conjuguées du verbe *être*, sont inclus dans le dictionnaire afin d'aider à reconnaître des verbes. De tels mots sont classifiés de la façon suivante: soit un substantif verbal (e.g., *étant*), soit un infinitif (e.g., *être*), soit un participe (e.g., *été*), soit un verbe actif avec un sujet singulier (e.g., *sera*), soit un verbe actif avec un sujet pluriel (e.g., *sont*). Finalement, les clitiques (e.g., *se*, *ne*, *y*) constituent le dernier ensemble syntaxique.

B) La liste des suffixes

Une liste de 50 suffixes communs attribués à certaines catégories syntaxiques sert de complément au dictionnaire. Si la terminaison d'un mot non reconnu se trouve dans cette liste, on lui attribue la catégorie syntaxique correspondante; e.g., un mot peut être classifié tentativement comme un adverbe s'il se termine en *-ment* et comme un substantif verbal s'il se termine en *-ant*.

Plusieurs adjectifs qui précèdent les noms qu'ils modifient y sont inclus parce qu'ils ont des terminaisons de mot différentes selon leur nombre et leur genre (e.g., *haut*, *haute*, *hauts*, *hautes*). Des mots se terminant en *-s* sont présumés être "pluriel" et sont examinés dans une deuxième passe afin d'assortir d'autres suffixes de la liste. Quelques terminaisons de mot qui suggèrent un nom féminin sont incluses (e.g., *-ance* et *-gst/ion* - où */gst/* signifie une lettre de l'ensemble [g, s, t]).

Quelques verbes ont des terminaisons de mot distinctives qui sont rarement trouvées dans d'autres classes de mot substantives (i.e., les noms et les adjectifs) avec lesquels ils se confondent. De telles terminaisons sont incluses dans la liste de suffixes, et des mots avec ces terminaisons sont tentativement identifiés comme des

verbes. Les terminaisons sont classifiées par personne (e.g., *-ons* – première, *-ez* – deuxième) et par nombre (e.g., *-ai* – singulier, *-aient* – pluriel). De plus, la plupart des infinitifs peuvent être bien reconnus dans la plupart des cas (e.g., *-/ei/r*).

C) Contractions

Des contractions sont des constructions simples qui omettent la lettre *-e* (ou *-a* dans le cas de *la*) à la fin d'un mot grammatical quand le mot suivant commence par une voyelle ou par la lettre *y* (e.g., *le + an* → *l'an*; *de + y* → *d'y*). Donc, ils ne causent aucune confusion pour une analyse syntaxique sauf une perte de l'information de genre dans le cas de *l'*. Des GNs débutant par *le* ou par *la* spécifient que le nom suivant est masculin ou féminin, respectivement, tandis que la contraction *l'* est ambiguë.

4) Fonctionnement des règles

Le programme procède de gauche à droite, mot par mot, d'un texte d'entrée. On examine le dictionnaire pour chaque mot; en cas de faute, on effectue une recherche dans la liste de suffixes pour une tentative d'attribution. De telles correspondances sont utilisées pour renouveler la tentative d'analyse syntaxique locale et globale, qui découle de l'analyse des mots précédants de la phrase.

A) Identifier des verbes

L'une des difficultés majeures de l'analyse syntaxique du texte est de trouver le verbe dans chaque proposition. Parce que les verbes constituent une classe ouverte, il n'est pas pratique dans la plupart des cas d'employer un dictionnaire complet pour tous les verbes parce que chaque verbe dans sa version infinitive possède 38-39 versions différentes conjuguées; de tels mots n'apparaissent pas dans la plupart des dictionnaires, mais se trouvent certainement dans des textes typiques. Normalement des noms se trouvent immédiatement après un mot d'introduction, et donc sont d'habitude identifiés correctement. Après un mot d'introduction, un mot avec une catégorie syntaxique inconnue est appelé un nom. Parce que le français permet à un nom d'être suivi d'un ou plusieurs adjectifs ou participes, il y a souvent une ambiguïté à savoir si un des mots suivant le nom est un adjectif ou un verbe.

La terminaison de mot la plus commune est la lettre *-e*. En position après un nom, un tel mot est fondamentalement ambigu entre un adjectif féminin et une version d'un verbe (e.g., le subjonctif ou la première conjugaison présent indicatif). Notre analyse syntaxique spécifie qu'un mot est un adjectif si les mots précédents forment un GN possiblement féminin, ou soit un verbe si aucun verbe antérieur ne se trouve dans la proposition. Des adjectifs normalement suivent des noms, des verbes, d'autres adjectifs, et des mots d'introduction, tandis que des verbes tendent à suivre des sujets pronoms, des mots interrogatifs, et des clitiques (*ne*, *se*, *y*). Donc, un mot inconnu suivant un mot de ces dernières classes est déclaré un verbe.

B) L'accord entre mots

Le français suit des règles fortes pour le genre, le nombre, et la personne, ce qui aide à séparer les verbes et les adjectifs. Dans un GN, le nombre et le genre doivent être consistents, et l'évidence se trouve dans les

terminaisons des mots composants. Un GN pluriel, spécifié par un mot d'introduction pluriel (e.g., *les*) ou par un nom se terminant en *-s* ou *-x*, exige que les adjectifs qui le modifient doivent aussi se terminer en *-s* ou en *-x*. Un GN féminin, spécifié par un mot d'introduction féminin (e.g., *sa*) ou par certains noms se terminant en *-e* ou en *-ion*, normalement exige que ses adjectifs modifiants se terminent en *-e*. Un désaccord de ces règles indique une frontière syntaxique, e.g., une frontière entre le GN et le verbe (e.g., le dernier mot dans *La boîte bleue tombait ...* ne fait pas l'accord avec le GN féminin précédant et est donc un verbe).

En plus de l'accord interne d'un GN en nombre et en genre, un sujet GN et son verbe suivant doivent être en accord en nombre et en personne. Une fois qu'un GN sujet est identifié, les mots suivants ambigus au niveau de la catégorie syntaxique peuvent être identifiés comme adjectifs sauf s'ils sont en accord avec le nombre et la personne du sujet.

Normalement des prépositions introduisent un GN qui modifie une phrase précédente, mais quelques prépositions (indiquées au dictionnaire) peuvent précéder un infinitif (e.g., *pour*) ou un substantif verbal (e.g., *en*). Parce qu'un infinitif ou un substantif verbal peut être suivi d'un GN objet, il y a des possibilités d'ambiguïté entre une frontière syntaxique majeure (i.e., entre des GNs dans différentes propositions, non séparés d'une virgule) et une frontière mineure (i.e., entre l'infinitif et son objet dans une phrase infinitive) (e.g., notant la frontière d'un '/': *Parce qu'on travaille pour vivre / un homme pense ...* et *Pour manger / une pomme il faut ...*).

Cette recherche était subventionnée par le FCAC.

Références

1. TAUM Group, Final Report, Projet de Traduction Automatique de l'Université de Montréal, 1973.
2. C. Fluhr, "Analyse syntaxico-sémantique de gros corpus," *Proc. Seminar "Syntax and Semantics in Speech Understanding"*, (GRECO-CNRS: Paris), pp. 41-53, 1980.
3. M. Gross, "Lexicon-grammar and the syntactic analysis of French," *COLING-84* (Proc. Conf. Computational Linguistics), pp. 275-282, 1984.
4. G. Bailly, "Multiparametric generation of French prosody from unrestricted text," *Proc. Intern. Conf. Acoustics, Speech, & Signal Processing*, pp. 2419-2422, 1986.
5. A-M. Derouault & B. Merialdo, "Natural Language Modeling for Phoneme-to-Text Transcription," *IEEE Trans. Patt. Anal. & Mach. Intell.*, vol. PAMI-8, no. 6, pp. 742-749, 1986.
6. C. Sorin et al, "Text-to-speech synthesis in the French electronic mail environment," *Proc. European Conf. on Speech Technology*, 1987.
7. R. Kager & H. Quené, "Deriving prosodic sentence structure without exhaustive syntactic analysis," *Proc. European Conf. on Speech Technology*, pp. 243-246, 1987.
8. D. O'Shaughnessy, "Specifying Intonation in a Text-to-Speech System using only a Small Dictionary," *IEEE Intern. Conf. on Acoustics, Speech & Signal Processing*, pp. 1430-1433, 1987.
9. D. O'Shaughnessy, "A Study of French Intonation for a Synthesis-by-rule algorithm," *Toronto Symposium on Intonation*, pp. 171-178, 1981.
10. O. Martin, "Phonetic realisations of prosodic contours in French," *Speech Communication*, vol. 1, pp. 283-294, 1982.

L'aspect temporel de la synthèse de parole par règles

Danièle Archambault

Département de linguistique

Université de Montréal

C.P. 6128, Succ. A

Montréal, Québec

Canada H3C 3J7

Depuis les dernières années, l'intérêt pour les systèmes de synthèse qui permettent de générer de la parole à partir d'un texte écrit est allé grandissant. Le champ d'application de tels systèmes s'est agrandi et les exigences des utilisateurs quant à la qualité et à la flexibilité des synthétiseurs a augmenté considérablement. La prosodie joue un rôle primordial dans la qualité de parole d'un synthétiseur, particulièrement l'aspect temporel. Cependant, la durée est un paramètre extrêmement difficile à investiguer car elle varie en fonction de nombreux facteurs. Parmi les facteurs de variation de durée les plus importants, on peut retenir la nature du segment, le contexte phonétique, l'organisation syllabique, la structure syntaxique (et donc accentuelle) et le débit. Dans cet article, je présenterai les résultats préliminaires d'une recherche ayant pour but de permettre à un synthétiseur de parole en français, développé à l'INRS-Télécommunications¹, de parler à différentes vitesses de débit. Cette option supplémentaire sur le synthétiseur permettra à l'utilisateur de choisir parmi trois vitesses de lecture: une vitesse normale, pour lecture courante une vitesse lente lorsque les textes sont difficiles, par exemple, et une vitesse rapide pour *feuilleter* le contenu d'un texte. L'INRS-Télécommunications possède deux systèmes de synthèse par règles qui permettent de générer un texte oral français à partir de n'importe quel texte écrit, chacun mettant à profit une technique différente: un synthétiseur à formants² et un synthétiseur par diphones³.

La première étape d'une telle recherche consiste à décrire les principales techniques d'accélération et de ralentissement du débit en français, à partir de l'analyse d'échantillons de parole naturelle. Dans la deuxième étape, les règles dégagées lors de l'analyse précédente seront implantées dans le synthétiseur. Le présent article porte sur la première étape de la recherche.

1. Méthodologie

Les deux principales composantes du débit, généralement défini par le nombre de syllabes prononcées par minute, sont la vitesse d'articulation et les pauses. La plupart des recherches sur le sujet ont eu pour but de cerner les différentes caractéristiques du débit normal à travers différents épreuves discursives (discours spontané, discours politique, lecture, etc.). On peut citer, entre autres,

¹ F. Gauthier-Ruel (1988), "The significance of changes in the rate of articulation".

² R. Lavoie et D. Deschamps, J.P. (1971) "Temporal patterns of rate variations in reading", *Acta Psychologica*, 11, 224-243.

³ F. Gauthier et Deschamps, A. (1972) "Analyse des variables temporelles du français spontané", *Phonetica*, 24.

¹ L'auteure est également professeure invitée à l'INRS-Télécommunications

² D. O'Shaughnessy (1984) "Design of a real-time French text-to-speech system", *Speech Communication*, vol. 3, pp. 233-243.

³ D.O'Shaughnessy, D.Archambault, D. Bernardi et L. Barbeau et al. (1987) "Diphone Speech Synthesis", *Speech Communication*. Vol. 7, No. 1. 55-65

les travaux de Goldman-Eisler⁴ et Lass⁵ pour l'anglais, Grosjean et Deschamps,⁶ et Duez⁷ pour le français. Peu de chercheurs, cependant, se sont penchés sur l'examen des variations imposées de débit (Lass⁸ et Gilbert⁹). Cette étude est la première à examiner les variations imposées de débit en vue de l'application à la synthèse par règles.

Dans cette recherche, j'ai eu recours à deux locuteurs ayant l'habitude de parler en public: un professeur (locuteur 2) et une annonceur de radio (locuteur 1). Les enregistrements ont été faits en chambre sourde au laboratoire de phonétique. Les locuteurs avaient à lire une série de textes de différentes longueurs. Pour chaque texte, les locuteurs devaient alterner leur vitesse de débit c'est-à-dire qu'ils devaient d'abord lire le texte à vitesse normale, ensuite à vitesse lente, relire le texte à vitesse normale, puis le lire à vitesse rapide. Ceci permet aux locuteurs de modifier leur vitesse de lecture en fonction d'une vitesse de référence (vitesse normale).

Pour l'examen des corpus, j'ai eu recours à deux types d'analyse: une analyse perceptuelle et une analyse acoustique. Les différents corpus ont été présentés à des sujets à l'intérieur de deux tests de perception (un pour chaque locuteur). Les sujets avaient pour tâche d'identifier la vitesse à laquelle parlait le locuteur (vitesse lente, normale ou rapide). Cette analyse perceptuelle a pour but de vérifier que le locuteur a réussi à produire la vitesse de débit désirée. Les analyses acoustiques ont été faites au laboratoire de phonétique à l'aide de programmes d'analyse spectrographique implantés sur un ordinateur Zenith AT. Il s'agit principalement de segmentation du signal afin de pouvoir dégager les données nécessaires à l'examen des vitesses d'articulation et des pauses.

2. Résultats: Paragraphe isolé-Locuteur 2

Ce corpus est composé d'un paragraphe constitué de trois phrases complexes (90 mots au total). Le locuteur 2 a réussi à produire trois vitesses de lecture distinctes (durée totale: 35.4 s, 28.6 s et 23.6 s) bien que le texte lent se distingue mieux du texte normal que le texte rapide. Ceci indique que le locuteur a eu plus de facilité à ralentir qu'à accélérer. Ce phénomène a déjà été relevé par Lass (1971) et peut être relié à des contraintes d'ordre articulatoire. En effet, il y a une limite à la vitesse à laquelle un locuteur peut parler et néanmoins conserver l'intégrité du message. Les vitesses d'articulation sont cependant assez semblables (6 syll./s, 6.8 syll./s et 7.5 syll./s, tableau 1). La distinction se fait au niveau des pauses. Il y a trois fois plus de pauses en débit lent qu'en débit rapide et deux fois plus en débit normal qu'en débit rapide. Le rôle privilégié des pauses dans le débit constitue une caractéristique très importante car elle permettrait de modifier le débit du synthétiseur beaucoup plus facilement. En effet, si, pour changer le débit d'un texte, il suffisait de

⁴ F. Goldman-Eisler (1966), "The significance of changes in the rate of articulation."

⁵ N. Lass et Deem, J.F. (1971) "Temporal patterns of rate alterations in oral reading", *Acta Symbolica*, 11, 254-263.

⁶ F. Grosjean et Deschamps, A. (1972) "Analyse des variables temporelles du français spontané", *Phonetica*, 26, 129-156.

⁷ D. Duez, (1982), "Silent and non-silent pauses in three speech styles", *Language and Speech*, Vol 25, Part 1, 11-28.

⁸ N. Lass et Deem, J.F. (1971) "Temporal patterns of rate alterations in oral reading", *Acta Symbolica*, 11, 254-263.

⁹ J.H. Gilbert et W.B. Kenneth (1969) "Rate alterations in oral reading", *Language and Speech*, 12, 192-201

Variations de débit. Par. 5. Loc. 2

	<u>Lent</u>	<u>Normal</u>	<u>Rapide</u>
Débit (syll./mn):	272.9	342	404
Durée totale:	35.4 s	28.6 s	23.6 s
Suites sonores:	20	14	7
Pauses:	19	13	6
Vitesse d'art.	6 syll./s	6.8 syll./s	7.5 syll./s
Rapport phonation/durée totale	76 %	83 %	91 %

Fig. 1 Variations de débit. Par. 5. Loc. 2

modifier le nombre de pauses sans toucher à la vitesse d'articulation, la tâche s'en trouverait non seulement facilitée mais il y aurait ainsi moins de risques de compromettre l'intelligibilité du texte.

Il reste cependant à définir les points d'occurrence des pauses. L'indice le plus important de l'apparition d'une pause est la présence d'un marqueur orthographique (point, virgule etc.). Il existe une hiérarchie à l'intérieur de cette catégorie: une pause en présence d'un point est obligatoire quelle que soit la vitesse de débit mais après une virgule, la pause peut tomber lorsqu'on accélère le débit. En l'absence de marqueurs orthographiques, si le texte est trop long, des pauses apparaîtront aux frontières syntaxiques. Cependant ces pauses ont un statut plus fragile et lorsqu'on accélère le débit, elles sont les premières à tomber.

3. Résultats: Phrases isolées-Locuteur 1

Pour ce test, j'ai utilisé une série de dix phrases isolées, phonétiquement balancées (Lennig¹⁰). Le locuteur (Locuteur 1) devait lire chaque phrase dans l'ordre suivant: vitesse normale, lente, normale, rapide. Au moment de l'enregistrement j'ai noté un problème pour les phrases à vitesse lente: elles semblaient déjà trop rapides. La locutrice a alors relu la série de phrases complète à vitesse lente seulement. Les résultats de l'analyse des deux séries sont présentées dans le tableau 2 (Lent 1 représente le deuxième essai et Lent 2 le premier). Ce tableau présente une analyse comparative des vitesses d'articulation pour les différentes vitesses de lecture. La présence d'une pause dans la phrase est indiquée par une virgule entre les chiffres représentant les vitesses d'articulation des suites sonores de la phrase.

Les différentes vitesses de lecture se distinguent par la présence ou l'absence de pauses. La vitesse lente est caractérisée par la présence d'au moins une pause dans la phrase et ce, pour les

¹⁰ M. Lennig (1981) "Phrases françaises phonétiquement balancées", *Revue d'acoustique*, 56, 31-42.

Phrases isolées: Loc. 1				
Vitesse d'articulation (syll./s)				
Phrase	Lent		Normal	Rapide
	Lent 1	Lent 2		
1	3.4, 3.6	4.9, 4.5	6.0, 5.1	7.5
2	3.4, 4.3	4.3, 6.4, 6.1	5.4	7.0
3	3.9, 3.7	2.9, 4.6	5.2	7.0
4	2.2, 2.9	3.0, 3.4, 4.9	4.7	5.9
5	3.5, 2.8, 3.3	4.0, 3.6, 3.9	4.8, 5.2	6.1, 9.6
6	2.3, 4.0	3.6, 4.9	4.0, 5.7	5.9
7	1.4, 3.4	1.8, 4.8, 4.9	5.3	7.0
8	4.6	4.8	5.6	7.2
9	2.9, 3.1	4.5, 3.4, 3.0	4.4	5.2
10	2.5, 5.1	4.3, 4.1	5.14	6.9

Fig. 2 Phrases isolées. Loc. 1

deux essais. Seule la phrase 8 ne présente pas de pauses. Au niveau perceptuel, cette phrase a, dans les deux cas, été reconnue par les auditeurs comme étant une phrase à vitesse normale. Du point de vue de la production, la présence d'une pause semble être un élément très important pour distinguer la vitesse lente de la vitesse normale. Cependant, au niveau perceptuel, environ la moitié des phrases à vitesse lente (premier essai, Lent 2) ont été jugées comme étant de vitesse normale dans les tests de perception, ce qui indique que la seule présence d'une pause ne suffit pas à indiquer la vitesse d'articulation. Il n'y a que peu de différence au niveau des pauses entre la vitesse normale et la vitesse rapide. Trois des phrases présentent une pause en vitesse normale et une seule phrase en vitesse rapide. L'auditeur ne peut donc, dans ce cas-ci, baser son jugement sur la vitesse de débit par la présence ou l'absence d'une pause.

La vitesse d'articulation semble donc ici jouer un rôle important. On peut d'ailleurs remarquer une plus grande distinction entre les vitesses d'articulation des trois débits (tableau 2). Les vitesses d'articulation moyennes sont 3.1 syll./s pour la vitesse lente (Lent 1), 5.1 syll./s pour la vitesse normale et 7.5 syll./s pour la vitesse rapide (tableau 3). La vitesse d'articulation semble jouer au niveau des phrases isolées un rôle beaucoup plus important que dans les autres types de texte (paragraphes et phrases). En fait, les phrases sont d'une complexité syntaxique assez simple et assez réduites quant au nombre d'éléments. Donc, pour changer la vitesse de débit à l'intérieur de phrases isolées, on ne peut se contenter des pauses. Les modifications de durées segmentales deviennent indispensables.

4. Discussion

Les résultats des analyses nous amènent à conclure que si les pauses jouent un rôle déterminant dans l'indication de la vitesse de débit pour les textes longs, il n'en va pas de même pour les

Comparaison des vitesses d'articulation			
	Lent	Normal	Rapide
Texte long	6.3	7	7.4
Par. 5 Loc. 2	6	6.8	7.5
Par. 5 Loc. 1	5.4	5.8	6.3
Phrases isolées	3.3	5.1	7.5

Fig. 3 Comparaison des vitesses d'articulation.

phrases simples isolées. Dans ce dernier cas, c'est la vitesse d'articulation qui semble l'élément déterminant. Donc, le modèle de modification de débit pour la synthèse de parole devra intégrer les deux composantes.

En ce qui concerne les pauses, le modèle doit prédire le nombre et les points d'occurrence des pauses en fonction des vitesses de débit. Ces pauses sont déterminées par la présence de marqueurs orthographiques ainsi que par des critères d'ordre syntaxique. Il faudra élaborer un corpus permettant de dégager les règles d'apparition des pauses en fonction des frontières syntaxiques et établir une hiérarchie dans l'apparition ou l'abandon de ces pauses. Pour être en mesure d'intégrer ces règles dans le synthétiseur, il sera probablement nécessaire d'élever le degré de sophistication de l'analyseur syntaxique.

Les modifications de vitesse d'articulation posent plus de problèmes. Idéalement, il faut, à partir de données établies pour la vitesse normale, élaborer un ensemble de règles d'allongement et d'abrègement des segments. Cependant, le modèle de durée actuel (O'Shaughnessy¹¹) produit une vitesse de parole souvent jugée trop lente. D'ailleurs, l'examen comparé des vitesses de débit produites par le synthétiseur (Loquax) et celles produites par le locuteur 1 pour les phrases isolées montre que le synthétiseur est très lent (Tableau 4). Les valeurs de vitesse d'articulation en lecture normale sont semblables à celles en vitesse lente chez le locuteur 1. Il faut vérifier à l'aide de tests de perception si ces valeurs sont acceptables comme indicatrices d'une vitesse normale pour une parole de synthèse. Dans l'éventualité d'une réponse négative, il faudra donc refaire un modèle de durée pour vitesse normale à partir duquel nous élaborerons les règles de changement de débit.

Il existe un autre modèle de durée pour la parole de synthèse par règles en français (Barkthova

¹¹ D. O'Shaughnessy (1981), "A Study of French Vowels and Consonants Durations", *Journal of Phonetics*, 9, 385-406.

Débit et synthèse

Phrases isolées: Loc. 1 et Loquax				
Vitesse d'articulation (syll./s)				
Phrase	Lent	Normal		Rapide
	Loc. 1	Loc. 1	Loquax	Loc. 1
1	3.4, 3.6	6.0, 5.1	4.4	7.5
2	3.4, 4.3	5.4	3.9, 3.8	7.0
3	3.9, 3.7	5.2	3.8	7.0
4	2.2, 2.9	4.7	3.4	5.9
5	3.5, 2.8, 3.3	4.8, 5.2	3.5	6.1, 9.6
6	2.3, 4.0	4.0, 5.7	3.9	5.9
7	1.4, 3.4	5.3	4.0	7.0
8	4.6	5.6	4.6	7.2
9	2.9, 3.1	4.4	4.3	5.2
10	2.5, 5.1	5.14	4.4	6.9

Fig. 4 Phrases isolées: Loc.1 et Loquax

et Sorin¹²). Ce modèle comme celui de O'Shaughnessy laisse de côté un aspect linguistique important dans les paramètres de modification de durée: la syllabe. En effet, aucun de ces modèles ne tient compte de la syllabe comme cadre d'application des règles. Pourtant, leur importance est reconnue. Prenons comme exemple le cas de la règle d'allongement des voyelles devant consonnes allongeantes. Cette règle ne s'applique que si la consonne qui suit est dans la même syllabe: *voyage* [vwaʒa:ʒ] mais *voyageur* [vwaʒa - ʒœr]. Aucun des modèles actuels ne peut rendre compte de ce phénomène. De plus, comme on peut le voir dans ce cas-ci, ce ne sont pas tous les éléments de la syllabe qui sont alors modifiés mais seulement ceux de la rime. En effet, quand une voyelle s'allonge sous l'effet d'une telle règle la consonne qui suit doit alors obligatoirement s'abrégier¹³. De plus, d'après l'examen sommaire de mes corpus, il semble que l'impression de débit soit principalement véhiculée par les modifications de durée des syllabes accentuées.

Il faudra donc probablement redéfinir un modèle de durée segmentale pour le français. Un autre problème se présentera alors: celui du corpus à utiliser. En effet, les modèles de variation de durée ont généralement pour corpus de base, du moins en partie, une série de mots isolés permettant d'examiner les variations de durée en fonction d'un ensemble complet de contextes phonétiques. Ce type de corpus est inutilisable dans les variations imposées de débit. En effet, plus un mot est réduit quant au nombre de segments et de syllabes, plus il est difficile de faire varier le débit en le prononçant. Un mot isolé laisse peu de place aux indices de débit. Il faudra donc trouver une méthode plus appropriée aux études sur le débit.

¹² K. Barkthova et C. Sorin (1987) "A Model of Segmental Duration for Speech Synthesis in French", *Speech Communication*, 6:3, 245-261

¹³ D. Archambault et al. (1986), "Problème de production ou problème de perception? Le cas du dévoisement dans la désintégration phonétique", Congrès de l'ACFAS, Université de Montréal.

Recherche subventionnée par les Fonds FCAR et l'Université de Montréal (CAFIR)

SYNTHESE DE LA PAROLE DANS LE DOMAINE TEMPOREL PAR SUPERPOSITION-ADDITION PITCH-SYNCHRONE

Francis CHARPENTIER et Eric MOULINES

Centre National d'Etudes des Télécommunications

22301 LANNION FRANCE

Résumé

Un algorithme temporel utilisant le procédé de synthèse par superposition-addition pitch-synchrone a été proposé récemment dans le cadre de la synthèse par diphtonges [1]. Cet algorithme permet de modifier la prosodie de la parole naturelle en conservant un très bon naturel de timbre. Or sur plan théorique, l'algorithme introduit deux types de distortions: (1) un élargissement des bandes passantes des formants si l'on utilise des fenêtres de synthèse courtes; (2) un effet de réverbération pour des fenêtres plus longues. En pratique, on utilise un compromis en prenant la fenêtre de longueur à deux périodes de pitch. Dans ce cas, la distortion formantique est presque imperceptible. L'algorithme est également applicable au signal résiduel dans le cadre de la synthèse par prédiction linéaire multiimpulsionnelle. Les distortions spectrales dans ce cas sont légèrement différentes, mais la qualité de synthèse reste très bonne.

1. Modifications de la prosodie par l'approche *PSOLA*

Plusieurs méthodes basées sur la synthèse par superposition-addition pitch-synchrone (Pitch-Synchronous Overlap-Add, ou *PSOLA* en abrégé), ont été proposées récemment afin de réaliser des modifications prosodiques de haute qualité de la parole naturelle. La principale motivation pour le développement de ces méthodes est la conception d'un système de synthèse à partir du texte fonctionnant par concaténation d'unités sonores et restituant une bonne qualité de timbre. Dans un tel système, il est indispensable de modifier simultanément le débit de la parole et la fréquence fondamentale de façon à appliquer un rythme et une intonation aussi naturels que possible tout en conservant le naturel du timbre.

Dans ces méthodes, le signal de parole original est transformé en une séquence de signaux à court-terme, appelés signaux élémentaires, obtenu par fenêtrage synchrone de la fréquence fondamentale. Ce flux de signaux élémentaires est alors traité pour produire un flux modifié de signaux élémentaires de synthèse. Le signal de synthèse final est obtenu par une procédure de synthèse par superposition-addition (overlap-add synthesis) [2,3].

Le signal de synthèse peut être interprété comme le résultat de modifications particulières effectuées sur le spectre à court terme du signal original. La nature de ces modifications dépend étroitement de la longueur de la fenêtre d'analyse utilisée pour engendrer les signaux élémentaires. Une fenêtre longue correspond à une condition spectrale d'analyse spectrale à bande étroite et une fenêtre courte correspond à une condition d'analyse à bande large.

Dans des conditions à bande étroite, le spectre à court terme possède une structure complexe, qui consiste d'une enveloppe spectrale superposée à une structure harmonique correspondant au pitch d'analyse. Une restitution correcte sur signal de synthèse exige de modifier la structure harmonique en la rendant compatible avec le pitch de synthèse. Cette opération nécessite d'adopter l'approche fréquentielle *FD-PSOLA* (Frequency-Domain *PSOLA*). Les signaux élémentaires de synthèse sont obtenus par modification dans le domaine fréquentiel des signaux élémentaires d'analyse [4,5]. Les spectres sont modifiés de façon à conserver l'enveloppe spectrale tout en rendant l'espacement entre les harmoniques de pitch compatible avec le pitch de synthèse.

Si l'on élimine les traitements fréquentiels de l'approche *FD-PSOLA*, on obtient la méthode temporelle *TD-PSOLA* (Time-domain *PSOLA*), présentée dans [1,6] et développée dans cet article. Cette méthode est également appelée *PSOLA/KDG* d'après des initiales en breton. Cet algorithme est très simple et ne fait intervenir que deux types d'opérations: (1) recopie, élimination ou répétition des signaux élémentaires, (2) ajustement des délais entre les signaux aussi la charge de calcul résultante est très faible.

Dans des conditions d'analyse à bande étroite, la méthode *TD-PSOLA* conduit à des distorsions de type réverbération en raison de l'incompatibilité entre la périodicité inhérente aux signaux élémentaires, qui constitue une trace du pitch original, et le pitch imposé au signal de synthèse.

En revanche, si l'on cherche à appliquer l'approche *PSOLA* à des conditions d'analyse à bande large, le spectre à court terme possède une structure simple et constitue une approximation acceptable de l'enveloppe spectrale. Il est donc inutile de recourir à une approche fréquentielle, comme pour le cas à bande étroite, et l'approche temporelle *TD-PSOLA* se révèle tout-à-fait suffisante. L'usage d'une fenêtre courte introduit toutefois un autre type de distorsions spectrales, qui peut s'interpréter comme un élargissement de la bande passante des formants, dépendant de la largeur de bande de la fenêtre d'analyse. Les conditions de synthèse optimales résultent d'un compromis entre les deux types de distorsions correspondant aux conditions à bande étroite et à bande large. En fait, pour une plage intermédiaire de longueur de fenêtre, entre deux à trois fois la période de pitch, les distorsions sont peu audibles et la qualité globale de la synthèse est considérée comme très bonne.

Alors que le traitement de type *TD-PSOLA* s'applique naturellement au signal de parole, on peut également l'appliquer au signal résiduel lorsqu'on utilise une technique de codage par prédiction linéaire (approche *LP-PSOLA*, ou Linear Predictive *PSOLA*). Pour des raisons pratiques nous avons développé cette approche en combinant un codage *LPC* à excitation multiimpulsionnelle (*MPLPC*) et un traitement *TD-PSOLA* de l'excitation [7,8], mais on pourrait également choisir un codage de type *CELP*. En fait, ce type de méthode revient à permuter le traitement *TD-PSOLA* pour les modifications prosodiques et le filtrage *LPC* à court-terme. Or, la transformation *TD-PSOLA* opérée sur le signal est équivalente à une combinaison linéaire de fenêtrages et de translations appliqués au signal original. En toute rigueur, elle ne peut pas la commuter avec une opération de type filtrage linéaire, de sorte que les distorsions introduites par les méthodes *TD-PSOLA* et *LP-PSOLA* ne sont pas équivalentes. Toutefois, sur le plan pratique, les deux méthodes sont capables de produire de la parole synthétique avec une bonne qualité de timbre et sont donc exploitables dans les applications de la synthèse à partir du texte.

2. Description de l'algorithme *TD-PSOLA*

Le signal de parole numérisé $x(n)$ est décomposé en une suite temporelle de signaux à court-terme $x_m(n)$ tels que des signaux successifs présentent un recouvrement mutuel important. Ces signaux élémentaires sont obtenus en multipliant le signal par une suite de fenêtres d'analyse $h_m(n)$:

$$x_m(n) = h_m(t_m - n) x(n)$$

Les instants consécutifs t_m , appelés pitch-marques d'analyse, sont déterminés selon une cadence synchrone de la fréquence fondamentale sur les portions voisées du signal. Le flux des signaux élémentaires de synthèse sert à produire un flux de signaux élémentaires modifiés $\tilde{x}_q(n)$ qui sont synchronisés sur une nouvelle suite d'instants f_q , appelées pitch-marques de synthèse.

L'algorithme détermine, en fonction des facteurs de modification du pitch et de la durée, les pitch-marques de synthèse f_q et la fonction $f_q \rightarrow t_m$ les appliquant sur les pitch-marques d'analyse, dont le rôle est de préciser pour chaque signal élémentaire de synthèse $\tilde{x}_q(n)$, le signal élémentaire $x_m(n)$ qui doit être utilisé pour l'engendrer. La suite des pitch-marques f_q définit les délais qui doivent être appliqués entre les signaux élémentaires de synthèse. Le traitement consiste donc à sélectionner certains signaux élémentaires d'analyse $x_m(n)$ et à les traduire selon la suite de délais $\delta_q = f_q - t_m$:

$$\tilde{x}_q(n) = x_m(n - \delta_q) = x_m(n + t_m - f_q)$$

On obtient ensuite le signal de synthèse $\tilde{x}(n)$ en superposant et additionnant le flux des signaux élémentaires de synthèse, en utilisant par exemple la formule de synthèse par superposition-addition (overlap_add) simple [3]:

$$\tilde{x}(n) = \frac{\sum_q \alpha_q \tilde{x}_q(n)}{\sum_q \tilde{h}_q(f_q - n)}$$

où $\tilde{h}_q$ désigne la suite des fenêtres d'analyse, réindexées en fonction de l'indice q des signaux de synthèse. Aussi, les appellent-on indifféremment fenêtre d'analyse ou de synthèse. Le facteur α_q et le dénominateur de la formule jouent le rôle de facteurs de normalisation énergétique.

3. Interprétation du traitement *TD-PSOLA*

Si l'on néglige les termes de normalisation, cette formule de synthèse se réécrit de façon simplifiée:

$$\tilde{x}(n) = \sum_q \tilde{x}_q(n) = \sum_q \tilde{h}_q(f_q - n) x(n - \delta_q)$$

Le signal de synthèse apparaît comme la combinaison linéaire de versions fenêtrées et translatées du signal original. Aussi, lorsqu'on combine la transformation de type *TD-PSOLA* et une opération de filtrage linéaire de type filtrage autorégressif (prédiction linéaire) ou filtrage passe-bas, une permutation de l'ordre dans lequel on effectue ces opérations modifie le comportement du système. En effet, à l'inverse des opérations de sommation et de translation de signaux, les opérations de fenêtrage ne sont pas commutables avec un filtrage linéaire. Cette propriété est facile à comprendre dans le cas d'un filtrage passe-bas idéal. Si le filtrage est effectué après le traitement *TD-PSOLA*, la bande d'affaiblissement sera de largeur nulle, et le filtrage aura une efficacité maximale. Dans le cas contraire, la bande d'affaiblissement est égale à la moitié de la bande passante de la fenêtre de synthèse, et le filtrage n'est plus idéal.

En fait, la permutation des opérations *TD-PSOLA* et filtrage linéaire donne lieu à deux applications intéressantes. La première est la méthode de synthèse *LP-PSOLA* décrite au paragraphe suivant. La seconde permet d'effectuer la synthèse *TD-PSOLA* en augmentant la résolution de la fréquence fondamentale. Ceci peut se réaliser en effectuant le traitement *TD-PSOLA* à une cadence supérieure (par exemple 16 kHz) à la fréquence d'échantillonnage de synthèse (par exemple 8 kHz), puis en effectuant un filtrage passe-bas de sous-échantillonnage. Cependant, il est avantageux dans un cas comme la synthèse par diphones, où le signal à traiter est préstocké en mémoire, d'effectuer le filtrage passe-bas avant la synthèse, car cela permet soit de l'effectuer hors ligne sur le dictionnaire

de diphtonges pour alléger la charge de calcul, soit de l'effectuer au fil de l'eau en suréchantillonnant le signal au préalable avant le traitement *TD-PSOLA*, ce qui permet de réduire la taille du dictionnaire de diphtonges.

Supposons que le signal est périodique de période P_0 . Il est normal dans ce cas de choisir une fenêtre de synthèse fixe $h_m(n) = h_0(n)$. Les signaux élémentaires sont tous égaux à un signal générique:

$$x_m(n) = x_0(n) = h_0(t_0 - n)s(n)$$

Le signal synthétique est obtenu par périodisation du signal générique $x_0(n)$ à une période de synthèse différente de P_0 . Une telle opération équivaut à échantillonner le spectre $X_0(\omega)$ de $x_0(n)$ aux fréquences des harmoniques de synthèse, différentes des harmoniques originales. En d'autres termes, l'enveloppe spectrale du signal de synthèse se trouve être identiquement le spectre $X_0(\omega)$ du signal original, dont la résolution spectrale est donnée par la fenêtre de synthèse $h_0(n)$. Il en résulte une distorsion de l'amplitude de chaque harmonique par rapport à l'enveloppe spectrale idéale, qui s'interprète différemment dans les cas d'analyse à bande large et étroite.

Dans le cas à bande étroite, comme le spectre présente des lobes étroits au voisinage des harmoniques d'analyse et des atténuations importantes entre les harmoniques, l'amplitude des harmoniques de synthèse est affectée de façon importante, voire complètement annulée. En fait, ce type de distorsion équivaut au filtrage du signal de synthèse idéal par un filtre en peigne, dont l'effet est d'introduire un écho parasite, de délai égal à la période du signal original.

En revanche, dans le cas à bande large, le spectre d'amplitude $|X_0(\omega)|$ constitue une estimation lissée de l'enveloppe spectrale véritable. La forme des résonances formantiques est altérée cependant, car la largeur de bande de la fenêtre est souvent supérieure à la largeur de bande véritable des formants. Dans des cas limites, on peut observer la fusion de deux formants proches en un seul maximum spectral. Ce type de distorsion, cependant, est moins gênant que la distorsion de type réverbération car les seuils différentiels pour perception des largeurs de bande formantiques est d'environ 40% pour les voyelles stables. Ainsi, en pratique, on choisit une condition de bande plutôt large, en prenant des fenêtres de synthèse égales à deux fois la période de pitch d'analyse.

Finalement, dans le cas de synthèse à bande large, l'approche *LP-PSOLA* possède un avantage théorique sur l'approche *TD-PSOLA*, dû au fait que les techniques de prédiction linéaire permettent une estimation plus précise de l'enveloppe spectrale que la transformée de Fourier à court terme. Dans l'approche *LP-PSOLA*, la séparation explicite entre la source et l'enveloppe spectrale fournit un degré de liberté supplémentaire en permettant une analyse à bande étroite en ce qui concerne l'enveloppe et une synthèse à bande large en ce qui concerne les modifications prosodiques. Comme

la synthèse à bande large s'applique au signal résiduel, dont le spectre est quasi-plat, ou tout du moins présente des oscillations résiduelles faibles, la distorsion d'élargissement de bande passante demeure sans effet.

Conclusion

Nous avons présenté dans ce papier l'algorithme *TD-PSOLA* permettant la réalisation de modifications prosodiques de la parole naturelle de haute qualité, ainsi qu'une combinaison de l'algorithme avec les techniques codage de la forme d'onde par prédiction linéaire (*RELPC*, *MPLPC*, *CELP*). Ces deux approches ont été intégrées dans le système de synthèse à partir du texte du CNET, et ont permis d'élever la qualité de la parole de synthèse de façon très sensible. Notre système de synthèse fonctionne actuellement par concaténation de diphones, ce qui semble imposer un autre type de limitation au naturel de la parole synthétique. Des recherches sont en cours afin de combiner la bonne qualité de restitution du timbre des méthodes de type *PSOLA* et l'utilisation d'unités sonores plus longues, de type polyphones ou mots.

Références

- [1] C. Hamon, E. Moulines, F. Charpentier, "A diphone synthesis system based on time-domain modifications of speech", *Proc. Int. Conf. ASSP, Glasgow, 1989*
- [2] D.W. Griffin, J.S. Lim, "Signal estimation from modified short-time Fourier transform", *IEEE Trans. ASSP*, 32(2), 236-243, 1984
- [3] J.B. Allen, L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", *Proc. IEEE*, 65(11), 1558-1564, 1977
- [4] F. Charpentier, E. Moulines, "Text-to-speech algorithms based on FFT synthesis", *Proc. Int. Conf. ASSP, New York, 667-670, 1988*
- [5] F. Charpentier, "Traitement de la parole par analyse-synthèse de Fourier: application à la synthèse par diphones", *Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, 1988.*
- [6] C. Hamon, "Procédé et dispositif de synthèse de la parole par addition-recouvrement de formes d'ondes", *dépôt de brevet No. 8811517, 1988.*
- [7] E. Moulines, F. Charpentier, "Diphone synthesis using multipulse linear prediction", *Proc. FASE Int. Conf., Edinburgh, 1988*
- [8] E. Moulines, *Thèse de doctorat, Ecole Nationale Supérieure des Télécommunications, en préparation.*

Un outil générateur d'applications de technologies vocales

Philippe Mabillean

Département de Génie Électrique
Faculté des Sciences Appliquées
Université de Sherbrooke
Sherbrooke, Québec J1K 2R1
Tél: (819) 821-7142

Centre Canadien de Recherche sur
l'Informatisation du Travail
1575 Bvd. Chomedey
Laval, Québec H7V 2X2
Tél: (514) 682-3400

E-mail: p_mabillean@ccrit.doc.cdn

Résumé

On présente ici des travaux sur la génération assistée par ordinateur d'applications dans le domaine des technologies vocales. En effet un des points cruciaux dans le développement de ces applications qui sont, rappelons le, très orientées vers l'utilisateur, est la facilité et par conséquent la vitesse avec laquelle elles vont pouvoir voir le jour ou être modifiées à partir de besoins spécifiques. Cet élément clef dans la stratégie des développeurs de services utilisant une interaction vocale a fait surgir le besoin d'outils facilitant, ou mieux automatisant, la production des logiciels impliqués dans ces applications. Les choix et les solutions proposés ici suivent une ligne pragmatique qui a pour objectif la mise à disposition, dans des délais raisonnables, de moyens et d'outils utilisant les concepts les plus avancés dans le domaine de l'ingénierie des logiciels.

Introduction

Le développement d'applications dans le domaine des technologies vocales nécessite en générale une flexibilité importante et doit être mené à bien dans un laps de temps relativement court. En effet l'adaptabilité aux besoins des utilisateurs de l'application et la possibilité de présenter un prototype très rapidement en réponse à une demande donnée sont des atouts majeurs pour conquérir le marché de l'interaction vocale.

Pour faire face à cette situation il est indispensable de disposer d'outils permettant la mise au point rapide et la modification simple des applications impliquant les technologies vocales. L'approche qui est proposée ici permet la construction,

l'évaluation et la modification interactive d'applications appartenant à une classe déterminée et utilisant une interaction vocale.

Structure et particularités des applications visées

Les applications dont la construction est visée avec le générateur proposé appartiennent à une classe bien déterminée. Malgré ses particularismes cette classe a été choisie et construite de façon à englober la plupart des applications rencontrées avec éventuellement un minimum d'adaptation.

Tout d'abord l'application doit être centrée sur un dialogue usager/ système impliquant une interaction vocale: utilisation du clavier téléphonique (DTMF) ou de reconnaissance de mots isolés de l'utilisateur vers le système, et lecture concaténée de parole pré-enregistrée ou synthèse à partir du texte du système vers l'utilisateur. En fait ce type d'interaction n'est pas le seul utilisable avec le générateur proposé mais les mises en œuvre disponibles, en particulier au niveau de la simulation des applications, sont orientées vers ces moyens de communication homme/ machine.

Ensuite la structure de l'application doit respecter certaines règles qui sont bien sûr compatibles avec la construction d'un dialogue interactif. La structure proposée permet la réalisation d'une machine à états qui va réagir aux événements extérieurs par des transitions d'un état vers un autre. Le niveau de raffinement qui est visible pour le concepteur est plus global que celui qui correspondrait réellement aux différents états de la machine. En effet de nombreuses chaînes d'états impliquant des transitions quasi automatiques ou certains regroupements d'états fermés ne vont apparaître que comme un seul état. Nous appellerons ces regroupements ou super-états "transactions".

Une application sera donc constituée d'un certain nombre de transactions qui apparaissent alors comme étant des briques élémentaires à partir desquelles on va bâtir les applications. Ces transactions seront choisies parmi un ensemble de transactions types constituant des gabarits de départ, et seront ensuite modélées pour s'insérer dans l'application finale. Pour permettre une construction de l'application directement en rapport avec son aspect fonctionnel, les transactions vont correspondre à des éléments charnières dans le déroulement du dialogue usager/ système.

Pour faciliter le raccordement et l'enchaînement des transactions, des règles précises définissent les liens entre une transaction et les autres:

- chaque transaction ne dispose que d'un seul point d'entrée auquel un nom symbolique peut être associé permettant de référer à cette transaction et
- la sortie d'une transaction avec aiguillage possible vers la ou les transaction suivantes se fait à un seul endroit dans la transaction; cet endroit est de fait la dernière étape d'une transaction.

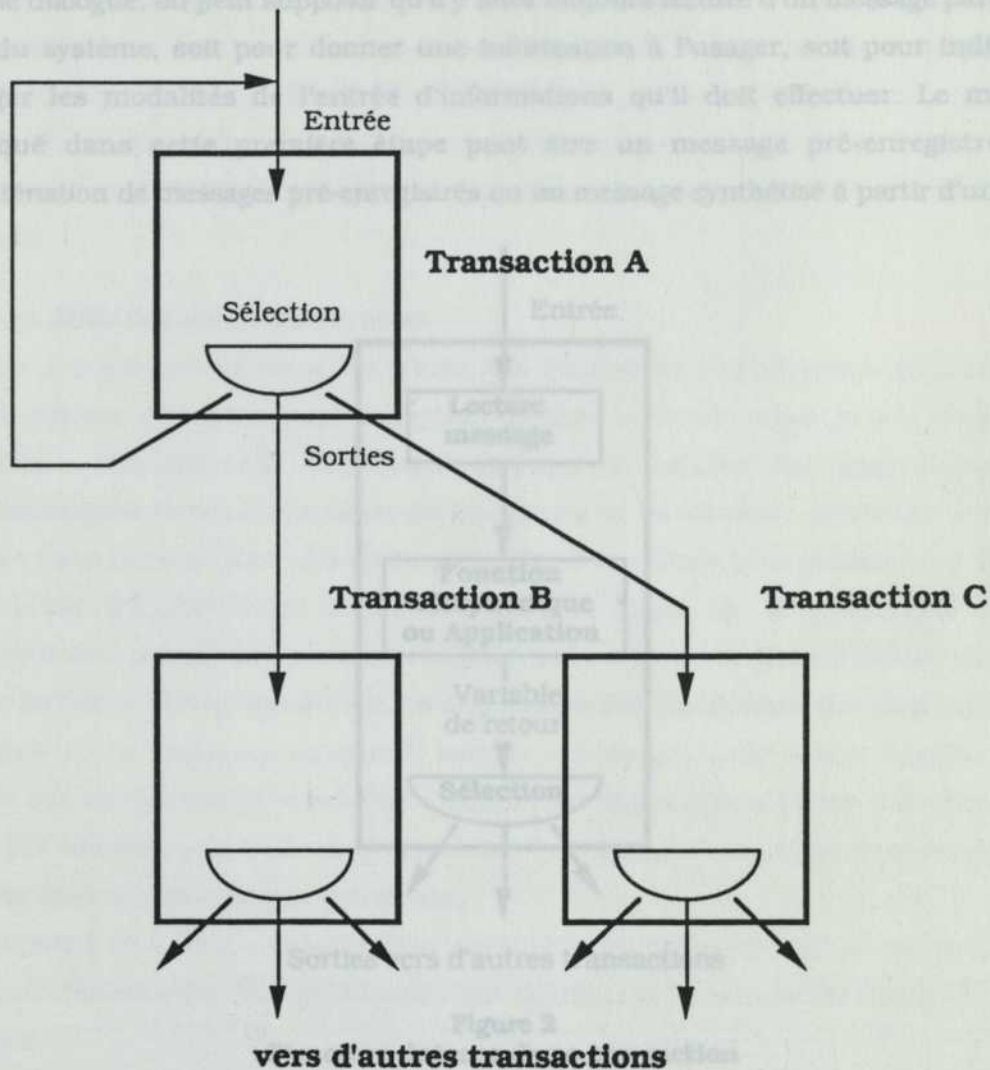


Figure 1
Exemple d'application montrant l'interconnexion des transactions

En ce qui concerne la structure interne de chaque transaction, on peut distinguer trois étapes principales:

- une première étape spécifiant quel est le message parlé qui va être joué à l'usager à l'entrée dans la transaction. Comme toute transaction correspond à un échange d'informations entre l'usager et le système pour permettre le passage à un état suivant dans le dialogue, on peut supposer qu'il y aura toujours lecture d'un message parlé de la part du système, soit pour donner une information à l'usager, soit pour indiquer à l'usager les modalités de l'entrée d'informations qu'il doit effectuer. Le message impliqué dans cette première étape peut être un message pré-enregistré, une concaténation de messages pré-enregistrés ou un message synthétisé à partir d'un texte.

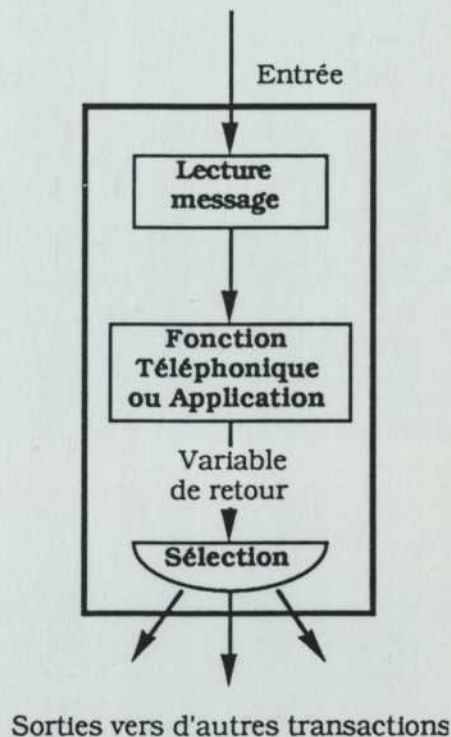


Figure 2
Structure interne d'une transaction

- une seconde étape constituant le cœur de la transaction et comportant une fonction d'interaction avec l'usager: entrée de DTMF ou reconnaissance de parole, ou bien une fonction d'interaction avec la source externe d'informations encapsulée ici sous le nom

de "fonction application". Ce concept de fonction application permet l'introduction dans l'application de tout le code qui ne concerne pas directement l'interaction vocale ni le dialogue qui est construit au-dessus. On va trouver en particulier ici tout ce qui concerne l'interrogation et la modification d'une base de données externe qui contient les informations réclamées ou mises à jour par les usagers.

- enfin une troisième et dernière étape dans la transaction va comporter la prise de décision associée à cette transaction: c'est le choix de la transaction suivante qui en fait détermine comment va évoluer le dialogue en cours. Cette prise de décision va pouvoir être basée sur le résultat de la fonction d'entrée d'information, parole ou DTMF, ou bien sur le résultat de la fonction application, par exemple validation d'un code ou d'un mot de passe.

Rôles et définition des outils proposés

Comme il a été précisé les outils d'aide à la génération d'applications en technologie vocale doivent permettre une conception facilitée et rapide ainsi qu'une mise à jour simplifiée. Ces objectifs impliquent un certain nombre de critères quant aux caractéristiques de ces outils. Nous allons dégager ici les éléments directeurs qui ont été utilisés dans la conception des outils proposés. D'une façon plus générale ces outils de conception d'applications se placent dans le cadre de la génération assistée d'applications, domaine où plusieurs méthodes et moyens ont déjà été définis (1) (2).

Au niveau de la conception initiale les exigences suivantes doivent être rencontrées:

- utilisation de dialogues interactifs avec le concepteur pour définir l'application et toutes ses composantes, c'est-à-dire définir les transactions et les interconnexions entre ces transactions. Cette programmation "par dialogue" exclut toute écriture directe de code dans quelque langage que ce soit.
- utilisation d'outils de visualisation graphique en particulier pour représenter les interconnexions entre les transactions qui définissent l'ossature du dialogue usager/système.
- simulation, dans des conditions aussi proches que possible de la situation réelle, du dialogue usager/ système correspondant à l'application développée. En particulier les moyens d'interaction entre l'usager et le système doivent être ceux utilisés lors du

fonctionnement normal de l'application. Par contre cette simulation ne doit pas nécessiter la construction en tant que telle de l'application.

- adaptabilité du générateur à différents langages de programmation au niveau du code construit ainsi qu'à différents environnements de programmation en particulier au niveau des bibliothèques disponibles pour contrôler l'interaction usager/ système. En fait, l'environnement présenté au concepteur pour l'application doit être calqué sur les ressources disponibles dans la bibliothèque utilisée mais doit être indépendant du langage et de l'environnement de programmation choisi.

- génération d'informations relatives à l'application construite et susceptibles d'être utilisées pour la réalisation de la documentation associée à cette application. En particulier des listes des transactions ou des messages utilisés seront indispensables.

En ce qui concerne la modification et la mise à jour des applications créées il faut également répondre à l'exigence suivante:

- possibilité de modifier les paramètres des transactions ainsi que les relations entre les transactions en tout temps et de façon réversible. Ceci signifie que la structure originale de la transaction doit être conservée et c'est sur cette structure que travaille l'outil de génération et non sur le code généré.

Mises en œuvre des outils de génération d'applications

A partir des spécifications proposées une mise en œuvre d'un générateur d'applications en technologie vocale est possible. En fait plusieurs phases ont été utilisées pour permettre une réalisation progressive d'un générateur d'applications en technologie vocale. Il a été ainsi possible de tirer profit des expériences de chacune des étapes pour avancer le projet.

On peut distinguer les trois réalisations suivantes:

- Mise au point d'un premier prototype de générateur qui utilise la technologie "Hypercard" sur Macintosh et qui est essentiellement orienté vers les applications de réponse vocale. Dans cette première phase le concept de transaction a été précisé. Les applications pouvaient être simulées dans l'environnement même de développement mais aucun code n'était effectivement produit. Cette étape a permis de mettre au point la structure générale de l'interface concepteur- générateur. La structure des applications générées ainsi que la nature et le contenu des transactions ont été définis, et ceci malgré

la gamme restreinte des applications visées: lecture de parole pré-enregistrée et entrée de DTMF.

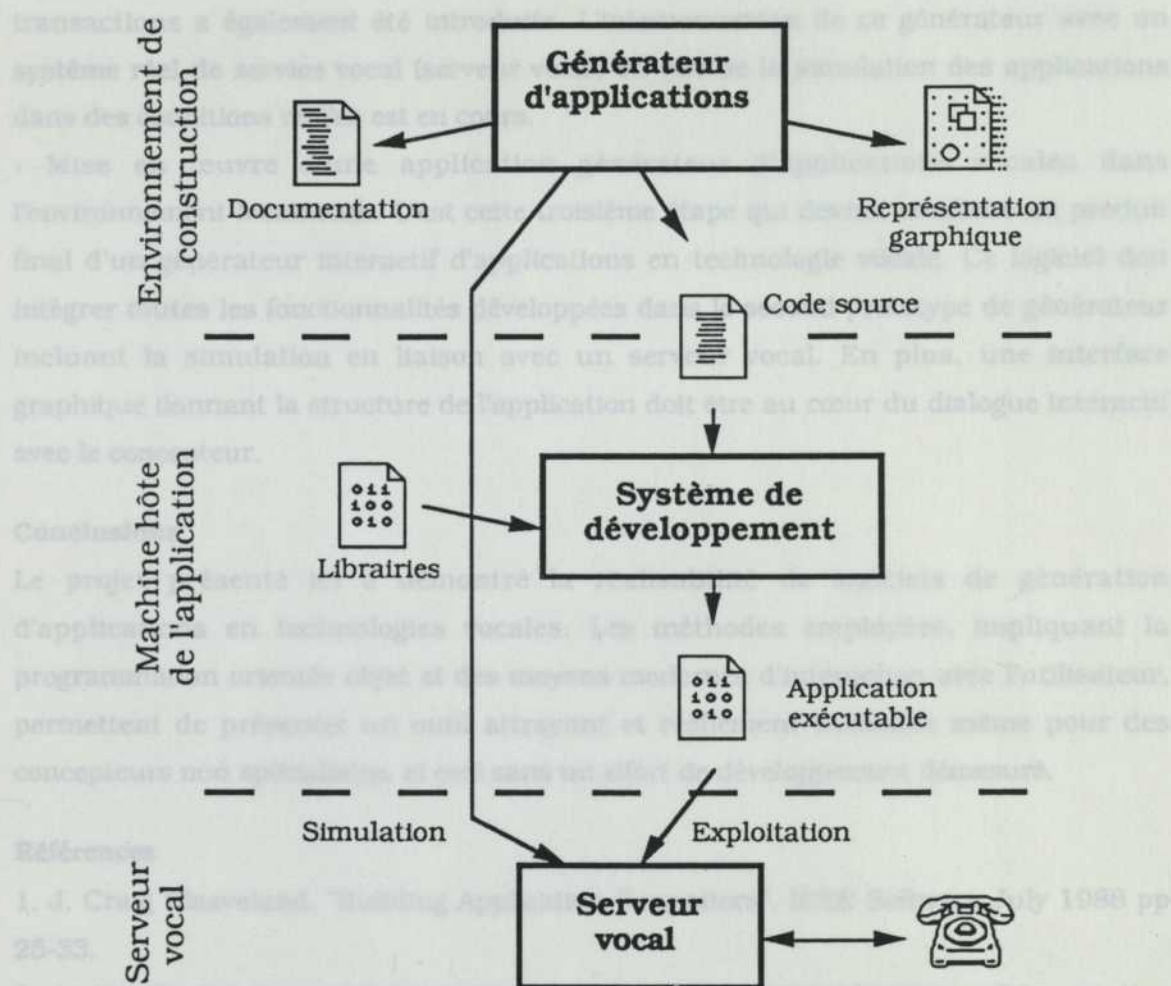


Figure 3
Présentation synoptique de la construction, du développement et de l'exploitation d'une application vocale

- Mise au point d'un second prototype utilisant la même technologie (Hypercard) que le premier mais plus orienté vers la structure interne et la conception du générateur. C'est à cette étape qu'ont été proposés les mécanismes qui permettent la génération effective du code par utilisation de gabarits relativement souples et modifiables par le générateur. Les concepts reliés aux listes d'objets manipulables par le générateur comme les transactions, les messages à jouer ou les fonctions téléphoniques et

applications ont été précisés. La production de données utilisables pour la composition de documentations ainsi que pour la représentation graphique de l'interconnexion des transactions a également été introduite. L'interconnexion de ce générateur avec un système réel de service vocal (serveur vocal) en vue de la simulation des applications dans des conditions réelles est en cours.

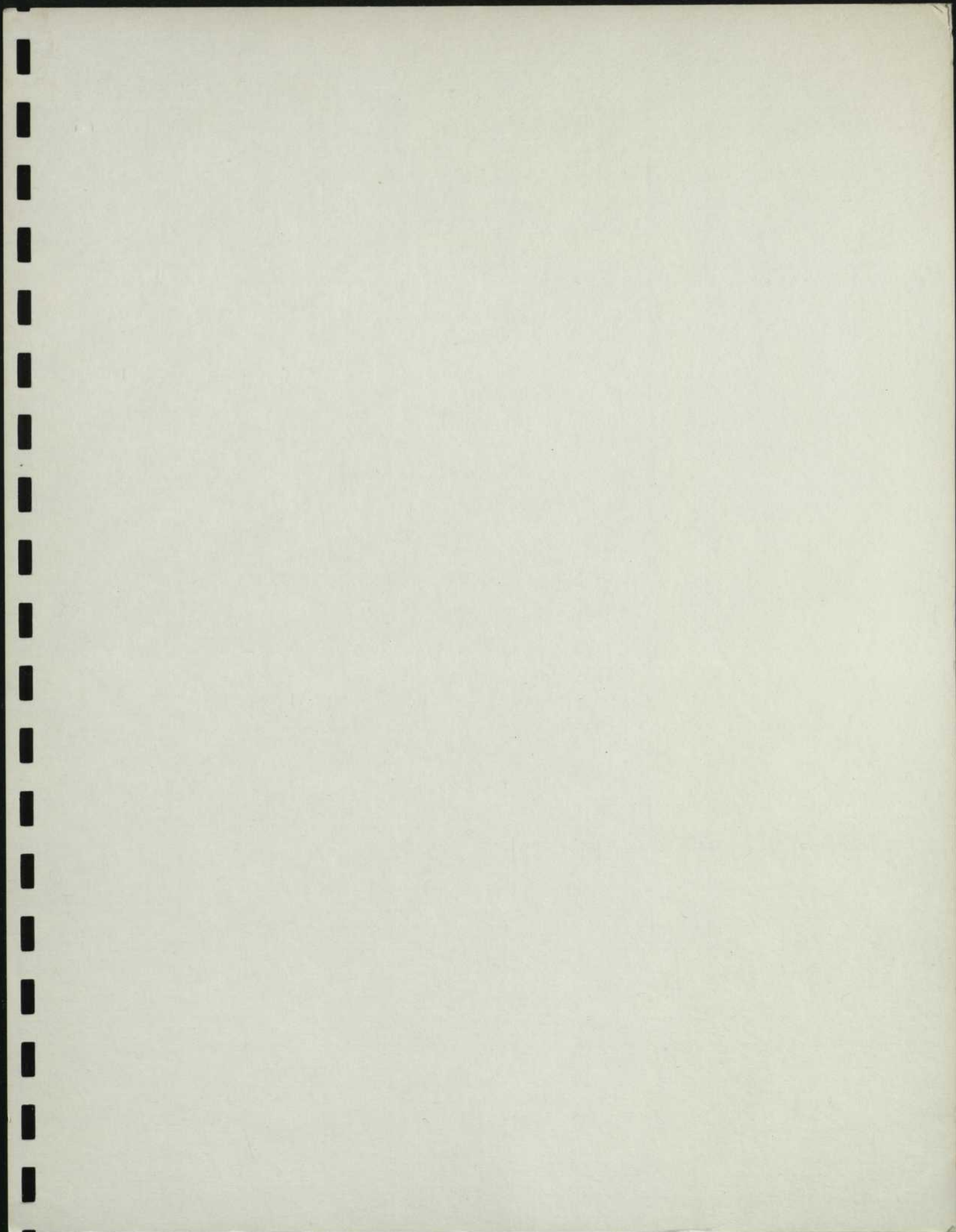
- Mise en œuvre d'une application générateur d'applications vocales dans l'environnement Macintosh. C'est cette troisième étape qui devrait conduire au produit final d'un générateur interactif d'applications en technologie vocale. Ce logiciel doit intégrer toutes les fonctionnalités développées dans le second prototype de générateur incluant la simulation en liaison avec un serveur vocal. En plus, une interface graphique donnant la structure de l'application doit être au cœur du dialogue interactif avec le concepteur.

Conclusions

Le projet présenté ici a démontré la réalisabilité de logiciels de génération d'applications en technologies vocales. Les méthodes employées, impliquant la programmation orientée objet et des moyens modernes d'interaction avec l'utilisateur, permettent de présenter un outil attrayant et réellement utilisable même pour des concepteurs non spécialistes, et ceci sans un effort de développement démesuré.

Références

1. J. Craig Cleaveland, "Building Application Generators", IEEE Software July 1988 pp 25-33.
2. Anders Rockström and Roberto Saracco, "SDL - CCITT Specification and Description Language", IEEE Transaction on Communications, Vol. COM-30, No.6, June 1982, pp 1310-1317.





Association 2730 Côte Ste-Catherine
canadienne-française Montréal, Québec
pour l'avancement Canada, H3T 1B7
des sciences

Pour commandes téléphoniques: (514) 342-1411