Reconnaissance automatique de phonèmes guidée par les syllabes

Olivier Le Blouch, Patrice Collen

Laboratoire TECH/IRIS
France Télécom R&D – 4 rue du Clos Courtel – 35510 Cesson Sévigné, France
Tél.: ++33 (0)2 99 12 48 51 – Fax: ++33 (0)2 99 12 40 98
Mail: olivier.leblouch@francetelecom.com & patrice.collen@francetelecom.com

ABSTRACT

This paper presents a phonetic transcription system of French speech. This recognizer is based on a phonetic transcription driven by syllables, a syllabic bigram language modelling and a HMM topology adapted to syllables. The phone error rate obtained is compared to basic, usual systems at phonetic level : once the resulting syllables have been converted to phones, the phone error rate on a 12 minute-part of BREF80 corpus is as low as 15.8% with 35 phones.

1. INTRODUCTION

La syllabe joue un rôle important dans la perception de la parole et son utilisation comme entité acoustique constitue un avantage certain vis-à-vis d'une approche par mot ou par phonème [5]. La syllabe est en effet une entité plus longue que le phonème, plus facile à indexer manuellement, et offrant la possibilité de générer un nombre de mots illimité avec un nombre de syllabes réduit, de l'ordre de 5000 [13]. En outre, des études en psycholinguistique et en phonologie suggèrent que l'information syllabique, avec une durée moyenne de 250ms, est cruciale au niveau de la perception et de la compréhension [7].

Cet article propose d'utiliser les syllabes de façon à améliorer la transcription phonétique. Notons que, contrairement au modèle syllabique décrit dans [5], le système décrit dans ce papier est conçu comme un système de transcription phonétique dont l'enchaînement est guidé par l'information syllabique.

De façon à se positionner vis-à-vis des systèmes existants, les systèmes implémentés ont été testés sur le corpus BREF80 [8], servant de base de test notamment aux systèmes décrits dans [9] et [3]. Le premier propose un système combinant des modèles de phonèmes en contexte, un modèle de langage bigram phonétique et une adaptation au genre du locuteur; ce système offre un taux d'erreur de 21.3%. Le second, quant à lui, propose une reconnaissance associant chaînes de Markov et réseaux neuronaux, et obtient un taux d'erreur de 25%.

Après une description des ressources utilisées, et un bref état de l'art sur les techniques de syllabification, nous détaillerons l'architecture générale des systèmes phonétiques et syllabiques développés. Avant de conclure, les performances sur BREF80 ainsi que la complexité des différents systèmes seront présentées.

2. RESSOURCES AUDIOS ET TEXTUELLES

2.1. Corpus Audio

Toutes les données audio sont échantillonnées à 16kHz.

Le corpus d'apprentissage, d'une durée de 4h30, est composé à 60% de données issues du corpus BREF80 et de 40% de données audio issues de corpus France Télécom. Il est au total constitué de 274 phrases prononcées par 40 locuteurs et de 310 phrases prononcées par 53 locutrices.

Le corpus de test de 12 minutes, issu du corpus BREF80, est quant à lui composé de 24 phrases prononcées par 8 locuteurs différents (4 hommes et 4 femmes), non présents dans le corpus d'apprentissage.

Les transcriptions phonétiques associées à toutes ces données ont été réalisées automatiquement sur la base des 35 phonèmes proposée dans [6], puis vérifiées manuellement afin de refléter au mieux le contexte acoustique.

2.2. Corpus textuel

Pour la création des modèles de langage bigram phonétique et syllabique, c'est-à-dire le calcul des probabilités de transition d'une entité phonétique ou syllabique vers une autre, les données textuelles présentes dans notre corpus audio d'apprentissage ont été enrichies par divers contenus en langue française issus du Web. Au total, ce corpus représente environ 300K mots, soit environ 600K syllabes et 1300K phonèmes.

3. DU TEXTE AUX SYLLABES

La décomposition du texte en entités syllabiques requiert un outil de syllabification. Pour le français, nous nous sommes inspirés des principes utilisés pour la syllabification des bases lexicales *Brulex* et *Lexique* [11].

Dans la suite de cet article, le formalisme de représentation des syllabes est un agglomérat de

phonèmes séparés par des "_", un mot de plusieurs syllabes étant quant à lui une suite de syllabes séparées par des blancs. Ex : *Syllabe* ➔ [S_I L_A_B].

Le découpage du français en syllabes obéit à certaines conventions [11]. Tout d'abord, chaque son vocalique, c'est-à-dire les voyelles et les semi-consonnes suivies d'une voyelle, constitue le noyau d'une syllabe alors que deux voyelles consécutives (Ex : *Agréable* ➔ [A G_R_**EI** **A**_B_L]) appartiennent à deux syllabes différentes : elles sont dites en hiatus. Ensuite, lorsque entre deux voyelles, une seule consonne est prononcée, elle est considérée comme formant une syllabe avec la voyelle qui la suit, et ce indépendamment du découpage en mot ; la phrase *Quelle heure est-il ?* est ainsi syllabée [**K**_AI **L**_OE **R**_AI **T**_I_L]. Enfin, dans le cas de plusieurs consonnes prononcées entre deux voyelles, il existe des règles inhérentes à chaque cas, comme par exemple les agrégats occlusives-liquides (Ex : R*empli* ➔ [R_AN **P**_**L**_I]) ou les suites d'occlusives (Ex : *Opter* ➔ [O_**P** **T**_EI]).

Néanmoins, le découpage du français en syllabes se heurte à différentes théories explicitées dans [11], comme par exemple la syllabification du mot *capsule* : [K_A_**P** **S**_U_L] ou [K_A **P**_**S**_U_L] ? Dans ce type de cas, et lorsque c'est possible, l'analyse acoustique sert de support pour le choix de segmentation : sachant que les occlusives sont précédées d'une courte période de silence, le choix s'est porté sur une segmentation avant l'occlusive et donc une syllabation en [K_A **P**_**S**_U_L].

L'algorithme implémenté suit donc ces conventions en convertissant dans un premier temps le texte en phonèmes, en prenant soin de substituer aux ponctuations des silences, puis en faisant appel à une soixantaine d'heuristiques de découpage des chaînes de phonèmes. Ceci aboutit à une segmentation du corpus textuel en 4352 syllabes différentes.

4. DESCRIPTION DES SYSTEMES

Après une description technique des paramètres utilisés et des algorithmes communs à toutes les expériences, les 3 systèmes sont présentés.

4.1 Paramétrage et description technique

Paramètres acoustiques

Le signal audio est converti en un jeu de 39 coefficients extraits toutes les 10ms sur des segments temporels de 32ms. Les vecteurs sont constitués de 12 coefficients MFCC, de la log-énergie, et des dérivées premières et secondes. Une normalisation par la moyenne des cepstres est finalement appliquée [1].

Modélisation

Nos unités sont modélisées par des chaînes de Markov cachées [12].

Modèles de langage

Un modèle bigram est appliqué à tous les systèmes : phonétique pour le premier système et syllabique pour les deux autres. Tous deux sont appris sur le corpus textuel à partir des modules HTK [16] dédiés.

Apprentissage des HMM

L'apprentissage est réalisé sous HTK selon l'algorithme de Baum-Welch [2] sur le corpus dédié, en multipliant par deux le nombre de gaussiennes par mixture toutes les 10 itérations. Au final, chaque état contient un mélange de 32 gaussiennes.

Décodage

Pour le décodage, on utilise une formulation alternative de l'algorithme de Viterbi, le *Token Passing Model* [14].

4.2. Système phonétique de base

Ce système met en œuvre des modèles de Markov cachés à trois états avec une topologie gauche-droite.

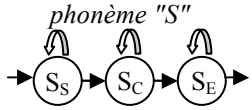

Figure 1 : Topologie des modèles de phonème

La topologie de ces modèles phonétiques utilise un formalisme simple pour décrire chaque état : pour chaque phonème X à 3 états, la notation X_S est donnée à l'état "START", X_C à l'état "CENTER" et X_E à l'état "END".

4.3 Premier système syllabique

A partir de cette base phonétique, chaque modèle de syllabe est construit par concaténation des modèles de phonèmes, comme illustré Figure 2.

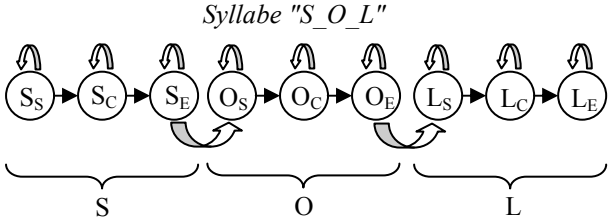

Figure 2 : Exemple de phonèmes concaténés en syllabe

Pour le décodage, on utilise le bigram syllabique décrit ci-dessus. Pour chaque trame, l'algorithme de décodage calcule ici les probabilités d'émission de 105 (35x3) états différents.

4.4 Second système syllabique

L'inconvénient majeur du système précédent est l'apprentissage de chaque modèle phonétique

indépendamment de leur contexte. Or les phonèmes, au sein d'un flux de parole, ne se suivent pas brutalement et des phénomènes de coarticulation apparaissent en fonction du contexte d'émission de chaque phonème. Il s'agit là d'un problème récurrent de la reconnaissance de parole auquel on répond le plus souvent par l'ajout d'informations contextuelles (diphones, triphones) [9].

Dans le cadre d'un système basé sur des unités syllabiques, il est souhaitable de travailler sur des modèles syllabiques cohérents à l'intérieur desquels le contexte est pris en compte et correctement modélisé. Dans cette optique, chacune des 4352 syllabes est associée à une chaîne de Markov spécifique, directement issue de la modélisation vue ci-dessus et présentée en Figure 3.

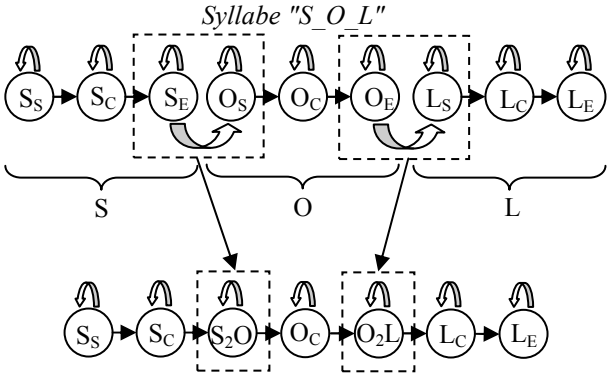

Figure 3 : Création d'un modèle syllabique

Cette modélisation conserve la topologie issue de la concaténation des chaînes de Markov phonétiques, à ceci près que les états extrémités des phonèmes contigus sont fusionnés en un seul, représentant le contexte de transition. La nouvelle notation consacrée pour ces nouveaux états transitoires entre deux phonèmes X et Y est X_2Y (X to Y). Ainsi, sur l'exemple présenté en Figure 3, les états S_E et O_S sont fusionnés en S_2O et O_E et L_S en O_2L. Cette fusion de deux états entraîne une réduction du nombre d'états à parcourir lors du décodage, car là où une concaténation de *n* phonèmes était modélisée par 3*n* états, le modèle syllabique correspondant n'en aura plus que 2*n*+1. Par ce formalisme, l'intégralité de l'espace de parole est donc couverte par un ensemble de 1295 états différents.

En outre, rappelons que ces syllabes ont été extraites d'un corpus textuel bien plus important que les seules données du corpus d'apprentissage audio. En effet, ce dernier n'en contenant que 2565, près de la moitié des modèles syllabiques ne seront pas appris directement sur le corpus audio. Malgré tout, la topologie de ces syllabes permet d'étendre l'apprentissage aux modèles non rencontrés en partageant les états [15], comme indiqué en Figure 4.

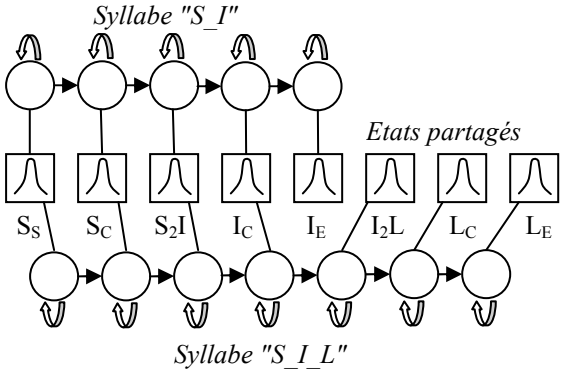

Figure 4 : Partage d'états entre syllabes

Au final, les 4352 syllabes se partagent donc 1295 états différents, c'est-à-dire près de 10 fois plus de probabilités d'émission à calculer lors du décodage que pour le système phonétique.

Une fois tous les modèles générés, l'apprentissage s'effectue suivant les mêmes étapes que pour le système phonétique.

5. PERFORMANCES ET COMPLEXITES

Le Tableau 1 reprend la description technique des trois systèmes décrit précédemment.

Tableau 1 : Désignation des systèmes

Systèmes	Désignation
A	Système phonétique avec bigram phonétique (§ 4.2)
B	Système phonétique avec bigram syllabique (§ 4.3)
C	Système syllabique avec bigram syllabique (§ 4.4)

Les performances et complexité de chacun des systèmes sur le corpus de test BREF précédemment décrit sont présentées Tableau 2. Le PER (Phone Error Rate) représente le taux d'erreur par phonème:

$$PER = 100 - Accuracy = 100 * \frac{S + I + D}{N}$$

Où N correspond au nombre total de phonèmes de la transcription manuelle de référence, I aux insertions, S aux substitutions et D aux omissions. Les silences ne sont pas pris en compte dans la mesure afin de ne pas accroître artificiellement les performances [9].
Notons qu'afin de rendre comparables les performances des différents systèmes proposés, les syllabes décodées ont été retranscrites en phonèmes.
La complexité, exprimée en fonction du temps réel (RT=Real Time), est ici calculée sur un Pentium Xeon 3.7 GHz avec 2Go de RAM.

Tableau 2 : Performances des différents systèmes

Système	Corr.	Subs.	Del.	Ins.	PER	Complexité
A	74.9	16.2	8.9	1.8	**26.9**	0.11 RT
B	83.65	10.8	5.5	1.7	**18**	1.9 RT
C	86.35	9.8	3.9	2.1	**15.8**	2.3 RT

6. DISCUSSION

L'apport du bigram syllabique sur les performances est évident. En effet, le taux d'erreur est réduit de près de 9 points en construisant les syllabes à l'aide des 35 phonèmes de base et en y appliquant le modèle de langage. L'explication provient du fait que les phonèmes sont dorénavant guidés sur un sous-espace plus représentatif de la parole, chaque syllabe (concaténation de *n* phonèmes) faisant office de *n*-gram au niveau phonétique. Le système C quant à lui, en réduisant encore le taux d'erreur de près de 3 points, montre l'intérêt d'inclure des états en contexte au sein des modèles syllabiques.

Les complexités liées à l'application du bigram syllabique dépendant du nombre élevé d'unités traitées, la durée de décodage atteint ici plus de 2 fois le temps réel. Notons que la complexité est moins élevée pour le système B que pour le système C, ce que l'on peut expliquer par le nombre de probabilités d'émissions différentes à calculer pour chaque système (10 fois moins pour B que pour C).
Notons toutefois qu'avec une simple stratégie d'élagage [10] dans l'algorithme de Viterbi, le temps de décodage du système C a été ramené à une fois le temps réel pour une même performance.

7. CONCLUSION ET PERSPECTIVES

Cette utilisation des syllabes comme guides de la reconnaissance phonétique permet donc de réduire significativement le taux d'erreur sur les phonèmes en proposant un système simple et performant, fonctionnant en temps réel et conservant une bonne liberté de généralisation pour des applications de détections de mots clés ou de reconnaissance grands vocabulaires. En outre, l'ajout de l'information contextuelle au sein même des entités syllabiques accroît également les performances grâce à une modélisation plus précise du signal sur des portions plus larges.

Cependant des améliorations sont envisagées, notamment au niveau des états transitoires X_2Y, pour lesquels une modélisation à un seul état n'est probablement pas suffisante étant donnée la dynamique de coarticulation. Il serait également utile d'enrichir les modèles syllabiques par l'information de durée moyenne des syllabes et par l'adaptation au genre du locuteur [9], ainsi que d'étendre le modèle de langage à plusieurs millions de mots. Finalement, une analyse plus poussée des stratégies d'élagage et des algorithmes de décodage permettrait de réduire la complexité de notre système.

BIBLIOGRAPHIE

[1] A. Acero, X. Huang. Augmented cepstral normalization for robust speech recognition. *Proc. of IEEE Automatic Speech Recognition Workshop*, 1995.

[2] L. E. Baum and J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology. *Bull. Amer. Math. Soc.,* 73:360--363, 1967.

[3] J.M. Boite et C. Ris. Development of a french speech recognizer using a hybrid HMM/MLP system. *ESANN*, 1999.

[4] B. Fisher. Syllabification Software. Tsylb2-1.1. *http://www.nist.gov/speech/tools/*

[5] A. Ganapathiraju et al. Syllable-Based Large Vocabulary Continuous Speech Recognition. *IEEE transactions on speech and audio processing*, vol. 9, N°4, May 2001.

[6] J.L. Gauvain, L.F. Lamel. Speaker-Independent Phone Recognition Using BREF. *ICASSP*, 1992.

[7] S. Greenberg. On the origins of speech intelligibility in the real world. *ESCARSR*, 97.

[8] L.F. Lamel, J.L. Gauvain, M. Eskénazi. BREF, a Large Vocabulary Spoken Corpus for French. *EUROSPEECH,* 1991.

[9] L.F. Lamel, J.L. Gauvain. High Performance Speaker-Independent Phone Recognition Using CDHMM. *EUROSPEECH*, 1993.

[10] B. Lowerre. The Harpy Speech Recognition System. *PhD Thesis, Carnegie-Mellon University,* 1976

[11] C. Pallier. Syllabation des représentations phonétiques de Brulex et de Lexique. 2004.

[12] L.R. Rabiner, B.H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, January 4-15, 1986.

[13] C. Schrumpf et al. Syllable-based language models in speech recognition for English spoken document retrieval. *Proceedings of the 7th International Workshop of the EU Network of Excellence DELOS on AVIVDiLib,* 2005.

[14] S. Young et al. Token Passing : a Simple Conceptual Model for connected Speech Recognition Systems. *Technical Report CUED/F-INFENG/TR38, Cambridge University Engineering Dept,* 1989.

[15] S. Young. The general use of tying in phoneme-based hmm speech recognisers. *IEEE*, 1992.

[16] S. Young et al. The HTK Book (for HTK version 3.3). *Cambridge University Engineering Department*, April 2005.