Optimisation d'arbres de décision pour la conversion graphèmes-phonèmes

Hubert Crépy, Christel Amato-Beaujard, Jean-Christophe Marcadet, Claire Waast-Richard

IBM France
Tour Descartes - 2, avenue Gambetta (La Défense 5) -- 92400 Courbevoie (France)
Tél.: +33 (0)1 49 05 71 22 Mél: crepy@fr.ibm.com

ABSTRACT

Extensive experiments on a data-driven decision-tree technique for French grapheme-to-phoneme conversion are dedicated to studying the effects of various tree-growing parameters as well as features and questions selection. Generated phonetic transcriptions of unknown words are used for speech recognition and synthesis. We report surprisingly good results, with recognition error rates better than with rule-generated transcriptions, and only slightly worse than with reference man-made transcriptions, and transcription phonetic error rates measuring as low as 1.56%, thanks in part to the introduction of POS tags into the context features.

INTRODUCTION

La *conversion graphèmes-phonèmes* (également appelée *phonétisation*) vise à la génération automatique de transcriptions phonétiques à partir d'entrées orthographiques. Elle trouve son application là où les lexiques rencontrent leurs limites, lorsque l'on a besoin de transcriptions pour des mots qui étaient inconnus lors de la conception du système (dits *mots hors vocabulaire*), aussi bien en *reconnaissance de la parole* (ajout de mots personnels à un système de dictée en grand vocabulaire, application à vocabulaire dynamique, …) qu'en *synthèse* (domaine lexical ouvert).

Les différentes techniques mises en œuvre pour générer ces transcriptions [Dam98] peuvent être schématiquement classifiées en deux grandes familles. Les *approches par règles* explicitement écrites ont généralement la faveur des experts linguistes et phonéticiens car elles sont plus facilement intelligibles, adaptables et contrôlables ; les *approches statistiques* basées sur l'apprentissage d'un lexique aligné de transcriptions présentent le grand avantage de requérir moins de travail expert (à la condition de disposer d'un lexique de transcriptions) et d'ainsi se prêter plus facilement à l'extension à de nouvelles langues. La dichotomie entre les deux écoles n'est pas aussi claire que cela, puisqu'en pratique l'expert rédige ses règles en faisant plus ou moins explicitement référence à un grand nombre de transcriptions connues. Damper et al. [Dam98] notent à ce sujet que cela rend l'indépendance de l'évaluation quelque peu hasardeuse, car s'il est possible d'exclure le jeu de test lors de l'apprentissage d'un système statistique, tel n'est pas le cas pour le "jeu d'apprentissage" implicite d'un expert écrivant des règles.

Parmi les différentes approches statistiques proposées, les expériences présentées ici reposent sur celle des *arbres de décision* implémentée dans les systèmes IBM, dont on trouve description dans [Luc84] et [Bah91]. Elle permet de transcrire des mots avec ou sans l'aide d'une comparaison acoustique avec un ou plusieurs enregistrements du mot par l'utilisateur. Deux circonstances nous incitent à revenir sur le sujet : le développement d'applications téléphoniques multi-utilisateurs en domaine relativement ouvert, qui renouvelle l'intérêt pour la conversion graphèmes-phonèmes pure sans enregistrement témoin, et l'accès à des lexiques de transcriptions de plus en plus massifs, qui permet l'entraînement de systèmes statistiques plus robustes.

Nous explorons les différentes dimensions d'optimisation du système pour étalonner et améliorer la performance de génération de transcriptions pour la reconnaissance aussi bien que pour la synthèse. Un problème relativement particulier à la langue française est celui des assez nombreux *homographes hétérophones* ("les poules du couvent couvent") dont l'ambiguïté ne peut être résolue en l'absence d'enregistrement témoin, alors qu'elle est cruciale en synthèse. On verra dans la suite comment la prise en compte de la *catégorie grammaticale* permet de résoudre ce problème.

Dans tous les cas, les caractéristiques recherchées sont la *compacité* (en particulier pour les systèmes embarqués), la *vitesse* de transcription, la *fidélité* (génération de transcriptions correctes pour des mots présents lors de l'apprentissage), mais surtout la capacité de *généralisation* (phonétisation correcte de mots inconnus).

DONNÉES D'APPRENTISSAGE

Le lexique de formes phonétiques utilisé contient approximativement 529.000 transcriptions, dont une majorité de formes fléchies des mots du français, 26.000 noms de famille et 12.000 prénoms (français et étrangers).

Alignement et sélection des données Dans une première phase, le lexique de transcriptions est aligné par une procédure semi-automatique, qui note pour chaque lettre du mot combien (0, 1 ou 2) de phonèmes sont produits par celle-ci. La procédure peut être totalement automatique [Luc84] [Bla98], mais à l'instar d'autres auteurs [Suo00], notre procédure actuelle est guidée par quelques règles prédéfinies. Les mots du jeu de test sont exclus du jeu d'apprentissage.

UTILISATION ET CALCUL DES ARBRES

Génération de transcription

Le processus fonctionne lettre par lettre de gauche à droite sur le mot à phonétiser. Le *contexte* pris en compte comprend 5 lettres à gauche, 5 lettres à droite, et les 5 phonèmes générés les plus récents. A chaque nœud de l'arbre, on pose une *question* sur l'un de ces 15 éléments (la lettre suivante est-elle un "a" ? Le phonème d'avant appartient-il à la classe des "plosives"?). On parcourt l'arbre jusqu'à une feuille qui donne une distribution de probabilités pour que cette lettre émette l'une des suites de phonèmes possibles, construisant ainsi une série d'hypothèses dont on garde les N plus probables.

Apprentissage de l'arbre

Questions suggérées L'arbre peut poser des questions spécifiques sur la valeur d'une lettre ou d'un phonème, mais également sur son appartenance à une classe prédéterminée (par exemple, consonnes ou voyelles). Il nous revient de suggérer un certain nombre de classes cohérentes et significatives. Ces classes pourraient être découvertes automatiquement par une procédure statistique telle que décrite dans [Luc84], reprise et améliorée par [Kei01], mais un choix "expert" à ce niveau nous semble plus susceptible de donner à l'arbre de bonnes performances de généralisation.

Développement de l'arbre, calcul des probabilités Pour chaque nœud de l'arbre, l'algorithme d'apprentissage (dont les bases sont posées dans [Luc84] et une version plus récente explicitée dans [Bah91]) examine toutes les questions suggérées et choisit la plus pertinente à ce niveau, celle qui partitionne le mieux les données présentes au regard d'un critère d'entropie. L'opération est répétée sur chaque branche, jusqu'à ce qu'un critère d'arrêt stoppe le partitionnement et donne lieu à la génération d'une feuille de l'arbre. A chaque niveau, les probabilités sont *lissées* par une combinaison linéaire avec celles du nœud précédent (10%), ce qui limite le sur-apprentissage.

Critères d'arrêt Un nœud n'est plus divisé si l'une des deux conditions suivantes est vérifiée : (1) *Taille de Nœud Minimum* TNM (en dessous d'un certain nombre d'échantillons d'apprentissage présents au nœud, on tend vers un sur-apprentissage des exceptions non-significatives) ; ou (2) *Gain d'Entropie Minimum* GEM (si aucune division envisagée ne permet de gagner suffisamment en information).

TESTS ET MESURES

Pour une utilisation en reconnaissance

Quand l'objectif est de générer des transcriptions phonétiques pour la reconnaissance, le critère de qualité final est bien le *taux d'erreur d'un système de reconnaissance* utilisant ces transcriptions. La mesure d'un taux d'erreur de transcription par comparaison à des transcriptions de référence, si elle est généralement plus facile, n'est qu'un pis-aller : elle ne prend pas aisément en compte les gradations de gravité des erreurs pas plus que le compromis à trouver sur le nombre de prononciations à proposer pour chaque mot.

Notre jeu de test de base est une *tâche "d'assistant téléphonique"* : 824 enregistrements de "prénom-nom" correspondant à une grammaire de 5.000 possibilités (6.451 mots différents, exclus du jeu d'apprentissage). Les enregistrements proviennent d'appels téléphoniques réels, spontanés, et contiennent une mixture d'hommes et de femmes ainsi que de qualité de ligne. Les expériences sont réalisées en faisant générer les transcriptions des 6.451 mots possibles de la grammaire par l'arbre.[1]

Pour vérifier que tous les tests et réglages successifs n'aboutissaient pas à un système sur-appris sur le jeu de test, on a utilisé un *jeu de contrôle* indépendant : enregistrements de 8 locuteurs lisant chacun une liste de 100 noms de *stations du métro* parisien, le vocabulaire de reconnaissance étant limité à ces 100 noms dont les transcriptions sont générées par l'arbre.

Pour une utilisation en synthèse

La meilleure mesure serait basée sur des tests d'écoute subjectifs, malheureusement prohibitifs pour l'évaluation du seul sous-système de phonétisation. On se rabat donc sur une mesure de taux d'erreur de transcription par rapport à des *transcriptions de référence*.

Pour cela, 10% du lexique initial (env. 47.000 mots) sont exclus du jeu d'apprentissage et constituent le jeu de test. On calcule un arbre sur le jeu d'apprentissage réduit. Pour chaque mot du jeu de test transcrit par l'arbre, on ne conserve qu'une seule hypothèse de transcription, et on mesure les erreurs phonétiques (insertions, suppressions, substitutions) par alignement avec toutes les transcriptions référence du mot (variantes légitimes de prononciation, les homographes hétérophones restant séparés). La transcription référence la plus proche de la transcription générée donne le *taux d'erreur phonétique* (TEP). Nous citons également le *taux d'erreurs par mots* (TEM), pourcentage de mots contenant au moins une erreur phonétique, bien qu'il nous paraisse moins approprié. Il s'agit de taux d'erreurs sur un lexique non pondéré par la fréquence des mots, donc a priori plus sévères que les mesures sur phrases [Yvo98] qui donnent un poids plus fort aux mots plus courants, mieux prédits par le système.

[1] En application réelle, une grande majorité d'entre eux (en particulier des prénoms) auraient eu une transcription connue du lexique.

EXPÉRIENCES, RÉSULTATS ET DISCUSSION

Paramètres de génération

Un des plus importants paramètres lors de l'utilisation de l'arbre est le *Score Relatif Minimum* qui permet de filtrer les hypothèses trop improbables par rapport à la meilleure. Au réglage optimal par rapport au taux d'erreur (en trait plein sur la figure), on génère assez de transcriptions par mot (1,7 en moyenne) sans toutefois créer trop de confusions (sachant que par ailleurs on limite le nombre maximal d'hypothèses à 3 par mot). La vitesse de traitement est figurée en pointillés sur l'échelle de droite, en mots/seconde sur notre machine de test.

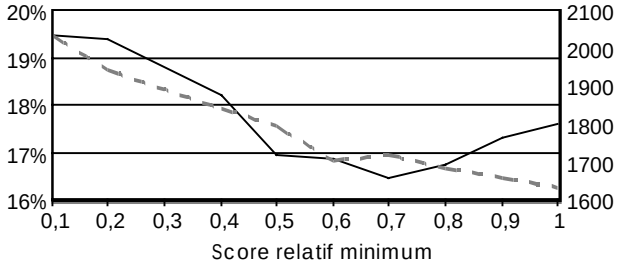

Figure 1 : Réglage du Score relatif minimum

Paramètres de construction

On contrôle l'apprentissage de l'arbre par les paramètres de critères d'arrêt : *Gain d'Entropie Minimum* (GEM) et *Taille de Nœud Minimum* (TNM). Plus ces seuils sont bas, plus l'arbre développe un grand nombre de feuilles représentant des spécificités locales du jeu d'apprentissage. Le lissage des probabilités avec les nœuds de rang supérieur permet de pallier le danger de sur-apprentissage dans une certaine mesure. Nous trouvons (Figure 2) *qu'il est souhaitable de développer l'arbre de façon assez extensive* (GEM=5,0 et TNM=5 pour 6.527 feuilles), sans toutefois aller trop loin (le développement maximum à GEM=0,0 et TNM=1 –limité toutefois à 50 niveaux de nœuds- donne 16.179 feuilles), ce qui nuirait aux capacités de généralisation. Ce résultat concorde avec celui de Black et al. [Bla98].

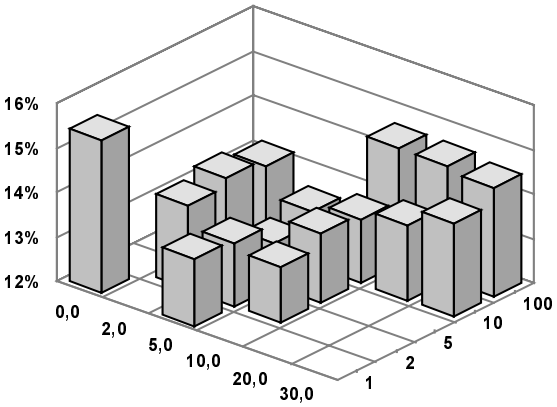

Figure 2 : Taux d'erreur en reconnaissance en fonction du GEM (de gauche à droite) et de la TNM (avant-arrière)

Sélection des questions

Dans une expérience comparative, deux arbres ont été construits, l'un utilisant des *questions singletons* (mono-phonème ou mono-lettre), l'autre ayant la possibilité de poser des *questions composites* sur des classes de lettres ou de phonèmes. Lorsque la possibilité en est offerte, on observe que les questions composites donnent un meilleur gain d'entropie et sont souvent préférées aux questions singleton (à 41% des nœuds). De fait, l'arbre à questions composites donne un taux d'erreur 11.6% relatif meilleur que celui à questions singleton. Conformément à l'intuition, et comme observé par Andersen et al. [And96], les *questions composites confèrent à l'arbre de meilleures capacités de généralisation.*

79% des questions choisies portent sur les lettres et non sur les phonèmes. Les questions sur les lettres portent plutôt sur le contexte droit (45%) que sur le contexte gauche (34%), mais les questions phonétiques (21%) sont (par force) sur le contexte gauche, et les conventions d'alignement des données (en particulier doubles consonnes) ne sont certainement pas sans effet ici.

Performances globales en reconnaissance

Le meilleur arbre que nous ayons obtenu donne un taux d'erreur de 12,98% sur la tâche "prénoms-noms". On peut comparer (Figure 3) ce résultat avec ceux d'autres méthodes de phonétisation : phonétisation manuelle par un expert (12,14%), phonétisation automatique par un système de règles écrites par un expert (16,11%).

Il peut être intéressant à ce sujet de noter que le nombre de règles expertes (environ 8.000, beaucoup plus que les 500 à 4.000 des systèmes de règles évalués par Yvon & al. [Yvo98]) est d'un ordre de grandeur comparable au nombre de nœuds de l'arbre optimal (environ 6.500).

La confirmation de résultats similaires sur le jeu de contrôle indépendant (stations de métro) montre que *la phonétisation obtenue par l'arbre est presque aussi performante qu'une phonétisation manuelle,* ce qui constitue un résultat tout à fait satisfaisant et quelque peu surprenant.

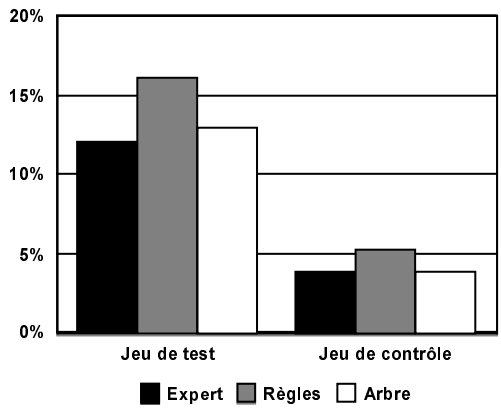

Figure 3 : Résultats en reconnaissance (taux d'erreur)

Prise en compte d'étiquettes grammaticales

La question des *homographes hétérophones* appelle une attention particulière : en synthèse, le choix de la mauvaise transcription est évidemment une erreur ; en reconnaissance, l'utilisation de nombreuses variantes pour couvrir l'ambiguïté est source d'erreurs de substitution.

Il apparaît qu'un *étiquetage grammatical* (par ailleurs disponible pour d'autres raisons en synthèse de la parole, comme noté par [Bla98]) permette de lever certaines ambiguïtés. Les exemples les plus courants sont les terminaisons en "-ent" ou "-tions" pour les verbes ou substantifs ("couvent", "formations"). De même, les règles de phonétisation des noms propres (en particulier d'origine étrangère) diffèrent des règles communes.

Plutôt que de construire des arbres séparés pour chaque catégorie grammaticale, on laisse les données elles-mêmes dicter l'utilisation de cette information aux endroits appropriés : la catégorie du mot est ajoutée aux éléments de contexte sur lesquels l'arbre peut poser des questions, au même titre que sur les lettres ou les phonèmes. En développant un tel arbre, on constate que la *question grammaticale* (nous utilisons uniquement des questions singleton, sur 12 classes grammaticales grossières) est posée à 9% des nœuds, à peine moins que les questions sur les lettres voisines (12% et 10%). D'après notre critère d'entropie pour la séparation des données, c'est donc bien une *information discriminante*.

Des mesures comparatives ont été réalisés avec deux arbres, construits avec et sans information grammaticale, en reconnaissance de noms propres et en synthèse. Les jeux de tests sont les mêmes que précédemment. Le jeu de test de synthèse est un tirage aléatoire de mots, et ne contient donc qu'une proportion "naturelle" d'homographes hétérophones.

Table 1 : Taux d'erreur avec et sans information grammaticale

	Sans	**Avec**
Reconnaissance	12,98%	11,94%
Synthèse (TEP)	1,64%	1,56%
Synthèse (TEM)	9,50%	9,25%

L'amélioration apportée par la prise en compte de la catégorie grammaticale du mot est donc détectable sur toutes les mesures disponibles. Elle est confirmée anecdotiquement en observant la phonétisation de quelques mots finissant en "-ent" ou "-tions", par exemple.

CONCLUSION

Ces expériences confirment que l'approche statistique à arbres de décision, relativement aisée à mettre en place quand on dispose d'un lexique d'apprentissage, donne des performances extrêmement satisfaisantes. En développant l'arbre de manière relativement extensive sur un gros volume de données d'apprentissage, les taux d'erreur de reconnaissance sur transcriptions générées sont comparables à ceux obtenus par des transcriptions manuelles. La prise en compte de la classe grammaticale du mot à phonétiser améliore encore les résultats et constitue une bonne solution au problème des homographes hétérophones en synthèse.

BIBLIOGRAPHIE

[Luc84] Lucassen J.M. et Mercer R.L. (1984), “An information-theoretic approach to the automatic determination of phonemic baseforms”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’84), pp. 42.5.1-42.5.4

[Bah91] Bahl L.R. et al. (1991), “Automatic phonetic baseform determination”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP’91), pp. 173-176

[And96] Andersen O. et al. (1996), “Comparison of two tree-structured approaches for grapheme-to-phoneme conversion”, ICSLP’96

[Bla98] Black A.W., Lenzo K., Pagel V. (1998), “Issues in building general letter to sound rules”, 3rd ESCA International Workshop on Speech Synthesis, pp 77-80

[Dam98] Damper R.I., Marchand Y., Adamson M.J., Gustafson K. (1998), “Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis”, 3rd ESCA International Workshop on Speech Synthesis, pp 53-58

[Yvo98] Yvon F. & al. (1998), "Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in French", Computer Speech & Language, Vol. 12, No. 4, pp. 393-410

[Suo00] Suontausta J., Häkkinen J. (2000), “Decision tree based text-to-phoneme mapping for speech recognition”, ICSLP 2000

[Man01] Mana F., Massimino P., Pacchiotti A. (2001), “Using machine learning techniques for grapheme to phoneme transcription”, EuroSpeech 2001

[Kei01] Keinappel A.K., Kneser R. (2001), “Designing very compact decision trees for grapheme-to-phoneme transcription”, EuroSpeech 2001