# Prédiction des Performances d'un Système de Reconnaissance Automatique de Parole

*Olivier Pietquin*[*]

Supélec – Campus de Metz
2 rue Edouard Belin – F-57070 Metz – France
olivier.pietquin@supelec.fr

*Richard Beaufort*

Multitel A.S.B.L.
1 Av. Copernic – B-7000 Mons –Belgique
richard.beaufort@multitel.be

## ABSTRACT

Pattern matching systems and especially automatic speech recognition systems are generally based on statistics theory. Because of this particular feature, performances of such systems never reach one hundred percent of correct recognition results. Those performances are linked to the environmental noise and to intra- and inter-speaker variability of course, but also to the vocabulary of allowed speech entries. In this paper, a method for predicting recognition task difficulty according to a given vocabulary is proposed.

## 1. INTRODUCTION

La reconnaissance automatique de parole est, depuis plusieurs dizaines d'années, le sujet de recherches intensives. Les techniques actuellement les plus performantes sont basées sur des modèles statistiques comme les Modèles de Markov Cachés (HMM : Hidden Markov Models)[1]. De part cette nature statistique, ces systèmes ne peuvent intrinsèquement pas atteindre des performances égalant les cent pourcents de reconnaissances correctes (tout comme l'être humain d'ailleurs). Les erreurs de reconnaissance sont souvent dues au bruit ambiant ou aux variabilités inter-locuteurs (timbre de voix, vitesse d'élocution etc.) et intra-locuteur (voix enrouée, stress, âge etc.). Néanmoins, même dans des conditions optimales, ces systèmes commettent des erreurs qui peuvent aussi être expliquées par les similitudes entre les différents mots que le système de reconnaissance doit traiter. Ces mots constituent un vocabulaire limité plus ou moins large.

L'estimation des performances d'un système de reconnaissances sur un vocabulaire donné peut être très importante afin de réaliser le design d'un vocabulaire plus performant quand cela est possible ou de prévoir des stratégies de confirmation dans le cas d'un système de dialogue homme-machine utilisant un vocabulaire sub-optimal. Il peut même être imaginé qu'un système de dialogue bien conçu choisisse une stratégie d'interaction faisant intervenir plus souvent des vocabulaires de reconnaissance offrant de meilleures performances si cela ne nuit pas à l'accomplissement de la tâche.

* Ce travail a été réalisé alors que le premier auteur travaillait à la Faculté Polytechnique de Mons (Belgique) et fut subventionné par le Réseau d'Excellence Européen SIMILAR.

## 2. PERFORMANCE DES SYSTÈMES DE RECONNAISSANCE VOCALE

Un système de reconnaissance automatique de parole (ASR : Automatic Speech Recognition) a généralement pour but de traduire en texte un signal acoustique de parole capturé par un microphone et transformé en signal numérique. On peut donc se représenter ce processus comme indiqué à la Figure 1.

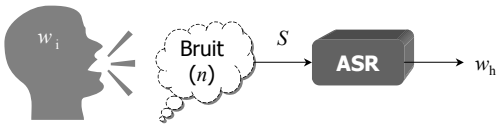


**Figure 1** : Processus de reconnaissance de la parole

Sur cette figure, un utilisateur prononce une séquence de mots $\boldsymbol{w_i} = \{w_{i1}, w_{i2},\ldots,w_{in}\}$ dans un environnent bruité. Au signal de parole s'ajoute donc le bruit $\boldsymbol{n}$ pour former le signal $\boldsymbol{S}$ reçu par le système de reconnaissance. Celui-ci transforme le signal reçu en une nouvelle séquence de mots $\boldsymbol{w_h} = \{w_{h1}, w_{h2},\ldots,w_{hm}\}$ (hypothèse). Il y a erreur de reconnaissance lorsque $w_h \neq w_i$. On peut rencontrer trois types d'erreurs :

- Les suppressions de mots : m < n
- Les insertions de mots : m > n
- Les substitutions de mots : $w_{ij} \neq w_{hj}$

On mesure alors les performances d'un système de reconnaissance vocale à son taux d'erreurs (WER : Word Error Rate) défini par :

$$\mathrm{WER} = \frac{N_{\mathrm{sup}} + N_{\mathrm{ins}} + N_{\mathrm{sub}}}{N} \times 100\%$$

Ici, $N$ est le nombre de mots que contient un ensemble utilisé pour tester le système et les $N_{xxx}$ sont les nombres d'occurrences des différents types d'erreurs durant ce test. La prédiction des performances d'un système passe donc par l'établissement d'une méthode permettant d'estimer cette mesure sans effectuer pour autant des tests longs et parfois impossibles à réaliser.

Comme cela a été dit auparavant, le processus de reconnaissance est stochastique. Il se base en fait sur l'équation suivante :

$$w_{\mathrm{h}} = \arg\max_{w} P(w \mid S) = \arg\max_{w} \frac{P(S \mid w) \cdot P(w)}{P(S)}$$

On peut donc déduire de cette équation que le processus est dépendant de la vraisemblance acoustique du signal $S$ étant donné la séquence de mots $w$ ($P(S|w)$) et de la probabilité absolue de rencontrer la séquence de mots $w$ ($P(w)$), plus communément appelée le *modèle de langage*. La probabilité absolue de rencontrer le signal $S$ ($P(S)$) n'intervient pas dans le processus de maximisation puisqu'elle est indépendante de $w$. Ceci veut dire que s'il est possible de confondre le signal $S$ avec la réalisation acoustique d'une séquence de mots $w_{\mathrm{j}} \neq w_{\mathrm{i}}$ ($P(S|w_{\mathrm{j}}) > 0$) mais que cette séquence possède une probabilité nulle d'être prononcée d'après le modèle de langage ($P(w_{\mathrm{j}}) = 0$), cette séquence ne sera pas produite par le système de reconnaissance. C'est donc lorsque des séquences de mots acoustiquement semblables sont autorisées par le modèle de langage que les erreurs sont les plus probables.

## 3. DISTANCE ACOUSTIQUE

Puisque les erreurs de reconnaissance semblent plus fréquentes entre des mots ou des séquences de mots dont la réalisation acoustique est semblable, il est indispensable de pouvoir identifier ces similarités pour prédire les performances d'un système de reconnaissance. Comme il existe une infinité de réalisations acoustiques pour chaque mot, il est impossible de les comparer directement. De plus, il serait plus pratique de pouvoir déterminer ces similitudes sur simple base de la transcription orthographique des mots. Pour cela, considérons qu'il est possible de dériver automatiquement une ou plusieurs transcriptions phonétiques (il peut exister plusieurs prononciations pour une même phrase) pour toute séquence de mots [[2]]. L'ensemble des transcriptions phonétiques de la séquences $w_{\mathrm{j}}$ sera notée

$$\Phi(w_{\mathrm{j}}) = \left\{ \varphi^{\alpha}(w_{\mathrm{j}}) \right\}_{\alpha=1}^{\mathrm{N_j}},$$

et chaque $\varphi^{\alpha}(w_{\mathrm{j}})$ représente une transcription phonétique possible. Chacune de ces transcriptions $\varphi^{\alpha}(w_{\mathrm{j}})$ est elle-même une séquence de phonèmes :

$$\varphi^{\alpha}(w_{\mathrm{j}}) = \left\{ \varphi_{\mathrm{k}}^{\alpha}(w_{\mathrm{j}}) \right\}_{\mathrm{k}=1}^{\mathrm{M}_{\alpha}}$$

Les phonèmes $\varphi_{\mathrm{k}}$ appartiennent à l'ensemble $A$ des phonèmes utilisés dans la langue étudiée (35 pour le français). Le problème revient alors à trouver une forme de distance acoustique entre deux séquences de phonèmes, ce qui peut s'exprimer comme la difficulté d'aligner deux séquences de symboles issus d'un alphabet donné. La méthode la plus populaire pour réaliser cet alignement est la programmation dynamique [[3]] qui permet, en associant un coût à chaque opération d'édition (insertion, suppression, substitution), de calculer le coût minimal d'alignement [[4]]. Pour utiliser cette technique, il nous faut donc définir un coût pour chaque opération d'édition qui reflétera au mieux la réalité acoustique qu'elle représente (insertion, suppression, substitution de sons).

Afin de déterminer le coût de substitution entre deux phonèmes $\varphi_{\mathrm{i}}$ et $\varphi_{\mathrm{j}}$, nous proposons de nous baser sur les propriétés articulatoires de ceux-ci. En effet, chaque phonème peut être caractérisé par un type (voyelle ou consonne), un point d'articulation (palatal, bilabial, etc.), la position des lèvres, le caractère voisé ou non et l'articulation (plosive, nasale, fricative, orale, liquide). Les voyelles peuvent, en plus, être caractérisées par un degré d'aperture (ou d'ouverture). En définissant l'ensemble des propriétés articulatoires du phonème $\varphi_{\mathrm{i}}$ ($\varphi_{\mathrm{j}}$) par $f_{\mathrm{i}}$ ($f_{\mathrm{j}}$), on peut définir le coût de substitution du phonème $\varphi_{\mathrm{i}}$ par le phonème $\varphi_{\mathrm{j}}$, comme la distance entre les deux ensembles de caractéristiques $f_{\mathrm{i}}$ et $f_{\mathrm{j}}$. Cette distance peut être dérivée du '*ratio model*' de Tversky [[5]] :

$$d(\varphi_{\mathrm{i}}, \varphi_{\mathrm{j}}) = \frac{F(f_{\mathrm{i}} | f_{\mathrm{j}})}{F(f_{\mathrm{i}} \cap f_{\mathrm{j}}) + F(f_{\mathrm{i}} | f_{\mathrm{j}}) + F(f_{\mathrm{j}} | f_{\mathrm{i}})}$$

Dans cette équation, $F(.)$ est une fonction appliquée à un ensemble permettant d'assigner un poids à chaque élément de celui-ci. On peut remarquer que cette distance est asymétrique.

Il nous reste donc à définir la fonction $F$. Si nous définissons cette fonction comme étant le cardinal de l'ensemble comme ça a pu se faire dans des études précédentes, la distance devient symétrique et $d \in [0,1]$. Elle possède alors l'allure décrite par la *Figure 2*, sur laquelle les phonèmes du français sont numérotés de 0 à 35 (le dernier étant le silence) en commençant par les 15 voyelles.

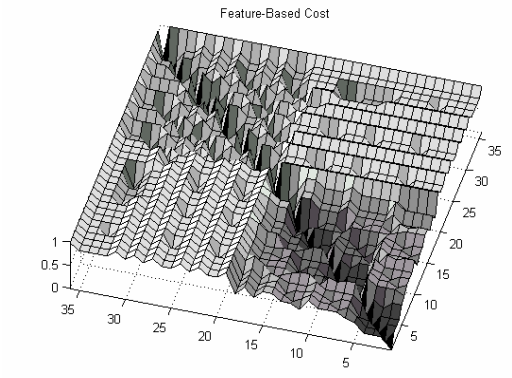


**Figure 2 :** Distance symétrique entre phonèmes

Néanmoins, il serait probablement judicieux de tenir compte de certains constats que la linguistique nous apprend, c'est à dire par exemple que :

- la substitution d'une voyelle par une consonne est très rare.

- la substitution entre deux consonnes qui ne diffèrent que par le point d'articulation est plus fréquente :

  ex: artère [a R t e r] → [a l t e r] (altère)

- la confusion entre des voyelles qui ne diffèrent que par le degré d'aperture est plus fréquente. Elle arrive plus souvent en augmentant l'aperture :

  ex: aimais [e m E] → [EmE]

- une consonne voisée peut devenir non-voisée lorsqu'elle est suivie d'une consonne non-voisée :

  ex: obturer [o b t y r E] → [o p t y r E]

- une voyelle peut être prononcée avec un degré d'aperture différent si elle est proche d'une autre voyelle ne divergeant que par cette caractéristique, spécialement si cette seconde voyelle porte l'accent tonique :

  ex: apaiser [a p E z 'e] → [a p e z 'e][1]

En incorporant ces données sous forme de poids dans la fonction *F*, nous obtenons une distance asymétrique dont la forme est montrée à la *Figure 3*.

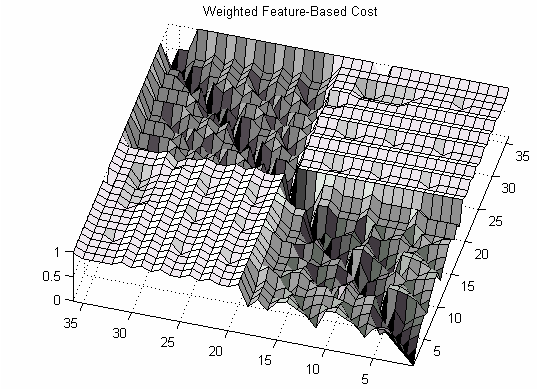


**Figure 3** : Distance asymétrique entre phonèmes

La linguistique nous donne aussi des indications quant aux coûts de suppression et d'insertion de phonèmes dans une séquence :

- les insertions ou suppressions de consones liquides (r ou l) sont plus fréquentes :

  ex: altère [a l t e r] → [a t e r] (à terre)

- l'insertion ou la suppression du *schwa* (@) est plus fréquente que pour les autres voyelles.
- l'insertion ou la suppression d'une consonne apparaît plus souvent devant une autre consonne :

  ex: à terre [a t e r] → [a l t e r] (altère)

- les suppressions apparaissent plus souvent au début ou à la fin d'un mot

Ces remarques permettent donc aussi d'associer des coûts différents aux opérations de suppression et d'insertion de phonèmes dans les séquences et donc de calculer la distance $d(\varphi^{\alpha},\varphi^{\beta})$ entre deux séquences grâce à la programmation dynamique sur base de propriétés articulatoires (fortement corrélées avec les propriétés acoustiques évidemment).

Comme cela a été mentionné auparavant, une séquence de mots *w* peut posséder plusieurs transcriptions phonétiques, c'est pourquoi nous définirons la distance entre deux séquences de mots comme la distance moyenne entre les différentes transcriptions :

$$d(w_i, w_j) = \sum_{\alpha=1}^{N_i}\sum_{\beta=1}^{N_j} d\left(\varphi^{\alpha}(w_i), \varphi^{\beta}(w_j)\right)\cdot P\left(\varphi^{\alpha}(w_i)\right)\cdot P\left(\varphi^{\beta}(w_j)\right)$$

Si nous admettons que toutes les prononciations sont équiprobables, nous obtenons :

$$d(w_i, w_j) = \frac{1}{N_i N_j}\sum_{\alpha=1}^{N_i}\sum_{\beta=1}^{N_j} d\left(\varphi^{\alpha}(w_i), \varphi^{\beta}(w_j)\right)$$

## EXPÉRIENCES ET RÉSULTATS

Afin d'utiliser la distance définie dans le paragraphe précédent dans le but de prédire les performances d'un système de reconnaissance vocale, nous avons effectué un test de reconnaissance vocale sur la base de données en français BDSons [[6]] comportant un vocabulaire *V* de 538 mots (|V| = 538) prononcés par des locuteurs masculins et féminins (plus de 8000 enregistrements) et nous avons tenté de corréler les erreurs de reconnaissance avec la distance entre les mots. Pour se faire, nous avons utilisé la distance afin de segmenter le vocabulaire des 538 mots en regroupant dans un même segment les mots dont la distance des uns aux autres était en-dessous d'un certain seuil (théoriquement les plus proches acoustiquement). Une partie des résultats est montrée dans le *Tableau 1*.

| Mot | Taille du Segment | WER | Mots dans le même segment |
|---|---|---|---|
| Barre | 23 | 95% | Bal, Balle, Dard, Gare, Par, Berre, Car, Jars, Tard, Parle, Dalle, Gale, Pal, Beurre, Bord, Bore, Gère, Guerre, Père, Char, Phare, Sar |
| Feinte | 16 | 75% | Faite, Sainte, Fente, Teinte, Peinte, Quinte, Tinte, Pinte, Geinte, Fonte, Fête, Sente, Vente, Chante, Faute |
| Express | 1 | 0% | |

**Tableau 1** : Résultats de reconnaissance sur BDSons

On peut se rendre compte que plus le segment contient de mots, plus le taux d'erreurs est élevé. Ceci pourrait nous conduire à définir un taux d'erreurs par mot du type :

$$E(w) = \alpha \,.\, \#(\text{segment}(w))$$

[1] Notons que sur la transcription de ce mot, les dictionnaires divergent. Certains notent la transcription phonémique [a p E z e], d'autres la phonétique [a p e z e].

Une estimation du taux d'erreurs global pourrait alors être donnée par :

$$WER = \alpha \cdot \frac{\sum_{w \in V} \#(\text{segment}(w))}{|V|}$$

D'autres possibilités peuvent être envisagées pour faire le lien entre la distance et le taux d'erreurs comme, par exemple, l'utilisation de cette distance afin de dériver une probabilité de confusion entre deux mots. Il faut alors choisir une fonction bijective de $\Re^+ \rightarrow [0,1]$. Par exemple, on peut utiliser la relation suivante :

$$P\left(\varphi^{\alpha}(w_h) | \varphi^{\beta}(w_i)\right) \propto e^{-\lambda \cdot d\left(\varphi^{\alpha}(w_h), \varphi^{\beta}(w_i)\right)}$$

Jusqu'ici, nous n'avons traité que de reconnaissance de mots isolés (lexique). Or, le problème de la reconnaissance vocale dans le cadre de systèmes de dialogue vocaux par exemple fait intervenir des modèles de langage plus complexes. Il est alors important de tenir compte des séquences de mots possibles et plus du lexique uniquement. Pour ce faire, nous avons utilisé la notion de perplexité acoustique définie dans [[7]] afin d'estimer un taux d'erreur de reconnaissance sur base d'une distance acoustique et d'un modèle de langage. La perplexité acoustique peut être approximée [8] par :

$$\hat{APP}_{LM} \approx \left( \prod_{w_h \in C} \frac{\sum_{\alpha=1}^{|\Phi(w_h)|} \sum_{\beta=1}^{|\Phi(w_h)|} P\left(s\left(\varphi^{\alpha}(w_h)\right) | \varphi^{\beta}(w_h)\right) \cdot P_{LM}\left(\varphi^{\beta}(w_h) | \ell_h\right)}{\sum_{w_i \in V} \sum_{\alpha=1}^{|\Phi(w_h)|} \sum_{\beta=1}^{|\Phi(w_i)|} P\left(s\left(\varphi^{\alpha}(w_h)\right) | \varphi^{\beta}(w_i)\right) \cdot P_{LM}\left(\varphi^{\beta}(w_i) | \ell_h\right)} \right)^{\frac{1}{|C|}}.$$

Dans cette relation, $P_{LM}(\varphi^{\beta}(w_h)|\ell_h)$ est la probabilité d'occurrence d'après le modèle de langage de la transcription phonétique $\varphi^{\beta}(w_h)$ précédée par le contexte gauche $\ell_h$. Le calcul de cette valeur pour différents modèles de langage et sa comparaison au taux d'erreurs d'un système de reconnaissance est présentée à la Figure 4.

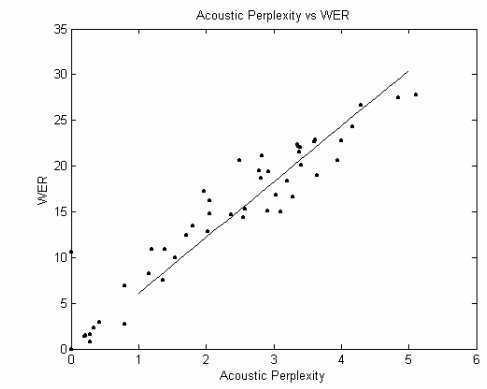


**Figure 4 :** Perplexité acoustique en fonction du taux d'erreurs

On peut se rendre compte sur cette figure qu'une bonne corrélation existe et que le calcul de la perplexité acoustique approximée peut donner lieu à une prédiction correcte du taux d'erreur sur un vocabulaire donné.

## CONCLUSIONS ET PERSPECTIVES

Cet article présente une méthode de définition d'une distance acoustique entre mots ou séquences de mots sur base des propriétés articulatoires des phonèmes qui les composent. Cette distance a permis de segmenter un vocabulaire et de montrer une certaine corrélation entre le nombre de mots semblables à un autre et le taux d'erreurs d'un système de reconnaissance vocale sur ce mot. L'utilisation d'une mesure approximée de la perplexité acoustique a aussi permis de donner une nouvelle mesure assez bien corrélée avec le taux d'erreurs.

Cela laisse entrevoir la possibilité de prédire les performances d'un système de reconnaissance automatique de parole sur un vocabulaire donné. Il serait alors possible de définir des vocabulaires plus performants ou des stratégies adéquates dans le cas de systèmes de dialogue utilisant de tels vocabulaires. Le système décrit ici a été comparé à d'autres méthodes de prédiction de performances des systèmes de reconnaissance vocale dans le cadre de l'apprentissage automatique de stratégies de dialogue et il a donné des résultats prometteurs [9].

## BIBLIOGRAPHIES


[1] L. R. Rabiner, '*A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.*' Proceedings of the IEEE, Vol.77, No.2, pp.257--286, 1989.

[2] J. Allen, S. Hunnicutt, D. Klatt, '*From Text to Speech: the MITalk System*.' Cambridge University Press, Cambridge, 1987.

[3] R. Bellman, '*Dynamic Programming*.' Princeton University Press, Princeton, NJ, 1957

[4] V. Levenshtein, '*Binary Codes Capable of Correcting Deletions, Insertions and Reversals*.' Soviet Physics Daklady, vol. 10, pp. 707-710, 1966.

[5] A. Tversky, '*Features of Similarity*.' In Psychological Review, vol. 84, no. 4, pp. 327-352, 1977.

[6] R. Carre, R. Descout, M. Eskenazi, J. Mariani, M. Rossi, '*The French Language Database: Defining, Planning, and Recording a Large Database*.' In Proceedings of ICASSP'84, pp. 42.10.1-42.10.4, 1984.

[7] H. Printz, P. Olsen, '*Theory and Practice of Acoustic Confusability*.' In Proceedings of ISCA ITRW ASR2000 Workshop, Paris, France, 2000, pp. 77-84.

[8] O. Pietquin, T. Dutoit, '*A Probabilistic Framework for Dialog Simulation and Optimal Strategy Learning*' IEEE Transactions on Audio, Speech and Language Processing, Volume 14, Issue 2, March 2006 Page(s):589 – 599.

[9] O. Pietquin, R. Beaufort, '*Comparing ASR Modeling Methods for Spoken Dialogue Simulation and Optimal Strategy Learning.*' In Proceedings of Eurospeech'05, Lisbon, Portugal, 2005.