# ETUDE DES VARIABILITÉS DE SYSTÈMES DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE UTILISANT DES PARAMÈTRES ACOUSTIQUES DIFFÉRENTS

*Loïc Barrault*

LIA CNRS
BP 1228 - 84911 Avignon Cedex 9 - France
loic.barrault@univ-avignon.fr

## RÉSUMÉ

Cet article décrit l'utilisation de deux systèmes de reconnaissance automatique de la parole (RAP) dont les paramètres d'entrée diffèrent. Le premier reconnaisseur effectue une Analyse à Résolution Multiple (MRA) alors que le second calcule les coefficients par Prédiction Linéaire Perceptive (JRASTAPLP). Les deux systèmes utilisent la même technique de débruitage mais effectuent des partitions différentes de leurs espaces acoustiques. Des expériences avec les parties italienne et espagnole du corpus Aurora3 montrent que les deux systèmes fournissent, dans une proportion signifiante de cas, des probabilités postérieures substantiellement différentes pour un même phonème et un intervalle de temps donné. Une règle de décision est proposée quand deux mots différents sont proposés par les deux reconnaisseurs. Elle est basée sur la probabilité qu'une hypothèse soit correcte étant donnée l'identité des hypothèses de mots en compétition. Une réduction significative du taux d'erreur mot (WER) a été observée pour la portion CH1 des parties italienne et espagnole du corpus Aurora3.

## 1. INTRODUCTION

L'utilisation de réseaux de neurones (ANN), d'arbres de décision et d'autres techniques d'apprentissage automatique ont été utilisées ([6], [5]) pour combiner les résultats de plusieurs systèmes de Reconnaissance Automatique de la Parole (RAP) afin de réduire le taux d'erreur mot (WER). Dans [7], une combinaison log-linéaire de modèles et des modèles de combinaison de paramètres acoustiques sont proposés pour améliorer les performances de RAP. Dans [4], une combinaison de modèles bayésiens des sorties des systèmes de RAP qui calcule la vraisemblance des phrases proposées par chaque système étant donné les hypothèses des systèmes et leurs score de confiance est proposée. L'indépendance entre les systèmes est supposée et les probabilités d'exactitude dépendent des performances du système global sans considérer quels phonèmes sont proposés par chaque système. D'autres travaux récents se concentrent sur les facteurs affectant le WER dans les systèmes de RAP. Dans [3], il est montré que la combinaison d'un rapport signal sur bruit (SNR) au niveau de la phrase avec ses variations locales fournissent des prédictions utiles du taux d'erreur de reconnaissance.

L'objectif de l'étude décrite dans cet article est de comprendre quand et comment différents jeux de paramètres affectent les performance de reconnaissance. Des probabilités spécifiques qu'un phonème soit correct étant donné les hypothèses de phonèmes générées par deux jeux de paramètres différents et leur probabilités postérieures sont introduites. De cette manière, les probabilités d'exactitude dépendent directement des hypothèses de mots et de phonèmes en compétition. Deux versions d'un reconnaisseur hybride Réseau de Neurones (ANN)/Modèles de Markov Cachés (HMM) sont utilisées. Les deux systèmes de RAP utilisent des paramètres acoustiques d'entrée différents, mais ont la même topologie. Les jeux de paramètres sont ceux obtenus par Analyse à Résolution Multiple (MRA) suivie d'une Analyse en Composante Principale (PCA) décrits dans [1], et par Prédiction Linéaire Perceptive (PLP) décrite dans [2] suivie d'un filtrage RASTA. Ces derniers paramètres seront appelés JRASTAPLP. La même technique de débruitage, décrite dans [1] est utilisée pour chacun des jeux de paramètres acoustiques. Les Réseaux de Neurones sont entraînés pour reconnaître des phonèmes et des transitions en utilisant un corpus de phrases phonétiquement équilibrées qui sont complètement indépendantes des données de test.

Des expériences ont été effectuées sur le corpus de test du corpus Aurora3 (parties italienne et espagnole). Les intervalles des probabilités postérieures d'un phonème calculés par les deux systèmes sont définis dans la section 2. Les statistiques sur les valeurs jointes des probabilités postérieures pour chaque classe de phonème sont présentées et montrent des différences importantes dans le comportement des deux systèmes. Une analyse basée sur le consensus des reconnaisseurs est présentée en section 3. La probabilité qu'ils génèrent une hypothèse correcte est haute quand leurs hypothèses sont les mêmes. Une stratégie de décision spécifique est proposée en section 4 dans les cas où il y a un désaccord entre les mots proposés. Des résultats expérimentaux sur cette stratégie sont présentés en section 5. Malheureusement, même si c'est peu probable, les reconnaisseurs peuvent générer la même hypothèse erronée. Le diagnostique dans ce cas est difficile, parce que les situations de ce type sont rares et les divergences entre les sorties ne peuvent pas être utilisées pour guider une analyse complémentaire. Cet aspect n'est pas traité dans cet article.

## 2. COMPARAISON DES PERFORMANCES DE CHAQUE JEU DE PARAMÈTRES

Un même signal échantillonné $S = \{s(n\tau)\}$, où $\tau$ est la période d'échantillonnage, est transformé par les deux reconnaisseurs en deux flux de paramètres acoustiques, à savoir $Y_m(nT)$ et $Y_j(nT)$ où $T$ est l'intervalle entre deux trames d'analyse successives et les indices $m$ et $j$ réfèrent respectivement aux paramètres MRA et JRASTA-

PLP. Les vecteurs $Y_m(nT)$ et $Y_j(nT)$ représentent deux observations différentes d'un segment de parole centré sur le même échantillon. La valeur de $T$ est 10 msecs et chaque jeu de paramètres contient sept trames d'analyse centrées sur la trame au temps $nT$.

Les deux reconnaisseurs possèdent des modèles acoustiques qui induisent les distributions de probabilités dans les espaces acoustiques $\Gamma_m$ et $\Gamma_j$ des deux types de paramètres. Les probabilités postérieures pour les phonèmes et les diphones dans un point d'un espace acoustique représentent la variabilité des paramètres acoustiques extraits. Les vecteurs $Y_m(nT)$ et $Y_j(nT)$ peuvent avoir des probabilités postérieures similaires ou très différentes pour la même suite de phonème $f$.

Soit $P\{f|Y_m(nT)\}$, la probabilité postérieure d'un phonème $f$ étant donné les paramètres MRA au temps $(nT)$ et $P\{f|Y_j(nT)\}$, la probabilité postérieure de ce même phonème étant donné les paramètres JRASTAPLP au temps $(nT)$. Considérons l'espace contenant les points ayant pour coordonnées $P\{f|Y_m(nT)\}$ et $P\{f|Y_j(nT)\}$. Un tel espace peut être partitionné comme dans la Figure 1. On peut alors comptabiliser l'occupation dee zones de cet espace pour chaque paire de phonèmes. Dans le but d'analyser les variabilités des paramètres acoustiques pour chaque phonème et chaque jeu de paramètres, nous avons considéré le cas $\{f_m = f_j\}$ et collecté des statistiques pour chaque phonème $f$. Ces résultats ont été obtenus en effectuant un alignement forcé des données sur le corpus d'apprentissage de Aurora3 (non utilisé pour entraîner les ANNs) en regardant l'intersection des intervalles proposés par les deux systèmes pour le même phonème $f$.

**FIG. 1:** Partition de l'espace des probabilités postérieures.

La zone H représente la cas dans lequel un phonème est l'hypothèse la plus probable pour les deux jeux de paramètres. Dans ce cas, le phonème devrait être correct. La zone M représente les cas dans lesquels les paramètres montrent une possibilité pour le phonème mais que cette décision n'est pas vraiment très fiable. la zone L correspond aux cas dans lesquels sélectionner le phonème revient à faire un choix aléatoire parmi des candidats en suivant une distribution uniforme (donc très peu fiable). Les zones A, B, et C montrent que les paramètres JRASTA-PLP représentent mieux un phonème que les paramètres MRA, et les zones D, E et F montrent le contraire. Les statistiques recueillies, concernant le fait que les probabilités pour un phonème sont représentées par des points dans les zones définies ci-dessus, peuvent montrer des confusions possibles dues à l'inadéquation des paramètres, ou des modèles, à représenter le phonème.

Une expérience a été mise en œuvre avec le corpus d'apprentissage du corpus Aurora3. Les résultats, groupés par classes de phonèmes, sont présentés dans la Table 1. Après un alignement forcé, des hypothèses de segments de phonèmes sont générées pour chaque flux de paramètres. La moyenne des probabilités postérieures a été calculée pour chaque segment et pour chaque jeu de paramètres, et la paire de valeurs a été représentée par une étiquette de la Figure 1.

Les résultats montrent clairement que les phonèmes non sonores sont plus difficiles à reconnaître que les sonores spécialement parce que les plosives ont une durée courte et les paramètres pour les fricatives sont souvent déformés par le débruitage. Les voyelles sont moins affectées par le débruitage puisque, de manière générale, leur segments ont un SNR assez haut. Les paramètres MRA fournissent une meilleure discrimination pour beaucoup de phonèmes, mais restent particulièrement faibles sur les semi-voyelles. La raison est que ces phonèmes sont souvent très courts et qu'une résolution en fréquence fine est nécessaire pour leur distinction alors que MRA a besoin d'une trame d'analyse longue pour obtenir une résolution en fréquence suffisante. Le SNR moyen par trame après débruitage varie entre 6.5 et 14.8 pour les voyelles, entre 3.6 et 10.4 pour les consonnes non sonores et entre 6.0 et 10.2 pour les consonnes sonores.

Le résultat le plus étonnant est que dans beaucoup de cas les probabilités avec un jeu de paramètres sont hautes alors qu'elles sont basses avec l'autre jeu. Dans les lignes étiquetées DEF, les paramètres MRA ont de meilleures performances, alors que pour les lignes étiquetées ABC, ce sont les paramètres JRASTAPLP. Il est peu probable que ce résultat soit juste dû aux limitations des techniques de modélisation, surtout parce qu'il y a beaucoup de cas dans lesquels une probabilité est supérieure à 0.5 et l'autre inférieure à 0.1. Une partie des divergences (cas A, B, D et F) est probablement due aux variabilités intrinsèques des paramètres acoustiques qui rendent difficile l'inférence des distributions appropriées conduisant à une bonne discrimination.

**TAB. 1:** Distributions des zones de probabilités pour les données CH1 des parties espagnole (37933 pho.) et italienne (34043 pho.) du corpus Aurora3.

| | Langue | Cons. non sonores | Cons. sonores | Voyelles |
|---|---|---|---|---|
| D∪E∪F | SPA | 20,74% | 13,39% | 23,12% |
| | ITA | 20,49% | 18,14% | 12,59% |
| A∪B∪C | SPA | 15,05% | 13,80% | 5,3% |
| | ITA | 16,01% | 11,09% | 14,85% |
| H | SPA | 40,36% | 63,94% | 66,4% |
| | ITA | 37,54% | 61,04% | 58,02% |
| M | SPA | 2,31% | 1,79% | 1,53% |
| | ITA | 8,29% | 3,87% | 7,01% |
| L | SPA | 21,55% | 7,08% | 3,64% |
| | ITA | 17,67% | 5,86% | 7,54% |

## 3. DÉFINITION DES SITUATIONS

Une séquence de mots $W$ génère un chemin dans $\Gamma_m$ et dans $\Gamma_j$. Ces deux chemins représentent le même signal. Les reconnaisseurs étiquettent chaque chemin avec une séquence de mots $W_m$ et $W_j$. Plusieurs situations peuvent alors être définies en fonction des types de consensus entre les mots dans $W_m$ et $W_j$. Les séquences peuvent soit être identiques ou différer d'un ou plusieurs mots. Dans ce dernier cas, il peut être utile d'identifier les segments temporels dans lesquels la différence apparaît et analyser les types de divergences.

Un ensemble d'*états de comparaison de résultats* est défini comme suit :

- $Q1 : W_m = W_j$ ;
- $Q2$ : le même mot ou deux mots différents sont proposés dans approximativement le même intervalle de temps même si $W_m \neq W_j$, *i.e.* :

$\exists(a,b)/\{w_{ma} = w_{jb}\} \wedge \{S(w_{ma}) \approx S(w_{jb})\}$
où $w_{ma} \in W_m$, $w_{jb} \in W_j$, et $S(x)$ fait référence à la segmentation de $x$.
– $Q3$ : deux segments de $W_m$ et $W_j$ ont approximativement les mêmes bornes temporelles mais sans aucun mot en commun.

Quand l'état du résultat de reconnaissance est $Q1$, la combinaison des scores des hypothèses générées par les deux systèmes de reconnaissance n'apporterait rien de plus. Le WER est très bas dans cet état et le diagnostic révèle que les erreurs de suppression dans $Q1$ sont essentiellement dues aux imprécisions du détecteur d'activité vocale (VAD) ou du débruitage et non à la capacité des jeux de paramètres à discriminer les phonèmes entre eux. Si une portion de signal appartenant a un mot est considéré comme un segment non parlé, alors la partie restante du mot est souvent attachée à un mot voisin qui sera probablement mal reconnu. Les erreurs de segmentation sont une cause d'erreur fréquentes dans $Q1$. Les insertions sont souvent dues au fait que le bruit de fond est considéré comme un segment de parole.

En utilisant les ensembles de test du corpus Aurora3, dans les cas de consensus sur la phrase entière, la couverture est de 72.66% pour l'espagnol avec un WER de 0.16% et 63.16% pour l'italien avec un WER de 2%.

Dans les cas correspondant à l'état $Q2$, le WER est également bas. Les erreurs sont essentiellement des substitutions. Certaines erreurs sont dues à la segmentation et au débruitage, mais les autres révèlent un faible pouvoir de discrimination entre les phonèmes. Cette faiblesse est commune aux deux jeux de paramètres dans certaines zones de leur espace acoustique, spécialement celles correspondant à un SNR bas.

Ces erreurs peuvent être évitées en utilisant un bon modèle de langage et un bon modèle lexical. Cet aspect n'est pas traité dans cet article dont l'objectif principal est l'étude comparative des performance des paramètres acoustiques. En absence de consensus, un WER oracle de 5,33% pour l'espagnol et 9,67% pour l'italien révèle que beaucoup d'erreurs pourraient potentiellement être évitées en utilisant une stratégie d'évaluation et de décision conçue pour maximiser la probabilité de sélectionner la bonne hypothèse quand celle-ci est proposée par l'un des deux reconnaisseurs.

## 4. TYPES DE DIVERGENCES ENTRE LES SORTIES DES SYSTÈMES DE RECONNAISSANCE

Quand il n'y a pas consensus entre les hypothèses générées par les deux reconnaisseurs, alors leurs hypothèses les plus probables sont alignées. Les expériences décrites en détails dans [1] montrent que le reconnaisseur utilisant les paramètres MRA a des meilleures performances dans les deux langues que l'autre système. Ses hypothèses sont donc considérées comme référence pour aligner les résultats des deux reconnaisseurs quand ils sont en désaccord. Soit $W_m(b,e)$, un mot ou un séquence de mots proposé par le système MRA dans l'intervalle de temps $(b,e)$ et $S_j(b_j,e_j)$ la séquence de phonèmes proposée en compétition par le système JRASTAPLP dans un segment de temps chevauchant fortement l'intervalle $(b,e)$.

Il est intéressant de considérer les situations définies dans la Table 2 décrivant les divergences entre les reconnaisseurs.

**TAB. 2:** Types de divergences entre les sorties de reconnaisseurs différents

| MRA : $W_m(b,e)$ | RPLP : $S_j(b_j,e_j)$ | TYPE DE DIVERGENCE |
|---|---|---|
| $w_i$ | $w_k$ | substitution $(sb)$ |
| $w_i$ | $w_i w_k$ | j-insertion $(i_j)$ |
| $w_i w_k$ | $w_i$ | m-insertion $(i_m)$ |
| $w_i$ | $w_q w_k$ | j-substitution + insertion $(si_j)$ |
| $w_q w_k$ | $w_i$ | m-substitution + insertion $(si_m)$ |
| Tous les autres cas | | plusieurs divergences $(md_{mj})$ |

La stratégie de décision suivante est proposée quand il n'y a pas consensus sur la phrase. Le mot proposé par le système MRA est sélectionné, excepté pour la situation $sb$ pour laquelle la règle de décision introduite ci-dessous est appliquée.

Soit $C(w,m,j)$ une fonction représentant le fait que $w$ est correct quand les hypothèses dont le meilleur score est obtenu avec les jeux de paramètres MRA et JRASTAPLP sont respectivement $w_m$ et $w_j$. La règle de décision est :

$$w^* = \underset{w \in w_m, w_j}{\text{argmax}} P\{C(w)|w_m, w_j, \sigma\} \quad (1)$$
$$= \underset{w \in w_m, w_j}{\text{argmax}} P\{\sigma|C(w,m,j)\}P\{C(w,m,j)\}$$

où $C(w)$ est un prédicat qui est vrai quand $w$ est correct et $\sigma$ :$[\sigma_1(w), \ldots, \sigma_n w, \ldots, \sigma_N]$ représente une séquence d'étiquettes définies dans la Figure 1. Étant donné que l'hypothèse $w$ est disponible avec sa segmentation, chaque segment qui le compose correspond à un phonème, de telle manière que $w$ peut être représentée par une séquence de phonèmes $w$ : $[f_1(w), \ldots, f_n w, \ldots, f_N(w)]$. Pour chaque phonème $f_n(w)$, il est possible de considérer son compétiteur proposé par l'autre reconnaisseur (celui qui a le plus grand nombre de trames en commun avec $f_n(w)$). Les probabilités postérieures des deux phonèmes sont représentées par un symbole $\sigma_n(w)$ en fonction de la grille définie dans la Figure 1.

La probabilité $P\{C(w,m,j)\}$ est calculée à partir du corpus d'apprentissage. Pour les larges vocabulaires, cette probabilité peut être obtenue comme un produit de probabilités *a-priori* de syllabes ou même de phonèmes. Notons que la probabilité que $w_m$ et $w_j$ soit tous les deux incorrects peut être obtenue en soustrayant de l'unité la somme des probabilités que $w_m$ ou $w_j$ soit correcte. Une procédure similaire peut être utilisée pour calculer la probabilité que l'hypothèse soit correcte en cas de consensus. Cette probabilité est aussi une mesure de confiance pour décider du rejet d'une hypothèse.

## 5. EXPÉRIENCES DE RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

Des expériences de reconnaissance ont été menées sur les données CH1 du jeu de test du corpus Aurora3. Les systèmes de reconnaissance ont été entraînés avec des données téléphoniques de la vie courante très différentes de celles du corpus d'apprentissage d'Aurora3.

Après alignement des meilleures séquences de mots générées par les deux reconnaisseurs, la stratégie de décision suivante a été utilisée. Si les deux systèmes proposent la

même hypothèse, alors elle est validée et conservée, sinon chaque hypothèse de mot $w_m(b,e)$ générée par le reconnaisseur utilisant les paramètres MRA est considérée pour validation. Ceci est motivé par le fait que ce système a un WER plus bas que celui utilisant les paramètres JRASTA-PLP.
La stratégie compare $w_m(b,e)$ avec les hypothèses générées par l'autre reconnaisseur. Trois cas possibles sont alors considérés, à savoir :

1. substitution : un seul mot $w_j(b_j,e_j)$ qui chevauche fortement $w_m(b,e)$,
2. cas spécial de j-suppression : une hypothèse de silence (segment non parlé) est générée par le reconnaisseur utilisant les paramètres JRASTAPLP dans l'intervalle de temps de $w_m(b,e)$,
3. cas spécial de j-insertion : une hypothèse de mot $w_j(b_j,e_j)$ est générée dans un intervalle de temps où le reconnaisseur utilisant les paramètres MRA a généré une hypothèse de silence.

La probabilité $P\{C(w,m,j)\}$ a été calculée a partir du corpus d'entraînement de Aurora3 et utilisé pour sélectionner une hypothèse parmi celles proposées. La règle de décision est l'équation 2. Les résultats, en terme de WER sont présentés dans la Table 3.

**TAB. 3:** Performance, en terme de WER, de la nouvelle stratégie de décision comparée au meilleur système

| Performances | Italien | Espagnol |
|---|---|---|
| Système MRA | 20.34% | 15.19% |
| Nouvelle strat. - sub | 6.2% | 5.71% |
| Nouvelle strat. - del | 9.79% | 1.99% |
| Nouvelle strat. - ins | 1.57% | 4.72% |
| Nouvelle strat. - WER global | 17.56% | 12.42% |

On observe une réduction relative du WER de 13.67% pour l'italien et 18.24% pour l'espagnol.

## 6. CONCLUSION ET TRAVAIL FUTUR

Deux systèmes de RAP utilisant deux jeux de paramètres acoustiques différents ont été utilisés pour effectuer le diagnostic et la combinaison de leurs résultats en vue d'améliorer la précision de reconnaissance. Grâce à l'alignement forcé de la phrase de référence, il a été observé que les probabilités postérieures des phonèmes obtenues avec les deux jeux de paramètres, pour le même phonème et le même intervalle de temps, peuvent être très différentes dans un nombre non négligeable de cas. Ceci peut être causé par la différence des modèles (qui génèrent ces probabilités), par l'effet du débruitage ou paramètres acoustiques. En effet, ces derniers ont des pouvoirs de discrimination plus opu moins effectif en fonction de la classe des phonèmes a reconnaître. Le travail futur consistera à vérifier si cette tendance est confirmée avec des techniques de modélisation différentes. Si la tendance est confirmée, notre attention se portera sur les paramètres acoustiques en essayant de caractériser les confusions dues aux variabilités intrinsèques de ces paramètres.
Une nouvelle stratégie pour combiner les résultats des systèmes de reconnaissance a été proposée. Elle est basée sur la probabilité qu'une hypothèse soit correcte, étant donné l'identité des hypothèses de mots en compétition. Des réductions significatives du WER ont été trouvée pour la portion CH1 des parties italienne et espagnole du corpus Aurora3. Nous explorerons par la suite comment cette approche peut être étendue à des tâches large vocabulaire.

## 7. REMERCIEMENTS

L'étude décrite dans cet article fait partie de l'effort de recherche mis en œuvre dans le projet DIVINES sur l'extraction et le diagnostic des paramètres pour la reconnaissance automatique de la parole. Ce projet est en partie financé par un programme de la division Technologie des Langues Humaines de la Communauté Européenne.

## RÉFÉRENCES

[1] R. Gemello, F. Mana, D. Albesano, and R. De Mori. Multiple resolution analysis for robust automatic speech recognition. *Computer Speech and Language*, 2004.

[2] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2 :578–589, October 1994.

[3] H. K. Kim and M. Rahim. Why speech recognizers make errors ? a robustness view. In *Proceedings of International Conference on Spoken Language Processing*, page ThA1703o1, Jeju, Korea, 2004.

[4] A. Sankar. Bayesian model combination (baycom) for improved recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 845–848, Philadelphia, PA, March 2005.

[5] T. Utsuro, Y. Kodama, T. Watanabe, H. Nishizaki, and S. Nakagawa. Confidence of agreement among multiple lvcsr models and model combination by svm. In *Proceedings of IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, volume I, pages 16–19, Hong Kong, China, 2003.

[6] R. Zhang and A.I. Rudnicky. Word level confidence annotation using combinations of features. In *Proceedings of European Conference on Speech Communication and Technology, Eurospeech 01*, pages 2105–2108, Aalborg, Denmark, 2001.

[7] A. Zolnay, R. Schluter, and H. Ney. Acoustic feature combination for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 457–460, Philadelphia, PA, March 2005.